Mechanistic Effect Models in the Environmental Risk Assessment of Chemicals

Special Focus on Plant Protection Products

A Report of the Model Acceptability Criteria and Scenario Development (MAD) Working Group, Part of the Effect Modeling SETAC Interest Group

Book Editors:

Andreas Focks, Udo Hommen, Thomas G. Preuss, and Alpar Barsi

SETAC®

Mechanistic Effect Models in the Environmental Risk Assessment of Chemicals

Special Focus on Plant Protection Products

A report of the Model Acceptability criteria and scenario Development (MAD) working group, part of the Effect Modeling SETAC Interest Group

Edited by

Andreas Focks, Udo Hommen, Thomas G. Preuss, Alpar Barsi

Published by the Society of Environmental Toxicology and Chemistry

Library of Congress Cataloging-in-Publication Data
Mechanistic effect models in the environmental risk assessment of chemicals: special focus on plant protection products / Edited by Andreas Focks ... [et al.].

Includes bibliographical references.

ISBN-13: 979-8-31781-192-1

SETAC Press is an imprint of the Society of Environmental Toxicology and Chemistry.

International Standard Book Number-13: 979-8-31781-192-1

Printed in the United States of America

14 13 12 11 105 4 3 2 1

The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences — Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Reference listing: Focks A, Hommen, U, Preuss TG, Barsi A, eds. 2026. Mechanistic effect models in the environmental risk assessment of chemicals: special focus on plant protection products. Society of Environmental Toxicology and Chemistry (SETAC).

Table of Contents

List of Figures

List of Tables

List of Acronyms

ABC	Approximate Bayesian Computation
ABM	Agent-Based Model
ADME	Absorption, Distribution, Metabolism, and Excretion
AF	Assessment Factor
AIC	Akaike Information Criterion
ALMaSS	Animal, Landscape and Man Simulation System
AOP	Adverse Outcome Pathway
ApisRAM	Apis (bee) Risk Assessment Model
AS	Active Substance
BBCH	Biologische Bundesanstalt für Land – und Forstwirtschaft, Bundessortenamt und Chemische Industrie (scale to identify the phenological development stages of plants)
BIC	Bayesian Information Criterion
BMP	Beekeeping Management Practice
CASM	Comprehensive Aquatic Systems Model
CDF	Cumulative Distribution Function
CRO	Contract Research Organization
CVD	Color Vision Deficiency
DaLaM	*Daphnia* Lake Model
DDD	Daily Dietary Dose
DEB	Dynamic Energy Budget
DEBtox	Dynamic Energy Budget models applied to (eco)toxicological problems
DIC	Deviance Information Criteria
DT50	Dissipation time to reach 50% of initial concentration
eCDF	empirical Cumulative Distribution Function
ECx	Effective Concentration for x% effect
EDx	Effective Dose for x%

EFSA	European Food Safety Authority
EMF	Exposure Multiplication Factor
EPx	Exposure Profile causing x% effect (defined as multiplication factor)
ERA	Environmental (Ecological) Risk Assessment
ERC	Ecotoxicologically Relevant type of Concentration
ERO	Ecological Recovery Option
ETO	Ecological Threshold Option
EU	European Union
FAIR	Findable, Accessible, Interoperable, and Reusable
FOCUS	FOrum for Co-ordination of pesticide fate models and their USe
FOCUSsw	FOCUS surface water
GAP	Good Agricultural Practice
GD	Guidance Document
GERDA	German Run-off, Erosion, and Drainage Risk Assessment
GLP	Good Laboratory Practice
GMP	Good Modeling Practice
GOF	Goodness of Fit
GSA	Global Sensitivity Analysis
GUTS	General Unified Threshold model for Survival
GUTS RED-SD GUTS	REDuced model based on Stochastic Death
HyLaM	Hydrodynamic Lake Model
IBC-grass	Individual-based Community model for grassland
IBM	Individual Based Model
IDamP	Individual based *Daphnia magna* Population (model)
IQR	Interquartile Range
KDE	Kernel Density Estimate
LDx	Lethal Dose for x%
LPx	Lethal exposure Profile causing x% mortality (defined as multiplication factor)
MACRO	Model of water flow and solute transport in macroporous soil
MCMC	Monte Carlo Markov Chains

MASTEP	Metapopulation model for Assessing Spatial and Temporal Effects of Pesticides
MEM	Mechanistic Effect Models (TKTD, population, community, and ecosystem models)
MRL	Maximum Residue Level
MS	Member State
N/A	not available, not applicable, or no answer
NAM	New Approach Methodologies
NHST	Null-Hypothesis Significance Testing
NOEC(L)	No Observed Effect Concentration (Level)
NOR	Normal Operating Range
NTA	Non-Target Arthropods
OAT	One At a Time
ODD	Overview, Design concepts, and Details
ODE	Ordinary Differential Equation
OECD	Organisation for Economic Co-operation and Development
PB(T)K	Physiologically Based (Toxiko-)Kinetic (model)
PDF	Probability Density Functions
PEARL	Pesticide Emission Assessment at Regional and Local scales
PEC	Predicted Environmental Concentration
PELMO	Pesticide Leaching Model
PERA	Partnership for next generation, systems-based Environmental Risk Assessment
PERSAM	PERsistence in Soil Analytical Model
PMoA	Physiological Modes of Action
POLARIS	POpulation-Level RISk assessment (modeling framework)
POM	Pattern-Oriented Modeling
PPC	Posterior Predictive Check
PPP	Plant Protection Product
PPR	Plant Protection Products and their Residues (panel)
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-analyses
PRZM	Pesticide Root Zone Model

PSRF	Potential Scale Reduction Factor
PT	Proportion of food taken from Treated fields
qAOP	quantitative Adverse Outcome Pathway
R^2	Coefficient of Determination
RAC	Regulatory Acceptable Concentration
RMM	Risk Mitigation Measure
ROV	Ratio of Variation
RPU-ED	Resource Providing Units and Environmental Drivers
RUD	Residue Unit Dose
SA	Sensitivity Analysis
SD	Standard Deviation
SE	Standard Error
SMC-ABC	Sequential Monte Carlo algorithm
SO	Scientific Opinion
SPG	Specific Protection Goal
StoLaM	Stoichiometric Lake Model
TER	Toxicity Exposure Ratio
TKTD	Toxicokinentic-Toxicodynamic
TRACE	Transparent And Comprehensive model Evalu(d)ation
TWA	Time Weighted Average
UA	Uncertainty Analysis
VPC	Visual Predictive Check
WAIC	Widely Applicable Information Criterion

Acknowledgements

A project like this requires the involvement of many people beyond their regular work to make it a reality. This was certainly true for this highly interdisciplinary book, which depended on contributions from various stakeholders and scientific disciplines. These included the editors and authors of the chapters, as well as a number of other people who are not named in the book. In particular, we would like to thank Alessio Ippolito, Vanessa Mazerolles, Roman Ashauer, Pernille Thorbek, and Michael Brauer for their thoughtful challenges and contributions during the discussion and writing process. We also extend our sincere thanks to the anonymous reviewers, who dedicated considerable time and effort to read the manuscript and provide detailed, constructive feedback. Special thanks go to Sabine Apitz and Jen Lynch, who managed the entire review process, carefully reviewing all the feedback from reviewers and verifying the authors' responses. They also assisted the authors with the linguistic revision of the manuscript. Erin Nelson handled the final processing of the book.

None of this would have been possible without the support of SETAC and the SETAC Effect Modeling Interest Group, which provided the platform for such important interdisciplinary discussions with many stakeholders. Finally, we would like to thank Bayer AG for hosting an in-person meeting in Monheim am Rhein, Germany, on 27-28 September 2021.

About the Editors

Andreas Focks

Andreas Focks is Professor of Environmental Systems Modeling at the Institute of Mathematics at the University of Osnabrück, Germany, and Managing Director of the Research Center for Environmental Systems Research (IUSF). He received his diploma in Applied Systems Science in 2005 and his PhD in 2009 on the topic "Risk assessment of veterinary pharmaceuticals in soils." Andreas moved to Wageningen, The Netherlands, for a PostDoc in the Marie Curie project CREAM (Mechanistic Effect Models for Ecological Risk Assessment of Chemicals) in 2011 and worked at Wageningen Environmental Research from 2013-2021. He is co-author of the EFSA scientific opinion on TKTD models (2018) and the EFSA bee guidance (2023). Andreas has been a SETAC member since 2006 and has been organizing symposia and sessions at SETAC meetings for many years. Since 2019 Andreas has been a member of the Steering Committee of the SETAC Europe Interest Group on Effect Modeling. He has coordinated the preparation of the MAD book as Chair from 2020.

Udo Hommen

Udo Hommen holds a master's degree in biology and a doctorate in ecology from RWTH Aachen University. Since 2000 he has been working at the Fraunhofer Institute for Molecular Biology and Applied Ecology in Schmallenberg, Germany, initially in the Department of Ecotoxicology, and since 2023 in the Department of Modeling and Bioinformatics. His work usually relates to the ecological risk assessment of chemicals, especially pesticides, in publicly or industry funded research projects. He has been GLP study director in many mesocosm studies in the context of chemical regulation. Udo was a member of the steering committee of the EU Marie Curie project CREAM (Mechanistic Effect Models for Ecological Risk Assessment of Chemicals) and of the founding committee of the SETAC Europe Interest Group on Effect Modeling. He was involved in the organization of SETAC workshops and special scientific symposia and serves in the editorial board of IEAM.

Thomas G. Preuss

Thomas G. Preuss studied biology at the University in Tübingen, Germany, and performed his PhD at the RWTH Aachen University, Germany, in 2007. After postdoc positions in Aachen and Sweden he built up a research group at the RWTH Aachen from 2010-2013. In that time, he supervised 8 PhD students and 12 master students in total. In 2013 he joined Bayer as an effect modeling expert, from 2016-2023 he was leading the new established effect modeling team. In 2020 he was nominated as a Distinguished Science Fellow for Bayer. Since 2023 he works as a scientific strategic advisor for modeling in the environmental risk assessment at Bayer. In his academic career he received several awards for his research as well as for his lectures. He was involved in three Marie-Curie projects (Aquabase, CREAM and QTox) and was member of the founding committee of the SETAC Europe Interest Group on Effect Modeling. He was involved in the organization of SETAC workshops and special scientific symposia and serves in the editorial board of ET&C.

Alpar Barsi

Alpar Barsi obtained a master's degree in ecology and environmental protection from the University of Novi Sad, Serbia, in 2009. After completing his studies, he moved to the Netherlands for an internship at Wageningen University and Research, where he gained first experience in ecological modeling. He was later awarded the Marie Curie ITN scholarship, allowing him to pursue doctoral studies jointly at the French National Research Institute for Agriculture, Food, and the Environment (INRAE) and the Free University Amsterdam (VU), the Netherlands. His research focused on advancing our understanding of chemical effects on freshwater snails by integrating various laboratory test protocols and effect models. Upon completing his PhD in 2015, Alpar briefly worked on a project at the Jagiellonian University in Kraków, Poland, before serving as a scientific assessor at the Dutch Board for the Authorisation of Plant Protection Products and Biocides (Ctgb) for eight years. In this role, he was responsible for evaluating dossiers in the context of the environmental risk assessment of pesticides. He has been an active participant in SETAC, contributing to working groups and organizing scientific sessions. Since 2023, Alpar's career focusses more on biodiversity issue, and he currently works as a policy officer for nature and monitoring at the Province of Utrecht in the Netherlands.

List of Authors

Johan Axelman
Swedish Chemicals Agency (KemI)
Sweden

Alpar Barsi
Former: Dutch Board for the Authorisation of Plant Protection Products and Biocides (Ctgb)
The Netherlands
Current: Province of Utrecht
The Netherlands

Jeremias Becker
German Environment Agency (UBA)
Germany

Eugenia Chaideftou
Benaki Phytopathological Institute (BPI)
Greece

Sandrine Charles
University of Lyon, CNRS
France

Sabine Duquesne
German Environment Agency (UBA)
Germany

Fabienne Ericher
Syngenta Crop Protection AG
United Kingdom

Andreas Focks
Osnabrück University
Germany

Michael Fryer
Chemical Regulations Directorate (CRD)
United Kingdom

Michail Gioutlakis
Former: Benaki Phytopathological Institute
Greece
Current: RIFCON GmbH
Germany

Benoit Goussen
ibacon GmbH
Germany

Simon Hansul
Former: Osnabrück University
Germany
Current: Research Institute for Ecosystem Analysis and Assessment (gaiac), Germany

Udo Hommen
Fraunhofer Institute for Molecular Biology and Applied Ecology IME
Germany

Oliver Jakoby
RIFCON GmbH
Germany

Mira Jakoby-Kattwinkel
Former: University of Koblenz-Landau
Germany
Current: Baader Konzept GmbH
Germany

Judith Klein
Fraunhofer Institute for Molecular Biology and Applied Ecology IME
Germany

Joachim Kleinmann
Former: WSC Scientific GmbH
Germany
Current: BASF SE
Germany

Josef Koch
Research Institute for Ecosystem Analysis and Assessment (gaiac)
Germany

Thomas G. Preuss
Bayer AG
Germany

Melissa Reed
Chemical Regulations Directorate (CRD)
United Kingdom

Stefan Reichenberger
knoell France SAS
France

Hanna Schuster
Cambridge Environmental Assessments
(RSK ADAS Ltd.)
United Kingdom

Tido Strauss
Research Institute for Ecosystem Analysis and Assessment (gaiac)
Germany

Sanne van den Berg
Wageningen University and Research
The Netherlands

Peter Vermeiren
Former: Radboud University
The Netherlands
Current: RIFCON GmbH
Germany

Magnus Wang
WSC Scientific GmbH
Germany

Johannes Witt
Bayer AG
Germany

Foreword

In 2019, the European Food Safety Authority (EFSA) published the "Scientific Opinion on Toxicokinetic/Toxicodynamic (TKTD) effect models for regulatory risk assessment of pesticides," which appeared as the first comprehensive reference to the use of mechanistic effect models (MEMs) in regulatory environmental risk assessment in Europe. Effect modelers and colleagues from regulatory authorities were excited to see MEMs being used in regulatory decision making, but also asking themselves how to build on this momentum. The community of scientists and regulators determined that there was a need to explore the scientific basis for the application and the assessment of MEMs in general, including not only TKTD models, but also population and community models, in the scope of regulatory environmental risk assessment (ERA). Such work still has high relevance and some urgency, because MEMs enable scientific understanding and methods to be potentially included into ERA procedures. Nevertheless, regulatory agencies in the European Union (EU) have limited means to properly assess the MEMs applied for risk assessment dossiers.

This book has drawn from the efforts of a Society of Environmental Toxicology and Chemistry (SETAC) working group focused on **M**odel **A**cceptability criteria and scenario **D**evelopment (MAD group), which organized a series of meetings, from its inception in 2020, to facilitate this scientific output. The chapters cover relevant topics for the regulatory use of mechanistic effect models, and touch on aspects of the general use of MEMs and considerations about criteria for model evaluation. Other chapters report on scenario development for the use of MEMs in regulatory ERA, and the evaluation of data underlying MEMs, as well as providing specific perspectives of modular model evaluation, model calibration and validation, and uncertainty and sensitivity analyses. In all chapters, the general aim is to provide the scientific grounds for future regulatory discussions and activities, without precluding regulatory decisions or guidance development. All chapters seek to support both model developers and model evaluators with their tasks; some explicitly by providing catalogues of questions or checklists.

The content of this book has been produced primarily in the context of risk assessment regulations in the European Union. Nevertheless, I strongly believe that many, if not most, aspects also will be immediately relevant for the development of risk assessment methodologies in other countries. A logical next step would be to adjust the outcome of the working group with the researcher and regulatory communities globally.

In the end, our goal is to provide some means to properly assess mechanistic effect models for their use in the environmental risk assessment of chemicals, here specifically of pesticides, because, if they are properly tested and applied, MEMs can improve ERA and make it safer and more relevant.

Osnabrück, August 2024
Andreas Focks

1. Mechanistic effect models in the environmental risk assessment of pesticides: Scope of the book and potential areas of application

Jeremias Becker, Magnus Wang, Alpar Barsi, Thomas G. Preuss, Udo Hommen, Oliver Jakoby, Benoit Goussen, Stefan Reichenberger, Melissa Reed, Johan Axelman, Sabine Duquesne

Mechanistic effect models (MEMs) dynamically simulate essential processes in ecological systems and how these processes are affected by chemicals (Grimm & Martin, 2013). The modeled systems can range from (sub)organisms, to populations, communities, or ecosystems. Recently, MEMs have been increasingly proposed for use in the regulatory environmental risk assessment (ERA) of pesticides[1], for example, in the European Union (EFSA PPR, 2015, 2017, 2018a, 2018b; European Commission, 2014; Hommen et al., 2016; Raimondo et al., 2018), and also on OECD level (OECD, 2006). This book will focus on the use of MEMs for the regulation of plant protection products in the European Union, but model developments and applications are considered more generally. So far, MEMs for regulatory purposes have been proposed mainly as refinement options, e.g., for assessing the effects of time-variable exposure on organisms or assessing effects on populations (EFSA PPR, 2013). However, MEMs can be useful exploratory tools to address various questions; this includes ranking the effectiveness of alternative risk mitigation measures or supporting the interpretation of ecotoxicological studies. In retrospective risk assessments, such models can be used to analyze mechanisms behind effects observed in (semi-) field studies and to support the establishment of appropriate assessment factors (AF) in ERA with an improved mechanistic understanding. These and other options are described in more detail in section 1.2 .

When compared with the simpler calculations of toxicity-exposure ratios (TER) or risk quotients, MEMs improve our mechanistic understanding of the potential risk of pesticides (see e.g., [EFSA, 2009, 2013; EFSA PPR, 2015]), and at the same time make uncertainties in the ERA explicit. However, the use of MEMs also contributes to the complexity of the ERA of pesticides. Part of this complexity stems from the considerable effort it takes to develop, validate, and evaluate a model, then generate suitable data. Thus, the benefit of implementing new prospective methods has to be thoroughly weighted against simpler calculations.

1 With pesticides, we refer both to plant protection products (PPP) and to the active substances (AS) therein.

MEMs must usually be validated, which is to say, model predictions should be tested with independent data or knowledge (pattern-oriented modeling). Such comparisons may not be feasible for all types of models, for example, spatially explicit population models are challenging in that context. Other open and controversial questions include who would collect these data and whether such data collection should be considered mandatory before corresponding MEMs can be used in regulatory ERA. In this context, one of the aims of this book is to collect knowledge around the application of MEMs. For example, the assessment of modular models (Chapter 5) may be useful in this context, and more detailed information on model calibration and validation is provided in Chapter 6.

In addition to the comparison of MEMs with empirical data, it is essential to benchmark the performance of MEMs against alternative (especially currently used) empirical methods in ERA. Here, the comparison to empirical methods, as well as between MEMs of different complexities, may be of use. For pesticide exposure models, Boström et al. (2019) provided an example in which the predictive power of models with varying degrees of complexity have been assessed compared with field data. For effect models, similar types of conceptual studies would appear relevant as well.

The application of MEMs requires extensive communication and thorough evaluation before a proposed model can be considered suitable for the regulation ERA of pesticides. Guidelines for model development and documentation have been laid out in several publications, such as the overview, design concepts, and details (ODD) protocol for the documentation and communication of individual-based models (Grimm et al., 2006; Grimm et al., 2010; Grimm et al., 2020), and the transparent and comprehensive model evaluation (TRACE) framework (Grimm et al., 2014), which includes the model description, and also requires the documentation of the model development and testing (e.g., comparison of different versions, sensitivity analyses, code verification, validation, etc.). Additionally, in Europe, the European Food Safety Authority Panel on Plant Protection Products and their Residues (EFSA PPR) published a "Scientific Opinion on good modeling practice in the context of mechanistic effect models for risk assessment of plant protection products" (GMP-SO; EFSA PPR, 2014), which includes a checklist for the evaluation of MEMs. In the US, Raimondo et al. (2021) published the "Pop-GUIDE" as a guidance for the development, use, and interpretation of population models in ERA.

So far, no clear quality criteria for the development and evaluation of MEMs in the ERA of pesticides have been established. The term "quality criteria" for the evaluation of MEMs is meant here as more than measures to compare outcomes of a MEM with observed data alone. The evaluation of MEMs starts earlier, following protocols such as those exemplified in the GMP-SO (EFSA PPR, 2014), and includes checks of the environmental scenario (Chapter 3), underlying ecological data (Chapter 4), equations and computer code, and other categories. Finally, quality criteria for the performance of MEMs need to be developed to address both (i) a list of measures applied to assess the quality or performance of a model or a model application to a specific risk assessment, and (ii) values or thresholds for these measures that should be met. Chapters 6 and 7 provide an overview on model calibration and validation, and uncertainty and sensitivity analyses.

Without operational acceptability criteria, models cannot be assessed in a consistent way, but would rely on expert judgment alone. As MEMs would not be acceptable for authorities such as EFSA and member states (MS) without appropriate acceptability criteria, expertise must be built for the development, use, and evaluation of MEMs in regulatory ERA in the future.

This book provides considerations and background on the development and evaluation of mechanistic effect models, compiled by a multi-partite consortium with stakeholders from industry, contract research organizations (CRO), academic research institutions, and a number of European regulatory authorities. Such a multi-stakeholder group cannot define binding acceptability criteria for MEMs in the ERA of pesticides. However, it can provide input for further regulatory development that may be used to promote the development of such criteria and to identify critical research questions for the further assessment of MEMs and their use in the ERA of pesticides.

1.1. Scope of the book

Throughout this book, MEMs are defined as mechanistic and dynamic models that can relate a given chemical exposure to the effects on organisms, populations, communities, or ecosystems.

Empirical (often referred to as statistical) models in ecotoxicology, such as classical dose-response models, are purely descriptive, and the model parameters can often only be determined by fitting the model to a dataset (e.g., EC50 and slope of a concentration-response model). In contrast, mechanistic models aim to describe processes using model parameters that, at least theoretically, can be measured directly. However, this differentiation is often not very strict and mechanistic models often include empirical components. In MEMs, predictions emerge from the physical, chemical, and biological mechanistic principles implemented in the models. For example, toxicokinetic-toxicodynamic (TKTD) models are mechanistic in the sense that the describe the course effect over time by considering uptake, elimination, damage, and repair processes, but on the other hand, the model parameters are often obtained by fitting to results of ecotoxicological tests.

Empirical models are typically calibrated to a given dataset such that the predicted effect size will meet the effect size in observational data, for example, a dose-response function to effects on reproduction at the end of the experiment. Therefore, predictions by means of empirical models are likely to be appropriate under the conditions used for model development and parameterization but may deviate from the real world in unpredictable ways when conditions change, for example, shorter or longer exposure times. In contrast, mechanistic models allow extrapolation outside the data range they were calibrated to, as long as they contain the relevant mechanisms (Baker et al., 2018), because typically the underlying processes are subject to parameterization. In that context, a mechanistic model might fail to explain data in the case that relevant processes have been missed or insufficiently implemented and parameterized (Pilkey & Pilkey-Jarvis, 2009).

Therefore, by systematically varying the degrees of freedom of mechanistic models, calibration data can be used to check whether relevant processes are included, and system understanding can be improved.

Mechanistic effect models are likely to perform better than empirical models when predicting effects under conditions that deviate from those used for model parameterization. Therefore, MEMs may be better suited for extrapolating effects to different environmental conditions, as long as one remains within the limits for which a model has been designed and set up (the domain of use of a model). This is discussed in detail in Chapters 5 and 6 in this book.

Because mechanistic effect models are dynamic, they simulate a system's development and predict effects explicitly over time. In contrast to static models, such as classical statistical dose-response models, they make use of temporal data obtained from experiments for model calibration and model testing (OECD, 2006). The explicit consideration of the temporal course of effects in MEMs, for example in dependency on the exposure profile over time, improves the realism, and is one of the important added values of MEMs to ERA. In addition, using MEMs can also reduce uncertainty or make uncertainties more explicit, for example, those concerning time-dependency and time varying exposure or the effects of realistic field conditions such as environmental temperatures. However, using MEMs also increases complexity and creates new challenges. Thus, it is desirable to critically assess the benefits of using (dynamic) MEMs in terms of the assessment of time-variable exposure and effects.

MEMs can be classified in different ways. In this book we use the level of biological organization that is the subject of the modeling as the main category (similar to Becker et al. [2024]; Galic et al. [2010]) and more technical criteria for subdivision (similar to Schmolke et al. [2010]).

1.2. Potential areas of application of MEMs in regulatory risk assessment

MEMs have the potential to support risk assessors with information on the risks of pesticides in various ways. The ERA of pesticides in the European Union is currently carried out following a tiered approach. For example, data requirements for the active substances of plant protection products are defined in the Commission Regulation (EU) No 283/2013. These data are used to conduct an initial Tier 1 assessment based on the uniform principles laid down in Commission Regulation (EU) No 546/2011.

Tier 1 aims at being conservative in the sense that potentially harmful substances for the environment should always fail the risk assessment. For the areas of concern (e.g., insects for insecticides, plants for herbicides), it is then possible to apply a less conservative and more realistic higher-tier assessment that requires additional data to refine the risk identified at Tier 1. One such example is outlined in the Aquatic Guidance Document (EFSA PPR, 2013). Tier 2 refinement options include the testing of additional species

under standard conditions to reduce uncertainty due to the unknown sensitivity distribution of species in the field. Additionally, testing under refined exposure conditions may be used to address more realistic time-variable exposure expected in the field. Finally, higher tier studies at population and community level, including (semi-)field studies and landscape level assessments, may be conducted at Tier 3 to assess direct and indirect effects, as well as the recovery of populations within their ecological context. Uncertainty in these various assessments is considered by applying assessment factors (AF) to the effect levels derived from the studies; these are defined for Tier 1 in the "Uniform Principles" (Commission Regulation (EU) No 546/2011[2]) and may be lowered at higher tiers if some of the uncertainties have been addressed.

The tiered effect assessment is mostly based on experimental approaches. However, MEMs may also be used and are explicitly mentioned as refinement tools, for example in the scheme of the tiered approach in the aquatic guidance document (EFSA PPR, 2013). In other documents, modeling is proposed as default part of the risk assessment, for example, in the Scientific Opinions on the risk assessment for soil organisms (EFSA PPR, 2017), non-target arthropods (EFSA PPR, 2015), amphibians and reptiles (EFSA PPR, 2018b), and, most recently, for bees and other pollinators (EFSA, 2023a). When MEMs are applied as refinement tools or as a default assessment, they are used in a prospective way to predict effects under non-tested environmental conditions (expected in the field), including time variable instead of constant concentrations, on higher levels of biological organization (e.g., populations) or on different species (e.g., focal species instead of surrogate test species). This way, MEMs can be used as virtual experiments to answer "what-if" questions. In all cases listed below where MEMs are used as tools in ERA, validation against independent data and benchmarking against existing risk assessment methods is needed. More details are given in Chapters 5, 6, and 7 of this book. This is a crucial area of research to ensure model acceptability and ultimately to safeguard the protectiveness of the ERA for PPP while allowing an appropriate level of realism.

In addition, MEMs can be used in a descriptive or retrospective way to describe and analyze experimental results or field observations. In that way, they lead to an improved understanding of mechanisms, processes, and traits behind effects observed in (semi-) field studies. They allow to better understand differences in observations between studies carried out under different conditions, for example the use of the individual-based model (IBM) BEEHAVE to explore winter starvation in large-scale colony feeding studies (Abi-Akar et al., 2020).

Additionally, MEMs can be used retrospectively to support the calibration of the different risk assessment tiers, for example, by adjusting the assessment factors (AF) representing the various uncertainties at different levels. For example, in the Scientific Opinion on the risk assessment for non-target arthropods spatially explicit population models are proposed for cases when an acceptable risk at the local scale has been shown (EFSA PPR, 2015). However, MEMs might be useful in many other ways, for example, for ranking of effectiveness of alternative risk mitigation measures (RMM). In the following sections, a number

2 https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32011R0546&from=EN

of potential uses of models acting at levels of organisms, populations, and communities or ecosystems are indicated, and an overview is given in Table 1.1.

Table 1.1: Overview of potential uses of models on the level of organisms, populations, and communities or ecosystems.

TKTD models	Population models	Community and ecosystem models
Analysis		
Improve understanding of mechanisms, processes, and traits behind effects observed in (semi-) field studies		
Better dose-response relationships by using observations over time	Improve ecotoxicological data analysis	Improve community-level data analysis
Explain differences between results from different studies	Quantify variation in population endpoints such as abundance; better understand drivers of natural variation observed in the field	Better understand drivers of natural variation observed in the field
Analyze interactions of multiple direct effects	Support the identification of focal species	
Prediction		
Inter – and extrapolate effects to additional observation times	Assess the protection level of established methods in risk assessment	Assess the protection level of established methods in risk assessment
Predict effects of dynamic exposure	Predict effect propagation from organisms to populations within their environmental context	Predict indirect effects
Extrapolate effects between species	Predict the potential for population recovery	Link effects on biota to effects on ecosystem services
Predict interactions of multiple direct effects	Predict the effectiveness of mitigation measures or other management options	Predict the effectiveness of mitigation measures or other management options
Simulate direct effects in population models	Predict effects in different environmental scenarios	
In vitro in vivo extrapolation, NAMs, qAOP[a]	Predict effects of dynamic exposure at the population level	
	Refine exposure based on individual behavior	

[a]NAM: New approach methodologies, see e.g., section 1.2.4, qAOP: quantitative adverse outcome pathways.

1.2.1. Organism-level models

MEMs for the effects of pesticides at the organism level include mainly TKTD models, which often model the time course of internal concentrations or the related damage from the dynamics of external concentrations or variable uptake of contaminated diet in space and/or time.

Models for chemical effects at the organism level can be summarized under the label "toxicokinetic-toxicodynamic" (TKTD) models. Most TKTD models follow either a damage-based modeling approach for lethal effects on individuals, or an energy-budget based modeling approach that can be applied to both lethal and sublethal effects on an individual organism. Various damage-based models have been developed to model survival following time-variable exposure to a pesticide. These models consider an abstract built-up of damage and damage repair. Models for survival have been united under a common framework called "General Unified Threshold models of Survival" (GUTS); (Jager et al., 2011; Jager & Ashauer, 2018). Other TKTD models simulate an energy budget for an organism that gains energy from food assimilation and spends energy for maintenance, growth, and reproduction. There are many ways of doing this (reviewed e.g., in Sibly et al. [2013]), but in ERA of PPPs most of the energy-budget models follow the principles of the dynamic energy budget (DEB) theory (Kooijman, 2009). Development of full DEB-TKTD models is complex, but simplified approaches for use in the ERA of toxicants have been developed (Jager, 2020; Sherborne et al., 2020). However, full DEB models are available nowadays for all standard laboratory species and a growing number of focal species in the so called "AddMyPet"-Database[3] including verification of the model on the calibration dataset, transparent link to the underlying data and version control. The possible use of TKTD models in the aquatic ERA of pesticides has been evaluated in a recent EFSA scientific opinion (EFSA PPR, 2018a); GUTS and a TKTD model for the duckweed *Lemna* were deemed to be fit for purpose for being used in the regulatory ERA, whereas algae and DEB-TKTD models were seen as candidates needing further development and experience. The most important potential uses of MEMs for the organism level in the ERA of pesticides in the EU are explained in the next sections.

1.2.1.1. Improve dose-response modeling for acute and chronic toxicity

Toxicokinetic-toxicodynamic models for effects of pesticides at the organism level can be advanced alternatives to classical dose-response models for the analysis of ecotoxicological data, see example in Nyman et al. (2012). Dose-response models are based on empirical relationships used to inter- and extrapolate observed effects to untested concentrations at the same observational time point, and hence same exposure duration. They have been established for the calculation of median lethal (LC50) and effective (EC50) concentrations, rates or doses (LD50, ED50) or for other effect magnitudes (e.g., EC10) from acute and chronic data in Tier 1 or 2. In contrast to empirical dose-response models, TKTD models describe the dynamics of the internal concentration of a toxicant in an organism and its relation to the effect on, e.g., survival, growth, or reproduction. TKTD models thus explain and predict effects based on a mechanistic theoretical foundation instead of a purely descriptive statistical regression. Additionally, TKTD models can be fit to all available observations over time in a test or multiple tests, and thus make use of a much higher number of data points than simple dose-response models. Usually, this gives a more robust fit and a higher statistical power to this mechanistic model approach. For that reason, TKTD models were included in the 2006 OECD document,

3 https://www.bio.vu.nl/thb/deb/deblab/add_my_pet/

"Current Approaches in the Statistical Analysis of Ecotoxicology Data: A Guidance to Application," as an alternative to the usual hypothesis testing and dose-response modeling (OECD, 2006).

If a TKTD model does not capture the time-to-effect dynamics well, the modeled mechanisms might not reflect the true toxicokinetics or the mode of action (e.g., physical mode of action of oil, leading to wrong assumptions about internal concentrations or effects), and this may point to potential issues with the data and trigger further assessment of the reliability of the test. It should be noted that this time-related uncertainty cannot be discovered with simple dose-response models.

One advantage of TKTD models is the possibility to assess whether acute and chronic effects on a given endpoint, such as mortality, are driven by the same mode of action. If a TKTD model can be fit well to both data from acute and chronic exposure duration separately, but not to a combined set of acute and chronic data, it may be concluded that the mode of action for acute and chronic effects is not the same. For example, (Gergs et al., 2021) demonstrated a shift in the mode of action for effects of imidacloprid over time, where acute mortality was driven by neurotoxicity, while chronic mortality was driven by feeding inhibition. This shift in the mode of action explains why the GUTS model cannot be calibrated jointly to the acute and chronic data for *Gammarus sp.* (Focks et al., 2018). Therefore, mechanistic models help to assess whether acute and chronic data could be analyzed together or should be considered separately due to different modes of action. Such conclusions cannot be drawn from data analyses with empirical time-to-effect models such as hyperbolic models (Sanchez-Bayo, 2009) and binomial generalized linear mixed-effects models (Becker & Liess, 2015), where the aim is simply to find the best fitting equation and parameter values irrespective of the underlying mechanisms. In summary, TKTD models can potentially perform better than dose-response models in predicting accurate and precise LCx or ECx values over time for use in Step 1 because they describe the processes mechanistically, take into account the mode of action, and use the whole dataset rather than just the data at a particular time point. A corresponding systematical evaluation for future use of TKTD models would appear beneficial.

1.2.1.2. Understand the difference between endpoints in different studies

TKTD models can be used to systematically compare and bridge between studies performed under various conditions (such as food, temperature, species), and by doing so, to understand differences within a larger dataset instead of doing study-by-study analyses. This might include data for one species from different assays (Baas et al., 2022; Gergs et al., 2021), or even data for different species (Gergs et al., 2015; Gergs et al., 2019). With this approach, one could move beyond the endpoint-to-endpoint evaluation toward a more holistic view considering more comprehensive datasets. In addition, TKTD models can be used to explain why certain effects seen in standard OECD biotests have been observed to increase or decrease under more environmentally relevant conditions (Zimmer et al., 2018). For example, a TKTD model could be fit to a full dataset that covers toxicity tests performed at different temperatures, and effects could be interpolated between the tested temperatures (Rakel et al., 2022). If the model reproduces the pesticide effects at each

temperature level in a satisfactory way, it could be concluded that the mechanisms leading to different responses at different temperatures have been well captured (good structural integrity). This might be considered also as a demonstration of certain realism in the model before it is used to make prospective predictions (see section 1.2.1.3), despite an ultimate validation would require testing of the temperature extrapolation with independent data. In any case, according to (EFSA PPR, 2018a), testing with an independent dataset is required before TKTD models might be used for prospective predictions.

1.2.1.3. Predict effects of dynamic exposure

Increasingly, TKTD models are used in a prospective way, such as to predict the effects of various dynamic (i.e., time variable) exposure patterns. While exposure in standard ecotoxicological tests at Tier 1 is usually kept constant as far as practically achievable, predicted exposure for field situations can be highly diverse and dynamic. In aquatic ERA, such exposure profiles in streams, ditches, and ponds can be predicted by the FOCUSsw step 3 or 4 models for different application patterns, soil and weather conditions (FOCUS, 2001). In Tier 1, the maximum predicted environmental concentration (PECmax) that occurs for any time period in such a profile is compared to Regulatory Acceptable Concentration (RAC) derived from the median lethal (or effective) concentrations such as LC50, EC50, EC10 or from the NOECs from standard tests by the use of assessment factors.

Because a refinement in Tier 2C of the aquatic risk assessment, TKTD models can be used to predict the direct effects on an organism that result from considered dynamic exposure profiles (EFSA PPR, 2013, 2018a). The model is repeatedly run with increasing exposure multiplication factors applied to the exposure profile (i.e., to all concentrations that occur during the exposure profile, possibly restricted to a relevant time window of application; [EFSA PPR, 2018a]). This way, the exposure multiplication factor can be derived that is predicted to result in 50% effect (often at the end of the profile; the length [duration] of the profile should be chosen to match the relevant risk assessment question). This factor is called LP50 (lethal profile) for lethal effects and EP50 or EP10 (effect profile) for sublethal effects such as decrease in growth rate, body size, or offspring number. Thus, the LP50 or EPx represents a margin of safety and is equivalent to an assessment factor (Ashauer et al., 2013). If the LP50 or EPx is larger than the Tier 1 assessment factor, the risk is considered acceptable.

Today, TKTD models have mainly been applied to assess effects based on refined exposure in aquatic risk assessment and proposed as Tier 2C method in parallel with similar experimental methods. Still, there are a number of limitations to the application of a Tier 2C approach, for example, it should address only

the effect threshold and not ecological recovery[4]. Apart from being used in Tier 2, TKTD models could also be considered as a Tier 1 approach under constant exposure conditions in the aquatic and terrestrial risk assessment to better describe effects on invertebrates and vertebrates. It should be noted, however, that exposure in terrestrial habitats is generally considered as more heterogenous and affected by the movement of the individuals, which makes the modeling of exposure more challenging than in aquatic habitats. In addition, more than one exposure route might be relevant. In a recent example for terrestrial habitats, different exposure routes were assessed with a TKTD model for honey bees. The BeeGUTS model was calibrated on contact and oral tests together, hence describing the full underlying dataset with one single TKTD model (Baas et al., 2022). For soil organisms, a discussion on the relevance of multiple exposure routes is found in Forbes et al. (2021).

Before effect predictions of a TKTD model can be considered reliable in the ERA of pesticides, the predictive power of the model needs to be tested. This requires a comparison of predicted and observed effects with independent data not used for model calibration (model validation). For aquatic organisms, the EFSA scientific opinion on TKTD models recommends testing profiles with two short-term exposure events with two different intervals in-between and three exposure levels each (EFSA PPR, 2018a). However, it is not yet clear whether this also applies for vertebrates, or whether for animal welfare reasons, requirements could be adapted for fish and amphibians. Guidance on the use of TKTD models is also needed for terrestrial organisms (see e.g., Trijau et al. [2023]).

1.2.1.4. Extrapolate effects between species

TKTD models with a physiologically-based toxicokinetic (PBTK) component (Baier et al., 2022; Cooper et al., 2019; Mavroudis et al., 2018) do not consider the organism as a single unit, but differentiate, for example, between blood and various organs. Thus, a species is characterized independently of a toxicant and its specific uptake and excretion rates by physiological parameters such as gill surface area, blood volume, etc., which are used to describe the toxicokinetic processes more precisely. Such models can potentially help to extrapolate the effects observed in one species to the effects expected in another species as well as to compare, for example, effects observed on individuals of different sizes. This extrapolation assumes that species vary in their sensitivity to a specific pesticide mainly due to differences in physiological traits for the absorption, distribution, metabolism, and excretion of pesticides (ADME, described by the TK part of the model). Examples for the extrapolation of toxicokinetics between species are known for fish (Brinkmann et

4 If used without coupling to a population model, TKTD models still simulate laboratory tests and their outputs are the endpoints addressed in the corresponding Tier 1 tests. For animals, these are organism level endpoints like survival, reproduction, development or growth. Under dynamic exposure, individuals can (partly) recover from sub-lethal stress (damage repair). In contrast, algae and the standard test macrophyte Lemna sp. are tested as populations and the endpoint used at Tier 1 is the inhibition of the population growth rate. Thus, dynamic exposure allows recovery of the population (ecological recovery). However, according to the EU Aquatic Guidance Document (EFSA, 2013), ecological recovery can only be assessed in a community context where species interactions can affect the recovery of a population (i.e. Ecological Recovery Option in Tier 3, micro – or mesocosm studies).

al., 2016) difficulties in extrapolating across freshwater invertebrate species have been described in Ashauer et al. (2016).

The sensitivity to internal exposure at the target site in the organism (described by the TD part) is considered much less variable among species than the empirical effect endpoints such as the EC/LC50 related to concentration in the test medium (Gergs et al., 2015; McElroy et al., 2011). Simple TKTD models, such as reduced GUTS, using scaled internal damage as a dose metrics could be used to extrapolate effects between different species (Gergs et al., 2015; Gergs et al., 2019). This kind of interspecies relationship can be used to assess whether a pesticide acts via a common mechanism of action on different species, as recently shown by fitting a reduced GUTS model jointly to a set of species (Singer et al., 2023). The joint GUTS model predicted effects on the different species even more precisely than the standard single-species GUTS and the authors suggest that this joint GUTS modeling can improve cross-species extrapolation.

There are several promising approaches available that can reduce animal testing (Escher et al., 2022) and also facilitate understanding of how pesticides affect different organisms using either *in vitro* data (proteins, homogenates, cells, organs) or using embryonic approaches. *In-vitro* to *in-vivo* extrapolation by use of simple or complex mechanistic models has been demonstrated for toxicokinetics (Krause & Goss, 2020; OECD, 2018) as well as for effects (Brinkmann et al., 2014; Scholze et al., 2020; Stadnicka-Michalak et al., 2015).

Ultimately, using MEMs for extrapolation of pesticide effects to untested species is a field of ongoing research and is currently not considered fit for regulatory ERA (EFSA PPR, 2018a). Thus, this would need to be scientifically scrutinized and tested for its performance in a regulatory ERA before its potential use (EFSA PPR, 2018a) while acknowledging clear advantages in animal welfare. Especially the combination of New Approach Methodologies (NAM) with MEMs to predict the effects on classical population level relevant endpoints like survival, growth, development, and reproduction could offer support for 3R strategies to improve animal welfare, while there is still much work to do.

1.2.1.5. Predict effects of multiple dynamic stressors

Some TKTD models can be used to predict how effects of multiple stressors interact in an organism. Simple static effect models for this task are limited to simultaneously acting stressors, such as the Concentration Addition and the Response Addition (= Independent Action) Model for pesticides with similar and with different modes of action (Hadrup et al., 2013). The Stress Addition Model works for a pesticide and other environmental stressors (Liess et al., 2016). In contrast, dynamic TKTD models such as GUTS have been adapted to fit dynamic effects of multiple chemicals (Bart et al., 2021) and may even handle effects of stressors with different exposure profiles over time (Goussen et al., 2020). Complex energy budget models such as those from the DEB framework (Kooijman, 2009) may be used to refine an estimated mixture toxicity in the aquatic risk assessment when active substances of a combined product enter a waterbody with different delay after application (Bart et al., 2021; Vlaeminck et al.). Additionally, a TKTD model that has been fit

to data from toxicity tests with different levels of a second stressor such as temperature or food limitation may be used to extrapolate combined effects from the pesticide and additional stressor at doses or levels that have not been tested before. However, TKTD models for effect interactions may require further data from a set of non-standard tests for parameterization and model validation.

In any case, the application of TKTD models for the assessment of effect interactions is a field of ongoing research and need to be scientifically scrutinized and tested for its performance in a regulatory ERA before its potential use; examples for testing multiple chemical stressor modeling by using TKTD models is given by Bart et al. (2021; 2022).

1.2.1.6. Simulate direct effects in population models

TKTD models can be coupled with or incorporated as building blocks for direct effects of pesticides at the organism level into higher-level MEMs for populations, communities, and ecosystems. For example, TKTD models can be used as a module in individual-based population models to propagate organism-level effects to the population level (Accolla, 2022; Gabsi et al., 2014; Galic et al., 2014; Martin et al., 2013; Vlaeminck et al., 2022).The higher-level models do not need to be individual-based: Higher-level models that aggregate some or all individuals of a population can convert the probability of death predicted by a TKTD model to a decline in abundance or biomass of a (sub)population (Jager & Zimmer, 2012). Similarly, the predicted loss in biomass of a model organism is considered as average across all individuals and can be converted to the loss in biomass of the whole population.

1.2.2. Population-level models

Population models comprise various forms that consider the population as one unit such as unstructured or scalar models based on differential equations (Barnthouse, 2004), or time-discrete equations (Calow et al., 1997), they can be composed of age or stage classes (structured or matrix models used in matrix algebra, for example Chandler et al. [2004]), or they can model the basic units of populations, individual organisms, in so-called individual- or agent-based models (IBMs, ABMs, overview in Railsback & Grimm [2019]). Individual-based models can be based on a set of coupled equations and/or rules (algorithms) describing individual life cycles and behavior. As a consequence, the population dynamics emerge from the developments of its individuals. Additionally, population models can be spatially explicit to address variability in exposure and movements of organisms in a heterogeneous landscape, e.g., Animal, Landscape and Man Simulation System (ALMaSS; Topping et al. [2016]); Metapopulation model for the Assessment of Spatial and Temporal Effects of Pesticides (MASTEP; Focks et al. [2014]; Van den Brink et al. [2007]); eVole (Kleinmann & Wang, 2017; Wang, 2013). Spatially explicit IBMs are often based on a set of behavioral rules rather than on coupled equations. There are different types of unstructured population models, in which a single organism is modeled as being representative for the population as a whole (Kooijman et al., 1996);

structured models where separate organisms are modeled that represent different ages or life stages (Billoir et al., 2007); or individual-based models, where each individual is modeled separately. Population models may include a TKTD submodel (module) to simulate physiological, toxicokinetic and/or toxicodynamic processes in organisms (e. g., in the IBMs of Galic et al. [2014]; Gergs, Gabsi, et al. [2016;] Johnston et al. [2014]; Martin et al. [2013]). The use of population models for regulatory risk assessment of chemicals has been reviewed and exemplified, amongst others by Dohmen et al. (2016) and Forbes et al. (2016).

Population models usually address the development of populations under realistic biotic and abiotic conditions in the field, or under semi-realistic conditions in a test facility. Biotic and abiotic factors and processes may be explicitly simulated or implicitly considered in the model parameterization (see Chapter 3). Typically, population models consider ecological realism by adding processes such as feeding, growth, reproduction, dispersal, and population-level processes such as intraspecific competition. However, it should be noted that population-level models, unlike (semi-)field studies, do not cover indirect effects mediated via trophic interactions (and thus are most relevant in support of such experimental Tier 3). To address this, a population model for one species may be extended to include one or more additional species, for example to consider their competitive or predatory effects on the model species (for an extended version of IDamP see Gabsi & Preuss [2014]). Such extended population models still focus on the effects on a single model species but improve the ecological realism required in ERA by considering that the species is embedded in a community context. More extensive consideration of species interactions, such as in food-chains, can be achieved with community and ecosystem-level models, which we consider here as MEMs for whole food webs that focus on effect propagation to indirectly affected species and to community composition (see section 1.2.3). Hybrid models are community or ecosystem models in which one or a few species of high interest are modeled in more detail (e.g., by an IBM) than the other taxa (usually modeled by differential equations). Examples are provided in Schmolke et al. (2019) and Strauss et al. (2017). The most important uses of population-level MEMs in the ERA of pesticides are explained in the following paragraphs.

1.2.2.1. Improve ecotoxicological data analysis

Population models can be used in a retrospective way to better understand and explain the results of higher tier studies such as mesocosm or field tests. For instance, predicted effects from a population model set up with standard toxicity data and run with different parameter settings or environmental scenarios (see Chapter 3) can be compared with observations from (semi-)field studies to better understand the mechanisms and environmental factors that drive effect propagation from individuals to populations. This way, models could be, for example, used to assess whether specific conditions in a mesocosm study such as climate/weather conditions represented rather a worst or best case. Concrete examples have not yet been published in the open scientific literature.

1.2.2.2. Assess the protection level of established methods in risk assessment

Population models can be used retrospectively to assess whether the (often pre-established) risk calculations and assessment factors (AFs) in the ERA of pesticides meet the required levels of protection, that is to say, to calibrate risk assessment tiers. If population modeling suggests that an exposure level that is considered acceptable in lower tiers will lead to unacceptable effects on the population level, this will be a warning that the AFs in lower tiers needs to be revisited.

1.2.2.3. Predict effect propagation from organisms to populations

Population models can be used to predict whether and to what extent direct (lethal and sublethal) effects on organisms observed in standard tests will propagate to effects on the overall abundance or biomass of a population or on the population growth rate. Thus, if effects at organism level cannot be excluded in lower tiers, population models can be used in higher tiers to assess the resilience of a population at risk. This refers to the capacity of a population to remain in its original state, for example, by compensating low levels of pesticide-induced mortality over an extended period of time through migration and/or an excess in reproduction. Such modeling studies address the Ecological Threshold Option (ETO), as explicitly laid out for aquatic organisms in EFSA (2013). The ETO accepts only negligible effects on the abundance or biomass of a population. It is the only refinement option in higher tiers available for vertebrates, for which the protection goal is "no visible mortality" and only negligible effects on populations (EFSA, 2009, 2013), and for other ecologically vulnerable species that are characterized by a low potential for population recovery (EFSA, 2013; EFSA PPR, 2010). Consequently, for vertebrates, assessing effect propagation with population models is often done when dealing with sublethal effects, e.g., on small mammals (Dalkvist et al., 2009; Liu et al., 2014; Liu et al., 2013; Topping et al., 2009; Wang, 2013) or skylarks (Topping & Odderskaer, 2004).

A reoccurring issue in this context is the challenge which comes with the definition of the significance of a population level effect that could be defined based on population modeling. According to the No Observable Effect Concentration or Level (NOEC/NOEL) in experimental studies, from a population modeling study for example, the highest concentration or dose that will result in negligible effects, at population level could be derived. However, a clear definition of what is a non-negligible effect has not been established yet. In that context, population models may provide useful information, also for supporting the setting of specific protection goals, as population models that have been designed to simulate the variability at population level can be used for this purpose. Variability of an endpoint has an impact on the (statistical) detection of non-negligible effects. Whether or not population models can simulate natural variability is depending on the objectives of model development, and not all models are designed to replicate that. Further work on this issue appears very relevant and useful.

In addition, modeling studies could also work with exposure modification factors to determine the margin of safety of an intended use according to good agricultural practice (GAP) before an unacceptable effect is expected to occur (Ashauer et al., 2013). This concept has for example been utilised for the derivation of the LPx/EPx values from TKTD models (EFSA PPR, 2018a).

In any possible use, as laid out above, population model predictions and related analyses can only be considered relevant in the regulatory ERA of pesticides, if validation against independent data and benchmarking against existing risk assessment methods is conducted. More details are given in Chapters 5, 6, and 7 of this book.

1.2.2.4. Predict the potential for population recovery

Population models can potentially be used for the assessment of population recovery after the temporally restricted impact of a pesticide and to assess the resilience of a population at risk, that is to say, its capacity to return back to its original state after disturbance by pesticide effects, either by population growth or recolonization. This is actually possible under the Ecological Recovery Option (ERO) in aquatic risk assessment (EFSA PPR, 2013) and for soil organisms and non-target arthropod in the treated field. It should be noted here, that models that can be used for simulations of population recovery need a community context. Without community context, for example, a less sensitive competitor, a population model can only demonstrate the potential of recovery. This was already acknowledged in the aquatic guidance document: "… it may be an option to also use population models (not described in this GD) to better address recovery potential of vulnerable invertebrates. These models, however, should consider how recovery is affected by possible interference with other populations (species interactions such as predation and competition)." (p. 55 in EFSA PPR, 2013).

In principle, temporary effects can be accepted for those non-target invertebrates and plants that are not considered highly vulnerable due to ecological traits such as long generation times, low reproductive output, or low re-colonization potential (EFSA PPR, 2010), as long as it can be demonstrated that the decline in abundance and/or biomass of local populations will be transient due to population recovery through growth, reproduction or recolonization.

However, the ERO of non-target organisms off-field is often not accepted in the context of regulatory decisions due to remaining uncertainty about the representativeness of the species tested for the most sensitive species in the field, the impact of ecological interactions and multiple stressors on recovery processes. Suitable MEMs could be used in future research to explore if and under which conditions recovery in an acceptable timeframe is possible. The relevance of ecological interactions such as competition or predation and multiple stressors have to be considered on a case-by-case basis.

1.2.2.5. Predict effects for different environmental scenarios

By running population models with different settings, effect propagation and population recovery could be extrapolated to various environmental scenarios, for example, climatic conditions, landscape structure or exposure. For example, IBMs have been used to study how population recovery of freshwater macroinvertebrates may change in the context of interacting species (Gabsi & Preuss, 2014; Kattwinkel & Liess, 2014) and in a spatial context (Van den Brink et al., 2007). When used in regulatory risk assessment, environmental scenarios should be realistic in terms of biotic and abiotic stressors (see Chapter 3). In this context, population models could be used to predict effects or calculate margin of safety for a specific intended use under various environmental scenarios. This could be very useful and provide relevant information for regulatory decisions, especially when ecologically less favorable scenarios are included in such analyses (for more details see section 3.3.3). In addition, population models could support (semi-)field studies by extrapolating observed effects to other, non-tested scenarios. Models for simulating population dynamics under the impact of pesticides and different scenarios should be critically assessed before model predictions are being considered relevant for informing the ERA.

When a model has been calibrated with organism-level effect sizes from the laboratory and population-level effect predictions have been successfully validated under (semi-)field conditions, this is a good indicator that the model captures the most relevant mechanisms that affect the predicted effect sizes and may do so also under different environmental conditions. However, data for validation of population dynamics under field conditions are not always available and experiments may not be possible with some taxa due to practical or animal welfare reasons. Pattern oriented modeling might be helpful in such situations (see section 6.2).

A population model can also be calibrated to population-level effects observed under a specific environmental scenario in a (semi-)field study (e.g. in mesocosms). It should be noted that in such a case, the parameters of the population model can also capture factors that result, for example, from ecological interactions. For such a model that has been calibrated using (semi-)field observations, the appropriate validation step would be to compare the predicted and observed effects under environmental conditions that were not used for calibration. This would provide confidence that the model can predict appropriate effect sizes under additional conditions. A population model validated in this way could then be used to extrapolate these effects to different scenarios. Again, it is possible that such ideal population dynamics datasets are not available for validation, in which case other approaches should be used (see e.g., section 6.2).

1.2.2.6. Predict effects of dynamic exposure at the population level

Population models can be used also to extrapolate effects observed in a mesocosm or other semi-field study performed under a certain exposure profile to different exposure profiles. Such extrapolations would in practice most likely need additional data from lab testing and the incorporation of a TKTD model for the direct effects at organism level (see section 1.2.2.3). The overall analysis of the combined experimental and modeling work could then address higher tier assessments, for example Tier 3 as described in the aquatic guidance document (EFSA PPR, 2013).

1.2.2.7. Refine exposure estimation by simulating individual behavior in a population model

Spatially explicit population models can be used to refine the exposure of individuals and, in consequence, the average exposure of a population. Such models refine the spatially implicit calculation of the overall exposure of a habitat, as usually resulting from common fate models. Instead, exposure in a spatially heterogeneous landscape is simulated, often operating on realistic landscape data or even GIS maps. The exposure of individuals is not driven by application and environmental fate of a pesticide alone; movement, dispersal and the feeding behavior of individuals (Kleinmann & Wang, 2017; Van den Brink et al., 2007; Wang & Grimm, 2007) or groups of individuals (Becher et al., 2014) in the modeled landscape also play a crucial role. Tin that way, the implemented behavioral rules and physiological and environmental settings translate into proportions of ingested contaminated food or the time spent in treated areas. With such information, the "proportion of an animal's daily diet that is obtained in habitat treated with pesticide" (PT) values as usually obtained from field studies might be refined, as it has been suggested by Liu et al. (2013) based on their wood mouse model. The PT parameter is used to estimate the daily dietary dose (DDD) as required for TER calculations in the higher-tier risk assessment for birds and mammals (EFSA 2009). Spatially explicit MEMs can simulate the movement of individuals and so include such an exposure refinement into the prediction of effects at population level. For soil risk assessment, the spatial heterogeneity in the soil column is already important at a given site. Behavior of soil organisms like earthworms can lead to different exposures for different species resulting in different effects, which can be modeled if the distribution of the pesticide as well as the behavior of the earthworms are known (Forbes et al., 2021; Roeben et al., 2020).

1.2.2.8. Assess the effectiveness of mitigation measures or other management options

Population models may be used to predict the impact of different risk management options on the risk of pesticides for a given species. This may be part of a general effort to evaluate the effectiveness of risk management options, or to identify the most efficient risk management option for a specific use-case. Effects of risk mitigation on exposure reduction, for example, a reduced application rate, drift reduction (Focks et al., 2014), buffer strips (Schmolke, Bartell, et al., 2021), or avoiding application during particularly sensitive times of the year, can be assessed by all types of effect models. In addition, spatially explicit population models allow us to assess other management options, like changes in the landscape structure (e.g., field size, diversity of structural elements etc.).

1.2.2.9. Support the identification of focal species

Pesticide effects on different life history traits could impact different species at different strengths. For example, r-strategists may be more vulnerable to decreased reproduction, whereas long-living species may be more vulnerable to decreased competitive strength or decreased escape from predators. Thus, suitable candidates for focal species in risk assessment may be identified with the help of population modeling. For this task, matrix models may be easier to parameterize for a number of different species than more detailed IBMs. Such matrix models propagate effects on the survival and fecundity of different age classes or stages to effects on the instantaneous population growth rate. Age-class models for various fish species expected to occur in edge-of-field water bodies in the EU were used to rank the species according to their vulnerability to effects on fecundity, juvenile and adult survival (Ibrahim, 2014). Rueda-Cediel and colleagues used matrix models to compare demographic traits and elasticity metrics to explore how similar these are between listed and non-listed terrestrial plants species in the US (Rueda-Cediel et al., 2019). It should be considered that a ranking of species according to their vulnerability may be affected by the environmental scenario, for example by the timing of pesticide application in relation to the timing of the main reproduction season of a species, which can vary across countries and climate zones. Thus, the use of population models for the preliminary identification of focal species should be further supported with empirical information on potential candidates for the most vulnerable species.

1.2.2.10. Link effects on biota to effects on ecosystem services

Population models could be used to link chemical or other stress to ecosystem services. This requires careful considerations of the specific service-providing units, for example, if single species populations are the service-providing units (such as specific fish for fisheries or sports fishing), and/or if there is a good understanding of the logic chains (Hayes et al., 2018), leading from exposure via simulated population effects to possible impacts on a specific ecosystem service. Forbes and colleagues applied the concept to the case of endocrine-disrupting effect on trouts and the ecosystem services they provide (Forbes et al., 2019). Maltby and colleagues provide some conceptual ideas of how such linking of population and ecosystem models to ecosystem services could be done, illustrated with examples on pest control, pollination, and soil fertility (Faber et al., 2021; Maltby et al., 2021; Van den Brink et al., 2021). However, more basic research on how various species contribute to ecosystem services is required before such use of population models could appear fit for purpose in the regulatory risk assessment.

1.2.2.11. Assess natural variation in abundance

To better formulate specific protection goals, it is crucial to have an idea about natural variability in population abundance or biomass, because it is a pre-condition for the definition of negligible, small, medium, and large effects. However, studies on (semi-) field populations can provide observation data only for specific, local environmental conditions and for a limited number of time points. Population

models can be used to assess the natural variation in population sizes by running simulations for a set of different control (non-contaminated) scenarios, if they were designed for such specific application. In this way, (semi-) field observations can be extrapolated to different environmental conditions and to longer observation times. Extended observation times may reveal seasonal or multi-year population cycles that might have not been monitored or detected in (semi-) field studies.

In general, for the assessment of natural variation in abundance, ecotoxicological models may not need to comply with all criteria for use in risk assessment (EFSA, 2020), specifically because exposure and effect modules (see Chapter 5) do not need to be implemented or used, respectively. However, because all models represent simplifications of reality and many models are designed to simulate a typical, or an average population, a comparison between the variability observed in a field and in a model should accompany the use of population models for quantifying variability of abundance or other characteristics to check whether such an analysis is within the model's domain of applicability.

For example, the BEEHAVE model (Becher et al., 2014) was recently used to identify negligible effects at the colony level based on the variability between colonies in the model (EFSA, 2021) to quantify the specific protection goal for honey bees. Colony dynamics in different regions across Europe were simulated and the normal operating ranges were defined as different percentiles of all simulated colony dynamics. The modeled variability of colony strength was used to determine how many field experiments would be needed to detect possible pesticide impacts for a certain magnitude of effects within the biological variability in a statistically reliable way (EFSA, 2020).

This application example shows that it is essential to select carefully the parameter variability, if the model aims to support the definition of the specific protection goals (SPGs). Nowadays, most models are developed to capture the mean response and not necessarily to include accurately the variability. In the case of honey bees within a field study, different types of variabilities are present and can be selected for this option. The variability of different developments of a single hive is considered as in-hive variability. This variability is currently implemented in BEEHAVE and explains the different colony trajectories that a single average hive could manifest with its queen having a specific egg-laying rate. This variability was reported to fail to explain the variability observed within semi-field or field studies (Schmolke et al., 2020). In contrast, the variability in field studies could be simulated with the BEEHAVE model if the different initial conditions of the hives are considered (Agatz et al., 2019).

The calculations for supporting the definition of the specific protection goal for honey bees (EFSA, 2021) were only including the variability of, for example, the abundances, so it would have been mandatory to transparently lay out which variability were included in the simulations and to consider the implications for the protection goal definition. By choosing reduced variability compared with field observations, the chosen approach led to conservative results.

1.2.3. Community and ecosystem level models

Models for communities and ecosystems are often based on a set of coupled ordinary differential equations and typically simulate the mass-balanced cycling of nutrients, contaminants, and biomass within a food web, e.g., AQUATOX, (Park & Clough, 2018); Streambugs (Kattwinkel et al., 2016); CASM (Bartell et al., 2020). Community models typically focus on the interaction between populations, without considering mass balance or nutrient cycling, for example in simple predator-prey models (Reeg et al., 2018; Reeg et al., 2017). Moreover, community and ecosystem models can also include other model types, ranging from few coupled discrete difference equations (Waage et al., 1985) to spatially explicit IBMs, such as Eco-SpaCE (Loos et al., 2010) and StoLaM (Strauss et al., 2017). Such models are often called hybrid models (Schmolke, Bartell, et al., 2021).

Many MEMs require detailed information on the spatio-temporal distribution of pesticides across various landscape segments or ecosystem compartments to calculate the actual exposure of modeled organisms. This information is not always available from established fate and exposure models. Therefore, some MEMs are technically implemented together with an additional specific fate module in a common software package. This fate model simulates the fate of pesticides within a modeled landscape or ecosystem and can be linked to an external fate model for the pesticide loading to that environment. In this book, we consider MEMs as the effect part only in such a combined mechanistic fate and effect modeling framework. Therefore, though the design and evaluation of associated fate models and of their connection with effect models is of high relevance, this is beyond the scope of this document and may be considered in future activities. Nevertheless, this book gives some recommendation on the evaluation of modular models in Chapter 5.

In contrast to population models, community and ecosystem MEMs address effects occurring on more than one species by interaction between the species (e.g., indirect effects). Community models focus on these interactions, e.g., via competition or predation. A clear differentiation to ecosystem models is not always possible. In general, ecosystem models also consider the effects of and interactions with abiotic factors and put more emphasis on turnover of food, elements (C, N, etc.), or biomass. Typically, in ecosystem models, biota do not respond to the modeled abiotic environment alone but can also affect the environment, for example, in terms of nutrients concentrations or oxygen levels. For example, in the ecosystem model AQUATOX, a pesticide-induced predatory fish kill may propagate through the overall food chain to decreased water quality due to lowered consumption of algae by zooplankton, changes in pH, nutrient release, and a secondary release of pesticide residues from decayed fish (Park & Clough, 2018). Nevertheless, in community models, consumption of abiotic food sources can also be considered.

So far, community and ecosystem models have rarely been applied in the context of the regulatory ERA of pesticides in Europe. This is related to the fact that while the general protection goal, as defined in Regulation (EC) 1107/2009, ultimately aims at protecting biodiversity and ecosystems, the specific protection goals have been established at the organism (lethal effects on vertebrates) and the population levels. Additionally, due to their complexity, community and ecosystem models are difficult to validate, which is

considered a prerequisite for prospective model use (EFSA PPR, 2014). Therefore, these models have been mainly used to retrospectively analyze observed effects of pesticides in the field (Park et al., 2008) or to assess effects of different restoration measures. Some examples for prospective assessments can be found in (Bartell et al., 2013; Bartell et al., 2018; Bartell et al., 2019; Bartell, 2000) using CASM (Bartell et al., 2020). The majority of community and ecosystem models available today are representing aquatic systems, the most prominent ones are probably AQUATOX (Park et al. 2008) and CASM (Bartell et al. 2020). IBM-grass (Reeg et al., 2018) modeling terrestrial plant communities and Eco-SpaCE modeling vertebrate communities (Loos et al., 2010) are a few exceptions.

In the following sections, a number of possible uses of community and ecosystem models for regulatory risk assessment are briefly mentioned.

1.2.3.1. Improve retrospective community-level data analysis

Community and ecosystem models can be used in a retrospective way to analyze and understand effects observed in complex (mesocosm) experiments or in the field, e.g., applications of AQUATOX to PCBs in lakes (Zhang et al., 2013) or of another aquatic model to analyze effects of insecticides and nutrients in aquatic microcosms (Rashleigh et al., 2009; Traas et al., 2004).

1.2.3.2. Assess the protection level of established methods in risk assessment

In the European framework for regulatory risk assessment of plant protection products, most SPGs have mostly been proposed at organism and population or metapopulation levels, for example, for the combination of the organism group "algae" and the ecosystem service (ES) "genetic resource," the ecological entity is the "metapopulation." But for some ES and organism groups, the entity "community" and/ or "functional group" can be also of relevance, for example, for algae and the ES "primary production, photosynthesis, nutrient cycling, water purification," the ecological entity is the "functional groups and communities" (EFSA PPR, 2010). It is assumed that by protecting the lower levels of biological organization (individuals and populations) and the structural endpoints, no unacceptable effects will occur on higher levels of organization (community structure and biodiversity) and the functional aspects. Community and ecosystem models may be used to assess whether and under which conditions this assumption holds.

For example, such models could be used to analyze data from laboratory studies and from field monitoring, to assess the impact of pesticides on an overall observed decline in species abundance and diversity, relative to confounding factors such as additional farming practices or land use. Additionally, community and ecosystem models can be used to mechanistically support the quantification of AFs that should be applied to single-species studies in the laboratory to cover propagation of direct effects and potential indirect effects in the field (De Laender et al., 2008).

1.2.3.3. Predict indirect effects

Community and ecosystem models can be used to predict how direct effects on one or more sensitive species may propagate to indirect effects on non-sensitive species via the food web. For example, the aquatic ecosystem model CASM was linked to an IBM of an endangered fish to assess the indirect effects of the herbicide atrazine on the fish population (Bartell et al., 2019). Similarly, to individual – and population-level MEMs, using community and ecosystem models for predictions in regulatory risk assessment requires thorough and appropriate model assessment and validation.

Only negligible direct and indirect effects should occur under the Ecological Threshold Option (EFSA PPR, 2013). Models could be used to explore whether direct effects considered to be negligible could lead to unacceptable indirect effects. The risk of unacceptable indirect effects is especially relevant under the Ecological Recovery Option, where pronounced but temporary effects on populations are acceptable, for example in-field effects on non-target arthropods.

1.2.3.4. Link effects on biota to effects on ecosystem services

Ecosystem models can be used to predict how pesticide effects on one or more populations may affect critical ecosystem services such as self-purification of a stream or good water quality as a pre-requisition for fish farming (Forbes & Galic, 2016; Galic et al., 2012; Park & Clough, 2018). In an exemplary case study, the AQUATOX model was used to predict the effects of a hypothetical insecticide on different ecosystem services provided by a lake (Galic et al., 2019).

1.2.3.5. Assess the effectiveness of mitigation measures or other management options

Like population models (see section 1.2.2.8), community or ecosystem models may be used to predict the impact of risk management options on affected populations, communities, or ecosystem services. For example, the CASM model was used to analyze the effects of a vegetative filter strip with the aim of mitigating the risks to an endangered fish species resulting from the use of atrazine on agricultural land (Schmolke, Galic, et al., 2021).

1.2.3.6. Assess natural variation in a community structure

Like population models, community and ecosystem models may be used to assess the variation in community structure that can be expected in different field situations. This information might be useful to establish specific protection goals.

1.2.4. Possibilities of using MEMs as tools in future ERA

The majority of the above potential applications of MEMs in regulatory ERA relates to the regulatory system of plant protection products as it is presently established in the EU, i.e., a tiered approach for assessments under the "single product – single use" paradigm. However, risk assessment legislation and guidance documents can undergo changes, and in future risk assessments MEMs might also become useful for other applications. While regulation EC 1107/2009 is not likely to change in the near future, EFSA guidance documents and national evaluation procedures are under continuous development. The aquatic guidance document (EFSA PPR, 2013) laid out the potential use of TKTD, as well as population and community modeling, which led to the EFSA scientific opinion on TKTD modeling for aquatic organisms (EFSA PPR, 2018a). The recently published guidance document for birds and mammals (EFSA, 2023b) also outlines the potential use of TKTD and population modeling, but without giving clear guidance how to implement it. Within the bee guidance document (EFSA, 2023a), the use of TKTD modeling is part of the lower-tier assessment and the use of colony and population models is roughly sketched. Other guidance documents are expected to be revised (or created) in the coming years, including guidance for terrestrial organisms (soil organisms and non-target arthropods and non-target terrestrial plants). In the regulatory context, the use of population models is already being proposed for applications other than just refinement. For example, population modeling at the screening level is proposed for soil organisms (EFSA PPR, 2017). Another example is non-target arthropods: A landscape-level assessment based on population modeling is proposed as an additional check if the local assessment indicates an acceptable risk; thus, an acceptable risk can only be established if it is indicated in both the local and the landscape assessment (EFSA PPR, 2015).

Next to development and improvement of guidance documents, more basic developments could change the ERA system in the coming years. Currently, EFSA and the EU are developing roadmaps for future risk assessment for a series of topics, among them new approach methodologies (NAMs; Escher et al., 2022) and systems-based risk assessment (Sousa et al., 2022). In the following bullets, we list a number of potential future uses of MEMs to support regulatory decisions and ERA.

- **Integrate multiple exposure pathways for risk assessment of groups of organisms such as pollinators, NTAs, birds, and mammals by means of organism level or TKTD models.** Currently, ERA for these organism groups is based on laboratory experiments that test explicitly for specific uptake routes, for example, oral uptake for bees, overspray or contact testing for bees, and NTA. By using organism-level modeling approaches such as TKTD models, test results that account for single-pathway exposure can be combined, and scenarios that consider the impact of exposure via multiple pathways on the estimated risk could be evaluated, and hence be more relevant and realistic for ERA as compared with the current approach.

- **Consider exposure and related potential effects in multiple environmental compartments by means of organism and population models.** Currently, ERA is strictly separated by compartments (aquatic, soil, groundwater, terrestrial) and organism groups (e.g., pollinators, non-target terrestrial arthropods, birds and mammals, non-target terrestrial plants, aquatic organisms, soil organisms). However, many species have life stages in more than one compartment, for example, with larval development in soil and terrestrial life stages after emergence or metamorphosis (e.g., many beetles). Amphibians have aquatic developmental stages and mostly terrestrial life stage as adult. Also, many insects have aquatic larval stages and terrestrial life as adults. For all these organism groups, MEMs, in combination with appropriate fate and exposure models, could provide a more relevant and realistic ERA that accounts for life-long exposure and possible impact. Such approach cannot be realized via experimental approaches alone.
- **Consider landscape context of pesticide application for regulatory ERA by means of population models.** The current ERA is based on generic scenarios that try to capture reality by considering static average or "realistic worst-case" conditions. Instead, landscape-level ERA can consider realistic landscape structures with crop fields, non-crop landscape elements, regional farming practices, and combine it with IBMs of individuals moving within such landscapes (Topping et al., 2020; Topping et al., 2009). Among other elements, they could consider real-world GIS data for an estimation of more realistic risks of pesticide application.
- **MEMs & NAMs: Using the power of in vitro data generation and integration and extrapolation via MEMs.** With perspective on a possible ban of all experimental toxicity testing that use animals, for several organism groups, especially vertebrates, but also soil organisms, NTA and others, one promising alternative for evaluating effects appears to be the combination of *in vitro* experiments and mechanistic models, currently emerging in the scientific field under the label "new approach methodologies" (NAMs). Early studies have investigated this combination, for example for fish, where in in vitro experiments toxicokinetics of chemical in fish cell lines were investigated, modeled, and extrapolated to whole organisms (Stadnicka-Michalak et al., 2015; Stadnicka-Michalak et al., 2014). MEMs, specifically PBTK (see section 1.2.1.4), TKTD, or other organism level models are here an integral part required for the *in vitro – in vivo* extrapolation. More discussion and overview are given in Di Nicola et al., 2023 and Forbes & Galic, 2016. However the implementation of such methods in ERA can only be foreseen as a potential long-term perspective, because there are a large number of issues to resolve, for example, the extrapolation of effects or responses measured in a cell line for the survival or growth of an individual or for effects on the population level, the extrapolation of an exposure of a cell line eliciting a response to the corresponding (internal or external) exposure of the organism and population. It is very likely that the new generation risk assessment concept using NAMs requires further explorations to be proven useful for the prospective ERA of data-rich substances such as PPPs.

- **Systems-based approaches** including elements as mentioned above, for example multiple exposure pathways or multiple environmental compartments, could potentially embed risk assessment into the context of economic analyses, risk-benefit analyses, or socio-economic considerations. What such approaches could look like, and what role models might take, need to be further worked out and specified, but some considerations have been suggested in the PERA roadmap (Sousa et al., 2022).

1.3. Bibliography Chapter 1

Abi-Akar, F., Schmolke, A., Roy, C., Galic, N., & Hinarejos, S. (2020). Simulating honey bee large-scale colony feeding studies using the BEEHAVE Model-Part II: Analysis of overwintering outcomes. *Environmental Toxicology and Chemistry*, *39*(11), 2286-2297. https://doi.org/10.1002/etc.4844

Accolla, C., Schmolke, A., Jacobson, A., Roy, C., Forbes, V. E., Brain, R., & Galic, N. (2022). Modeling pesticide effects on multiple threatened and endangered cyprinid fish species: The role of life-history traits and ecology. *Ecologies 3*, 183-205. https://doi.org/10.3390/ecologies3020015

Agatz, A., Kuhl, R., Miles, M., Schad, T., & Preuss, T. G. (2019). An evaluation of the BEEHAVE Model using honey bee field study data: Insights and recommendations. *Environmental Toxicology and Chemistry*, *38*(11), 2535-2545. https://doi.org/10.1002/etc.4547

Ashauer, R., Albert, C., Augustine, S., Cedergreen, N., Charles, S., Ducrot, V., Focks, A., Gabsi, F., Gergs, A., Goussen, B., Jager, T., Kramer, N. I., Nyman, A. M., Poulsen, V., Reichenberger, S., Schafer, R. B., Van den Brink, P. J., Veltman, K., Vogel, S., . . . Preuss, T. G. (2016). Modelling survival: Exposure pattern, species sensitivity and uncertainty. *Scientific Reports*, *6*, 11, Article 29178. https://doi.org/10.1038/srep29178

Ashauer, R., Thorbek, P., Warinton, J. S., Wheeler, J. R., & Maund, S. (2013). A method to predict and understand fish survival under dynamic chemical stress using standard ecotoxicity data. *Environmental Toxicology and Chemistry*, *32*(4), 954-965. https://doi.org/10.1002/etc.2144

Baas, J., Goussen, B., Miles, M., Preuss, T. G., & Roessink, I. (2022). BeeGUTS-A toxicokinetic-toxicodynamic model for the interpretation and integration of acute and chronic honey bee tests. *Environ Toxicol Chem*, *41*(9), 2193-2201. https://doi.org/10.1002/etc.5423

Baier, V., Paini, A., Schaller, S., Scanes, C. G., Bone, A. J., Ebeling, M., Preuss, T. G., Witt, J., & Heckmann, D. (2022). A generic avian physiologically-based kinetic (PBK) model and its application in three bird species. *Environment International*, *169*, 107547. https://doi.org/10.1016/j.envint.2022.107547

Baker, R. E., Pena, J.-M., Jayamohan, J., & Jérusalem, A. (2018). Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biology letters*, *14*(5), 20170660.

Barnthouse, L. W. (2004). Quantifying population recovery rates for ecological risk assessment. *Environmental Toxicology and Chemistry*, *23*(2), 500-508. https://doi.org/10.1897/02-521

Bart, S., Jager, T., Robinson, A., Lahive, E., Spurgeon, D. J., & Ashauer, R. (2021). Predicting mixture effects over time with toxicokinetic-toxicodynamic models (GUTS): Assumptions, experimental testing, and predictive power. *Environ Sci Technol*, *55*(4), 2430-2439. https://doi.org/10.1021/acs.est.0c05282

Bart, S., Short, S., Jager, T., Eagles, E. J., Robinson, A., Badder, C., Lahive, E., Spurgeon, D. J., & Ashauer, R. (2022). How to analyse and account for interactions in mixture toxicity with toxicokinetic-toxicodynamic models. *Science of the Total Environment*, *843*, 12, Article 157048. https://doi.org/10.1016/j.scitotenv.2022.157048

Bartell, S. M., Brain, R. A., Hendley, P., & Nair, S. K. (2013). Modeling the potential effects of atrazine on aquatic communities in midwestern streams. *Environ Toxicol Chem*, *32*(10), 2402-2411. https://doi.org/10.1002/etc.2332

Bartell, S. M., Nair, S. K., Galic, N., & Brain, R. A. (2020). The Comprehensive Aquatic Systems Model (CASM): Advancing computational capability for ecosystem simulation. *Environmental Toxicology and Chemistry*, *39*(11), 2298-2303. https://doi.org/10.1002/etc.4843

Bartell, S. M., Nair, S. K., Grant, S., & Brain, R. A. (2018). Modeling the effects of thiamethoxam on Midwestern farm ponds and emergent wetlands. *Environ Toxicol Chem*, *37*(3), 738-754. https://doi.org/10.1002/etc.4010

Bartell, S. M., Schmolke, A., Green, N., Roy, C., Galic, N., Perkins, D., & Brain, R. (2019). A Hybrid Individual-Based and Food Web-Ecosystem Modeling Approach for Assessing Ecological Risks to the Topeka Shiner (Notropis topeka): A Case Study with Atrazine. *Environ Toxicol Chem*, *38*(10), 2243-2258. https://doi.org/10.1002/etc.4522

Bartell, S. M. C., Kym Rouse; Lovelock, Cynthia M.; Nair, Shyam K.; Shaw, Jennifer L. (2000). Characterizing aquatic ecological risks from pesticides using a diquat dibromide case study III. Ecological process models. *Environmental Toxicology and Chemistry*, *19*(5), 1441-1453. https://doi.org/10.1002/etc.5620190529

Becher, M. A., Grimm, V., Thorbek, P., Horn, J., Kennedy, P. J., & Osborne, J. L. (2014). BEEHAVE: a systems model of honey bee colony dynamics and foraging to explore multifactorial causes of colony failure [Article]. *Journal of Applied Ecology*, *51*(2), 470-482. https://doi.org/10.1111/1365-2664.12222

Becker, J. L., M.; Kramer-Schadt S.; Franz, M.; Jager, T. (2024). *Critical Evaluation of Effect Models for the Risk Assessment of Plant Protection Products* (Texte Issue. G. E. Agency. https://www.umweltbundesamt.de/en/publikationen/critical-evaluation-of-effect-models-for-the-risk

Becker, J. M., & Liess, M. (2015). Biotic interactions govern genetic adaptation to toxicants. *Pproceeding of the Royal Society B-Biological Sciences*, *282*(1806), 8, Article 20150071. https://doi.org/10.1098/rspb.2015.0071

Billoir, E., Péry, A. R. R., & Charles, S. (2007). Integrating the lethal and sublethal effects of toxic compounds into the population dynamics of *Daphnia magna*: A combination of the DEBtox and matrix population models. *Ecological Modelling*, *203*(3), 204-214. https://doi.org/10.1016/j.ecolmodel.2006.11.021

Brinkmann, M., Eichbaum, K., Buchinger, S., Reifferscheid, G., Bui, T., Schaffer, A., Hollert, H., & Preuss, T. G. (2014). Understanding receptor-mediated effects in rainbow trout: In vitro-in vivo extrapolation using physiologically based toxicokinetic models. *Environmental Science & Technology*, *48*(6), 3303-3309. https://doi.org/10.1021/es4053208

Brinkmann, M., Schlechtriem, C., Reininghaus, M., Eichbaum, K., Buchinger, S., Reifferscheid, G., Hollert, H., & Preuss, T. G. (2016). Cross-species extrapolation of uptake and disposition of neutral organic chemicals in fish using a multispecies physiologically-based toxicokinetic model framework. *Environmental Science & Technology*, *50*(4), 1914-1923. https://doi.org/10.1021/acs.est.5b06158

Calow, P., Sibly, R. M., & Forbes, V. (1997). Risk assessment on the basis of simplified life-history scenarios [Article]. *Environmental Toxicology and Chemistry*, *16*(9), 1983-1989. https://doi.org/10.1002/etc.5620160931

Chandler, G. T., Cary, T. L., Bejarano, A. C., Pender, J., & Ferry, J. L. (2004). Population consequences of fipronil and degradates to copepods at field concentrations: An integration of life cycle testing with Leslie matrix population Modeling. *Environmental Science & Technology*, *38*(23), 6407-6414. https://doi.org/10.1021/es049654o

Cooper, A. B., Aggarwal, M., Bartels, M. J., Morriss, A., Terry, C., Lord, G. A., & Gant, T. W. (2019). PBTK model for assessment of operator exposure to haloxyfop using human biomonitoring and toxicokinetic data. *Regulatory Toxicology and Pharmacology*, *102*, 1-12. https://doi.org/10.1016/j.yrtph.2018.12.004

Dalkvist, T., Topping, C. J., & Forbes, V. E. (2009). Population-level impacts of pesticide-induced chronic effects on individuals depend more on ecology than toxicology. *Ecotoxicology and Environmental Safety*, *72*(6), 1663-1672. https://doi.org/10.1016/j.ecoenv.2008.10.002

De Laender, F., De Schamphelaere, K. A. C., Vanrolleghem, P. A., & Janssen, C. R. (2008). Comparison of different toxic effect sub-models in ecosystem modelling used for ecological effect assessments and water quality standard setting. *Ecotoxicology and Environmental Safety*, *69*(1), 13-23. https://doi.org/10.1016/j.ecoenv.2007.08.020

Di Nicola, M. R., Cattaneo, I., Nathanail, A. V., Carnesecchi, E., Astuto, M. C., Steinbach, M., Williams, A. J., Charles, S., Gestin, O., Lopes, C., Lamonica, D., Tarazona, J. V., & Dorne, J. L. C. M. (2023). The use of new approach methodologies for the environmental risk assessment of food and feed chemicals. *Current Opinion in Environmental Science & Health*, *31*, 100416. https://doi.org/10.1016/j.coesh.2022.100416

Dohmen, G. P., Preuss, T. G., Hamer, M., Galic, N., Strauss, T., van den Brink, P. J., De Laender, F., & Bopp, S. (2016). Population-level effects and recovery of aquatic invertebrates after multiple applications of an insecticide. *Integrated Environmental Assessment and Management*, *12*(1), 67-81. https://doi.org/10.1002/ieam.1676

EFSA (European Food Safety Authority). (2009). Risk assessment for birds and mammals. *EFSA Journal*, *7*(12), 1438. https://doi.org/10.2903/j.efsa.2009.1438

EFSA (European Food Safety Authority). (2013). Guidance on the risk assessment of plant protection products on bees (Apis mellifera, Bombus spp. and solitary bees). *EFSA Journal*, *11*(7), 3295. https://doi.org/10.2903/j.efsa.2013.3295

EFSA (2020). Supporting document for risk managers consultation on specific protection goals for bees – Analysis of background variability of honey bee colony size. *Preliminary report*. https://www.efsa.europa.eu/sites/default/files/topic/review-guidance-document-bees-specific-protection-goals.pdf

EFSA (European Food Safety Authority, Alessio Ippolito, Andreas Focks, Maj Rundlöf, Andres Arce, Marco Marchesi, Franco Maria Neri, Agnès Rortais, Csaba Szentes, Domenica Auteri). (2021). Analysis of background variability of honey bee colony size. *EFSA Supporting Publications*, *18*(3), 6518E. https://doi.org/10.2903/sp.efsa.2021.EN-6518

EFSA (European Food Safety Authority). (2023a). Revised guidance on the risk assessment of plant protection products on bees (Apis mellifera, Bombus spp. and solitary bees). *EFSA Journal*, *21*(5), e07989. https://doi.org/10.2903/j.efsa.2023.7989

EFSA (European Food Safety Authority). (2023b). Risk assessment for Birds and Mammals. *EFSA Journal*, *21*(2), e07790. https://doi.org/10.2903/j.efsa.2023.7790

EFSA PPR (Panel on Plant Protection Products & their Residues). (2010). Scientific opinion on the development of specific protection goal options for environmental risk assessment of pesticides, in particular in relation to the revision of the Guidance Documents on Aquatic and Terrestrial Ecotoxicology (SANCO/3268/2001 and SANCO/10329/2002). *EFSA Journal*, *8*(10), 1821. https://doi.org/10.2903/j.efsa.2010.1821

EFSA PPR (EFSA Panel on Plant Protection Products and their Residues). (2013). Guidance on tiered risk assessment for plant protection products for aquatic organisms in edge-of-field surface waters. *EFSA Journal*, *11*(7), 3290. https://doi.org/10.2903/j.efsa.2013.3290

EFSA PPR (Panel on Plant Protection Products & their Residues). (2014). Scientific opinion on good modelling practice in the context of mechanistic effect models for risk assessment of plant protection products. *EFSA Journal*, *12*(3), 3589. https://doi.org/10.2903/j.efsa.2014.3589

EFSA PPR (Panel on Plant Protection Products & their Residues). (2015). Scientific opinion addressing the state of the science on risk assessment of plant protection products for non-target arthropods. *EFSA Journal*, *13*(2), 3996. https://doi.org/10.2903/j.efsa.2015.3996

EFSA PPR (Panel on Plant Protection Products & their Residues). (2017). Scientific opinion addressing the state of the science on risk assessment of plant protection products for in-soil organisms. *EFSA Journal*, *15*(2), e04690. https://doi.org/10.2903/j.efsa.2017.4690

EFSA PPR (EFSA Panel on Plant Protection Products and their Residues). (2018a). Scientific opinion on the state of the art of Toxicokinetic/Toxicodynamic (TKTD) effect models for regulatory risk assessment of pesticides for aquatic organisms. *EFSA Journal*, *16*(8), 5377-5565. https://doi.org/10.2903/j.efsa.2018.5377

EFSA PPR (EFSA Panel on Plant Protection Products and their Residues). (2018b). Scientific opinion on the state of the science on pesticide risk assessment for amphibians and reptiles. *EFSA Journal*, *16*(2). https://doi.org/10.2903/j.efsa.2018.5125

Escher, S. E., Partosch, F., Konzok, S., Jennings, P., Luijten, M., Kienhuis, A., de Leeuw, V., Reuss, R., Lindemann, K.-M., & Bennekou, S. H. (2022). Development of a roadmap for action on new approach methodologies in risk assessment. *EFSA Supporting Publications*, *19*(6), 7341E. https://doi.org/10.2903/sp.efsa.2022.EN-7341

European Commission (European Commission, Directorate-General for Health and Consumers,). (2014). Addressing the new challenges for risk assessment. https://doi.org/10.2772/37863

Faber, J. H., Marshall, S., Brown, A. R., Holt, A., van den Brink, P. J., & Maltby, L. (2021). Identifying ecological production functions for use in ecosystem services-based environmental risk assessment of chemicals. *Science of the Total Environment*, *791*, 146409. https://doi.org/10.1016/j.scitotenv.2021.146409

Focks, A., Belgers, D., Boerwinkel, M.-C., Buijse, L., Roessink, I., & Van den Brink, P. J. (2018). Calibration and validation of toxicokinetic-toxicodynamic models for three neonicotinoids and some aquatic macroinvertebrates. *Ecotoxicology*, *27*(7), 992-1007.

Focks, A., ter Horst, M., van den Berg, E., Baveco, H., & van den Brink, P. J. (2014). Integrating chemical fate and population-level effect models for pesticides at landscape scale: New options for risk assessment. *Ecological Modelling*, *280*, 102-116. https://doi.org/10.1016/j.ecolmodel.2013.09.023

FOCUS (FOCUS Working Group on Surface Water Scenarios, J. Linders, P. Adriaanse, R. Allen, E. Capri, V. Gouy, J. Hollis, N. Jarvis, M. Klein, P. Lolos, W.-M. Maier, S. Maund, C. Pais, M. Russell, L. Smeets, J.-L. Teixeira, S. Vizantinopoulos, D. Yon). (2001). Focus surface water scenariosin the EU evaluation process under 91/414/EEC. *EC Document Reference SANCO/4802/2001-rev.2*. https://esdac.jrc.ec.europa.eu/public_path/projects_data/focus/sw/docs/FOCUS_SWS_Final_Report.doc

Forbes, V. E., Agatz, A., Ashauer, R., Butt, K. R., Capowiez, Y., Duquesne, S., Ernst, G., Focks, A., Gergs, A., Hodson, M. E., Holmstrup, M., Johnston, A. S. A., Meli, M., Nickisch, D., Pieper, S., Rakel, K. J., Reed, M., Roembke, J., Schäfer, R. B., . . . Roeben, V. (2021). Mechanistic effect modeling of earthworms in the context of pesticide risk assessment: Synthesis of the FORESEE Workshop. *Integrated Environmental Assessment and Management*, *17*(2), 352-363. https://doi.org/10.1002/ieam.4338

Forbes, V. E., & Galic, N. (2016). Next-generation ecological risk assessment: Predicting risk from molecular initiation to ecosystem service delivery. *Environment International*, *91*, 215-219. https://doi.org/10.1016/j.envint.2016.03.002

Forbes, V. E., Galic, N., Schmolke, A., Vavra, J., Pastorok, R., & Thorbek, P. (2016). Assessing the risks of pesticides to threatened and endangered species using population modeling: A critical review and recommendations for future work. *Environmental Toxicology and Chemistry*, *35*(8), 1904-1913. https://doi.org/10.1002/etc.3440

Forbes, V. E., Railsback, S., Accolla, C., Birnir, B., Bruins, R. J. F., Ducrot, V., Galic, N., Garber, K., Harvey, B. C., Jager, H. I., Kanarek, A., Pastorok, R., Rebarber, R., Thorbek, P., & Salice, C. J. (2019). Predicting impacts of chemicals from organisms to ecosystem service delivery: A case study of endocrine disruptor effects on trout. *Science of the Total Environment*, *649*, 949-959. https://doi.org/10.1016/j.scitotenv.2018.08.344

Gabsi, F., Hammers-Wirtz, M., Grimm, V., Schaffer, A., & Preuss, T. G. (2014). Coupling different mechanistic effect models for capturing individual – and population-level effects of chemicals: Lessons from a case where standard risk assessment failed. *Ecological Modelling*, *280*, 18-29. https://doi.org/10.1016/j.ecolmodel.2013.06.018

Gabsi, F., & Preuss, T. G. (2014). Modelling the impact of the environmental scenario on population recovery from chemical stress exposure: A case study using *Daphnia magna*. *Aquatic Toxicology*, *156*, 221-229. https://doi.org/10.1016/j.aquatox.2014.09.002

Galic, N., Ashauer, R., Baveco, H., Nyman, A. M., Barsi, A., Thorbek, P., Bruns, E., & Van den Brink, P. J. (2014). Modeling the contribution of toxicokinetic and toxicodynamic processes to the recovery of gammarus pulex populations after exposure to pesticides. *Environmental Toxicology and Chemistry*, *33*(7), 1476-1488. https://doi.org/10.1002/etc.2481

Galic, N., Hommen, U., Baveco, J., & van den Brink, P. J. (2010). Potential application of population models in the European ecological risk assessment of chemicals II: Review of models and their potential to address environmental protection aims. *Integrated Environmental Assessment and Management*, *6*(3), 338-360. https://doi.org/10.1002/ieam.68

Galic, N., Salice, C. J., Birnir, B., Bruins, R. J. F., Ducrot, V., Jager, H. I., Kanarek, A., Pastorok, R., Rebarber, R., Thorbek, P., & Forbes, V. E. (2019). Predicting impacts of chemicals from organisms to ecosystem service delivery: A case study of insecticide impacts on a freshwater lake. *Science of the Total Environment*, *682*, 426-436. https://doi.org/10.1016/j.scitotenv.2019.05.187

Galic, N., Schmolke, A., Forbes, V., Baveco, H., & van den Brink, P. J. (2012). The role of ecological models in linking ecological risk assessment to ecosystem services in agroecosystems. *Science of the Total Environment*, *415*, 93-100. https://doi.org/10.1016/j.scitotenv.2011.05.065

Gergs, A., Gabsi, F., Zenker, A., & Preuss, T. G. (2016). Demographic toxicokinetic-toxicodynamic modeling of lethal effects. *Environmental Science & Technology*, *50*(11), 6017-6024. https://doi.org/10.1021/acs.est.6b01113

Gergs, A., Hager, J., Bruns, E., & Preuss, T. G. (2021). Disentangling mechanisms behind chronic lethality through toxicokinetic-toxicodynamic modeling. *Environmental Toxicology and Chemistry*, *40*(6), 1706-1712. https://doi.org/10.1002/etc.5027

Gergs, A., Kulkarni, D., & Preuss, T. G. (2015). Body size-dependent toxicokinetics and toxicodynamics could explain intra – and interspecies variability in sensitivity. *Environmental Pollution*, *206*, 449-455. https://doi.org/10.1016/j.envpol.2015.07.045

Gergs, A., Rakel, K. J., Liesy, D., Zenker, A., & Classen, S. (2019). Mechanistic effect modeling approach for the extrapolation of species sensitivity. *Environmental Science & Technology*, *53*(16), 9818-9825. https://doi.org/10.1021/acs.est.9b01690

Goussen, B., Rendal, C., Sheffield, D., Butler, E., Price, O. R., & Ashauer, R. (2020). Bioenergetics modelling to analyse and predict the joint effects of multiple stressors: Meta-analysis and model corroboration. *Science of the Total Environment*, *749*, 10, Article 141509. https://doi.org/10.1016/j.scitotenv.2020.141509

Grimm, V., Augusiak, J., Focks, A., Frank, B. M., Gabsi, F., Johnston, A. S. A., Liu, C., Martin, B. T., Meli, M., Radchuk, V., Thorbek, P., & Railsback, S. F. (2014). Towards better modelling and decision support: Documenting model development, testing, and analysis using TRACE. *Ecological Modelling*, *280*, 129-139. https://doi.org/10.1016/j.ecolmodel.2014.01.018

Grimm, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., Goss-Custard, J., Grand, T., Heinz, S. K., Huse, G., Huth, A., Jepsen, J. U., Jorgensen, C., Mooij, W. M., Muller, B., Pe'er, G., Piou, C., Railsback, S. F., Robbins, A. M., . . . DeAngelis, D. L. (2006). A standard protocol for describing individual-based and agent-based models. *Ecological Modelling*, *198*(1-2), 115-126. https://doi.org/10.1016/j.ecolmodel.2006.04.023

Grimm, V., Berger, U., DeAngelis, D. L., Polhill, J. G., Giske, J., & Railsback, S. F. (2010). The ODD protocol A review and first update. *Ecological Modelling*, *221*(23), 2760-2768. https://doi.org/10.1016/j.ecolmodel.2010.08.019

Grimm, V., Johnston, A. S. A., Thulke, H. H., Forbes, V. E., & Thorbek, P. (2020). Three questions to ask before using model outputs for decision support. *Nature Communications*, *11*(1), 3, Article 4959. https://doi.org/10.1038/s41467-020-17785-2

Grimm, V., & Martin, B. (2013). Mechanistic effect modeling for ecological risk assessment: Where to go from here? *Integrated Environmental Assessment and Management*, *9*. https://doi.org/10.1002/ieam.1423

Hadrup, N., Taxvig, C., Pedersen, M., Nellemann, C., Hass, U., & Vinggaard, A. M. (2013). Concentration addition, independent action and generalized concentration addition models for mixture effect prediction of sex hormone synthesis in vitro. *Plos One*, *8*(8), 13, Article e70490. https://doi.org/10.1371/journal.pone.0070490

Hayes, F., Spurgeon, D. J., Lofts, S., & Jones, L. (2018). Evidence-based logic chains demonstrate multiple impacts of trace metals on ecosystem services. *J Environ Manage*, *223*, 150-164. https://doi.org/10.1016/j.jenvman.2018.05.053

Hommen, U., Forbes, V., Grimm, V., Preuss, T. G., Thorbek, P., & Ducrot, V. (2016). How to use mechanistic effect models in environmental risk assessment of pesticides: Case studies and recommendations from the SETAC Workshop MODELINK. *Integrated Environmental Assessment and Management*, *12*(1), 21-31. https://doi.org/10.1002/ieam.1704

Ibrahim, L. P., T. G.; Schaeffer, A.; Hommen, U. (2014). A contribution to the identification of representative vulnerable fish species for pesticide risk assessment in Europe-A comparison of population resilience using matrix models. *Ecological Modelling*, *280*, 65-75. https://doi.org/10.1016/j.ecolmodel.2013.08.001

Jager, T. (2020). Revisiting simplified DEBtox models for analysing ecotoxicity data. *Ecological Modelling*, *416*, 108904. https://doi.org/10.1016/j.ecolmodel.2019.108904

Jager, T., Albert, C., Preuss, T. G., & Ashauer, R. (2011). General unified threshold model of survival – a toxicokinetic-toxicodynamic framework for ecotoxicolog. *Environmental Science & Technology*, *45*(7), 2529-2540. https://doi.org/10.1021/es103092a

Jager, T., & Ashauer, R. (2018). *Modelling survival under chemical stress. A comprehensive guide to the GUTS framework*. Available from Leanpub: https://leanpub.com/guts_book.

Jager, T., & Zimmer, E. I. (2012). Simplified dynamic energy budget model for analysing ecotoxicity data. *Ecological Modelling*, *225*, 74-81. https://doi.org/10.1016/j.ecolmodel.2011.11.012

Johnston, A. S. A., Hodson, M. E., Thorbek, P., Alvarez, T., & Sibly, R. M. (2014). An energy budget agent-based model of earthworm populations and its application to study the effects of pesticides. *Ecological Modelling 280*, 5-17. https://doi.org/10.1016/j.ecolmodel.2013.09.012

Kattwinkel, M., & Liess, M. (2014). Competition matters: Species interactions prolong the long-term effects of pulsed toxicant stress on populations. *Environmental Toxicology and Chemistry*, *33*(7), 1458-1465. https://doi.org/10.1002/etc.2500

Kattwinkel, M., Reichert, P., Ruegg, J., Liess, M., & Schuwirth, N. (2016). Modeling macroinvertebrate community dynamics in stream mesocosms contaminated with a pesticide. *Environmental Science & Technology*, *50*(6), 3165-3173. https://doi.org/10.1021/acs.est.5b04068

Kleinmann, J. U., & Wang, M. (2017). Modeling individual movement decisions of brown hare (lepus europaeus) as a key concept for realistic spatial behavior and exposure: a population model for landscape-level risk assessment. *Environmental Toxicology and Chemistry*, *36*(9), 2299-2307. https://doi.org/10.1002/etc.3760

Kooijman, B. (2009). *Dynamic Energy Budget Theory for Metabolic Organisation* (3 ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511805400

Kooijman, S., Hanstveit, A. O., & Nyholm, N. (1996). No-effect concentrations in algal growth inhibition tests. *Water Research*, *30*(7), 1625-1632. https://doi.org/10.1016/0043-1354(96)00056-5

Krause, S., & Goss, K. U. (2020). Comparison of a simple and a complex model for BCF prediction using in vitro biotransformation data. *Chemosphere*, *256*, 8, Article 127048. https://doi.org/10.1016/j.chemosphere.2020.127048

Liess, M., Foit, K., Knillmann, S., Schafer, R. B., & Liess, H. D. (2016). Predicting the synergy of multiple stress effects. *Scientific Reports*, *6*, 8, Article 32965. https://doi.org/10.1038/srep32965

Liu, C., Bednarska, A. J., Sibly, R. M., Murfitt, R. C., Edwards, P., & Thorbek, P. (2014). Incorporating toxicokinetics into an individual-based model for more realistic pesticide exposure estimates: A case study of the wood mouse. *Ecological Modelling*, *280*, 30-39. https://doi.org/10.1016/j.ecolmodel.2013.09.007

Liu, C., Sibly, R. M., Grimm, V., & Thorbek, P. (2013). Linking pesticide exposure and spatial dynamics: An individual-based model of wood mouse (*Apodemus sylvaticus*) populations in agricultural landscapes. *Ecological Modelling*, *248*, 92-102. https://doi.org/10.1016/j.ecolmodel.2012.09.016

Loos, M., Ragas, A. M. J., Plasmeijer, R., Schipper, A. M., & Hendriks, A. J. (2010). Eco-SpaCE: An object-oriented, spatially explicit model to assess the risk of multiple environmental stressors on terrestrial vertebrate populations. *Science of the Total Environment*, *408*(18), 3908-3917. https://doi.org/10.1016/j.scitotenv.2009.11.045

Maltby, L., Brown, R., Faber, J. H., Galic, N., Van den Brink, P. J., Warwick, O., & Marshall, S. (2021). Assessing chemical risk within an ecosystem services framework: Implementation and added value. *Science of the Total Environment*, *791*, 148631. https://doi.org/10.1016/j.scitotenv.2021.148631

Martin, B. T., Jager, T., Nisbet, R. M., Preuss, T. G., Hammers-Wirtz, M., & Grimm, V. (2013). Extrapolating ecotoxicological effects from individuals to populations: a generic approach based on Dynamic Energy Budget theory and individual-based modeling. *Ecotoxicology*, *22*(3), 574-583. https://doi.org/10.1007/s10646-013-1049-x

Mavroudis, P. D., Hermes, H. E., Teutonico, D., Preuss, T. G., & Schneckener, S. (2018). Development and validation of a physiology-based model for the prediction of pharmacokinetics/toxicokinetics in rabbits. *Plos One*, *13*(3), e0194294. https://doi.org/10.1371/journal.pone.0194294

McElroy, A., Barron, M., Beckvar, N., Driscoll, S., Meador, J., Parkerton, T., Preuss, T., & Steevens, J. (2011). Use of the tissue residue approach for organic and organometallic compounds. *Integrated Environmental Assessment and Management*, *7*, 50-74. https://doi.org/10.1002/ieam.132

Nyman, A. M., Schirmer, K., & Ashauer, R. (2012). Toxicokinetic-toxicodynamic modelling of survival of Gammarus pulex in multiple pulse exposures to propiconazole: model assumptions, calibration data requirements and predictive power. *Ecotoxicology*, *21*(7), 1828-1840. https://doi.org/10.1007/s10646-012-0917-0

OECD. (2006). *Current Approaches in the Statistical Analysis of Ecotoxicity Data*. https://doi.org/10.1787/9789264085275-en

OECD. (2018). *Test No. 319B: Determination of in vitro intrinsic clearance using rainbow trout liver S9 sub-cellular fraction (RT-S9)*. https://doi.org/10.1787/9789264303232-en

Park, R. A., & Clough, J. S. (2018). *AQUATOX (Release 3.2) – Modeling Environmental Fate and Ecological Effects in Aquatic Ecosystems. Volume 2: Technical Documentation*. https://www.epa.gov/ceam/aquatox-32-supporting-documents.

Park, R. A., Clough, J. S., & Wellman, M. C. (2008). AQUATOX: Modeling environmental fate and ecological effects in aquatic ecosystems. *Ecological Modelling*, *213*(1), 1-15. https://doi.org/10.1016/j.ecolmodel.2008.01.015

Pilkey, O. H., & Pilkey-Jarvis, L. (2009). *Useless Arithmetic – Why Environmental Scientists Can't Predict the Future*. Columbia University Press.

Railsback, S. F., & Grimm, V. (2019). *Agent-based and individual-based modeling: a practical introduction*. Princeton university press.

Raimondo, S., Etterson, M., Pollesch, N., Garber, K., Kanarek, A., Lehmann, W., & Awkerman, J. (2018). A framework for linking population model development with ecological risk assessment objectives. *Integrated Environmental Assessment and Management*, *14*(3), 369-380. https://doi.org/10.1002/ieam.2024

Raimondo, S., Schmolke, A., Pollesch, N., Accolla, C., Galic, N., Moore, A., Vaugeois, M., Rueda-Cediel, P., Kanarek, A., Awkerman, J., & Forbes, V. (2021). Pop-GUIDE: Population modeling guidance, use, interpretation, and development for ecological risk assessment. *Integrated Environmental Assessment and Management*, *17*(4), 767-784. https://doi.org/10.1002/ieam.4377

Rakel, K., Becker, D., Bussen, D., Classen, S., Preuss, T., Strauss, T., Zenker, A., & Gergs, A. (2022). Physiological dependency explains temperature differences in sensitivity towards chemical exposure. *Arch Environ Contam Toxicol*, *83*(4), 349-360. https://doi.org/10.1007/s00244-022-00963-2

Rashleigh, B., Barber, M. C., & Walters, D. M. (2009). Foodweb modeling for polychlorinated biphenyls (PCBs) in the Twelvemile Creek Arm of Lake Hartwell, South Carolina, USA. *Ecological Modelling*, *220*(2), 254-264. https://doi.org/10.1016/j.ecolmodel.2008.09.007

Reeg, J., Heine, S., Mihan, C., Preuss, T. G., McGee, S., & Jeltsch, F. (2018). Potential impact of effects on reproductive attributes induced by herbicides on a plant community. *Environmental Toxicology and Chemistry*, *37*(6), 1707-1722. https://doi.org/10.1002/etc.4122

Reeg, J., Schad, T., Preuss, T. G., Solga, A., Korner, K., Mihan, C., & Jeltsch, F. (2017). Modelling direct and indirect effects of herbicides on non-target grassland communities. *Ecological Modelling*, *348*, 44-55. https://doi.org/10.1016/j.ecolmodel.2017.01.010

Roeben, V., Oberdoerster, S., Rakel, K. J., Liesy, D., Capowiez, Y., Ernst, G., Preuss, T. G., Gergs, A., & Oberdoerster, C. (2020). Towards a spatiotemporally explicit toxicokinetic-toxicodynamic model for earthworm toxicity. *Science of the Total Environment*, *722*, 137673. https://doi.org/10.1016/j.scitotenv.2020.137673

Rueda-Cediel, P., Brain, R., Galic, N., & Forbes, V. (2019). Comparative Analysis of Plant Demographic Traits Across Species of Different Conservation Concern: Implications for Pesticide Risk Assessment. *Environmental Toxicology and Chemistry*, *38*(9), 2043-2052. https://doi.org/10.1002/etc.4472

Sanchez-Bayo, F. (2009). From simple toxicological models to prediction of toxic effects in time [Article]. *Ecotoxicology*, *18*(3), 343-354. https://doi.org/10.1007/s10646-008-0290-1

Schmolke, A., Abi-Akar, F., Roy, C., Galic, N., & Hinarejos, S. (2020). Simulating Honey Bee Large-Scale Colony Feeding Studies Using the BEEHAVE Model-Part I: Model Validation. *Environmental Toxicology and Chemistry*, *39*(11), 2269-2285. https://doi.org/10.1002/etc.4839

Schmolke, A., Bartell, S. M., Roy, C., Desmarteau, D., Moore, A., Cox, M. J., Maples-Reynolds, N. L., Galic, N., & Brain, R. (2021). Applying a hybrid modeling approach to evaluate potential pesticide effects and mitigation effectiveness for an endangered fish in simulated oxbow habitats. *Environmental Toxicology and Chemistry*, *40*(9), 2615-2628. https://doi.org/10.1002/etc.5144

Schmolke, A., Bartell, S. M., Roy, C., Green, N., Galic, N., & Brain, R. (2019). Species-specific population dynamics and their link to an aquatic food web: A hybrid modeling approach. *Ecological Modelling*, *405*, 1-14. https://doi.org/10.1016/j.ecolmodel.2019.03.024

Schmolke, A., Galic, N., Feken, M., Thompson, H., Sgolastra, F., Pitts-Singer, T., Elston, C., Pamminger, T., & Hinarejos, S. (2021). Assessment of the vulnerability to pesticide exposures across bee species. *Environmental Toxicology and Chemistry*, *40*(9), 2640-2651. https://doi.org/10.1002/etc.5150

Schmolke, A., Thorbek, P., DeAngelis, D. L., & Grimm, V. (2010). Ecological models supporting environmental decision making: a strategy for the future. *Trends in Ecology & Evolution*, *25*(8), 479-486. https://doi.org/10.1016/j.tree.2010.05.001

Scholze, M., Taxvig, C., Kortenkamp, A., Boberg, J., Christiansen, S., Svingen, T., Lauschke, K., Frandsen, H., Ermler, S., Hermann, S. S., Pedersen, M., Lykkeberg, A. K., Axelstad, M., & Vinggaard, A. M. (2020). Quantitative in vitro to in vivo extrapolation (QIVIVE) for predicting reduced anogenital distance produced by anti-androgenic pesticides in a rodent model for male reproductive disorders. *Environmental Health Perspectives*, *128*(11), 17, Article 117005. https://doi.org/10.1289/ehp6774

Sherborne, N., Galic, N., & Ashauer, R. (2020). Sublethal effect modelling for environmental risk assessment of chemicals: Problem definition, model variants, application and challenges. *Science of the Total Environment*, *745*, 141027. https://doi.org/10.1016/j.scitotenv.2020.141027

Sibly, R. M., Grimm, V., Martin, B. T., Johnston, A. S. A., Kulakowska, K., Topping, C. J., Calow, P., Nabe-Nielsen, J., Thorbek, P., & DeAngelis, D. L. (2013). Representing the acquisition and use of energy by individuals in agent-based models of animal populations. *Methods in Ecology and Evolution*, *4*(2), 151-161. https://doi.org/10.1111/2041-210x.12002

Singer, A., Nickisch, D., & Gergs, A. (2023). Joint survival modelling for multiple species exposed to toxicants. *Science of the Total Environment*, *857*, 159266. https://doi.org/10.1016/j.scitotenv.2022.159266

Sousa, J. P., Aldrich, A., Axelman, J., Backhaus, T., Brendel, S., Dorronsoro, B., Duquesne, S., Focks, A., Holz, S., Knillmann, S., Pieper, S., Schmied-Tobies, M., Silva, E., Topping, C., Wipfler, L., & Williams, J. (2022). Building a European Partnership for next generation, systems-based Environmental Risk Assessment (PERA). *EFSA Supporting Publications*, *19*(8), 7546E. https://doi.org/10.2903/sp.efsa.2022.EN-7546

Stadnicka-Michalak, J., Schirmer, K., & Ashauer, R. (2015). Toxicology across scales: Cell population growth in vitro predicts reduced fish growth. *Science Advances*, *1*(7), 8, Article e1500302. https://doi.org/10.1126/sciadv.1500302

Stadnicka-Michalak, J., Tanneberger, K., Schirmer, K., & Ashauer, R. (2014). Measured and modeled toxicokinetics in cultured fish cells and application to in vitro-in vivo toxicity extrapolation. *Plos One*, *9*(3), e92303. https://doi.org/10.1371/journal.pone.0092303

Strauss, T., Gabsi, F., Hammer-Wirtz, M., Thorbek, P., & Preuss, T. G. (2017). The power of hybrid modelling: An example from aquatic ecosystems. *Ecological Modelling*, *364*, 77-88. https://doi.org/10.1016/j.ecolmodel.2017.09.019

Topping, C. J., Aldrich, A., & Berny, P. (2020). Overhaul environmental risk assessment for pesticides. *Science*, *367*(6476), 360-363. https://doi.org/10.1126/science.aay1144

Topping, C. J., Dalby, L., & Skov, F. (2016). Landscape structure and management alter the outcome of a pesticide ERA: Evaluating impacts of endocrine disruption using the ALMaSS European Brown Hare model. *Science of the Total Environment*, *541*, 1477-1488. https://doi.org/10.1016/j.scitotenv.2015.10.042

Topping, C. J., Dalkvist, T., Forbes, V. E., Grimm, V., & Sibly, R. M. (2009). The Potential for the Use of Agent-Based Models in Ecotoxicology. In J. Devillers (Ed.), *Ecotoxicology Modeling* (pp. 205-235). Springer US. https://doi.org/10.1007/978-1-4419-0197-2_8

Topping, C. J., & Odderskaer, P. (2004). Modeling the influence of temporal and spatial factors on the assessment of impacts of pesticides on skylarks. *Environmental Toxicology and Chemistry*, *23*(2), 509-520. https://doi.org/10.1897/02-524a

Traas, T. P., Janse, J. H., Van den Brink, P. J., Brock, T. C., & Aldenberg, T. (2004). A freshwater food web model for the combined effects of nutrients and insecticide stress and subsequent recovery. *Environ Toxicol Chem*, *23*(2), 521-529. https://doi.org/10.1897/02-524

Trijau, M., Goussen, B., Brain, R., Maul, J., & Galic, N. (2023). Development of a mechanistic model for analyzing avian reproduction data for pesticide risk assessment. *Environmental Pollution*, *327*, 121477. https://doi.org/10.1016/j.envpol.2023.121477

Van den Brink, P. J., Alix, A., Thorbek, P., Baveco, H., Agatz, A., Faber, J. H., Brown, A. R., Marshall, S., & Maltby, L. (2021). The use of ecological models to assess the effects of a plant protection product on ecosystem services provided by an orchard. *Science of the Total Environment*, *798*, 149329. https://doi.org/10.1016/j.scitotenv.2021.149329

Van den Brink, P. J., Baveco, J. M., Verboom, J., & Heimbach, F. (2007). An individual-based approach to model spatial population dynamics of invertebrates in aquatic ecosystems after pesticide contamination. *Environmental Toxicology and Chemistry*, *26*(10), 2226-2236. https://doi.org/10.1897/07-022r.1

Vlaeminck, K., Viaene, K. P. J., Van Sprang, P., & De Schamphelaere, K. A. C. (2022). Predicting Combined Effects of Chemical Stressors: Population-Level Effects of Organic Chemical Mixtures with a Dynamic Energy Budget Individual-Based Model [Article; Early Access]. *Environ Toxicol Chem*, *41*(9), 2240-2258. https://doi.org/10.1002/etc.5409

Waage, J. K., Hassell, M. P., & Godfray, H. C. J. (1985). The dynamics of pest parasitoid insecticide interactions. *Journal of Applied Ecology*, *22*(3), 825-838. https://doi.org/10.2307/2403232

Wang, M. (2013). From home range dynamics to population cycles: Validation and realism of a common vole population model for pesticide risk assessment. *Integrated Environmental Assessment and Management*, *9*(2), 294-307. https://doi.org/10.1002/ieam.1377

Wang, M., & Grimm, V. (2007). Home range dynamics and population regulation: An individual-based model of the common shrew *Sorex ayaneus*. *Ecological Modelling*, *205*(3-4), 397-409. https://doi.org/10.1016/j.ecolmodel.2007.03.003

Zhang, L. L., Liu, J. L., Li, Y., & Zhao, Y. W. (2013). Applying AQUATOX in determining the ecological risk assessment of polychlorinated biphenyl contamination in Baiyangdian Lake, North China. *Ecological Modelling*, *265*, 239-249. https://doi.org/10.1016/j.ecolmodel.2013.06.003

Zimmer, E. I., Preuss, T. G., Norman, S., Minten, B., & Ducrot, V. (2018). Modelling effects of time-variable exposure to the pyrethroid beta-cyfluthrin on rainbow trout early life stages. *Environmental Sciences Europe*, *30*(1), 1-13.

2. Acceptability of effect models for use in regulatory risk assessment: Considerations and conditions

Jeremias Becker, Magnus Wang, Alpar Barsi, Udo Hommen, Thomas G. Preuss, Oliver Jakoby, Benoit Goussen, Stefan Reichenberger, Melissa Reed, Sabine Duquesne

According to Grimm et al. (2020), demonstration, understanding, and prediction are three principal purposes that MEMs could serve in making regulatory decisions. Specific ways how MEMs could support the risk assessment of pesticides are outlined in Chapter 1. For the application of MEMs, it is required that a proposed model is fit for its purpose. The purpose of a model within risk assessment will guide which evaluation steps are considered necessary for the assessment of a model's suitability.

In this chapter, we first summarize the suggestions of the EFSA scientific opinion on good modeling practice (GMP-SO; EFSA PPR, 2014) for the assessment of MEMs, before we make two suggestions concerning how future assessments of MEMs could be organized efficiently (sections 2.2 and 2.3). After that, this chapter will briefly touch on all aspects relevant for regulatory model evaluation in the GMP-SO. Many of these aspects, including scenario definition (Chapter 3), documentation and evaluation of data (Chapter 4), handling modular MEMs (Chapter 5), model calibration and validation (Chapter 6), and model sensitivity and uncertainty analyses (Chapter 7), are laid out in detail in other book chapters.

2.1. The EFSA scientific opinion on good modeling practice as basis for evaluation of MEMs

The evaluation of a MEM follows the typical steps in the development and application of a model, as they are often laid out in a modeling report and described, for example, in the TRACE framework (Grimm et al., 2014). The EFSA GMP-SO provides suggestions including a checklist for the evaluation of the model itself and its application for a specific risk assessment.

The development and application of models is often illustrated as a modeling cycle (Grimm et al., 2014). Based on the modeling cycle, the GMP-SO identified four basic stages in the development and

evaluation of a MEM for the ERA of pesticides, called the conceptual, the formal, the computer, and the regulatory model. The conceptual model describes the model design or structure that is developed after an initial phase of problem formulation and collection of background information. Thus, the conceptual model defines the entities and processes, as well as the abiotic and biotic environmental factors to be considered, their linkages, and how the processes are considered to work mechanistically. The formal model translates the conceptual model to a set of coupled mathematical equations and/or algorithms. Software implementation of the formal model, i.e., coding and debugging (the latter also called model output verification), results in the computer model. Finally, the computer model needs to be parameterized and set up for a specific environmental scenario, that describes the abiotic, biotic, and agronomic conditions. Altogether, this results in what is termed the regulatory model according to GMP-SO (Figure 2.1). Thus, the regulatory model covers everything that is required to run a control or reference simulation (i.e., without exposure to the chemical to be assessed). The regulatory model also includes the equations and computer code for simulating the effects of a pesticide and the underlying exposure (fate model), along with a specific environmental scenario. However, the regulatory model does not include the associated substance – and use-specific parameterization, including chemical properties, ecotoxicological data, and application patterns.

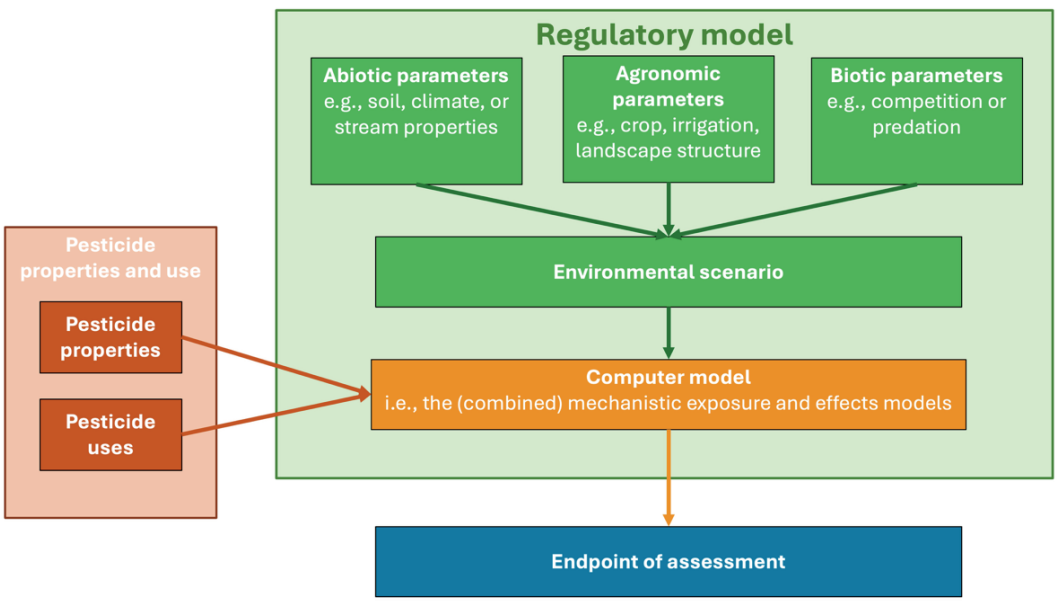


Figure 2.1: Concept of the regulatory model redrawn from the EFSA scientific opinion on good modeling practice in the context of mechanistic effect models for risk assessment of plant protection products. We consider EFSA's "regulatory model" the components (the light green box) as a "general regulatory model" and this together with information in the pesticide (the light red box) as a "specific regulatory model." Adapted with permission from EFSA PPR (2014).

For example, for a terrestrial higher-tier risk assessment, a regulatory model may be defined as a specific population model (computer model), for a specific focal species, in a specific crop, and in a specific environmental scenario that is considered representative as a realistic worst case for a specific country or EU zone. Only the case-specific parameterization regarding use and properties of the pesticide to be assessed needs to be added and evaluated afterwards.

Thus, the regulatory model can be assessed independently from its application for the risk assessment of a specific pesticide use.

The evaluation of a regulatory model according to the GMP-SO includes a review on the model concept and its formalization and software implementation. Additionally, the data used for model development and parameterization, the calibration procedures (if necessary) and the environmental scenario are reviewed. A sensitivity analysis is required to identify influential input parameters with potentially high impact on relevant modeling endpoints. Finally, the ecological realism and predictive power of a model needs to be tested by comparing model predictions with independent observations (model validation). To aid model documentation and evaluation, the GMP-SO provides questionnaires for modelers and risk assessors (appendices B and C in EFSA PPR [2014]).

When the regulatory model has been evaluated and accepted, it is considered ready for use to address specific questions in regulatory risk assessment. Acceptance of a regulatory model, as described in the GMP-SO, is therefore restricted to a particular domain of use for which it has been evaluated. Incorrect use of a regulatory model, for example, by using a wrong pesticide related parameterization or by applying the model outside its specific domain, does not affect the reliability and acceptability of the regulatory model itself.

The schemes are suitable for a regulatory model that is developed from scratch for a given use in ERA. However, because model development takes considerable resources, models are typically not developed for a single application in regulatory ERA. When using a regulatory model multiple times, the evaluation of a full MEM for each of its applications in regulatory risk assessment would each time require considerable efforts, irrespective of how close the protocol suggested in the GMP-SO was followed previously. So far, no fit-for-purpose regulatory models other than GUTS and the *Lemna* macrophyte model have been acknowledged by EFSA. This means in consequence, that currently a rapporteur member state has to fully evaluate a regulatory model each time a new application of a MEM is proposed in a dossier for risk assessment. This is inefficient and has the potential to lead to inconsistency in the evaluation and acceptance of MEMs in the ERA of pesticides. In addition, MEMs are often combined of sub-models, often termed "modules" (see section 5.2) that have probably already been evaluated earlier, so that a repeated evaluation of such modules appears redundant.

If a MEM (or module) is likely to be applied often, e.g., for many crop types, countries or zones, etc. it would appear reasonable to assess and evaluate it in a generic way, for example as it was done for TKTD models in the EFSA scientific opinion on TKTD models (EFSA PPR 2018a). When a regulatory model that has been accepted is applied for a specific risk assessment, the overall evaluation effort could be reduced to the steps required to set it up for the specific application in a dossier.

The evaluation schemes according to the GMP-SO consider only a differentiation between a regulatory model and its use. In the present book, we consider the regulatory model as described in Figure 2.1 as a general regulatory model, because it is non-specific in terms of pesticide use; after setup for a specific pesticide and use-pattern we consider it a specific regulatory model (see Figure 2.2 and 2.3). However, a regulatory model as outlined in (EFSA PPR, 2014) refers to a full framework of combined fate and effect models. We consider this being outside our definition of MEMs, since fate and effect models are often developed and evaluated separately. Additionally, even a general regulatory model is already specific in terms of the environmental and agronomic conditions, which in practice may often be adjusted case-by-case for new model applications in ERA. Therefore, even a general regulatory model may be too large and specific for a general evaluation that could provide a basis for future model applications.

To overcome this situation, we suggest a modified way to evaluate and assess MEMs for regulatory ERA. First, we suggest identifying purpose and modules of a MEM, to break down their evaluation to smaller elements and to identify possibly already evaluated building blocks, as further discussed in section 2.2. Second, we suggest separating the evaluation of a MEM in general and its specific application for a risk assessment question. This separation would allow the assessment of MEMs intended to be used in regulatory ERA only once and establish them within their domains of applicability as general effect model ("GEM"; further outlined in section 2.3); afterwards, the regulatory model for a specific risk assessment question could be evaluated based on such GEM separately per application. In this way, redundant evaluations can be avoided, while the assessment of the regulatory model for a specific regulatory risk assessment question remains obligatory.

2.2. What to evaluate – identify purpose and components of MEMs

A first step within the process of evaluating a MEM is to identify its purpose (Grimm et al., 2020) in the context of the specific anticipated task of the model, within regulatory risk assessment. This purpose determines the focus for the evaluation, because the suitability of a model may vary depending on the question to be addressed.

The arrow from the ecology module to the exposure means, that the actual exposure of the modeled organisms is also affected by the development, distribution, and behavior of organisms as simulated in the ecology module.

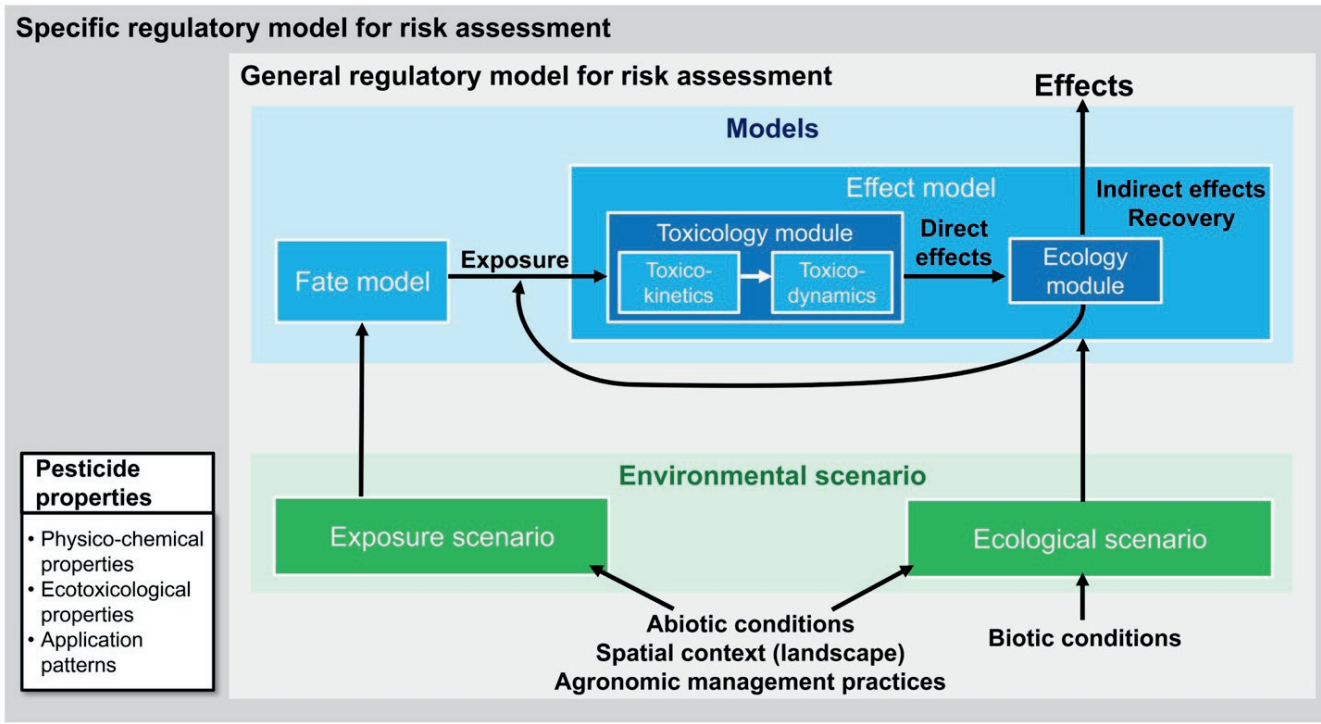


Figure 2.2: Proposed scheme of a combined fate and effect modeling framework and of the associated environmental scenario for risk assessment.

MEMs can be built of several sub-models or modules. It can be beneficial to break down a full MEM into the modules from which it is assembled, and to individually evaluate the suitability of these modules for use in risk assessment (further details in Chapter 5.). Afterwards, a specific combination of accepted modules, together with an environmental scenario may be evaluated as a full GEM. Such modular evaluation approach could be based on the identification of generic or static components of a module that need to be evaluated only once, and specific or variable components that need to be re-evaluated when being applied in a new MEM. The documentation and evaluation of models and environmental scenarios could be structured accordingly.

To aid the identification of modules for model documentation and evaluation, we propose a scheme of the main components that typically make up a full mechanistic fate and effect modeling framework for the ERA of pesticides (Figure 2.2). The scheme separates the framework by functionality into a fate model and an effect model. The effect model (GEM) is separated into an ecology module and a toxicology module, which may be further structured into submodules as discussed in section 5.4.2.

The ecology module covers the biological and environmental processes required to simulate the modeled entities (individuals, populations) in a non-contaminated "control" scenario. The complexity of the ecology module can range from a simple rate of (background) mortality in simulations of survival by a GUTS model to a full population or community model with respective processes.

The fate model provides the level and duration of environmental exposure to the effect model. It is assumed here, that all parameters being required by the fate model to calculate the exposure for use in the effect model are summarized in the exposure scenario. It should be noted, however, that in many MEMs the actual exposure of the modeled organisms is also affected by the development, distribution, and behavior of organisms as simulated in the ecology module (shown in Figure 2.2). This means that whether organisms are exposed during critical life stages depends not only on the timing of pesticide application and its fate in the environment, but also on the timing of organism development. Moreover, in spatially explicit models, individuals might experience different exposure over time depending on their spatial distribution and behavior. The toxicology module relates exposure to immediate effects at the suborganism, organism, or population level. In the ecology module, these effects mechanistically propagate to mediate effects at the biological organization level for which the MEM has been developed (Figure 2.2). The mediate effects emerge as differences in state variables of the ecology module across control and treatment runs, e.g., as a decrease in body mass and survival (in organism-level MEMs) or in abundance (in population-level MEMs) following pesticide exposure.

In an example for a TKTD model for organism-level effects used at Tier 2C of the aquatic ERA, the toxicology module consists of the TK and TD component. The TK part may be connected to a model for the environmental fate providing the time course of the external concentration. The TK part models the internal concentration, representing processes of absorption, distribution, metabolism, and excretion (ADME) in more or less detail. The TD part relates internal pesticide concentrations to immediate effects on the physiological processes modeled in the ecology module. This may include damage build-up (GUTS), or impairment of feeding, growth, and maturation in a DEB-TKTD model, but also processes like damage repair. In a DEB-TKTD model, the ecology module is represented by the DEB part, in which the immediate effects may propagate to mediate effects on survival, body size, and the number of offspring. In a GUTS-TKTD model, the ecology module is limited to a simple survival rate that can be affected when sufficient damage has been built up.

In a second example, a DEB-TKTD model could be integrated in an individual-based population model. Again, the toxicology module is represented by the TD part, whereas the DEB part and the population-level processes form the ecology module. In this ecology module, effects at the (sub-) organism level will mechanistically propagate to effects at the population level.

The overall (mediate) effects predicted at the level of biological organization that is considered relevant for the regulatory question (e.g., population level) are the model output of primary interest for risk assessment. From this output, regulatory-relevant endpoints, as for example threshold concentrations for effects, time to population recovery or margins of safety for specific effects can be derived. Consequently, these overall effects appear, together with the performances of the undisturbed "control," as the most relevant modeling output for risk assessment and should be in the focus of analysis for model evaluation (e.g., uncertainty and sensitivity analyses, validation studies). Parameterizing the ecology module of a

given GEM (and the pesticide-independent parameters of a fate model) for a given environmental scenario leads to a regulatory model as defined in the GMP-SO (EFSA PPR, 2014). More details on scenario development are given in section 2.4.3, and, more comprehensively, Chapter 3.

2.3. How to evaluate – separating MEMs and their specific applications

The modular concept outlined in Figure 2.2 may help to evaluate models in a more general way. Models can, according to their function as module in a larger mechanistic fate and effects modeling framework, be evaluated separately (see details in Chapter 5). Accepted modules can be assembled into a general effect model (GEM). Such a GEM is developed in relation to a general problem formulation and should be evaluated according to the typical steps (conceptual – formal – computer model; left hand side in Figure 2.3). Together with a default environmental scenario, such a GEM can be evaluated and assessed independently from any specific use and compound-specific information. Such assessment would be only required once per GEM, and any refinement of such a GEM and associated assessments should be documented in a version control system to ensure that older versions are still available so that calculations can be reproduced if necessary.

Following a general problem formulation, the development and evaluation of a mechanistic model is leading to an accepted and well-defined general effect model (GEM) (left). For a given application in ERA, following the specific problem formulation, the GEM is coupled with A fate model and set-up for a specific environmental scenario to form the general regulatory model. Adding data on pesticide use and properties that concern toxicology, fate, and exposure creates the specific regulatory model that, after evaluation and acceptance, is ready to use for a case-specific risk assessment (right). Vertical labels illustrate the processes that lead to the different stages of a model and its application for a specific risk assessment. Feedback loops between stages are possible and common but omitted for clarity.

For any concrete application of the general effect model to a specific risk assessment question, the GEM may be coupled with a fate model and set up for a specific environmental scenario to form the general regulatory model (GRM). The definition of environmental scenarios requires consideration of any specific aspects that may arise from the specific problem formulation and the regulatory questions to be answered (see Chapter 3). A general regulatory model is specific for a certain risk assessment question (including, e.g., a certain use, application modes, protection goals), but still not specific for a certain pesticide. In that sense, a given GRM might be used repeatedly in ERA if for multiple pesticides the same regulatory question applies, and the same scenarios are considered relevant. When, in a last step, further case-specific information on

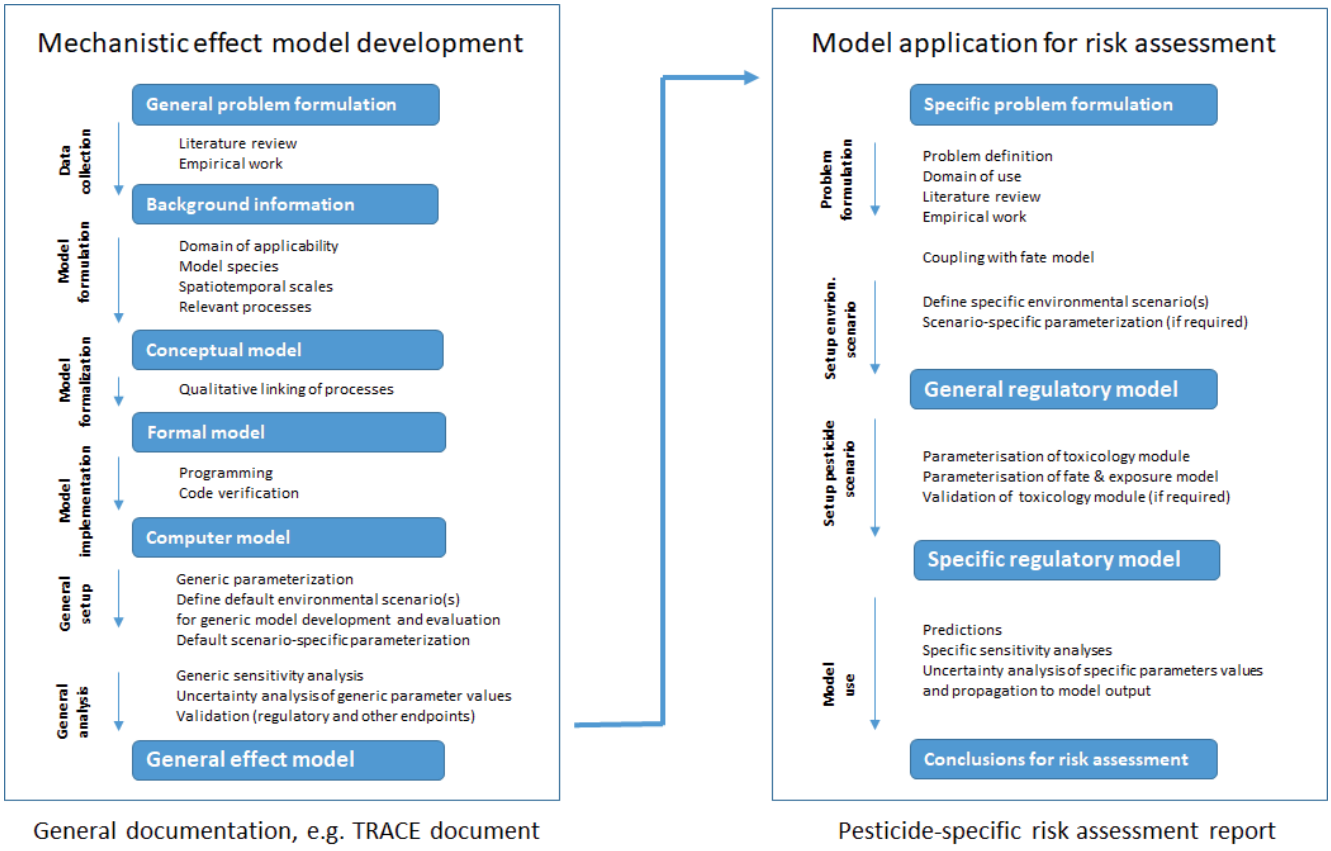


Figure 2.3: Proposed concept for the evaluation of mechanistic effect models and their application for regulatory risk assessment.

pesticide use and properties is added, the specific regulatory model (SRM) is built (right hand side of Figure 2.3). An evaluation and assessment of this SRM would need to be done for each new model application. However, this evaluation would not consider the whole general effect model but could rather focus on the case-specific aspects. This approach to the regulatory evaluation of effect models in two steps is thought to result in more efficient and consistent evaluations. In the following, we provide some considerations for the proposed additional evaluation of a formerly evaluated GEM when being applied in ERA.

When (sub-)models have successfully passed the evaluation as modules for risk assessment, they may be assembled to a complete modeling framework and applied for a specific risk assessment. For that, it should be assessed whether the domains of use of the modules address the specific risk assessment question, and whether the selected modules work together well. Then, the specific environmental scenario can be defined, and the model set up accordingly. This setup includes the parameterization of (i) those parameters in the fate and the effect module that are specific for an environmental scenario, and (ii) of the pesticide-specific parameters in the toxicology and fate module. This flexibility and modularity would allow for a high number of different regulatory models that can be assembled from a relatively small number of fate and effect models and/or modules. Because mechanistic effect modeling is under ongoing development,

no standards for the combination of models in a modular modeling framework have been established yet (further discussion in Chapter 5).

Specific regulatory models are designed for a given application of a certain pesticide in risk assessment and consist of the general regulatory model parameterized for a given pesticide. The parameterization of the toxicology module is typically obtained from calibration to toxicity test data for a specific pesticide. It should be evaluated whether all types of immediate effects (e.g., decreased survival, growth, reproduction, etc.) have been considered in the toxicology module that are potentially relevant for the risk assessment question to be addressed. If a toxicology module is used to predict immediate effects for conditions others than those used for parameterization, the realism in these predictions should ideally be experimentally tested, or at least be critically discussed. For example, when a TKTD model is used to predict effects of different exposure profiles in Tier 2C, realism in these predictions needs to be validated by comparing predictions for a different exposure profile with independent observations (EFSA PPR, 2018a). Ideally, the same should apply when TKTD models are integrated in population models for higher tiers, although it is hard to imagine such strict experimental proof for a large number of environmental field conditions including not only temperature, but also pH, organic matter content, biotic conditions, etc. The validation of complex, modular models is discussed in section 2.4.5, and, more thoroughly, in Chapter 6.

Additionally, when MEMs are used in higher tier studies, the toxicology module predicts immediate effects in a field situation, in which the individuals may be more (or less) susceptible than under standardized test conditions. The parameterization may be adjusted to be conservative; alternatively, realism may be increased by adjusting conditions in the tests used for parameterization or by using more complex toxicology modules that explicitly simulate the effects of additional stressors.

Based on the experience of research projects on the effects of environmental factors on TKTD processes, the uncertainty of using TKTD models within population models applied under dynamic and diverse conditions needs to be addressed in more detail before concrete suggestions can be made for testing.

Finally, the specific regulatory model should be subjected to a sensitivity analysis with scenario-specific parameters, initial conditions and driving variables, that investigates their influence on the model output (see Figure 2.3). Additionally, uncertainty in the values of specific parameters should be analyzed. Then, the specific regulatory model may be applied to address the specific risk assessment question, and model predictions should be accompanied with an uncertainty analysis.

Taken together, the outlined scheme suggests first an evaluation of a generic effect model (or modules) for potential use in addressing risk assessment questions. This is followed by an evaluation of a specific regulatory model application for a given risk assessment question. The scheme may be considered as an interim approach to pave the way for a consistent development of full regulatory models as outlined in EFSA PPR (2014).

2.4. Evaluation of effect models for use in regulatory risk assessment

The following subsections aim at giving a short overview of aspects of the evaluation of MEMs from a regulatory point of view. Most of these aspects are discussed more extensively in following book chapters.

2.4.1. Model suitability and applicability in relation to problem formulation and regulatory question

A scientifically excellent model may not be suitable for a risk assessment question if the model does not match the conditions that need to be addressed in the risk assessment. Such a mismatch can be for example the selection of the modeled species, the model output, or the design of the conceptual model or other aspects. In addition, it is important that a model is used within its domain, that is to say, that a model is used within the boundaries for which it was designed, and that this domain corresponds to the domain addressed in the risk assessment.

The following questions allow the identification of key requirements for the use of MEMs in regulatory risk assessment are given, that can help evaluate whether MEMs are "fit for purpose." Many of these aspects have been identified already in the EFSA scientific opinion on GMP (EFSA PPR, 2014).

- Has a relevant species been selected?
- Is the model type appropriate for addressing the protection goal (e.g., spatially vs. non-spatially explicit model)?
- Is the model appropriately designed for the temporal and spatial scale at which it is intended to be applied? Does it cover all processes that are relevant at the intended spatial and temporal scale and are relevant for the risk assessment?
- Does it include all relevant ecological or physiological traits, i.e., those traits that have an influence on either exposure or effects (e.g., survival, reproduction, movement, regulatory mechanisms)?
- Is there any major structural uncertainty? (This could be addressed, e.g., by adding or removing a process such as density dependence; see also Chapter 7.)
- Is there a risk of bias toward underestimation of the real risk?
- Is the selection of model parameters and their values justified (Chapter 4)?
- Is the model able to detect effects with a given certainty to address the protection goals (for example, demonstration with a reference item)?
- Is a comprehensive model documentation available (e.g., in the TRACE format)?

The Pop-GUIDE (Raimondo et al., 2021) addresses in more detail the development and evaluation of conceptual models. The paper focuses on population modeling, especially for species relevant under the Federal Insecticide Fungicide and Rodenticide Act and Endangered Species Act in the USA, but the main concept and the recommendations are relevant for MEMs in general. A very relevant point in the Pop-GUIDE concept is the emphasis on communication of the conceptual model, and the idea to use conceptual model diagrams as communication toll for information and inclusion of stakeholders in the early model development phases.

The selection of the relevant species (first point above) is a regulatory question and not related to the reliability of a model. The species assessed in the lower tiers are surrogates for the species that could be exposed in the environment, so the standard test species are not automatically the species that should be considered for population modeling. The choice of a species to be modeled might be given already in the formulation of the regulatory question, and needs to be justified, including considerations whether the modeled species should be the one for which an unacceptable risk was concluded in lower tiers, or a most vulnerable species. Also, aspects of representativeness of modeled species should be considered.

The aim of using a model in relation to the specific protection goal (considering also the temporal dimension) should be very clear. A model should clearly address the risk assessment problem and should ideally relate to the protection goal (problem formulation step). The model's purpose in the risk assessment must be well explained (e.g., which refinements apart from modeling have been applied, what exactly is the added value of the model, whether it is used in an explorative way, etc.). In the case the model application does not address the relevant protection goal, this should be well justified and documented (Raimondo et al., 2018).

There are several decisions regarding the evaluation of model output that need to be taken: the specific protection goal (SPG), if recovery is an acceptable endpoint, how the control population is defined and assessed. The general protection goals are usually formulated rather vaguely (EC 1107/2009), and also SPGs are not often very precisely expressed (Nienstedt et al., 2012). Hence, it may be challenging to define model output that enables an assessment of the SPGs. To clarify these points, communication between the different relevant stakeholder groups early in the model development process could be useful, as also suggested in the Pop-GUIDE framework (Raimondo et al., 2021).

The requirements listed above are related to the use of models for the registration of plant protection products. However, as indicated in earlier sections of this chapter, the level of detail of evaluation and assessment of MEMs may depend on their use. Besides being used in the risk refinement, MEMs can be used in the regulatory risk assessment also in more explorative ways, for example, to identify margins of safety or quantify uncertainties related to SPGs (see section 1.2). In this context, criteria for the assessment of model quality can depend on the intended use of a model. A model that is agreed as being reliable (enough) in an academic context is not automatically appropriate for making real-world decisions, a fact that was painfully experienced by some authors of epidemiological models (Grimm et al., 2020) during the Covid-19 pandemic.

2.4.2. Parameterization and data selection

The parameterization of a model (i.e., giving values to model parameters) is preceded by data collection, usually by means of a literature review and the generation of experimental data. The quality of a model parametrization relies on the quality, and sometimes also quantity, of available data. The collection of data needs to be done systematically, as described by EFSA (2011) for public literature review. Most importantly, all steps of the review need to be properly documented, and information needs to be provided on which studies were used and which studies were omitted. The evaluation of studies should be endpoint-related and not study-related, because a study can be invalid for one purpose but useful and valid for another one. Key criteria for data selection are (i) reliability in the sense of considering measurement methods and a discussion of uncertainty if needed; and (ii) representativeness and relevance, including the experimental scenario, geographical regions, etc. (represented by data).

In a final step of a literature review, potential data gaps need to be identified, and possible solutions to be addressed, for example, whether missing data can be taken from a related species or if they can be generated by means of calibration. For example, parameters of models focusing on organism-level effects (such as TKTD models) are often calibrated using data from ecotoxicological tests in the laboratory. Likewise, ecological parameters of more complex models (e.g., population or community models) may need to be calibrated (Evans et al., 2013). This can involve fitting to observed data (e.g., number of litters per female per year), in case no direct observations can be used. In any case, calibration must be properly documented and discussed, and criteria for assessing the quality of the fit, comparable to the ones presented in the EFSA scientific opinion on TKTD modeling (EFSA PPR, 2018a), need to be defined and documented (see also Chapter 6 for more information). Data collection and model parameterization is more comprehensively discussed in Chapter 4.

In any mechanistic effect model beyond simple TKTD models, the parameterization of the ecological module (and accordingly, of the environmental scenario) is very important and often very laborious. In general, without a toxicological component a MEM simply forms an ecological model; and it is not considered that it should be changed when being applied for different risk assessment questions or for case studies that support model evaluation (e.g., sensitivity analysis, validation). This refers particularly to coefficients that quantify basic physiological and ecological processes such as the temperature-dependency of a growth rate. These parameters may be considered *generic*, that is to say, independent from a specific environmental scenario. The built-in, default parameterization of generic parameters is an important factor determining the domain of use of an ecology module. Other parameters and settings may be *specific* for an environmental scenario and therefore variable across applications of the ecology module to different environmental scenarios (e.g., information on crop type, farming practices, landscape structure, climate, food availability, and interactions with additional species). Thus, it is beneficial to identify generic and specific parameters of a module and to evaluate their parameterization separately. The evaluation of a MEM in general should include an analysis of the uncertainty in values of the generic model parameters. The available data for parameterization and validation is often heterogeneous because studies may have been conducted in various environments and under

different conditions. As a result, the environmental domain of an ecology module may be broader than the domain of a regulatory model as described in EFSA PPR (2014). For example, the physiological parameterization and validation of a population model of a small mammal may have been performed with field studies in several crops. Therefore, the default generic parameterization may be related to a given EU zone but not a particular crop. Similarly, the physiological parameterization and validation of the energy budget model in a TKTD model under varying conditions in the laboratory may be not specific for a particular crop or EU zone.

2.4.3. Environmental scenarios

The environmental scenario is a part of the regulatory model and combines aspects of exposure and ecological scenarios (see Chapter 3). They comprise abiotic, biotic, and agronomic factors, and spatial components (e.g., landscape configuration and connectivity characteristics of the ecosystem). Specific eco(toxico)logical information and physico-chemical properties of the pesticide are in a strict sense not considered as part of environmental scenarios for application in regulatory risk assessment, but certainly have an impact on some of the components of those. The environmental scenario has a large influence on the outcome of a risk assessment performed with a model, or, to put it more extremely, a model could not be run without a defined environmental scenario. This holds not only for ecological models applied for risk assessment, but any ecological model that takes into account environmental parameters as influencing factors, despite this might not always be documented in an explicit way. Hence, a well justified and appropriate definition of an environmental scenario and its documentation is of huge importance in the process of evaluating an effect modeling approach.

The environmental scenario is often separated in an exposure scenario for the fate model and an ecological scenario for the effect model (Figure 2.2), both of which should be consistent. It should be noted here, that selecting a scenario that maximizes exposure (worst case exposure scenario) does not necessarily maximize the possible effects (worst case ecological scenario). For example, a population may be less vulnerable at the time with highest predicted exposure than at another time during the year when lower exposure could therefore result in higher effects. This means, that exposure scenarios, which might have been defined in the past aiming at maximizing (peak) exposure, cannot automatically be used in the context of using MEMs but would need to be assessed and potentially refined considering also ecological aspects.

There are several aspects that play a role for the scenario definition. Abiotic environmental factors determine the "environmental matrix" that provides the grounds for chemical fate and transport processes, as well as for biological and ecological dynamics, and are hence part of the exposure scenario. The choice of the values of such factors, for example, temperature, soil composition and structure, water quality parameters, must be consistent between exposure and ecological scenarios. Abiotic factors are therefore often part of both the ecological and exposure scenario. Specific values for abiotic factors will

only in rare cases be clearly definable; in most cases, ranges rather than specific values will be available from data collections. Sensitivity analyses (Chapter 7) with such factors are useful to find out the impact especially of uncertain environmental factors on the model output. Other factors are part of the specific scenario for the application of a MEM in regulatory risk assessment, including application conditions as laid down within the good agricultural practice (GAP), for example, crop type and application method and times.

The environmental scenario also requires definitions of the spatial scale and resolution on which a model operates. For landscape models, this is linked to the definition of certain landscape structures, for example, mixtures of crops, non-crop-elements, habitats for the modeled species of different value, etc. The structural definition of a landscape can have a tremendous influence on the model outcome and hence on the level of conservatism in the model predictions (Topping et al., 2014; Wang et al., 2022). Therefore, a set of different landscape scenarios is recommended rather than a single choice.

The ecological scenario contains biotic factors, which define interactions with the biotic environment, for example, ecological interactions such as predation, competition, parasitism, etc. That does not always mean an explicit modeling approach but can also mean an explanation for why certain interactions were ignored, or how they might be considered via factors.

For the definition of an environmental scenario, two overarching principles are recommended (see Chapter 3). First, a combination of sensitivity analyses and consequent definitions of certain scenario factors should be applied. This means, a sensitivity analysis of environmental parameters that are considered by a specific model can be performed to identify which parameters are most sensitive, and which values result in certain levels of conservativeness. For example, the habitat quality can be a very sensitive parameter for a landscape-scaled population parameter, and high habitat quality can lead to high resilience and short recovery times of a population. This does not mean that global sensitivity analyses of a model using all possible determinants of an environmental scenario always need to be performed, because this may often not be feasible. A clear and justified explanation on factors being included or excluded from the scenario definition is required in any case. Second, especially for a number of most sensitive factors, single values are not defined but rather a range of possible values, ideally together with an estimation of the level of conservativeness and realism. Based on this, a set of environmental scenarios representative of the diversity of situations in the field should be addressed and used for risk assessment simulations. This diversity of field conditions could range from optimal to less favorable conditions, including realistic worst-case scenarios, to provide information on margins of safety for situations of different habitat qualities, or for different risk mitigation options. The definition of more or less favorable conditions could practically be done by the analysis of real landscape information, and could require considerable data analyses.

The specific chapter on the definition of environmental scenario (Chapter 3) works out these aspects in more details and provides a list of questions to support the definition of environmental scenarios.

2.4.4. Implementation and model verification

In the context of being used as tool in regulatory risk assessment, a MEM needs to be evaluated to verify that it performs according to the theoretical model description. There is no dedicated chapter on model verification in this book, so in this section we give only a few basic recommendations. Notably, code verification and software testing are broad fields (Sommerville, 2018), hence here only some points are summarized, which are also discussed in the EFSA scientific opinion on GMP (EFSA PPR, 2014):

- Have the model equations been implemented correctly?
- Check against analytical solutions for special cases
- Identify central parts of code (e.g., translation of exposure into effects)
- Modular verification
- Visual debugging

In the EFSA GMP opinion (EFSA PPR, 2014), it is stated: "Verification of a model is the process of checking that the computer code correctly represents the conceptual and mathematical model (AIAA, 1998)". The TRACE format (Grimm et al. 2014) also differentiates between "Implementation verification" and "model output verification." These two elements relate to separate and relevant parts of model verification. Verification does not check the realism of the model." The latter process, checking the realism of the model output, is usually referred to by the term "validation" and is handled in section 2.4.5 below.

Implementation verification: The process of implementation verification needs to capture a number of aspects around the use of effect models for use in regulatory risk assessment, including checking correct implementation of mathematical equations, and checking in particular central elements of effect models (e.g., calculation of exposure and related effects), both elements at levels of source code. Ideally, in the conceptual and the formal model definition equations, flow diagrams or pseudo-code give a more or less detailed insight in how the modelers have implemented these aspects. Examples could be provided to demonstrate the correct implementation, for example, the correct calculation of the daily dietary dose could be checked for some selected situations by external calculations, or a flowchart could demonstrate how home range behavior is implemented. Source code can be checked based on this information, provided that the model code is accessible, and readable, at least the respective parts. In addition to verifying the implementation of code lines or formulae, it is also useful to verify functions (components or modules) of the model separately. Also, groups of components can be tested as units at higher hierarchical levels.

Depending on the nature of a model, for example whether it is an IBM, or a model based on differential equations such as DEB-TKTD, it might also be a relevant option to test the stability of numerical solutions of differential equations and the impact of rounding errors in calculations.

In case models have constrained complexity, it might be possible to reimplement a model or module independently in another software and to use a check of consistence in model outputs as indication of correct implementation.

Model output verification: Within the TRACE format (Grimm et al., 2014), model output verification is one of the elements of model testing. The review of the model outputs does not need to be quantitative, but rather it needs to focus on general patterns and principles. For example: Do the state variables remain in realistic ranges, for example, no negative or too high abundances? Is the seasonal timing abundance maxima or, more generally, the pattern of population dynamics, realistic? Do populations go extinct without food? If nutrient cycles are modeled, is the mass balance constant? Such checks can be carried out with initial model parameters and before fine-tuning the model by calibrating the parameters to a set of observational data. Thus, model output verification is different from model output corroboration in the TRACE terminology (Grimm et al. 2014), where model predictions are compared to independent data and patterns that were not used for model development before. The latter is often called validation, and is, together with calibration, discussed in detail in section 2.4.5 and Chapter 6.

2.4.5. Validation

Historically, many different terms have been used to describe the process of systematically testing whether a model adequately predicts the "real world." Often, the term "validation" is used as defined by Goodall (1972), who referred to validation as the process of testing to determine the degree of agreement between model predictions and the real system. This term has also been used in the EFSA scientific opinion on good modeling practice (EFSA PPR, 2014). Many other terms have been used, such as model testing or "evaluation." The latter term was introduced to cover also additional parts of model evaluation such as sensitivity analyses (Augusiak et al., 2014; Beven & Lane, 2019; Mankin et al., 1977; Oreskes et al., 1994; Rastetter, 1996; Rykiel, 1996). It is important to acknowledge that model validation can only falsify (prove wrong) but never prove that a model adequately reflects all reality. This is because our knowledge in general is limited, and referring to Karl Popper, a hypothesis can never be proven but only be falsified. In other words, we cannot test whether a model performs realistically under all possible conditions. Moreover, a model is, by definition, a simplified description of reality and thus always wrong to a certain extent. Hence, the requirement that a model should be "true" to be considered valid is not achievable. A more useful definition of a valid model would be that a model shall be "good enough" (Jarvis & Larsbo, 2012) or "fit for purpose." Comprehensive validation experiments can show this. However, also our knowledge of the ecology of a species is always limited. Hence, the best validation analysis can only reveal whether a model accurately reflects our knowledge of a species.

Before a model can be used to support regulatory decisions, its ability to predict the relevant aspects of the "real world" sufficiently well should be tested. For population models, such testing can include

population abundance dynamics, habitat preferences, and behavioral aspects of the model species such as spatial behavior. For DEB-TKTD models, testing could include the correct prediction of individual growth (weight, length), maturation, and/or reproduction as observed in an ecotoxicological experiment that has not been used for model calibration. This way, the performance of a MEM in predicting the development and/or behavior of a modeled organism or population in a (possibly non-contaminated) reference scenario can be assessed, and the MEM may be validated, indirectly assessing structural integrity and ecological realism of a model. Additionally, in the context of risk assessment it is of particular importance to assess the performance of a MEM in predicting effects of pesticides on the modeled organism or population, hence to validate the MEM also concerning the simulation of the pesticide impact. This can be done for example, by checking if predicted differences between the reference and a pesticide treatment scenario match experimental or monitoring observations.

To demonstrate the structural integrity of an ecology module, validation may address various states or variables even if they are not directly relevant from a regulatory point of view. However, validation in the context of using MEMs for regulatory purposes should put focus on regulatory relevant endpoints to demonstrate the realism of mediate effects, that is to say, effects that emerge on higher levels of organization in the ecology module. Of course, the evaluation of the ecological model or module as such should show that the modeling results in realistic abundances, dynamics, and patterns. However, the use of an ecology module to be used in the ERA of chemicals is adding another aspect for assessment, that is to say, whether the model propagates immediate effects to effects on respective higher entities (individual, population, community) in a realistic way. Ideally, this is tested with observations in or close to natural environments. Studies for such validation are not necessarily limited to pesticides: For example, populations may be subject to known levels of hunting or harvesting (immediate effect of increased mortality) or eggs may be removed (decreased reproduction) to assess whether the predicted potential for compensation and recovery in a population model matches observation under realistic biotic and abiotic conditions. Similarly, organisms may be subjected to food limitation to assess whether the energy budget of a TKTD model predicts realistic consequences of decreased energy intake on growth and reproduction.

The next subsections shed some light from the regulatory point of view on relevant aspects for the performance of validation studies, on the selection of model components that are subject for validation, and the special role of effect predictions. In more detail and with a more technical view, validation aspects are discussed in detail in Chapter 6.

2.4.5.1. How validation can be done

When conducting a validation study (e.g., reproducing a field study with a model to compare model results with field measurements), some general considerations are necessary. It needs to be clear for which domain (or area of applicability) a model is to be validated. For example, a model may be perfectly validated for central Europe, but that does not allow conclusions about the validity of the model for other regions. A key

aspect for validation is that the validation cannot use the same data that were used for model calibration. Nevertheless, this does not mean that data cannot be used if originating from a study that has also been considered for parameterization, because even when data were taken from the same study, they can still be independent from each other. For example, non-contaminated control ponds in a mesocosm study may be used to parameterize basic physiological and ecological processes in the ecology module of a MEM; then the direct effects at organism-level may be parameterized with laboratory toxicity tests, and the predicted effect propagation to the population level can be tested with observed effects in the treated test units of the mesocosm studies.

Validation is suitable and required to assess realism in the model structure and the associated model output, not in model input. Input parameters and datasets that have been parameterized with experimentally derived values already reflect reality, or at least a snapshot of it. It may of course be debatable whether measured parameters are representative for conditions not covered in experiments, but in principle it is not necessary to validate measured data, which were *imposed* in a model. In contrast, the "behavior" of a mechanistic model output *emerges* from the interaction of one or various parameters and state variables. These emerging model behaviors or emerging "patterns" are very useful for validation. Particularly useful for validation are "patterns" (pattern-oriented modeling; [Grimm et al., 1996; Wiegand et al., 2003; Wiegand et al., 2004]; see also Chapter 6). Because such patterns often emerge from the interaction of several parameters or processes, they give a particularly meaningful insight into the realism of a model. For example, in a mammal population model, reproductive parameters such as the duration of the breeding season, gestation length, lactation length, estrus, litter size, will affect many different outputs of a model such as recruitment, age structure, population dynamics, number of offspring per female and season. Only a correct combination of parameters and processes together with a correct mechanistic implementation of reproduction will result in correct reproduction of these patterns in a model. For instance, if the breeding season is artificially prolonged, then the timing of the seasonal population peak will be affected, the number of offspring per female and season will increase and age structure will change, resulting in an unrealistic model prediction. Models with a large number of parameters, processes and complex internal structures allow the conduct many types of validation experiments, because a wide range of emerging output behaviors are available. This ultimately makes it easier to detect errors and then invalidate a model. For this, appropriate observation data is required, which might be available from field observations or field studies. It might also happen that appropriate observation data is not available or unrealistic to be collected (e.g., for large-scale landscape population models), which leaves the questions: To what extent this can be done in practice, who would collect these data, and, last but not least, whether such data collection should be seen mandatory before respective MEMs could be used in regulatory ERA. As already stated in the introduction of Chapter 1, these questions are still open and controversially discussed.

2.4.5.2. What needs to be validated?

For complex models, it may be very time consuming to conduct validation experiments for each part of a model. On the other hand, when using models in the ERA of pesticides, it is important to include the parts of a model that affect the outcome of a risk assessment (Wang & Luttik, 2012). Here, the model type determines which parts of the model can affect the risk assessment outcome. For population models, this can include spatial behavior (influence on exposure), or survival and reproduction (influence on recovery). Examples of processes that might not be relevant for the purpose of a risk assessment might be migratory behavior in birds or vertical movement of aquatic organisms, as long as this behavior will not affect exposure to a pesticide. For DEB-TKTD models, validation could include growth patterns observed with or without the presence of a toxicant at different or variable food or temperature conditions.

For complex models, a modular validation approach could be useful, that is to say, different submodels (modules) could be validated separately to cope with complexity. When conducting validation in a modular way, this implicitly also includes a structural review of the model, which could reveal whether important processes may be missing. The assessment and validation of modular models is further discussed in Chapter 5.

From a more technical view, an entire species model (e.g., in the case of population models) or an organism-level effect model can be seen as modules. Here, the validation of a species model can be rather complex, including the collection and use of ecological and environmental information, while the validation of an organism-level effect model might be simpler, because only individual effects need to be predicted correctly (e.g., effects seen in laboratory studies in individual animals). In this sense, a modular approach can make the model evaluation process more efficient. For example, once a species model has been validated, only the effect model may need to be re-validated for different substances.

For validation, it is crucial to take into account reliability and methodological bias of empirical data. Hence a detailed review of literature (EFSA, 2011; Gergs, Classen, et al., 2016) is not only required for model parameterization but also for validation.

Ideally, validation of a population model would show that a model reflects our complete knowledge about a species. This includes species interactions with the environment and responses to exposure to a pesticide. A comparison between modeling results and empirical observations would require empirical data that were considered at least "reliable." Nevertheless, even a partially validated model might be helpful to answer specific questions, as long as model parts that are critical for conclusions in risk assessment were validated carefully. More concrete suggestions and recommendations how TKTD models for birds and mammals, population models, or community/food-web models can be validated are beyond the scope of this book and remain to be done by a working group with a mandate.

2.4.5.3. Ability to predict effects

The prediction of effects plays a crucial role for the assessment with regard to a possible regulatory use of a MEM. In general, the realism in the predicted effects of pesticides should be assessed by validation studies. In many cases, such validation may be difficult due to the absence of data (e.g., effect data at large spatial scales). However, whenever information can be provided that indicates whether the effect part of a model reflects reality, it should be used. In addition, analogous to laboratory or field studies, data from toxic references (e.g., earthworm or non-target arthropod studies) could be used for model simulations (Johnston et al., 2014). This could be done, for example, if a toxic reference is used in field trials. However, if the model includes modules that are restricted to a mode of action that is different from that of the toxic reference, this approach cannot be used. If toxic reference data are not available, for example in field studies on birds and mammals, the application rates could be increased until they trigger effects in the modeling results, ideally in a stepwise manner to understand the mechanisms and determine the onset of effects. Such analyses, also referred to as exposure modification factors (EMFs, see also section 3.3.3), could also be considered as a mandatory part of the risk assessment. The ability to predict effects through the use of toxic references or increased application rates should be part of the dossier when a model is submitted.

In the future, it would be valuable if data from incidences or misuse could be collected systematically and made available for validation of effect models. These datasets would not need to be limited to pesticide use but might include other stressors that lead to clear effects on a population level. An example of such a collection of observational data regarding different stressors in aquatic systems is provided by Gergs, Gabsi, et al. (2016).

In the field of environmental risk assessment, a validated model is not necessarily a model that gives precise predictions of the observations in the validation studies. It can be sufficient for using the model if the predictions were conservative. For example, if TKTD modeling is used to assess effects of short-term exposure events, a model that overestimated the effects in the validation experiments may still be used for a refined risk assessment. This is, because the TKTD model might conservative but more realistic than a Tier 1 approach, where a maximum PEC is compared to an acceptable effect level derived from an experiment with constant exposure over the test duration. There might be cases as well, where a model is over-predicting some treatments, while under-predicting others, in that case it holds that a model should not be used for conditions where it has been found to be not protective.

2.4.6. Model sensitivity and uncertainty analysis

Sensitivity analysis and uncertainty analysis are the two main tools used in exploring the uncertainty of mathematical models (Saltelli et al., 2019). There is definitely the need to elaborate on sensitivity and uncertainty analyses in context of regulatory risk assessment using MEMs, but more concrete perspectives on what exactly regulators need to evaluate and what is required to have an acceptable model analysis would

require a dedicated working group and was beyond the scope of the MAD group discussions. This section gives a short overview about these concepts, while in in Chapter 7 they are introduced and discussed in more detail, including a simple example for the procedure of uncertainty and sensitivity analysis.

2.4.6.1. Sensitivity analysis

According to Saltelli (2002), one definition of sensitivity analysis (SA) is "the study of how the uncertainty in the output of a model (numerical or otherwise) can be apportioned to different sources of uncertainty in the model input." As such, it is very much related to – but distinct from – uncertainty analysis (UA), which, as we define it here, characterizes the uncertainty in model prediction, without identifying which assumptions are primarily responsible. In other words, SA is backward-oriented (tracing the uncertainty back to the sources), while UA is forward-oriented (propagation of uncertainty). Sensitivity analysis is the analysis of how different sources of uncertainty affect the uncertainty in the model output.

While model parameters are the most common items to include in a SA, apart from parameters one can also vary many other assumptions: for example, initial conditions, user input variables like pesticide application rates and dates, model structure (e.g., switch on/off certain processes) or even driving data such as temperature. Saltelli et al. (2019) therefore use the more general term "input factors" for the items that are varied in a SA or UA.

The EFSA SO on the state of the art of TKTD effect models (EFSA PPR, 2018a) uses a slightly different definition of SA: "Sensitivity analyses are a tool to understand how model outputs respond to changes in parameter values. They allow the modeler to distinguish less influential parameters (that could be fixed at point values without any major change in model outputs) from the most influential parameters. The most influential parameters should receive most attention when calibrating the model." Influential parameters are those for which small variations significantly affect the model results (Ciric et al., 2012). Non-influential parameters can be fixed at a nominal value without significantly reducing the variance of the output, thereby making the calibration step less complex (Saltelli et al., 2005).

Both for Saltelli et al. (2019) and EFSA PPR (2018a) the aim of an SA is to rank the model inputs (parameters, initial conditions, driving data, etc.) with respect to their influence on the model output, and to identify non-influential ones that could be fixed (i.e., whose uncertainty can be ignored) when calibrating the model or making predictions. For example, a population model on a bird species may include many different parameters regarding reproduction (time of breeding, clutch size, no. of broods, etc.), survival (density-dependent) and territorial behavior. A very detailed quantification of the uncertainty of all parameters may be very time consuming. A sensitivity analysis may show, for example, that the single most influential parameter determining population density is density-dependent survival. In this case one could conclude that the uncertainty regarding density-dependent survival really needs to be addressed in detail for each prediction.

For risk assessment purposes, SA are usually carried out for the following two different types of target variables:

- The raw model output (e.g., % survival, biomass, population density, other population parameters)
- The aggregated result of the risk assessment (e.g., no adverse effects at the population level)

Methods for SA can be broadly divided into local and global methods. A local SA is a sensitivity analysis around a fixed-base scenario (i.e., parameter set and initial conditions), that is to say, a single point in the parameter space. A local SA consists in varying only one input parameter value from one simulation to the next (one at a time) and then returning to the base scenario. Hence, all simulations (parameter sets) differ from the base scenario only by one parameter value. Sensitivity is then usually computed as an approximation of the first partial derivative, that is to say, change of the target model output variable divided by change of the value of the varied parameter (Brylinsky, 1972).

Lauvernet and Muñoz-Carpena (2018) define Global Sensitivity Analysis (GSA) as follows: "The 'global' term denotes that GSA studies output variability when all input factors vary globally, within their validity domain defined by probability distribution functions (PDFs), as opposed to locally, (one at a time), that is to say, around an arbitrary range from a base value. GSA allows for simultaneous estimation of the factors' individual importance (first order effects) and interactions (higher order effects)." In a global sensitivity analysis, the input factors (parameters, initial conditions, model switches, etc.) are usually sampled from probability distributions, using different sampling strategies. In contrast to local SA, GSA allows taking parameter interactions and non-linearity into account. GSA can thus be more complex but is considerably more useful than local SA. However, while GSA methods are common in exposure modeling, they have not been sufficiently adopted in effect and ecological modeling. An illustrative example is given, however, by Accolla et al. (2022) at for IBMs including a DEB module. Further details on sensitivity analysis can be found in section 7.3.2.

2.4.6.2. Uncertainty analysis

Uncertainty can refer to parameter uncertainty, structural uncertainty and other types of uncertainty (for an overview see Li & Wu [2006]). Uncertainty is covered in more detail in Chapter 7, however, here we address uncertainty already in the general context of its relevance for using mechanistic models.

Parameter uncertainty describes how reliably parameters have been measured or estimated. Uncertainty must not be confused with variability. While variability refers to the natural variation present in the real world (e.g., body weight naturally varies), parameter uncertainty refers to the degree of certainty with which we know a parameter.

Structural uncertainty refers to the degree of confidence one can have in the model structure. For example, when important processes are not considered in an ecological model, then the model output may not be reliable. In other words, uncertainty is not only "visible," uncertainty that can be seen in empirical data or quantified in model output, for example, using of confidence bands, but it may also be hidden (due to a lack of knowledge) when relevant processes are not included. However, the non-inclusion of processes does not necessarily mean that a model is less reliable. In some cases, the non-inclusion of processes in a model may be worst-case, in others it may be best-case.

Often, when evaluating mechanistic models, uncertainty becomes explicit, that is to say, it becomes visible which parameters have been well measured and which have not. This sometimes leads to the subjective conclusion that uncertainty is introduced in the risk assessment by using models. However, although such uncertainty becomes more explicit when evaluating models, it is not new. For example, default body weights used for the risk assessment of birds and mammals (EFSA, 2009) include considerable uncertainty. But this is not of concern in a standard risk assessment, because it is agreed that these weights are used for the risk assessment. The same applies for food intake rate and many other values. By explicitly showing uncertainties, models can help to understand risk, and they can be used to address or sometimes even remove uncertainties. This can be done by asking "what if" questions, such as "what if a feeding rate was higher or body weight was lower?" or "what would happen in another landscape?" This capacity of mechanistic models makes them particularly useful for the risk assessment and for other types of applications, for example, exploration of mitigation options, exploration of effects occurring at other time scales than those usually tested experimentally. When input parameters include uncertainties, models can be run with different values to explore the consequences of this uncertainty.

Finally, uncertainties and benefits of using a model need to be balanced: A model will never remove all uncertainties in the risk assessment, and a model will only be a representation of the reality and not reality (by contrast with the results of a semi-field study that provides real results specific for the specific set of conditions that were represented in the study). Hence validation is important for modeling. Even if not all parts of a model can be validated in a particular case, using the model may still help to address uncertainties in risk assessment that cannot be addressed otherwise (for example, extrapolation from the laboratory to the field; extrapolation to different field conditions such as different climatic zones, as empirical studies cannot be done for every situation). For further details, see also Chapter 6.

To assess uncertainty in a modeling endpoint of interest, Monte-Carlo simulations can be used to propagate known or expected variation in all (or at least in relevant) model parameters to variation in that endpoint. Uncertainty in a modeling endpoint can be visualized then by showing for example, the deterministic predictions together with a prediction band (or interval, if only a single timepoint is considered) calculated from a number of Monte-Carlo simulations. In Chapter 7**,** approaches for UA are introduced and discussed in more detail, and in section 8.5.2 related examples are given.

2.5. Bibliography Chapter 2

Accolla, C., Schmolke, A., Jacobson, A., Roy, C., Forbes, V. E., Brain, R., & Galic, N. (2022). Modeling pesticide effects on multiple threatened and endangered cyprinid fish species: The role of life-history traits and ecology. *Ecologies 3*, 183-205. https://doi.org/10.3390/ecologies3020015

AIAA. (1998). *Guide for the verification and validation of computational fluid dynamics simulations*. www.aiaa.org/StandardsDetail.aspx?id=3853

Augusiak, J., Van den Brink, P. J., & Grimm, V. (2014). Merging validation and evaluation of ecological models to 'evaludation': A review of terminology and a practical approach. *Ecological Modelling*, *280*, 117-128. https://doi.org/10.1016/j.ecolmodel.2013.11.009

Beven, K., & Lane, S. (2019). Invalidation of Models and Fitness-for-Purpose: A Rejectionist Approach. In C. Beisbart & N. J. Saam (Eds.), *Computer Simulation Validation: Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives* (pp. 145-171). Springer International Publishing. https://doi.org/10.1007/978-3-319-70766-2_6

Brylinsky, M. (1972). Steady-State Sensitivity Analysis of Energy Flow in a Marine Ecosystem. In B. Patten (Ed.), Mathematical modelling of water quality: streams, lakes and reservoirs (Vol. Volume II, pp. 591). Academic Press.

Ciric, C., Ciffroy, P., & Charles, S. (2012). Use of sensitivity analysis to identify influential and non-influential parameters within an aquatic ecosystem model. *Ecological Modelling*, *246*, 119-130. https://doi.org/10.1016/j.ecolmodel.2012.06.024

EFSA (European Food Safety Authority). (2009). Risk assessment for birds and mammals. *EFSA Journal*, *7*(12), 1438. https://doi.org/10.2903/j.efsa.2009.1438

EFSA (European Food Safety Authority). (2011). Submission of scientific peer-reviewed open literature for the approval of pesticide active substances under Regulation (EC) No 1107/2009. *EFSA Journal*, *9*(2), 2092. https://doi.org/10.2903/j.efsa.2011.2092

EFSA PPR (Panel on Plant Protection Products & their Residues). (2014). Scientific opinion on good modelling practice in the context of mechanistic effect models for risk assessment of plant protection products. *EFSA Journal*, *12*(3), 3589. https://doi.org/10.2903/j.efsa.2014.3589

EFSA PPR (EFSA Panel on Plant Protection Products and their Residues). (2018a). Scientific opinion on the state of the art of toxicokinetic/toxicodynamic (TKTD) effect models for regulatory risk assessment of pesticides for aquatic organisms. *EFSA Journal*, *16*(8), 5377-5565. https://doi.org/10.2903/j.efsa.2018.5377

Evans, M. R., Grimm, V., Johst, K., Knuuttila, T., de Langhe, R., Lessells, C. M., Merz, M., O'Malley, M. A., Orzack, S. H., Weisberg, M., Wilkinson, D. J., Wolkenhauer, O., & Benton, T. G. (2013). Do simple models lead to generality in ecology? *Trends in Ecology & Evolution*, *28*(10), 578-583. https://doi.org/10.1016/j.tree.2013.05.022

Gergs, A., Classen, S., Strauss, T., Ottermanns, R., Brock, T. C. M., Ratte, H. T., Hommen, U., & Preuss, T. G. (2016). Ecological Recovery Potential of Freshwater Organisms: Consequences for Environmental Risk Assessment of Chemicals. In P. DeVoogt (Ed.), *Reviews of Environmental Contamination and Toxicology, Vol 236* (Vol. 236, pp. 259-294). Springer. https://doi.org/10.1007/978-3-319-20013-2_5

Gergs, A., Gabsi, F., Zenker, A., & Preuss, T. G. (2016). Demographic toxicokinetic-toxicodynamic modeling of lethal effects. *Environmental Science & Technology*, *50*(11), 6017-6024. https://doi.org/10.1021/acs.est.6b01113

Goodall, D. W. (1972). Building and testing ecosystem models. In J. N. J. Jeffers (Ed.), *Mathematical Models in Ecology* (pp. 173-194). Blackwell.

Grimm, V., Augusiak, J., Focks, A., Frank, B. M., Gabsi, F., Johnston, A. S. A., Liu, C., Martin, B. T., Meli, M., Radchuk, V., Thorbek, P., & Railsback, S. F. (2014). Towards better modelling and decision support: Documenting model development, testing, and analysis using TRACE. *Ecological Modelling*, *280*, 129-139. https://doi.org/10.1016/j.ecolmodel.2014.01.018

Grimm, V., Frank, K., Jeltsch, F., Brandl, R., Uchmanski, J., & Wissel, C. (1996). Pattern-oriented modelling in population ecology. *Science of the Total Environment*, *183*(1-2), 151-166. https://doi.org/10.1016/0048-9697(95)04966-5

Grimm, V., Johnston, A. S. A., Thulke, H. H., Forbes, V. E., & Thorbek, P. (2020). Three questions to ask before using model outputs for decision support. *Nature Communications*, *11*(1), 3, Article 4959. https://doi.org/10.1038/s41467-020-17785-2

Jarvis, N., & Larsbo, M. (2012). MACRO (v5.2): Model use, calibration, and validation. *Transactions of the ASABE*, *55*(4), 1413-1423. https://doi.org/10.13031/2013.42251

Johnston, A. S. A., Hodson, M. E., Thorbek, P., Alvarez, T., & Sibly, R. M. (2014). An energy budget agent-based model of earthworm populations and its application to study the effects of pesticides. *Ecological Modelling 280*, 5-17. https://doi.org/10.1016/j.ecolmodel.2013.09.012

Li, H.-t., & Wu, J. (2006). Uncertainty analysis in ecological studies: An overview. In (pp. 45-66). https://doi.org/10.1007/1-4020-4663-4_3

Mankin, J., O'Neill, R., Shugart, H., & Rust, B. (1977). The importance of validation in ecosystem analysis. *New directions in the analysis of ecological systems, part*, *1*, 63-71.

Nienstedt, K. M., Brock, T. C. M., van Wensem, J., Montforts, M., Hart, A., Aagaard, A., Alix, A., Boesten, J., Bopp, S. K., Brown, C., Capri, E., Forbes, V., Köpp, H., Liess, M., Luttik, R., Maltby, L., Sousa, J. P., Streissl, F., & Hardy, A. R. (2012). Development of a framework based on an ecosystem services approach for deriving specific protection goals for environmental risk assessment of pesticides. *Science of the Total Environment*, *415*, 31-38. https://doi.org/10.1016/j.scitotenv.2011.05.057

Oreskes, N., Shraderfrechette, K., & Belitz, K. (1994). Verification, Validation, and confirmation of numerical-models in the earth-sciences. *Science*, *263*(5147), 641-646. https://doi.org/10.1126/science.263.5147.641

Raimondo, S., Etterson, M., Pollesch, N., Garber, K., Kanarek, A., Lehmann, W., & Awkerman, J. (2018). A framework for linking population model development with ecological risk assessment objectives. *Integrated Environmental Assessment and Management*, *14*(3), 369-380. https://doi.org/10.1002/ieam.2024

Raimondo, S., Schmolke, A., Pollesch, N., Accolla, C., Galic, N., Moore, A., Vaugeois, M., Rueda-Cediel, P., Kanarek, A., Awkerman, J., & Forbes, V. (2021). Pop-GUIDE: Population modeling guidance, use, interpretation, and development for ecological risk assessment. *Integrated Environmental Assessment and Management*, *17*(4), 767-784. https://doi.org/10.1002/ieam.4377

Rastetter, E. B. (1996). Validating models of ecosystem response to global change. *Bioscience*, *46*(3), 190-198. https://doi.org/10.2307/1312740

Rykiel, E. J. (1996). Testing ecological models: The meaning of validation. *Ecological Modelling*, *90*(3), 229-244. https://doi.org/10.1016/0304-3800(95)00152-2

Saltelli, A. (2002). Sensitivity analysis for importance assessment. *Risk Analysis*, *22*(3), 579-590. https://doi.org/10.1111/0272-4332.00040

Saltelli, A., Aleksankina, K., Becker, W., Fennell, P., Ferretti, F., Holst, N., Li, S. S., & Wu, Q. L. (2019). Why so many published sensitivity analyses are false: A systematic review of sensitivity analysis practices. *Environmental Modelling & Software*, *114*, 29-39. https://doi.org/10.1016/j.envsoft.2019.01.012

Saltelli, A., Ratto, M., Tarantola, S., & Campolongo, F. (2005). Sensitivity analysis for chemical models. *Chemical Reviews*, *105*(7), 2811-2828. https://doi.org/10.1021/cr040659d

Sommerville, I. (2018). Software engineering.

Topping, C. J., Kjaer, L. J., Hommen, U., Høye, T. T., Preuss, T. G., Sibly, R. M., & van Vliet, P. (2014). Recovery based on plot experiments is a poor predictor of landscape-level population impacts of agricultural pesticides. *Environ Toxicol Chem*, *33*(7), 1499-1507. https://doi.org/10.1002/etc.2388

Wang, M., & Luttik, R. (2012). Population level risk assessment: practical considerations for evaluation of population models from a risk assessor's perspective. *Environmental Sciences Europe*, *24*(1), 3. https://doi.org/10.1186/2190-4715-24-3

Wang, M., Park, S.-Y., Dietrich, C., & Kleinmann, J. (2022). Selection of scenarios for landscape-level risk assessment of chemicals: case studies for mammals. *Environmental Sciences Europe*, *34*(1), 35. https://doi.org/10.1186/s12302-022-00612-4

Wiegand, T., Jeltsch, F., Hanski, I., & Grimm, V. (2003). Using pattern-oriented modeling for revealing hidden information: a key for reconciling ecological theory and application. *Oikos*, *100*(2), 209-222. https://doi.org/10.1034/j.1600-0706.2003.12027.x

Wiegand, T., Revilla, E., & Knauer, F. (2004). Dealing with uncertainty in spatially explicit population models. *Biodiversity & Conservation*, *13*(1), 53-78. https://doi.org/10.1023/B:BIOC.0000004313.86836.ab

3. Development of environmental scenarios for the application of mechanistic effect models in environmental risk assessment

Sanne van den Berg, Sabine Duquesne, Oliver Jakoby, Udo Hommen, Thomas Preuss, Magnus Wang, Alpar Barsi, Stefan Reichenberger, Andreas Focks

"A model without a scenario is like a fish without water"

– Theo Brock

3.1. Aims and scope of this chapter

Mechanistic effect models (MEMs) offer a platform for comprehensive analyses of the environmental risk of pesticides at different biological, spatial, and temporal scales. An extensive introduction into MEMs and their potential use in regulatory risk assessment for pesticides is provided in Chapter 1. In short, MEMs provide predictions of the environmental effects of pesticides, or other chemical or non-chemical stressors based on mechanistic principles implemented in the model. These mechanistic principles are based on physical, chemical, and biological rules. Because many physical, chemical, and biological rules are immediately influenced by environmental conditions like temperature, pH, and food availability, environmental scenarios describing these environmental conditions are required when using MEMs to assess the effects of chemicals on the environment in a meaningful way.

An environmental scenario is defined as a combination of abiotic, biotic, and agronomic conditions and may also include landscape properties (Figure 3.5). We can further differentiate the environmental scenario into an ecological and an exposure scenario, with some environmental factors or conditions being part of only exposure or ecological scenarios, and some others (mainly abiotic factors, like temperature, precipitation, or landscape conditions) being part of both. Specifically, the fact that some environmental factors or conditions have an influence on both exposure and ecological scenarios indicates the need for a consistent definition of environmental scenarios. Consider, for instance, that exposure scenarios are often intended to represent realistic worst-case conditions, but that a realistic worst-case exposure

scenario does not necessarily result in a realistic worst-case environmental scenario with respect to the resulting effects. It might be that the exposure-scenario setting is not suitable for the species in focus or that other environmental conditions result in more harmful effects. For example, a mismatch between the occurrence of the most sensitive species (or life stage) and the worst-case exposure scenario could lead to an incorrect assessment of potential effects. In addition, variation in environmental conditions can cause less than worst-case exposure scenarios result in more severe effects, for instance due to factors like food availability or temperature.

Environmental scenarios can be defined in many different ways, because the way to set up such scenarios is not (yet) standardized. In addition, no extensive guidance on the formulation and evaluation of environmental scenarios as part of the model evaluation is available. Therefore, this chapter reviews the ways in which exposure, ecological, or environmental scenarios have been defined in the past in the context of environmental risk assessment (described in section 3.2 and summarized in Table 3.2), and describes our vision on how environmental scenarios could be defined and evaluated in the future (section 3.4). For the latter, we suggest and explain a catalogue of questions that can serve two purposes: (i) to help the modeler set up a well-founded, comprehensive, and structured set of environmental scenarios to be used in evaluating a MEM for risk assessment, and (ii) to help the risk assessor conduct a comprehensive evaluation of the environmental scenarios used to perform the risk assessment (section 3.4). Finally, we provide three examples of how environmental scenarios can be defined when assessing chemical risks using MEMs.

3.2. Strengths and weaknesses of existing approaches for environmental scenario development

The main aim of this section is to obtain an understanding on why we need environmental scenarios, and how environmental, ecological, and exposure scenarios, or aspects of each, have been defined in the context of environmental risk assessment. In this context, we present Table 3.2, which provides an overview of the most relevant documents currently available that touch on the topics of environmental, ecological, or exposure scenarios in the context of chemical risk assessment. We shortly summarize each document and extract the main lessons we can learn from them, focusing on aspects like (i) how are the scenarios defined, (ii) which factors have been considered in the derivation of scenarios, (iii) which links exist between fate and ecology in these scenarios. To avoid ambiguity, relevant documents are discussed in chronological order. Finally, this chapter serves as a starting point to describe how environmental scenarios should ideally be defined and evaluated in the future.

3.2.1. FOCUS surface water exposure scenarios

The FOCUS (*FOrum* for the Coordination of pesticide fate models and their Use) surface water exposure scenarios have been developed because of a need for standardized exposure scenarios for the European Union (FOCUS, 1995). The background of the 10 FOCUS Step 3 exposure scenarios is summarized by FOCUS (2001) in section 3 as follows:

> *In developing a set of scenarios for Step 3, the aim of the working group was to produce a limited number of "realistic worst-case" surface water scenarios, which were broadly representative of agriculture as practiced in the major production areas of the EU. These scenarios should take into account all relevant entry routes to a surface waterbody, as well as considering all appropriate target crops, surface water situations, topography, climate, soil type and agricultural management practices. The lack of comprehensive databases that characterize most of these agro-environmental parameters at a European level meant that it was not possible to select representative worst-case scenarios in a rigorous, statistically based manner. The group therefore adopted a pragmatic approach to selection, using very basic data sources together with expert judgement.*

Section 4.6.5 is decidedly cautious about the overall worst-case-ness of the exposure scenarios:

> *The various assumptions and "worst-case" assessments summarized above show that, for many of the scenario factors that determine the magnitude and duration of pesticide residues in water bodies, a 90th+ percentile worst-case has been adopted. In order not to create scenarios where worst-case conditions are unrealistically combined, other scenario factors are less severe and represent 50th to 70th percentile worst-cases. The FOCUS Surface Water Scenarios Group does not consider it statistically valid to attempt to integrate the various worst-case assessments into a single value. However, it considers that the 10 Step 3 scenarios […] will provide a realistic range of PECsw estimates that represent significant agricultural areas within Europe. The highest PECsw estimates from the ten scenarios are likely to represent at least a 90th percentile worst-case for surface water exposures resulting from agricultural pesticide use within the European Union.* (FOCUS, 2001)

In brief, there is no statistical substantiation that the FOCUS surface water exposure scenarios represent an overall 90th percentile worst-case for the entire European Union.

3.2.2. Scientific opinion on good modeling practice

The EFSA opinion on good modeling practice (EFSA PPR, 2014) recommends that when applying a model for regulatory purposes, relevant environmental scenarios need to be defined. In this context, EFSA defines environmental scenarios as, "representative combination of abiotic, biotic and agronomic parameters for the purpose of modeling" (Figure 2.1). Pesticide properties and use data directly feed the computer model and are therefore not considered a part of the environmental scenarios. Furthermore, EFSA defines that while the model should be realistic, the level of conservatism (defined in the specific protection goals, or SPGs) should be reflected in the (environmental) scenarios.

The opinion suggests a systematic approach to environmental scenario development (Figure 3.1). First, the SPGs and the level of conservatism should be both decided on in the problem definition step. The SPGs determine the appropriate spatial and temporal scales to be considered, as well as the suitable unit for each dimension considering the group of species of interest (e.g., aquatic vertebrate or invertebrate). However, the application of SPGs in risk assessment are still under development and are discussed in more detail in section 3.2.6 of this chapter. Next, a database outlining the relevant spatial and temporal scales (i.e., area and time frames) at the required resolution should be compiled. These data should subsequently be used to construct model inputs. An environmental scenario describing the biotic and abiotic conditions in the absence of pesticides is then selected. Once the environmental scenarios have been developed, pesticide exposure can be added to the model according to the chemical's characteristics and application schedule.

The opinion (EFSA PPR 2014) points out the difficulty of accounting for the spatial aspect of exposure for mobile species, which can move across large areas. It also highlights the complexity of applying a percentile-based approach when model inputs describe biotic, abiotic, and agronomic factors, which then interact with organism behavior. Due to the vast number of potential combinations, deriving distributions and selecting a percentile would be a challenging task. Therefore, approaches from fate modeling (e.g., FOCUS surface water exposure scenarios) are assumed to be most easily transferrable to models that consider individuals or stationary populations in a homogenous environment. For mobile species, such approaches are not found suitable because their ecology and behavior need to be considered together with the definition of spatio-temporal dynamic exposure profiles.

Overall, there is no set of fixed environmental scenarios that could be used for a broad range of models. Some general characteristics of an environmental scenario are given in the EFSA opinion, but the explicit environmental scenario needs to be tailored separately for each specific risk assessment question and requires a careful justification that it represents a realistic worst-case scenario.

3.2.3. Scientific opinion on non-target arthropods

According to the EFSA opinion on non-target arthropods (NTA: EFSA PPR, 2015), biodiversity must be supported to provide important ecosystem services. The opinion pragmatically defines NTA as (life stages of) terrestrial invertebrates that dwell primarily on the soil surface and/or the vegetation, whereas (life stages of) terrestrial invertebrates that move primarily in the soil are called "in-soil organisms." This distinction was made because the main exposure routes for these two groups of organisms differ and consequently, so does their risk assessment. Therefore, the opinion does not discuss the exposure assessment of in-soil organisms, which is instead addressed in the "EFSA scientific opinion addressing the state of the science on risk assessment of plant protection products for in-soil organisms" (EFSA PPR, 2017) in section 3.2.8.

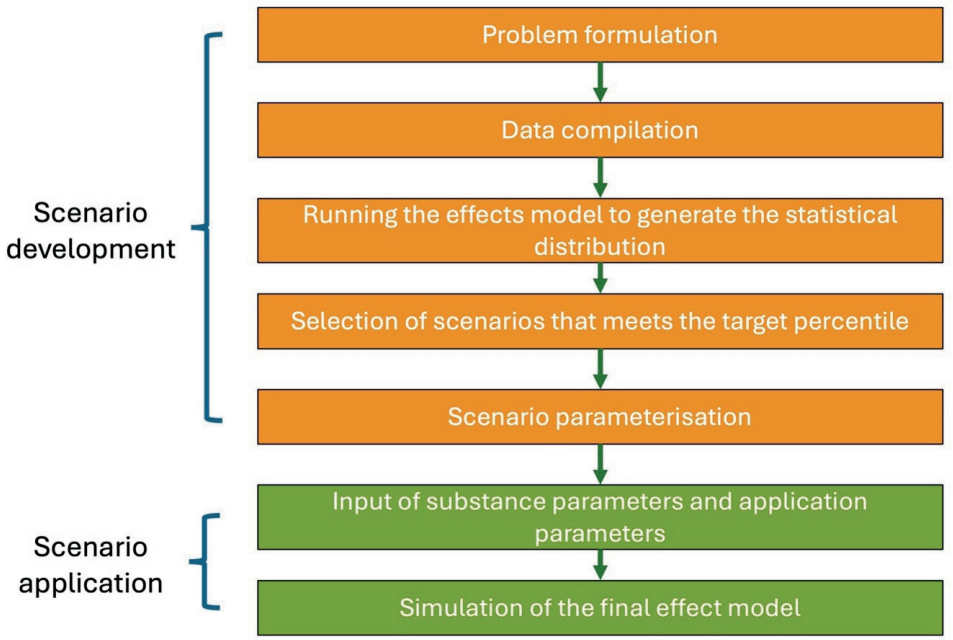


Figure 3.1: Illustration of the process to develop and apply environmental scenarios based on percentiles of the model results. The process should be carried out for each relevant effect endpoint (for example, baseline and toxic standards as well as for the product to be tested). Adapted with permission from EFSA PPR (2014).

The EFSA opinion on NTA contends that the assessment of effects of pesticides on NTA biodiversity must focus on the landscape scale – in addition to the local scale – to enable the implementation of an effective mitigation policy in the future. Risk assessment at the landscape scale is recommended to be conducted with population models, while effect modeling is not recommended for a risk assessment at a lower scale. Effect modeling at the landscape scale is recommended because spatio-temporal patterns of exposure can be critical in determining population-level impacts for (highly) mobile species that live in different exposure compartments throughout their life cycle. Examples are given such as a pollinator choosing a sprayed field

for foraging, or a carabid beetle spending larval and pupal stages in soil in-field, but as an adult moving between in-field and off-field.

The EFSA opinion on NTA emphasizes that exposure needs to be dynamically calculated in space and time and linked directly to simulation of phenology and spatial dynamics of the NTA. This allows the spatio-temporal variability of a stressor to become a component of a dynamic environmental scenario for effect models. Moreover, the dynamic simulation could be used to generate the range of exposure profiles predicted for real landscapes, and thus be able to place any field experiment in the context of this distribution (e.g., identify if the field experiment represents the 90th percentile of exposure concentrations). Nevertheless, it was also recognized that such models are scarce, likely due to the computation power they require.

Furthermore, SPGs for field exposure should ideally be defined in terms of percentiles of effect distributions (EFSA PPR, 2015). The EFSA opinion suggests that this can be achieved by evaluating a large number of environmental scenarios representing the range of conditions across the geographical scale relevant for the risk assessment and ranking them according to their effects. This would allow direct calculation of percentiles of effect distributions. However, an agreed methodology for developing such environmental scenarios is not yet available. For the time being, the EFSA opinion proposed to separately define SPGs, and to base the field exposure assessment of NTA on the 90th percentile concentration in each of the three regulatory geographical zones (North, Center, South).

3.2.4. Conceptual framework for the development and implementation of environmental scenarios

Rico and colleagues (2016) suggest a conceptual framework for the development and implementation of environmental scenarios in environmental risk assessment (ERA)(Figure 3.2a). This framework is composed of six steps. In the first step (problem definition), the problem is defined in terms of the pesticide that is selected for the risk assessment. In the second step (habitat definition), based on an example for aquatic risk assessment, the aquatic habitats (potentially) exposed to these pesticides (i.e., streams, ditches, and ponds) in the different climatic regions of the European Union are taken as a starting point for deriving a set of ecological scenarios. In the third step, the SPGs are addressed, describing (i) the different ecological entities and their attributes that must be protected, (ii) the acceptable effect magnitude with its associated degree of certainty, and (iii) the spatial-temporal scale of the acceptable effect. In the fourth step, the environmental scenarios are defined, starting with the definition of the ecological scenarios. For this 4th step, a stepwise approach is followed, as illustrated in Figure 3.2b. This results in ecological scenarios, consisting of (i) the focal species, (ii) the environmental characteristics, and (iii) the spatial-temporal frame. For the exposure scenarios, Rico et al. decided to use already existing scenarios. They were aware of a potential mismatch between the ecological and the exposure scenarios, both in spatial-temporal scales and in terms of their environmental characteristics. Therefore, they were careful

in selecting exposure scenarios that matched the spatio-temporal scale of the focal species, and if possible and applicable, they fed the environmental characteristics of the selected exposure scenarios directly into the environmental characteristics of the ecological scenarios. Finally, in the 5th and 6th steps of the conceptual framework, the defined environmental scenarios are implemented in the model(s), and the environmental risk assessment is performed.

Although we consider the approach of Rico et al. (2016) reasonable, which starts with the chemical under consideration, followed by habitats potentially exposed, and ends with species or communities likely inhabiting these habitats, the six steps they distinguished are usually covered in the problem definition phase of the risk assessment. Therefore, most of the steps discussed by Rico et al. are not relevant for the development of environmental scenarios and should only be considered for refining the problem definition phase. We consider the environmental scenario to be a set of environmental conditions required as an input for a given regulatory model for a focal species, community, or ecosystem (Figure 3.5).

3.2.5. Supporting publication on a mechanistic model to assess risks to honey bee colonies

The EFSA supporting publication (EFSA, 2016) describes the development of a conceptual model for the risk assessment of pesticides on a single bee colony under different environmental scenarios. The conceptual model described in the publication has resulted in the development of the ApisRAM model (EFSA SC, 2021). However, what makes the supporting publication (EFSA, 2016) so interesting is that a lot of attention is given to scenario development during the development of the conceptual model.

In the EFSA supporting publication (EFSA, 2016), scenarios are defined as "a representative combination of crop, soil, climate, and agronomic parameters to be used in modeling; representative means in this context that the selected scenarios should represent physical sites known to exist." Furthermore, they say that "A scenario is intended as any plausible combination of state variables and their dynamics describing a process, which is designed to evaluate the implications of alternative possible situations to be assessed." With both statements, they highlight the importance of defining realistic scenarios and representing the range of variability that is found in the environment.

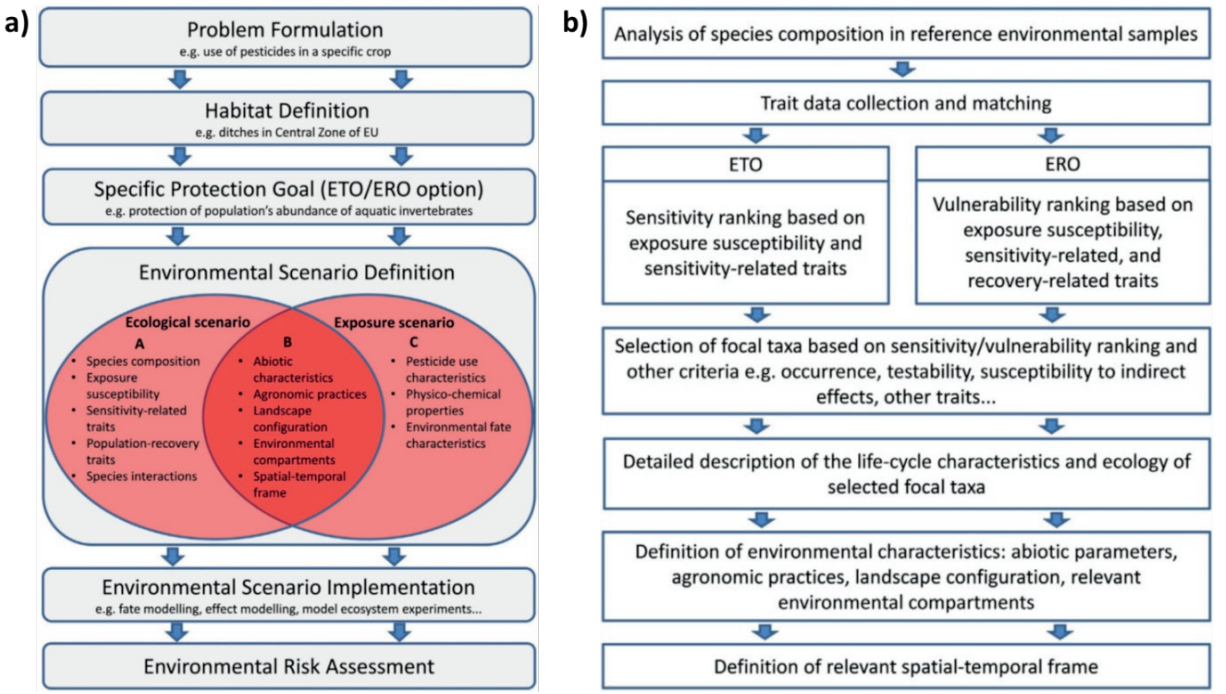


Figure 3.2: Conceptual framework for the development and implementation of environmental scenarios in ERA, a) conceptual framework, and b) diagram showing the proposed steps for the definition of the ecological scenarios. Reprinted with permission from Rico et al. (2016).

Overall, the conceptual model is made up of six modules. Three of these six modules are described by process-based mechanistic processes. The other three modules are described using a scenario-based modeling approach: (i) the resource providing units and environmental drivers (RPU-ED), (ii) the biological agents (e.g., *Varroa, Nosema*), and (iii) the beekeeping management practices (BMP). This scenario-based approach signifies that time-dependent effects of these three modules on target processes of the model are described by external forcing variables (i.e., driving variables).

These forcing variables can be incorporated in forcing functions, which are defined as "A function that appears in an equation and is only a function of time, not of any other variables." These scenario variables are, therefore, described in terms of the spatial and temporal pattern of forcing variables. In the EFSA supporting publication (EFSA, 2016), the population abundance of the parasitic *Varroa* mites is given as an example of a forcing variable, and the mortality of larvae and pupae is the target process.

Finally, because three of the six modules are described by scenarios, EFSA suggests that future research should focus on redeveloping the modules into process-based dynamic modules. For instance, introducing a separate module for the population dynamics of the parasitic *Varroa* mites, and their interaction with the bee population.

3.2.6. Guidance document on specific protection goals

In some of the documents described in the preceding paragraphs, it is mentioned that environmental scenarios should be in line with SPGs. For instance, the spatial scale of the SPG defines the minimum scale of the environmental scenario. To get a clearer view on which information results from the derivation of SPGs, we briefly review the EFSA guidance document on SPGs (EFSA SC, 2016).

Table 3.1: Non-exhaustive list of potential options for each of the five dimensions used to describe specific protection goals (from EFSA [2016], Table 8).

Dimension	Options
Ecological entity	Individual – (meta)population – functional group – community – ecosystem – habitat
Attribute	Behavior – survival – growth – reproduction – abundance – biomass – process – within and between species diversity – landscape or habitat structure
Magnitude	Negligible – small – medium – large
Temporal scale	Days – weeks – months – seasons – years – decades – generations – rotations
Spatial scale	In crop or field – edge of field or field margin – nearby off-crop – protected area – watershed – landscape – region – continent

The EFSA guidance document suggests that the development of SPGs consists of three steps: (i) identifying relevant ecosystem services for ERA, (ii) identifying relevant service providing units (SPUs) for relevant ecosystem services (SPU is synonymous to "key driver" [EFSA PPR, 2010]), and (iii) specifying the level or parameters of protection. Steps (i) and (ii) are relatively straight-forward. A table with ecosystem service categories and examples of the SPUs providing them is given in Table 1 in EFSA SC (2016). An example is pollination as an ecosystem service, with pollinators such as bees as SPU. In step (iii), however, more detailed information must be provided to further define the SPG. It is this information that is important for defining the model and the environmental scenarios that will be tested to perform the environmental risk assessment.

When specifying the level or parameters of protection, the EFSA suggests five dimensions to be considered when describing SPGs: (i) ecological entity, (ii) attribute, (iii) magnitude, (iv) temporal scale, and (v) spatial scale. These dimensions have been proposed and used by EFSA PPR (2010) to structure and focus the procedure for making protection goals operational. Table 3.1 lists these five dimensions and provides a non-exhaustive list of options that are possible in each dimension.

Concrete SPGs remain absent from most chemical risk assessment guidance and need to be developed further.

3.2.7. Environmental scenarios for ecological risk assessment of down–the–drain chemicals in freshwater environments

The work of (Franco et al., 2017) focuses on down-the-drain chemicals. Because some of the processes relevant to down-the-drain chemicals are not relevant for pesticides, we limit the review of the Franco et al. (2017) study to concepts deemed relevant to pesticides.

Franco and colleagues (2017) adopt the definition of an environmental scenario from Rico et al. (2016), focusing on two important aspects of the scenario. Namely obtaining (i) spatially explicit exposure scenarios, and (ii) vulnerability-based ecological scenarios. In addition, much emphasis is placed on using SPGs as guidance to select the appropriate biological entities and spatio-temporal scales that the environmental scenarios should address.

Key factors are incorporated at increasing spatiotemporal resolution (exposure scenarios) and taxonomic resolution (ecological scenarios) toward integrated exposure and ecological scenarios (environmental scenarios) for specific combinations of realistic worst-case catchment and vulnerable taxa (from Franco et al. 2017, Figure 3.3). We note here that Franco et al. (2017) describe a potential bottom-up method for defining relevant and environmentally realistic SPGs by "the empirical characterization of scenarios with representative ecological community structures and functions derived from biomonitoring data."

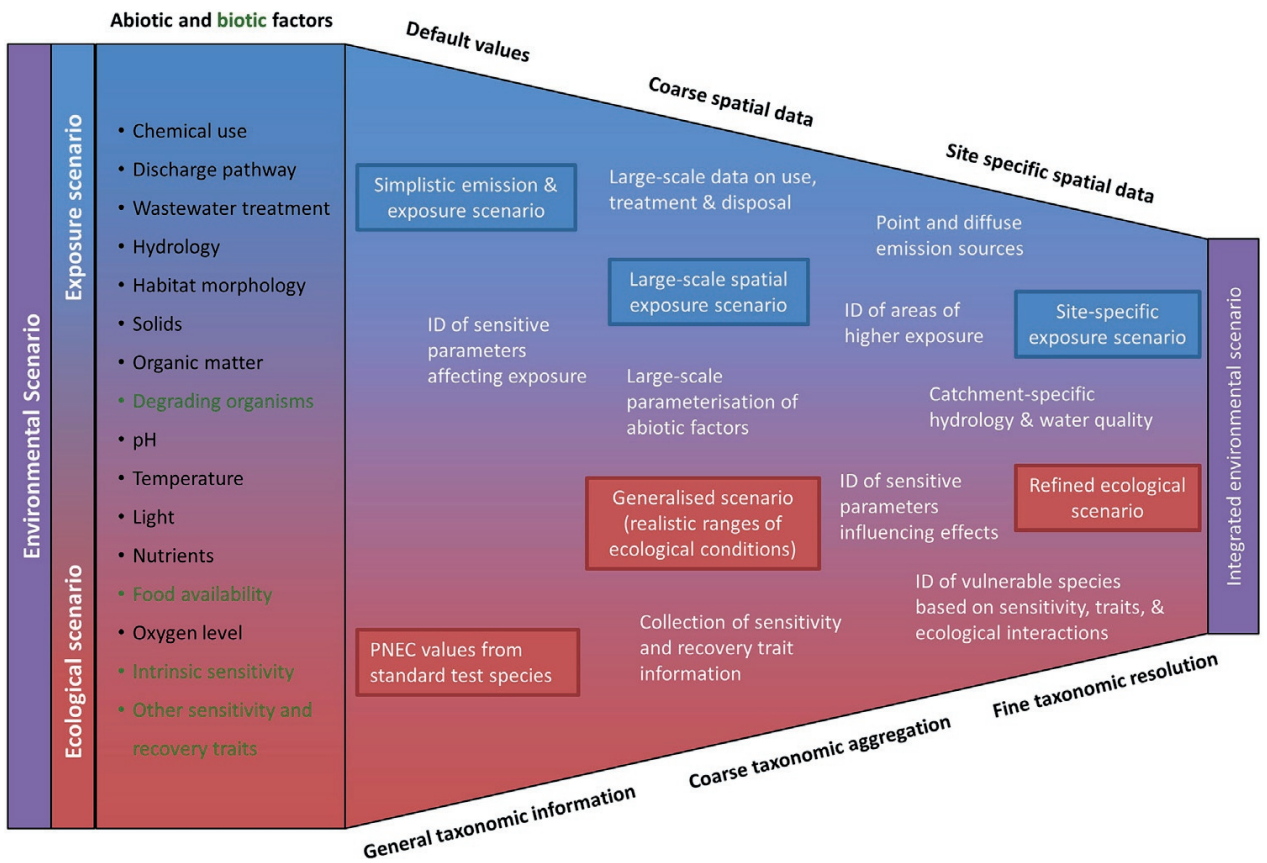


Figure 3.3: Development of environmental scenarios from lower – to higher-tier risk assessment. Reprinted with permission from Franco et al. (2017)

Because Franco and colleagues (2017) give a lot of attention to SPGs, they see much potential in trait-based analyses and suggest two crucial applications. First, for the selection of the most vulnerable species, they envision that the most vulnerable species can be filtered from the complete species pool by combining taxonomic and trait analysis with habitat filtering models. Because we consider the selection of focal species as being a part of the problem definition (as explained in section 3.2.4), we do not deem this relevant for the environmental scenario development. Second, for incorporating vulnerability-related traits into ecological models, Franco et al. recognize that knowledge gaps in traits specifically affecting population vulnerability constitute the main constraint in our current ability to target the most vulnerable species. They suggest that the use of ecosystem models can assist in identifying the main factors affecting the response of real food webs to chemical stress.

Franco et al. (2017) suggest a tiered approach for the development of environmental scenarios (Figure 3.3). They describe that "Environmental scenarios developed at different scales and levels of resolution can be applied at a given tier of assessment according to need for refinement and data availability. The degree of integration between exposure and effect assessment increases at higher tiers because the matching of the abiotic parameter values and the spatial-temporal scales." Finally, they suggest that scenario-based probabilistic assessment should be used to obtain effect prevalence plots that can be used for risk assessment.

3.2.8. Scientific opinion on soil risk assessment

In the EFSA scientific opinion addressing the state of the science on risk assessment of plant protection products for in-soil organisms (EFSA PPR, 2017), population modeling is a fixed part of the proposed risk assessment scheme. The scientific opinion proposes to use "conservative population modeling scenarios, pre-run and tabulated as look-up tables using the toxicity data from standard tests applied to models of vulnerable focal species" as a screening tool for the assessment of long-term effects. If this screening indicates unacceptable risk, a higher-tier assessment can be done by running the dynamic model using refined model inputs for toxicity and exposure.

With respect to exposure scenarios, the EFSA PPR panel explains that "In all cases when population modeling is used, the development of suitable baseline scenarios against which to evaluate the effect is critical. However, depending on the SPG, it is not always easy to determine which baseline will provide the most sensitive outcome (see section 3.4). For this reason, we recommend that in all cases a representative range of baselines should be used from intensive agricultural systems to extensive sustainable systems, and natural conditions in the case off-field or boundary-scale scenarios are needed."

In contrast to terrestrial and aquatic animals, the relevant spatial scale for most in-soil organisms is smaller due to their low mobility. Thus, often no or only vertical gradients of factors and vertical movement must be considered. Thus, an ecological scenario for in-soil organisms is mainly driven by soil type,

abiotic conditions like temperature and moisture, and impacts of agricultural management. In addition, according to the soil scientific opinion, the scenario includes the selection of the focal species. However, in the context of this chapter, we consider the selection of the focal species as a part of the problem (see last paragraph of section 3.2.4).

There are no further recommendations to define the environmental scenarios in the soil scientific opinion nor are example scenarios suggested.

3.2.9. Scientific opinion on on pesticide risk assessment for amphibians and reptiles

In the EFSA scientific opinion on the state of the science on pesticide risk assessment for amphibians and reptiles (EFSA PPR, 2018b), a population model of the great crested newt is presented to demonstrate the potential use of such models in the risk assessment of amphibians and reptiles. In addition to the newt, two other amphibian and three reptile species are proposed as focal species for their potential uses in amphibian ERA. The newt model used as an example is a spatially explicit individual-based model (IBM) where the life-cycles of the animals are modeled in 10 km * 10 km landscapes. Thus, the environmental scenario is defined by landscape structure, such as number and distribution of ponds and agricultural fields, the farming practice, weather conditions, and the use of the pesticide to be assessed. The authors acknowledge that "a number of steps and improvements to the model and scenario are needed before the model could be used in practice." With respect to the scenarios, it is recommended that:

- Landscape, weather, and farming scenarios need to be considered carefully to be representative of the region under consideration.
- Ideally, scenario and model development should be part of an interactive model cycle, fed by data from the real world, for example via monitoring.
- Development of landscape and farming simulations for the regulatory zones in the European Union is needed to support the model in these zones, and similarly for any individual country that may wish to use the model (EFSA PPR panel 2018).

3.2.10. Scientific opinion on toxicokinetic-toxicodynamic models

In EFSA's toxicokinetic-toxicodynamic (TKTD) opinion document (EFSA PPR, 2018a), the environmental scenario consists of agronomic and abiotic parameters (Figure 3.4). An ecological scenario is not explicitly mentioned in the description of environmental scenarios used in the TKTD opinion. The exposure scenario is almost always fixed by laboratory test conditions. However, most TKTD approaches allow the consideration of more environmental factors and their impact (e.g., temperature-dependent

bioavailability), as long as calibration and validation data for multiple environmental conditions are available. Other ecological aspects, like interactions between species, are not part of the environmental scenario; TKTD models are restricted to the organism level. Often, TKTD models are used as a refinement tool when considering dynamic exposure and thus, the Tier 1 species or the sensitive species identified in Tier 2 are modeled. Nevertheless, TKTD models are usually used with the equivalent of Tier-1 or Tier-2 assessment factors to cover for the extrapolation from laboratory to field. Therefore, aspects like fixing the abiotic conditions from laboratory conditions and ignoring species interactions are considered appropriate.

In the TKTD opinion, the factors considered in the derivation of the environmental scenarios differ between model type. Specifically, the General Unified Threshold model of Survival (GUTS) (Jager et al., 2011), the Dynamic Energy Budget (eco)toxicological (DEBtox) models (Jager, 2017), and primary producer models (Heine et al., 2014; Heine et al., 2015; Schmitt et al., 2013; Weber et al., 2012) are addressed separately. The GUTS models do not require further definition of environmental conditions, because model parameters are calibrated with data obtained from experiments performed under standard laboratory conditions. Also, pesticide concentrations are generated using relevant FOCUS exposure scenarios, which in turn account for factors such as soil type, rainfall, and agronomic practices. It is considered appropriate to fix abiotic factors of the environmental scenarios to the laboratory conditions of the experiments used for calibrating the TKTD model, and also to ignore ecological factors such as species interactions, because model application is restricted to Tier-1 or Tier-2 risk assessment. Hence, equivalent assessment factors are applied to extrapolate from laboratory to field conditions.

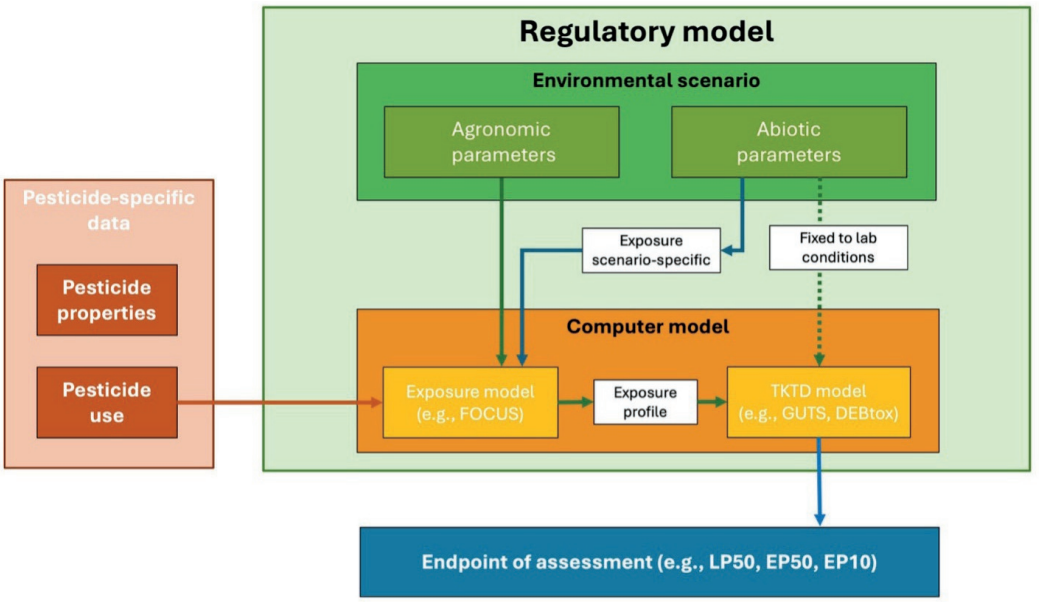


Figure 3.4: Schematic representation of a TKTD regulatory model (orange boxes). The environmental scenario feeds into the exposure model in combination with information on pesticide properties and uses. The exposure model in turn delivers the exposure profile, which is used as input by the TKTD models. Altogether, the regulatory model delivers an output, that is, the endpoint of the assessment. For DEBtox, GUTS, and primary producers' models, as indicated by the dotted arrow and the associated box, the recommendation is, for the time being, to fix the abiotic parameters to the laboratory conditions of the experiment used for calibrating the model. If in the future the relationship between toxicity and environmental conditions will be better understood and described, such recommendations can be revised. Adapted with permission from EFSA PPR (2018a) and EFSA PPR (2014).

For DEBtox models, a basic environmental scenario would be required, including temperature and food conditions, because the physiological part of the model has the potential to extrapolate predictions of growth and reproduction to time-variable environmental conditions, in terms of, for example, temperature and food availability. In fact, accounting for variable energy intake is one of the main ideas behind DEB modeling. In the application of DEB theory to predict toxicological effects, however, environmental scenarios are usually fixed to the laboratory conditions of the experiments used for calibrating the TKTD part of the model, and therefore it is complicated to assess whether or not the extrapolation of toxicity to other environmental conditions would be accurate. In cases where appropriate validation data for variable temperature or food conditions is provided, explicit extrapolation to variable environmental conditions and hence the definition of different environmental scenarios would be facilitated. This has been successfully demonstrated (Cedergreen et al., 2016; Gergs et al., 2014; Goussen et al., 2020).

For primary producer models applied under Tier-2 refinement, the environmental conditions of the experiment usually define the environmental scenario, and the control growth rate (which serves as the baseline scenario) is determined by the test conditions. In this case, the model does not need to define the dependencies of model parameters from abiotic factors. In higher tiers, it is naturally possible to define the dependency of growth parameters on environmental conditions. Although, experiments must be performed using ranges of realistic light conditions and nutrient levels to perform model calibration.

Table 3.2: Overview of relevant documents that mention the topic of exposure, ecological, or environmental scenarios and descriptions of how each document defines (exposure/ecological/environmental) scenarios, the factors considered in scenario derivation, as well as the strengths and weaknesses of the used approach.

Relevant document(s)	Definition of environmental scenario	Factors considered in scenario definition	Strengths	Weaknesses
FOCUS surface water exposure scenarios	Restricted to surface water exposure scenarios, defined as a set of "realistic worst-case" scenarios broadly representative of agriculture as practiced in the major production areas of the EU	Different entry routes; Different target crops; Three surface water situations; Topography, climate, soil type; Agricultural management practices	Standardized Limited number of scenarios	Based on basic data sources and expert judgement Rudimentary approach to conservativeness to define "realistic worst-case" (combination of probability of occurrence for main driving factors of exposure) Restricted to exposure scenarios; therefore ignores (potentially) relevant ecological factors
EFSA scientific opinion on good modeling practice (EFSA PPR, 2014)	Representative combination of abiotic, biotic, and agronomic parameters for the purpose of modeling	Specific protection goal and level of conservatism considered in problem definition; Species ecology and behavior; Biotic factors (competition, predation); Abiotic factors (habitat quality or availability, weather, land management); Agronomic factors (crop species, application rate, irrigation); Exposure route; Spatial and temporal scale and resolution	General – applicable to a wide range of model types	Difficulty in defining worst-cases when biotic, abiotic, and agronomic factors are considered Reliant on carefully argued, realistic worst-case environmental scenarios No set of fixed environmental scenarios is provided Too vague – no specific support for the definition of environmental scenarios given

EFSA scientific opinion on non-target arthropods (EFSA PPR, 2015)	Landscape scale	Entry routes: Overspray, contact and oral exposure; Environmental, agricultural and physiochemical variables	Dynamically simulated exposure	Complex Calculation intensive due to fine level of details
Rico et al. (2016)	Combination of biotic and abiotic parameters (including agronomic practices and properties of agronomic landscapes) providing a realistic worst-case representation of pesticide exposure, effects and recovery for the ecological entity under evaluation	Species composition; Exposure susceptibility; Sensitivity-related traits; Population-recovery traits; Species interactions; Abiotic characteristics; Agronomic practices; Landscape configuration; Environmental compartments; Spatial-temporal frame; Pesticide use characteristics; Physico-chemical properties; Environmental fate characteristics	Ready to use, as exemplified	Identification of model species part of environmental scenario formulation, hence potentially new model required for each chemical
EFSA publication on a mechanistic model to assess risks to honey bee colonies (EFSA, 2016)	A representative combination of crop, soil, climate, and agronomic parameters to be used in modeling; Representative means in this context that the selected scenarios should represent physical sites known to exist	Environmental drivers such as weather and climate; Resource-providing unit including resource availability, land use and cropping practice; Beekeeping management practices	Allow for the future development of scenario-components into process-based dynamical modules	Implementation remains unclear

Franco et al. (2017)	The result of combining an exposure and an ecological scenario, where the "exposure scenario is defined by its spatial and temporal scale and by a qualitative description of the environmental context it represents," while the ecological scenario can be loosely described as "the entire pool of species potentially present in a given geographical context"	Chemical use; Discharge pathway; Wastewater treatment; Hydrology; Habitat morphology; Solids, organic matter; Degrading organisms; pH; Temperature; Light; Nutrients; Food availability; Oxygen level; Intrinsic sensitivity; Sensitivity – and recovery-related traits	Different levels of complexity following a tiered approach	Potentially new models required for each chemical Scenario-based probabilistic assessments lend themselves to the creation of effect prevalence plots
EFSA scientific opinion on soil organisms (EFSA PPR, 2017)	Baseline exposure scenarios against which to evaluate the effects, from intensive to extensive agricultural systems and natural conditions for off-field scenarios. Focal species as part of environmental scenario	Soil exposure scenarios include arithmetic mean annual air temperature; Mean annual precipitation; Organic-matter content of the top 30 cm of the soil; pH of the topsoil; Soil textural class	Exposure scenarios based on spatial distributions	No examples of ecological scenarios yet Exposure scenarios might provide conditions not suitable for focal species

EFSA opinion on TKTD models (EFSA PPR, 2018[a])	Environmental scenario is determined by the exposure scenario; Consists of agronomic and abiotic parameters (respectively, the ecological scenario is defined by the test conditions)	Differs per model type (GUTS, DEBtox, plant models): (i) same as FOCUS scenarios, (ii) abiotic conditions of experiments, (iii) various abiotic conditions, depending on availability of validation experiments	Large flexibility possible	No ecological interactions Requires the use of assessment factors to extrapolate to other environmental scenarios Increased flexibility (e.g., additional environmental scenario components) requires increase in validation experiments
EFSA opinion on reptiles and amphibians (EFSA PPR 2018b)	Landscape scale; Focal species as part of environmental scenario	Landscape structure (based on real maps); Weather; Regional farming practice; Pesticide use	Systems-based approach General framework (ALMASS) applicable for different species Integration of fate and effect modeling	Calculation intensive due to fine level of details. Large number of scenarios possible (e.g., regions, farming practice)

[a] The EFSA guidance on specific protection goals (EFSA SC, 2016) is addressed in this chapter, because they provide crucial information that is required as an input when deriving environmental scenarios.

3.3. General aspects of regulatory environmental scenario development

3.3.1. Distinction between model input, model parameters, and environmental scenario components

It was challenging to clearly differentiate which factors depicted in the grey boxes in Figure 3.5 should be considered as model parameters, model inputs, or environmental scenario components. In fact, it turns out that this largely depends on the risk assessment question and the model used to answer that question. This means that all factors depicted in the grey boxes in Figure 3.5 can potentially be considered as model parameters, model input data, or environmental scenario components, depending on the context of the situation under consideration, and the precise question formulation to be described with a modeling

approach. We think that the distinction between model inputs and environmental scenario components is not always meaningful, because some factors can be considered for both, without causing any confusion in the modeling exercise. Consider, for example, the physico-chemical properties of an agrochemical. On the one hand, they can determine which environmental factors might be of interest for the risk assessment and should therefore be incorporated into the environmental scenarios. On the other hand, some physico-chemical characteristics can also be a part of the fate model (or module) and will thus be considered as model inputs. The same is true for abiotic factors that determine application or emission patterns.

The distinction between model parameters and environmental scenario components, however, can be considered more important, because model parameters will be fixed during model application, while environmental scenario components are often varied during model application to resemble realistic environmental conditions for a specific risk assessment question. Therefore, we suggest defining the difference between model parameters and scenario components by questioning *whether a factor is likely to differ between the different scenarios that are used for answering one risk assessment question*. In this sense, a model parameter is, for example, a scenario-independent physiological or ecological quantity that is likely to stay constant across all tested scenarios, irrespective of the risk assessment question and the model application. An explicit example is the maximum feeding rate constant at reference conditions (without a stressor), which would stay constant for a given species irrespective of a concrete risk assessment question. A scenario component is then an environmental factor, used or required as input into the model, that can take different values across different environmental scenarios, for example food availability or habitat quality. Sometimes, a clear differentiation into model parameters and scenario components can be difficult, for instance when model parameters contain aspects of the environmental scenario, for example, having scenario-specific growth rates, but this will be the exception rather than the rule. Any potential issues with this distinction between model parameters and environmental scenario components can be prevented by ensuring an appropriate sensitivity analysis of all these parameters, so that their importance to the risk assessment question can be assessed.

Other factors that are commonly part of MEMs might also be considered as part of the environmental scenarios if they include aspects of environmental variability. Typical regulatory questions want to know whether or how ecological interactions and/or processes, for example, predation competition pressure or dispersal, are considered and implemented in the risk assessment. Such interactions can be accounted for via explicit modeling, for example, in a community modeling approach where competition or predation can emerge from explicitly simulated interactions. In this case, they may not be part of an environmental scenario. Alternatively, ecological interactions might be represented as components of an environmental scenario, for instance, by introducing a fixed factor for predation and/or competition depending on the different landscapes.

3.3.2. Components of environmental scenarios

In our definition, an environmental scenario includes both an ecological and an exposure scenario (Figure 3.5), among which a number of environmental factors and properties are shared.

Indeed, all factors mentioned in the grey boxes in Figure 3.5, (e.g., abiotic and biotic conditions) can generally both occur as part of the exposure or the ecological scenario. However, some aspects are more frequently found to be important for exposure scenarios (e.g., application patterns), whereas other factors are more frequently found to be important for ecological scenarios (e.g., biotic conditions). Nevertheless, all potential combinations are theoretically possible. If a factor is indeed relevant for both the exposure and the ecological scenarios, it needs to be consistently considered in both, or even better, identically defined. In this section we briefly discuss all categories of environmental factors mentioned in Figure 3.5.

Abiotic conditions, for example, weather, temperature, pH, oxygen, hydrology, light, and habitat morphology, are all environmental factors that often have relevance for both the exposure and the ecological scenarios. Pesticide fate models usually consider a set of abiotic conditions for the definition of the environmental scenarios, for example, temperature or precipitation, but other factors such as particulate organic matter in water or soil can also play a role in the fate and transport processes that PPP undergo after application. In such cases, variability of these abiotic conditions must be considered in exposure scenarios. Some of these abiotic conditions also directly influence the ecology of the model system under study. For example, temperature can determine the degradation kinetics of a compound in the environment, but also whether a chemical substance is metabolically active, or how fast chemical uptake might be. It is also important to consider the relevance of biogeochemical cycles on the growth and development of most species, and therefore their relevance to the ecological scenarios.

Biotic conditions, for example, food availability, food quality, predation, competition, and habitat quality, are often most relevant to ecological scenarios. However, if the main exposure route occurs through diet, (e.g., chemical binding to organic matter), biotic conditions such as availability (and contamination) of different food items can also be accounted for in an exposure scenario. Some ecological aspects, including possible interactions with other species, can influence the ecological scenario development, for example competition for resources or the predation pressure a specific population might be exposed to. In that context, full consideration of dynamic competition or predation in form of explicitly simulated predator or competitor populations can often not be achieved in practice, the use of factors to account for such interactions seem more reasonable to consider. Such factors are not meant to be simple assessment factors to cover uncertainty, but instead act for example, as rate constants in a dynamic model context. Such factors, for example the intensity of predation pressure, could then be applied in a MEM, and even potentially be defined in a dynamic way, meaning that predation pressure is modulated to account for variation in different time periods within a year. An important aspect is the consideration of food availability and food quality, because these biotic factors often have a strong influence on the resilience of a population.

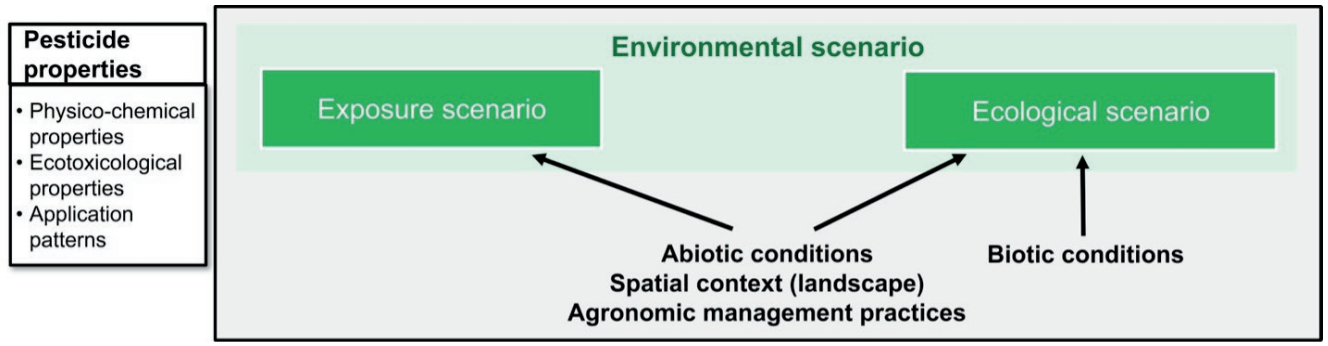


Figure 3.5: Main components of environmental scenarios.

It should be noted that pesticide properties are mentioned here as important influencing factors, but they are not considered as part of the scenario. More details and explanations are given in the text. It should also be noted that the definition of an environmental scenario for a specific application might depend on pesticide properties, but for simplicity and communication reasons this potential dependency is left out here.

Additional potentially relevant components of an environmental scenario are spatial context (landscape), including landscape configuration and connectivity characteristics of locations within a given ecosystem (e.g., aquatic habitat connections, availability, and/or distance between refugia, heterogeneity). The spatial context often has a major influence on both the exposure and the ecological scenarios. Consider for example the potential exposure of bees when their hive is surrounded by large strips of uncontaminated forests and meadows, compared with when the hive is surrounded by purely (contaminated) agricultural lands and urban areas. On the ecological side, the recovery of a population might be present or absent depending on the connectivity of the modeled system to other uncontaminated systems. Therefore, the spatial context can be essential for both the exposure and the ecological scenarios.

Considering eco(toxicological) information, for example, sensitivity of different life stages, survival dynamics over time, and interspecific variation in intrinsic sensitivity, can also have a strong impact on simulations and might therefore be a significant part of ecological scenarios. On the exposure side, it might be crucial to know exactly which life stage is sensitive, and which seasons or environmental compartments this life stage usually occurs in. On the ecological side, it can be crucial to know at which level of the food chain sensitive species can be found, because this can, for instance, determine food availability for organisms that are higher up in the food chain. Ecotoxicological information can also serve as model input, for instance, to model survival patterns over time in the toxicology module (Figure 2.2).

In a strict sense, however, ecotoxicological data are substance – and species-specific and thus, would not be part of an environmental scenario, which should be applicable to other substances or species. In this

view, ecotoxicological information is needed to parameterize the effect model, but it may affect the definition or selection of environmental scenarios.

Application patterns result from timing, frequency, and application rates of PPP as defined in the Good Agricultural Practice (GAP). Application patterns and their specific parameters might be of particular importance regarding which life stages are exposed to the chemical, or which seasons should be included in the fate module. Usually, application pattern parameters serve as direct inputs for a fate model, which translates emission patterns into exposure patterns in time and space or can be used as a direct input for a MEM that has an incorporated fate module (e.g., the ALMaSS model, Topping & Odderskaer [2004]).

The physicochemical properties of a chemical might strongly determine which abiotic conditions are important for the exposure (and, therefore, for the ecological) scenario. For example, when a chemical has a high octanol-water partition coefficient it has a large potential to bind to organic matter, making the amount of organic matter present in the system of crucial importance for the exposure. On the ecological side, the physicochemical properties of the chemical might dictate how fast a chemical degrades, and therefore which species (competitors and/or predators) and/or life stages are potentially exposed, potentially impacting ecological processes such as competition and/or predation. Because physico-chemical parameters do not usually vary between scenarios, in a strict sense they would not be considered as part of a (general) environmental scenario but as part of the fate model. However, different environmental scenarios may still require the consideration of different values of a model parameter (e.g., different half-life [DT50] values in colder or warmer regions). Thus, some physicochemical properties may become relevant for the scenario selection, if for example, they drive factors that should be considered. Therefore, we have still referred to them in this chapter for comprehensiveness.

Last but not least, agronomic management practices, such as dredging of drainage ditches, weed removal, soil tillage, crop rotation, or pesticide risk mitigation measures (e.g., buffer strips), are components that can influence exposure or habitat quality and indirectly influence the outcome of the risk assessment. The practices can vary across scenarios, so should be considered as part of the environmental scenarios. Agronomic management practices related to some species-specific aspects, for example bee-keeping practices for honey bee scenarios, also belong to potentially relevant aspects of environmental scenarios.

3.3.3. Using environmental scenarios to derive the margin of safety

Recently, effect modeling approaches have been developed and applied to increase the mechanistic understanding of ecotoxicological effects and to overcome some of the limitations of experiment-based methods. A well-known limitation of experiment-based methods is the practical limitation of experiments to test a few environmental scenarios only. Indeed, such experimental test designs are representations of one site only, consisting of real systems with their own complexity and environmental conditions. Mechanistic effect models (MEMs), however, are always a simplification of reality. Acknowledging this, a modeling study can,

in our view, not be used in the same way a field study is traditionally used: Test a (low) number of selected situations and consider the outcome immediately in the overall regulatory risk assessment. Instead, ideally a model is used by considering its simplification bias and at the same time exploiting its main advantage. In this manner, a large number of environmental scenarios can be screened, and a deeper understanding of basic mechanisms can be achieved by using MEMs, allowing reduced uncertainty in the ERA.

Therefore, it is generally recommended to define environmental scenarios to span across ranges of relevant environmental factors (e.g., food, habitat, temperature, and exposure patterns), resulting in a set of environmental scenarios ranging from ecologically favorable to unfavorable (but still realistic) conditions. In section 3.4, more detailed steps for the assessment of the impact of environmental scenario components are explained. Here, we rather intend to emphasize that, for the application of MEMs in regulatory risk assessment, the provision and testing of only one environmental scenario is largely insufficient. Instead, a set of environmental scenarios, representing realistic variations of the most sensitive and uncertain components of environmental scenarios, should be used for the provision of model input data for an assessment of the regulatory question. An exploration of minimum requirements for the most crucial environmental factors (e.g., food, habitat, temperature, and exposure patterns) can indeed provide useful information, and is often already part of model development and transparent and comprehensive ecological modeling documentation (TRACE). Screening of environmental scenario settings for identification of thresholds for the transition into non-sufficient conditions (e.g., local population extinction), or model tests at boundary conditions, such as extra stressors (e.g., cold winter or deterioration of landscape quality) can add to the identification of less favorable environmental scenario settings. Nevertheless, it is important to stay realistic enough in such scenario developments, because environmental scenarios should always allow for population persistence.

In itself, variation of the exposure scenarios can provide very useful information for the risk assessment. By modulating the exposure level, the margin of safety of a certain planned application of a PPP can be determined, for example, by using exposure modification factors (EMFs) that increase until an onset of effects is detected. This is not a check of the model performance, nor an indication of a realistic field situation, but a support for decision making by estimating how far away from having an unacceptable impact a certain exposure situation is. By combining a defined set of environmental scenarios across a range of ecological conditions with exposure multiplication factors, a comprehensive picture of the onset of effects and respectively the general margin of safety of an application of a PPP across a wide range of environmental conditions can be established.

If available, observations from field or mesocosm studies can be used to compare observations with model predictions for study-specific environmental scenario conditions and to corroborate the realism of the model predictions. Such a model validation based on field test data could be used to extrapolate to other environmental conditions.

3.3.4. Matching the spatio-temporal scale of exposure between fate and ecological modeling

Although the general spatial and temporal scale of the risk assessment is usually defined in the problem definition phase and should match with the selected model, some decisions are still required during the environmental scenario definition. For instance, the exact size and resolution of the simulated area, and the simulated time period and resolution must be fixed. The spatio-temporal context of the environmental scenarios needs to account for the life cycle and mobility characteristics of the selected species, as well as for the landscape (e.g., crop types, non-crop elements) and exposure characteristics. The spatio-temporal context is partly determined by the choice of the species to be modeled, which depends in turn on the risk assessment questions and problem definition, and therefore also on the specific compound and its intended use. This builds a set of dependencies, which makes it difficult to develop a set of generally applicable environmental scenarios and suggests the need for compound or species-specific environmental scenarios. A typical definition of spatio-temporal aspects (e.g., edge of field), and a limited list of typically chosen species may lead to practical handles for model developers.

Consistency of timing and location of exposure between fate and ecological modeling needs special attention, because it is crucial for a meaningful analysis of possible effects to which the timing of application, fate dynamics, and population dynamics in space and time are realistically connected. This is not automatically the case (see example below), and inconsistency can have a major influence on the realism of the individual species exposure and subsequently on the simulated impacts of PPP application. An example is given by the use of a fixed application time point for a PPP as defined in a GAP, together with a MEM where population dynamics emerge in time based on environmental factors, for instance, specific seasonal weather scenarios. In this example, the GAP and simulated population dynamics should be consistent, or synchronized, for example according to the specific weather scenario. Otherwise, based on a GAP, an application of the PPP is simulated at a time when population dynamics have not yet started due to the weather conditions, for example, a later start of vegetation period in a specific year, hence minimizing the exposure. In an ideal situation, consistency would be reached by a befitting definition of the environmental factors shared by the exposure and ecological scenarios. The harmonization of spatial and temporal scales in the exposure also needs attention in probabilistic approaches, where distributions of time-relevant factors and timing in the exposure modeling require synchronization. In addition, for the use of MEMs that simulate spatial dispersal, it must be checked to what extent simulated landscape dynamics (e.g., by crop harvest or ploughing, which can change attractiveness of certain parts of a landscape for a species) is realistically synchronized with the simulated PPP application timing.

3.4. A catalog of questions for environmental scenario development

The aims of this catalog of questions are two-fold: (i) to help the modeler set up and document a well-founded, comprehensive, and structured set of environmental scenarios to be used in applying a MEM and analyzing its results for risk assessment, and (ii) to help the risk assessor conduct a comprehensive evaluation of the environmental scenarios used to perform the risk assessment. For a meaningful evaluation of an environmental scenario, it is recommended that a concise justification for each point is provided.

The questions for environmental scenario development start with a section on the general regulatory problem definition and aspects of the selected model, because this information is required to set up the environmental scenarios. The environmental scenario development can eventually be embedded in a report describing the model development as a whole (e.g., TRACE documentation) and/or a model application for a specific application case. Hence, the integration of questions for environmental scenario development into the GMP checklist should be sought. In the case that this integration is done, section 3.4.1 would become redundant.

The suggested questions below have been devised with the intent to provide a reasonable structure for a comprehensive and useful documentation of environmental scenarios. However, it is likely that in practice this catalog of questions will reveal gaps or shortcomings in its current form and should not be expected to resolve all issues surrounding the development of consistent environmental scenarios. Future adaptations, extensions, and editions of these questions are therefore likely.

3.4.1. Impact of the problem definition on the environmental scenario

Each model application process starts with formulating the problem definition. The problem definition describes how the model will be used within the risk assessment and includes the selection of the focal species or species group as well as a description of the (specific) protection goals. In our definition of environmental scenarios, the selection of the species to be modeled is usually done within the problem formulation step, unless the model application is specifically intended to identify focal or vulnerable species. This is in contrast to where the consideration of aspects of sensitivity and vulnerability (sensitivity and vulnerability-related traits) of a species is suggested as one of the first steps in the ecological scenario derivation (EFSA PPR, 2017; Franco et al., 2017; Rico et al., 2016). However, here we consider the development of environmental scenarios when the decision about the species to be modeled has already been made. Therefore, the scenario definition here focuses on the question of which environmental factors are to be considered, and how they are to be defined to enable a sufficiently protective assessment.

Naturally, the selection of the focal species and the definition of the regulatory question has a major impact on environmental scenario development, and a thorough analysis of the specific life-cycle characteristics and ecology of the selected species (e.g., life cycle duration and home-range of the specific population) and abiotic factors influencing internal and external recovery, are required during the problem definition. In this way, it is ensured that the most appropriate species is used in the modeling exercise, and that all environmental factors that are considered relevant and important for this species are incorporated, first, into the conceptual model, and later into the computer model. During the development of the environmental scenario these factors can then vary over a specific range of interest that meets the risk assessment.

The problem definition needs to explain in sufficient detail how the modeling supports the overall risk assessment, how it can assess the protection goals, and how exposure and effects are eventually integrated. The problem definition should therefore clearly detail:

a. What is the regulatory context in which the model will be used (e.g., which tier of the risk assessment)?

b. Which species or species group is going to be modeled, and why?

c. What are the protection goals, giving all necessary SPG attributes, such as ecological entity, attribute, magnitude, and temporal and spatial scale?

d. Which model outputs are required to answer the risk assessment question(s), including the required performance criteria?

e. How was the appropriate model selected (e.g., for what reason(s) have other models, if (publicly) available, been discarded, how does the model address the risk assessment question)?

f. What is the domain of applicability of the chosen model?

g. How will exposure and effects be integrated (i.e., how will the risk be determined, and at what spatio-temporal resolution)?

3.4.2. The environmental scenario layout

Once the regulatory problem has been clearly defined, as well as the MEM chosen for the analysis, more specific questions need to be addressed to further specify the environmental scenario layout. These steps could be part of the development of the conceptual model (namely, the model formulation phase of the modeling cycle). During this phase, provide an overview of which environmental scenario components should be considered in more detail. During this phase of the environmental scenario development,

clearly defend the selection of relevant environmental scenario components, either from biologically based reasoning and/or from literature review. Finally, establish an overview of environmental scenario components that require a detailed consideration by answering the following questions:

a. Is it necessary to define crop-specific environmental scenarios?

b. Is it necessary to define country or zone-specific environmental scenarios?

c. Which biotic factors are potentially relevant for the environmental scenarios used in the risk assessment (e.g., interspecific variability in intrinsic sensitivity, food availability, food quality, dispersal, reproduction, predation, competition, and habitat quality)?

d. Which abiotic factors are potentially relevant for the environmental scenarios used in the risk assessment (e.g., weather, temperature, pH, oxygen, hydrology, light, and habitat)?

e. Which spatial scale and/or spatial structures are possible to be implemented or considered in the model?

3.4.3. Assessment and analysis of the most relevant environmental scenario components

In this phase of the environmental scenario development, potentially relevant components identified during the previous scenario layout should be assessed based on their actual impact on the modeling effort and the model outputs. The main aim of this phase is to end up with a more concise list of environmental scenario components that will eventually be included in the environmental scenarios.

Similar to how a sensitivity analysis is performed to check for the sensitivity of model parameters on model outputs, the impact of changes in environmental scenario components needs to be analyzed. Due to computational and model complexity limitations, we do not expect that a full sensitivity analysis combining both model parameters and scenario components is practical. A suggestion is to cover the combination of model parameters and scenario components by performing a sensitivity analysis of the scenario components after the model parameters have been reduced to the most influential ones. For models that require only a low number of environmental scenario components as input for the model, it might be possible to perform sensitivity analyses with a full factorial design. For more complex MEMs that need a large number of environmental scenario components as input, we suggest using a combination of simulation-driven and expert-based selection of (potentially) important environmental scenario components and documenting this selection process transparently. It is recommended at this stage to check for the possibility of supporting expert-based decisions and dismissing some potentially critical scenario components from sensitivity analyses by small modeling exercises, for example, by using extreme values at the boundaries of parameter ranges.

The main aim of this step of the environmental scenario development phase is to ensure that variability of environmental scenario components that might have a large influence or impact on the model outcome will be covered in a set of environmental scenarios. Environmental factors with high influence on the model outcome need to be analyzed with regards to their environmental variability and/or uncertainty. Below is a list of question that can be answered to support such an analysis:

a. Has the influence and relevance (combination of influence or impact and variability or uncertainty) of each environmental scenario component (section 3.4.2) been assessed (either by expert-opinion, literature research, model simulation, or a combination of methods)?

b. Has the natural range of each environmental scenario component relevant for the risk assessment been considered?

c. Which spatial scale and structure is chosen, and is this choice appropriate to the problem definition (e.g., does the landscape structure influence the outcome of the risk assessment by the presence or absence of off-crop elements, connectivity, and/or landscape management)?

d. Are there biotic or abiotic components that need to be standardized between the exposure and the ecological scenarios? If so, are the exposure and the ecological scenarios consistent with each other (e.g., do the data used to define the exposure scenario and the ecological scenarios originate from the same country or zone)? If the exposure and ecological scenarios are not consistent, has the environmental scenario been demonstrated as conservative?

e. What are the agronomic practices considered in the risk assessment? How has the timing of the application within the GAP been chosen (e.g., was the worst-case application day chosen or were a number of time points for application screened)? What are the other agronomic practices that should be taken into account? How are agronomic practices and application spread into the landscape (e.g., are all fields treated identically or are differences in agronomic practices included in the environmental scenarios)?

Any possibly remaining open questions could be discussed between modelers and evaluators and possibly answered by additional exploratory simulations based on refined environmental scenarios.

3.4.4. Selection of components for the environmental scenarios, and check for representativeness and consistency

Based on previous steps of the environmental scenario development, scenario components that are relevant to the risk assessment question are clearly identified, as well as their impact on model outputs, in terms of their variability and/or uncertainty. In this last phase of the scenario development, we incorporate those components into the environmental scenarios that will be used for the risk assessment simulations.

Once these scenario components have been combined into a final set of environmental scenarios, a check for their representativeness and consistency is necessary. Ideally, model developers and users should check together with risk assessors whether a set of environmental scenarios is realistic and covers the potential realistic worst-case, for instance regarded as appropriately reflecting the characteristics of the agricultural area under focus given the regulatory question. In practice, such processes will need time, so it is likely that an exchange based on suggested environmental scenarios, comments by reviewers of risk assessment reports, and further requirements from authorities will develop. Nevertheless, it might be possible in specific cases for the applicant and authority to consult prior to the submission of a modeling study with the aim of agreeing on critical aspects of the environmental scenarios.

To ensure that the final environmental scenarios are sufficient to answer the risk assessment problem definition, the following questions should be answered:

a. Has a set of environmental scenarios been developed? A short overview about reasons and choices should be given here.

b. Do the defined environmental scenarios provide a range of environmental conditions from favorable to unfavorable characteristics regarding the resilience of the modeled population?

c. Have most uncertain and sensitive environmental scenario components been addressed either in their natural and/or reasonable ranges or have conservative values been chosen (e.g., are the landscape and climatic conditions appropriate for the organism(s) in question)?

d. Have most relevant exposure conditions and exposure pathways been assessed and considered in the environmental scenario development?

e. For spatially explicit models: Do the chosen environmental scenarios result in realistic (spatial) behavior of individuals (e.g., demonstrated by model simulations)?

f. Has an estimation of conservatism of the environmental scenarios in ecological and exposure dimensions been derived and provided (related only to environmental scenario components, not to, for example, model parameters including ecotoxicological aspects)?

g. Are the environmental scenarios representative for the risk assessment under consideration (e.g., are the spatio-temporal scales of the environmental scenarios in line with the problem definition; does the conservatism of the environmental scenarios match with the conservatism of the problem definition)?

3.5. Conclusions and outlook

With this chapter, we aimed at contributing to the ongoing development of (environmental) scenarios and provided an approach that can help model developers with the construction of representative and protective environmental scenarios, and can also help risk assessors with the evaluation of the representativeness and protectiveness of environmental scenarios. Nevertheless, during the development of this chapter, it became apparent that we have not reached the final solution. Further discussion and harmonization on the definitions of some of the terms, the process of scenario development, and evaluation of conservativeness will be required. The checklist and considerations suggested in this chapter might help to start and structure these discussions and will likely lead to a revision or an extension of the work performed in the future.

To keep the development of environmental scenarios ongoing, we encourage the development of some real examples of environmental scenario development and evaluation in the context of an actual risk assessment. For this, we made a start by means of three examples, applying the checklist with its considerations to the model application of a vole, honey bee, and macrophyte model (Appendix 8.1). However, these examples need to be developed in more detail, and potential modifications to the checklist must be considered. In addition, further discussions among and between model developers and risk assessors are required, for instance to clarify how the process of environmental scenario development can be streamlined or integrated better with model development and application.

3.6. Bibliography Chapter 3

Cedergreen, N., Norhave, N. J., Svendsen, C., & Spurgeon, D. J. (2016). Variable temperature stress in the nematode caenorhabditis elegans (Maupas) and its implications for sensitivity to an additional chemical stressor. *Plos One*, *11*(1), 21, Article e0140277. https://doi.org/10.1371/journal.pone.0140277

EFSA (European Food Safety Authority). (2016). A mechanistic model to assess risks to honey bee colonies from exposure to pesticides under different scenarios of combined stressors and factors. *EFSA Supporting Publications*, *13*(7), 1069E. https://doi.org/10.2903/sp.efsa.2016.EN-1069

EFSA PPR (Panel on Plant Protection Products & their Residues). (2010). Scientific opinion on the development of specific protection goal options for environmental risk assessment of pesticides, in particular in relation to the revision of the Guidance Documents on Aquatic and Terrestrial Ecotoxicology (SANCO/3268/2001 and SANCO/10329/2002). *EFSA Journal*, *8*(10), 1821. https://doi.org/10.2903/j.efsa.2010.1821

EFSA PPR (Panel on Plant Protection Products & their Residues). (2014). Scientific opinion on good modelling practice in the context of mechanistic effect models for risk assessment of plant protection products. *EFSA Journal*, *12*(3), 3589. https://doi.org/10.2903/j.efsa.2014.3589

EFSA PPR (Panel on Plant Protection Products & their Residues). (2015). Scientific opinion addressing the state of the science on risk assessment of plant protection products for non-target arthropods. *EFSA Journal*, *13*(2), 3996. https://doi.org/10.2903/j.efsa.2015.3996

EFSA PPR (Panel on Plant Protection Products & their Residues). (2017). Scientific opinion addressing the state of the science on risk assessment of plant protection products for in-soil organisms. *EFSA Journal*, *15*(2), e04690. https://doi.org/10.2903/j.efsa.2017.4690

EFSA PPR (EFSA Panel on Plant Protection Products and their Residues). (2018a). Scientific opinion on the state of the art of Toxicokinetic/Toxicodynamic (TKTD) effect models for regulatory risk assessment of pesticides for aquatic organisms. *EFSA Journal*, *16*(8), 5377-5565. https://doi.org/10.2903/j.efsa.2018.5377

EFSA PPR (EFSA Panel on Plant Protection Products and their Residues). (2018b). Scientific opinion on the state of the science on pesticide risk assessment for amphibians and reptiles. *EFSA Journal*, *16*(2). https://doi.org/10.2903/j.efsa.2018.5125

EFSA SC (EFSA Scientific Committee,Diane Benford, Thorhallur Halldorsson, Anthony Hardy, Michael John Jeger, Katrine Helle Knutsen, Simon More, Alicja Mortensen, Hanspeter Naegeli, Hubert Noteborn, Colin Ockleford, Antonia Ricci, Guido Rychen, Josef R. Schlatter, Vittorio Silano, Roland Solecki, Dominique Turck). (2016). Guidance to develop specific protection goals options for environmental risk assessment at EFSA, in relation to biodiversity and ecosystem services. *EFSA Journal*, *14*(6), e04499. https://doi.org/10.2903/j.efsa.2016.4499

EFSA SC (EFSA Scientific Committee, Simon More, Vasileios Bampidis, Diane Benford, Claude Bragard, Thorhallur Halldorsson, Antonio Hernández-Jerez, Susanne Hougaard Bennekou, Kostas Koutsoumanis, Kyriaki Machera, Hanspeter Naegeli, Søren Saxmose Nielsen, Josef Schlatter, Dieter Schrenk, Vittorio Silano, Dominique Turck, Maged Younes, Gerard Arnold, Jean-Lou Dorne, Angelo Maggiore, Stephen Pagani, Csaba Szentes, Simon Terry, Simone Tosi, Domagoj Vrbos, Giorgia Zamariola, Agnes Rortais). (2021). A systems-based approach to the environmental risk assessment of multiple stressors in honey bees. *EFSA Journal*, *19*(5), e06607. https://doi.org/10.2903/j.efsa.2021.6607

FOCUS (FOrum for the Coordination of pesticide fate models and their USe, J Boesten, A Helweg, M Businelli, L. Bergstrom, H Schaefer, A Delmas, R Kloskowski, A Walker, K Travis, L Smeets, R Jones, V Vanderbroeck, A Van Der Linden, S Broerse, M Klein, R Layton,O-S Jacobsen, D Yon). (1995). Leaching Models and EU Registration. *European Commission Document 4952/VI/95*. https://esdac.jrc.ec.europa.eu/public_path/projects_data/focus/docs/Leaching%20models%20and%20EU%20registration.pdf

FOCUS (FOCUS Working Group on Surface Water Scenarios, J. Linders, P. Adriaanse, R. Allen, E. Capri, V. Gouy, J. Hollis, N. Jarvis, M. Klein, P. Lolos, W.-M. Maier, S. Maund, C. Pais, M. Russell, L. Smeets, J.-L. Teixeira, S. Vizantinopoulos, D. Yon). (2001). Focus surface water scenarios in the EU evaluation process under 91/414/EEC. *EC Document Reference SANCO/4802/2001-rev.2*. https://esdac.jrc.ec.europa.eu/public_path/projects_data/focus/sw/docs/FOCUS_SWS_Final_Report.doc

Franco, A., Price, O. R., Marshall, S., Jolliet, O., Van den Brink, P. J., Rico, A., Focks, A., De Laender, F., & Ashauer, R. (2017). Toward refined environmental scenarios for ecological risk assessment of down-the-drain chemicals in freshwater environments. *Integrated Environmental Assessment and Management*, *13*(2), 233-248. https://doi.org/10.1002/ieam.1801

Gergs, A., Preuss, T. G., & Palmqvist, A. (2014). Double trouble at high density: Cross-level test of resource-related adaptive plasticity and crowding-related fitness. *Plos One*, *9*(3), 13, Article e91503. https://doi.org/10.1371/journal.pone.0091503

Goussen, B., Rendal, C., Sheffield, D., Butler, E., Price, O. R., & Ashauer, R. (2020). Bioenergetics modelling to analyse and predict the joint effects of multiple stressors: Meta-analysis and model corroboration. *Science of the Total Environment*, *749*, 10, Article 141509. https://doi.org/10.1016/j.scitotenv.2020.141509

Heine, S., Schmitt, W., Gorlitz, G., Schaffer, A., & Preuss, T. G. (2014). Effects of light and temperature fluctuations on the growth of Myriophyllum spicatum in toxicity tests-a model-based analysis. *Environmental Science and Pollution Research*, *21*(16), 9644-9654. https://doi.org/10.1007/s11356-014-2886-8

Heine, S., Schmitt, W., Schaffer, A., Gorlitz, G., Buresova, H., Arts, G., & Preuss, T. G. (2015). Mechanistic modelling of toxicokinetic processes within *Myriophyllum spicatum*. *Chemosphere*, *120*, 292-298. https://doi.org/10.1016/j.chemosphere.2014.07.065

Jager, T. (2017). Making sense of chemical stress. *Application of Dynamic Energy Budget Theory in Ecotoxicology and Stress Ecology. Version*, *2*.

Jager, T., Albert, C., Preuss, T. G., & Ashauer, R. (2011). General unified threshold model of survival – a toxicokinetic-toxicodynamic framework for ecotoxicology. *Environmental Science & Technology*, *45*(7), 2529-2540. https://doi.org/10.1021/es103092a

Rico, A., Van den Brink, P. J., Gylstra, R., Focks, A., & Brock, T. C. M. (2016). Developing ecological scenarios for the prospective aquatic risk assessment of pesticides. *Integrated Environmental Assessment and Management*, *12*(3), 510-521. https://doi.org/10.1002/ieam.1718

Schmitt, W., Bruns, E., Dollinger, M., & Sowig, P. (2013). Mechanistic TK/TD-model simulating the effect of growth inhibitors on Lemna populations. *Ecological Modelling*, *255*, 1-10. https://doi.org/10.1016/j.ecolmodel.2013.01.017

Topping, C. J., & Odderskaer, P. (2004). Modeling the influence of temporal and spatial factors on the assessment of impacts of pesticides on skylarks. *Environmental Toxicology and Chemistry*, *23*(2), 509-520. https://doi.org/10.1897/02-524a

Weber, D., Schaefer, D., Dorgerloh, M., Bruns, E., Goerlitz, G., Hammel, K., Preuss, T. G., & Rattey, H. T. (2012). Combination of a higher-tier flow-through system and population modeling to assess the effects of time-variable exposure of isoproturon on the green algae Desmodesmus subspicatus and Pseudokirchneriella subcapitata. *Environmental Toxicology and Chemistry*, *31*(4), 899-908. https://doi.org/10.1002/etc.1765

4. Documentation and evaluation of data used in mechanistic effect models

Benoit Goussen, Michael Fryer, Melissa Reed, Thomas G. Preuss, Josef Koch, Magnus Wang, Joachim Kleinmann, Peter Vermeiren

Acknowledgment: Vanessa Mazerolles

4.1. Introduction

Pesticide regulation in the European Union requires an evidential basis to support decision making. Therefore, where an applicant relies on a mechanistic effect model to demonstrate that no unacceptable effect will occur following use of a plant protection substance or product, it is important to clearly demonstrate how the model has been used to address a particular risk assessment question. How and why the model has used the available empirical data in answering the risk assessment question is an important aspect of this. The quality and appropriateness of the input data used in the model will in turn influence the quality and appropriateness of the model outputs. To this end, the regulatory risk assessors will need to review the underlying studies providing the data used in the model implementation.

Because the use of mechanistic effect models in pesticide risk assessment is still relatively novel, clear and transparent documentation of the quality and appropriateness of input data is of particular importance. The documentation must demonstrate the fitness for purpose of a mechanistic model for risk assessment to regulators who might be less familiar with this method. Clarity and transparency in presenting how a model is developed and used in risk assessment will increase confidence in the suitability of that model, provide a solid foundation for future applications of the model, and increase familiarity to promote greater uptake of mechanistic models in the future.

Development and use of mechanistic effect models to support risk assessment may be described by the modeling cycles (see Figure 4.1). This illustrates how data are required both for model development, to setup and validate the model, and for the case-specific use of a model to address a particular risk assessment problem.

Currently, the primary use for mechanistic effect models in pesticide risk assessment is as part of a higher-tier modeling approach. As such, this chapter covers documenting and evaluating the use of data for modeling environmental risk assessment of pesticides. It is proposed that this documentation should follow a sequential two-step process, part A as described in section 4.2 and part B, as described in section 4.3 of this chapter.

In general, data used in the model should be considered in view of the regulatory problem formulation for which they are used (i.e., the higher-tier risk assessment question under consideration). This part requires a detailed description of the relevant data that were used for a specific parameter or aspect of the model. It should be clear for what purpose the data are used and how the parameters were derived from the study or studies, including any transformation of the data.

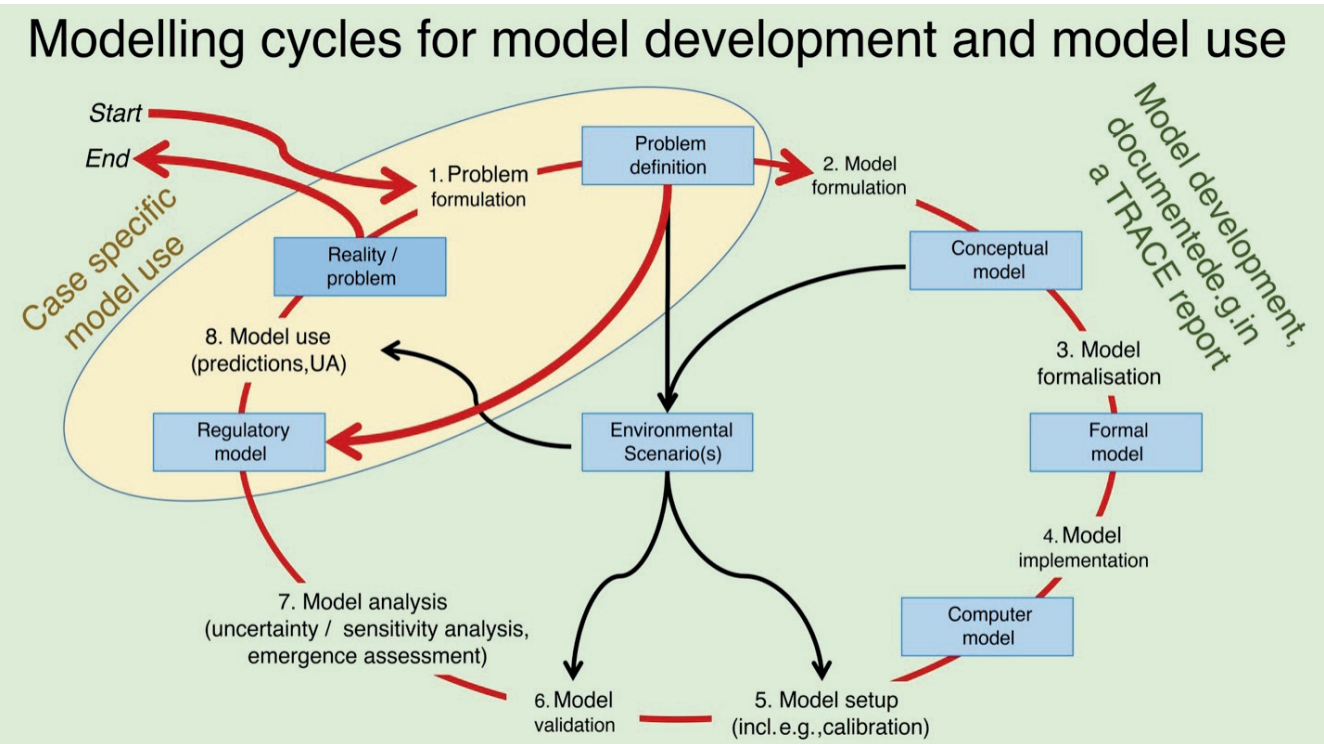


Figure 4.1: Modeling cycles as used during model development and a case-specific model use in support of regulatory risk assessment.

The first part of this chapter in section 4.2 (part A of the data appraisal) covers the documentation of data available for both the development and case-specific application of the model. As such, this part focuses on the literature search for all data supporting the model, and on presenting and identifying the relevant and reliable studies. It should be clear how data were searched for, and which data were selected or discarded for model development, calibration, or validation. The reasoning behind the decisions over selecting or discarding data should also be clear and justified. The identification of the most sensitive parameters for the model and risk assessment question is also included.

The second part of this chapter in section 4.3 (part B of the data appraisal) focuses on providing a more detailed description and evaluation of the data used to inform the most sensitive parameters used in the model. This more detailed consideration of the key data is needed to allow the regulatory risk assessor to transparently evaluate the use of data in the model and to understand the associated uncertainties.

4.2. Documentation of data used in the model (Part A)

4.2.1. Overview of available guidance

Guidance on the use of models for risk assessment of pesticides is provided by the EFSA scientific opinion on good modeling practice in the context of mechanistic effect models for risk assessment of plant protection products (EFSA PPR, 2014), as well as the scientific opinion on the state of the art of toxicokinetic/toxicodynamic (TKTD) effect models for regulatory risk assessment of pesticides for aquatic organisms (EFSA PPR, 2018a). These scientific opinions provide recommendations on how to apply and document the use of mechanistic effect models.

In particular, the scientific opinion on good modeling practice presents the steps to focus on for model development in general. These steps focus on the modeling cycle (Figure 4.2). As such, the documentation of the (i) problem definition, (ii) conceptual model, (iii) formal model, (iv) model implementation in a computer code, (v) model analysis, including sensitivity, uncertainty analysis, and comparison with data, and (vi) model evaluation with regard to the specific protection goals are the main steps to be considered. The scientific opinion on TKTD modeling refined and applied these steps for TKTD models in particular.

To support transparent and reproducible model development and application, the steps taken to search, review, select, and use data throughout this modeling cycle should be documented.

In addition to these guidance documents, several authors have been tackling the challenge of model documentation. Among these, TRACE is one of the most advanced and comprehensive standards for documenting a model (Augusiak et al., 2014; Grimm et al., 2014; Schmolke et al., 2010).

TRACE documents are designed to be supplementary material that supply evidence that the model was thoughtfully designed, correctly implemented, thoroughly tested, well understood, and appropriately used for its intended purpose.

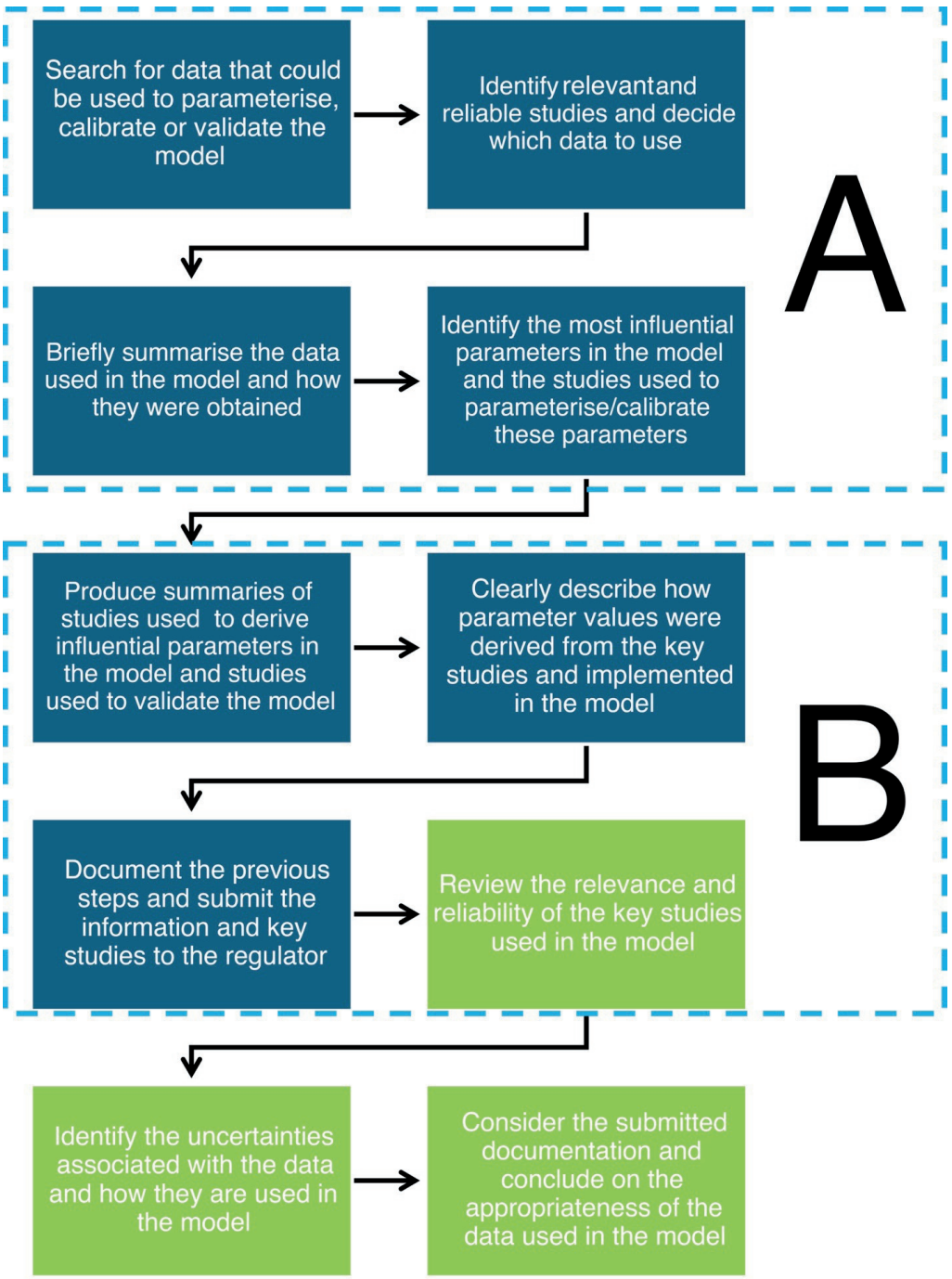


Figure 4.2: Flowchart of the tasks involved in documenting and evaluating the use of data in a mechanistic effect model for the purposes of addressing a risk assessment problem. A refers to section 4.2, and B refers to section 4.3.

4.2.2. Literature research

During modeling, a systematic search of available literature is conducted. Systematic literature reviews are critical, they ensure that the necessary information to build, parameterize and validate models has been identified. This is often a first step after defining the modeling problem but can also be conducted iteratively throughout the modeling cycle when additional data and knowledge are needed. Documenting these reviews in a clear and exhaustive way allows the user of the model to evaluate how critical information has been identified and considered when developing the model.

Several best practice guidelines and manuals have been developed regarding systematic literature reviews (see review of Cooper et al. [2018]), including guidance relevant to literature reviews that support applications for approval of pesticide active substances (EFSA, 2011). Eight key stages shared between different guidance documents have been identified: (1) identifying who should conduct the literature search, (2) setting aims, (3) preparing for the literature search, (4) designing the search strategy, (5) bibliographic database searching, (6) supplementary search methods, (7) managing references, and (8) reporting the search, details are given in (Cooper et al., 2018). The last step is particularly critical to ensuring model users have transparency on the relevant data used to inform model development and application. Specific guidance on how to report systematic reviews is provided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement (Page et al., 2021). The PRISMA statement has been broadly taken up in the scientific community including extensions into ecology, evolution, and toxicology (Munoz & Vermeiren, 2020; O'Dea et al., 2021).

The long history of systematic reviews in medical fields has led to guidance documents that are exceedingly comprehensive and provide a solid reference basis for conducting systematic literature reviews (Cooper et al., 2018). Nevertheless, the level of detail and broad scope also makes guidance documents such as PRISMA less applicable in some of its proposed steps. By contrast, the regulatory basis and closer link to ecotoxicology in the EFSA guidance on literature review provides a more directly applicable basis of the evaluation of systematic reviews underlying the development of mechanistic effect models, although it might not cover all possible data needs that a modeler might encounter.

In this chapter, we will build on the guidance provided by EFSA, as demonstrated in the example in the Appendix 8.2. Briefly, the minimum amount of information to provide is two-fold. In a first step, the research procedure itself should be documented. The main criteria that are considered relevant for reporting include: The aim of the literature search and eligibility criteria for identified studies, the combination of search terms and the search engine or databases used, search strategies employed, and the use of limits (e.g., date, language, filters; Cooper et al. [2018]). The resulting number of hits is also documented. In a second step, the selection and extraction procedure should describe how the search hits were further considered and how the final selection was made. The final number of articles and datasets should be recorded.

4.2.3. Documentation of the selection of used or dismissed studies

Data may be needed at different stages of the modeling cycle, for the purposes of calibrating, parameterizing or validating the model. The selection of appropriate studies for use in parameterization and/or calibration and validation should be endpoint-based. An *a priori* definition of criteria for data relevance and reliability for the selected endpoint should be documented. It means that studies with no data on endpoints or containing no data on relevant endpoints should be rapidly excluded. For the excluded studies, a brief description of the reason is sufficient: no data, no relevant data, etc. If a study contains relevant data, but is deemed to be not reliable, a short justification should be given. By contrast, for the studies identified in the literature search as having the potential to be used in the model, their reliability (or internal validity) and relevance (or external validity) should be considered.

Study reliability concerns methodological quality and refers to the extent to which a study is free from bias and its findings support claims about cause and effect. A reliable study follows a sound scientific methodology, and its conduct and results are transparently reported. The reliability of a study is largely independent of how the data are subsequently used in the model or risk assessment.

An additional consideration in study reliability is the availability of data, and the efforts taken to access raw source data and clarify methodological issues. Regarding industry studies, it is the case that the underlying raw data from studies conducted according to Good Laboratory Practice (GLP) are available and are also made available to the regulatory risk assessor. At present such data may not be routinely available to other users, though some companies do make their data available to academia and/or may provide them on request. This is changing under the new EU General Food Law (Transparency Regulation from March 2021), which requires all studies assessed for registration of a pesticide to be made available by the authorities.

Regarding scientific publications, it can be difficult to obtain the raw data needed from a study to allow for an intensive analysis. With increasing attention in the scientific community to open and FAIR (findable, accessible, interoperable and reusable) practices (Stall et al., 2019), data are increasingly being archived in repositories that support FAIR principles such as Zenodo or Figshare, thereby promoting greater reuse of scientific data. Nevertheless, it often remains difficult to access original data and extract all necessary information to interpret and use these data, especially for older publications and those that have not made their data FAIR. For example, to reconstruct the tissue distribution of organic pollutants among internal tissues in sea turtles, a systematic search identified 26 studies (Munoz et al., 2021). To obtain access to data and to clarify issues regarding research methodologies, the authors contacted researchers for 18 of these studies. The final compiled literature dataset contained data that were made available by the contacted researchers, as well as data extracted directly from the publications or supplement tables or figures after digitization (Munoz et al., 2021). The latter extracted data often consisted of aggregated values such as means across a group of individuals or chemicals. This example illustrates that

even if highly reliable data were found, the eventual usefulness of these data for use in modeling might be affected by the way the data are made accessible. Consequently, to provide full transparency regarding the reliability of data used in the modeling process, documentation of data extracted from a literature search should provide proper citation to the publication as well as where and how data were extracted.

Relevant studies contain data that inform the model and the case-specific use of the model in the risk assessment. Therefore, when considering the relevance of a study, both the applicability of the data for use in the model and the applicability of the data for use in the particular risk assessment should be considered. For example, a study may contain information on a parameter used in the model, but this information may only be representative of a particular geographical area or set of conditions.

A study containing no reliable information should not be used in model development or risk assessment. A study that does not contain relevant information should not be used in model development or risk assessment; however, it could be used in the future to develop a model to address a different risk assessment question. There is further guidance on how to assess the relevance and reliability of studies that are used in the model in section 4.3.

The model documentation should include information about what data were used to decide which endpoints are relevant (e.g., during parameterization, validation), and specify which studies have been utilized. Additionally, where conflicting information is available, information should be provided on why data from one study was chosen over another or how information contained in different studies were pooled.

When documenting the use of data, both relevant studies used in model development and irrelevant studies dismissed for model development and case-specific models use should be listed. It is important to document and justify why, for instance, a certain set of values or parts of a curve were selected, but also why certain data or information were discarded or not considered relevant. Pay particular attention to review or meta-data articles, where the reference of the original article and data should be listed. The level of description of the relevant studies will depend on their reliability and their contribution to the sensitivity analysis (see also Chapter 7). The reason why a study is considered not relevant should be reported (for example: "Does not contain data relevant for the intended uses of the pesticide") together with its level of reliability as it will allow the use of this data in another context in future (i.e., to address a different risk assessment question).

The information on the studies used in the model could be presented in a tabular format with entries as proposed in the template below. If using this format, it is recommended that a separate table is produced for each endpoint used in the model. This template is followed by examples to facilitate understanding.

4.2.4. Template proposal for presenting data selected for modeling

Ref: Reference of the study (Author, date, reference where the data could be found).

Habitat and environmental conditions: Details of the geographical location on the study (for example the country) and the habitat (type of crop, natural area...). Lab test and/or water media.

4.2.4.1. Methodology: Explain briefly what have been measured and how

Values: Precisely state the value considered in modeling. If several values, please indicate the difference between them (for example month of sampling, stage of individuals...).

Reliability: Indicate if reliable or less reliable or not reliable. Add a brief explanation on reliability status (cross reference to part B is possible for studies and/or data that will be further detailed). For estimation of the level of reliability some indication on how to assess reliability is presented in part B of this document.

Relevance: Indicate if relevant or less relevant or not relevant. Add a brief explanation on relevance status (cross reference to part B is possible for studies and/or data that will be further detailed). For estimation of the level of relevance some indication on how to assess reliability is presented in part B of this document.

4.2.5. Examples of selecting data for modeling

Disclaimer: The examples provided below do not reflect applicant or risk assessor official points of view regarding the studies referenced in these examples. They only provide examples of the level of detail that is expected for summaries according to section 4.2.

4.2.5.1. Honey bee (adult pollen requirement)

Information is summarized on pollen consumption by worker bees. This is relevant for the daily pollen need per adult model parameter. In the model the adults consume this amount of pollen from the pollen stores per day.

Table 4.1: An example of study summaries considered for modeling (honey bee [adult pollen requirement]).

Ref	Habitat and environmental conditions	Methodology	Values	Reliability	Relevance
(Pernal & Currie, 2000)	Laboratory study under controlled conditions (30 ± 1 °C and 70% relative humidity).	Newly emerged honey bees were fed sucrose syrup and seven different pollen diets. Pollen consumption per bee was measured.	1.5 mg pollen/ bee/d	Fully reliable. Clearly reported, replicated study including statistical analysis.	Relevant. Contains information on pollen consumption by worker honey bees.
(Yang et al., 2021)	Laboratory study under controlled conditions (34°C and 60 ± 10% relative humidity)	Newly emerged bees were fed with sucrose syrup and pollen. Polyfloral bee pollens were collected with a pollen trap and contained at least 5 different pollen types. Pollen and nectar consumption was measured.	0.67 mg pollen/bee/d	Fully reliable. Clearly reported, replicated study including statistical analysis.	Relevant. Contains information on pollen consumption by worker honey bees at different ages.

* Common shrew (litter size)

Table 4.2: An example of study summaries considered for modeling (common shrew [litter size]).

Ref	Habitat	Methodology	Values	Reliability	Relevance
(Skarén, 1973); (Tab. 1)	Trapped in bushes/ wood, Central Finland	Although no details about the determination of litter size were given it can be assumed that embryos in utero were determined because animals were snap trapped.	First litter: 7.7 (± 0.29 SE[a], n = 47, range = 5-11, in May) Second litter: 8.3 (± 0.08 SE[a], n = 12, range = 6-11, late May or early June)	Less reliable. Probably determined by dissection, but not stated in paper. Data for the second litter are less reliable, due to the small sample size.	Relevant. Contains information on the number of embryos per trapped female
(Schmidt et al., 2009); (Tab. 1)	Meadows, Denmark	Number of foetuses determined for autopsied females (which died in live traps) over about two years (summer 1998 to spring 2000).	Site / Grazing / N Fetuses / SD / N East / No / 7.67 / 0.82 / 156 Low / 5.50 / 2.38 / 144 High / 6.33 / 1.54 / 56 West / No / 8.00 / - / 49 Low / 6.50 / 0.71 / 60 High / - / - / 0	Very reliable. Large sample size collected over 2.5 years.	Relevant. Contains information on the number of embryos per trapped female

[a] In the original paper it was stated that this is the standard deviation (SD), but own calculations indicated that this is the standard error (SE).

4.2.6. Justify which studies need more detailed descriptions in Part B

Part A of the data appraisal outlines the studies that have been determined for use of the parameterization, calibration and validation of a mechanistic effect model. The key endpoints relevant for the application of the model to a particular risk assessment question should be determined next. This also should include a more detailed assessment of the critical studies providing these endpoints. This will enable the regulatory risk assessor to make their own evaluation of the reliability and relevance of the studies used in the model for the specific risk assessment question. It can be identified whether a study requires more detailed consideration based on the reliability of the study and the results of the sensitivity analysis (see Figure 4.3). For reliable studies that provide data on sensitive endpoints, a more detailed summary of the study and a more transparent consideration of the reliability and relevance of the study will be needed. By focusing on the detailed consideration on the more reliable studies that provide data on sensitive endpoints, the effort of the model developer and/or user in documenting the selection of data can be focused on areas that have the greatest potential to impact the use of the model in addressing the regulatory risk assessment problem.

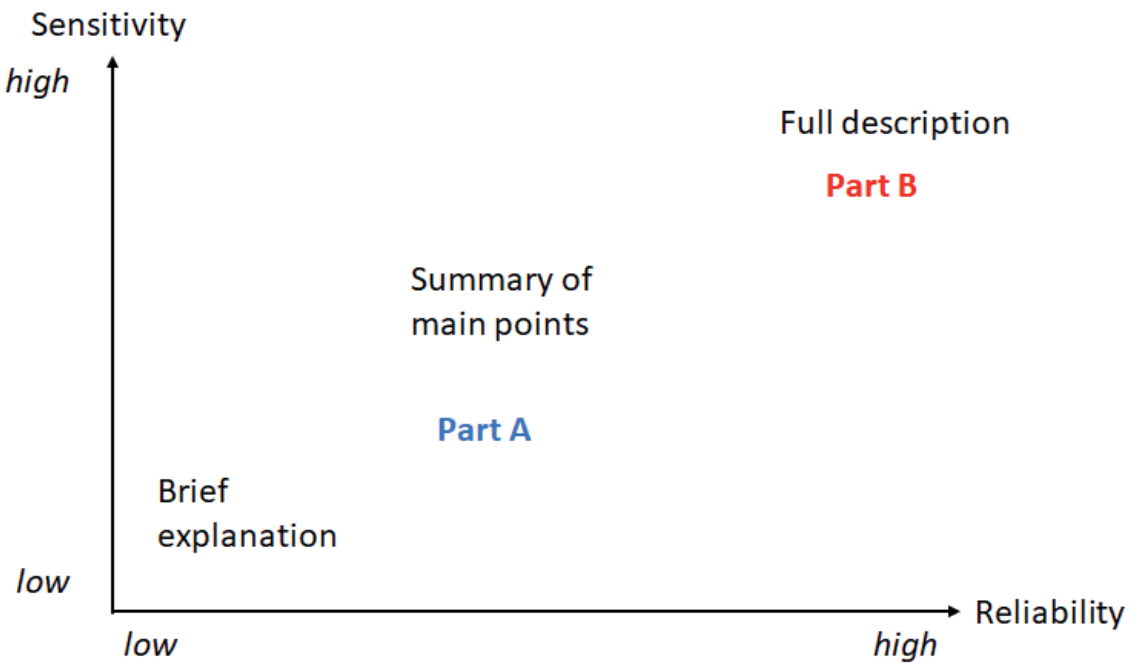


Figure 4.3: Level of detail required in the consideration of studies used in the development and/or use of a model according to the sensitivity of the endpoint and reliability of the study.

4.3. Evaluation of key data (Part B)

4.3.1. Relevance and reliability criteria in more detail

When a model is used as part of a risk assessment, for the regulatory authority to be confident in the use of that model and to understand the associated uncertainties, it is necessary for the regulatory risk assessor to assess the relevance and reliability of the data used in the model. Where data from a study are used in the model, it may be necessary to consider the whole study or only parts of that study. This assessment of the data used in the model will be particular to the case specific use of the model. This needs to be clear when describing the relevance and reliability of a study to avoid unnecessary duplication (where common aspects are reliable) or rejection of suitable information in the future (where a study was not relevant for the specific use). It is also necessary to separate toxicity studies from other information used to describe the life history of the species. In general, toxicity data should adhere to all the standard regulatory requirements (for example be conducted according to GLP, follow standard protocols where possible, and have concentrations of the test substance confirmed by analysis).

Studies on toxicity of the chemical being assessed should be reported in the usual way using a standardized format for study summaries, such as the OECD template. They will be assessed using the standard quality criteria required under the regime of plant protection products regulation (1107/2009 and related guidance documents) and are not discussed further in this document.

For other studies, an assessment of reliability is necessary, but a scoring system, such as given in Klimisch et al. (1997), that stresses the importance of GLP is not suitable as most information used in a model is unlikely to be obtained from GLP studies. Moermond et al. (2016) include more detailed reliability and relevance criteria which may be useful for informing an assessment of studies used to obtain information for modeling, but they also aimed at regulatory toxicity studies so it is not directly applicable. The information is also likely to come from studies not specifically designed for the model, so what has been measured may not be exactly what is required.

It should be noted, that relevant work on literature data extraction was initiated by EFSA (EFSA 2015; Lahr et al., 2023), where the use of critical appraisal tools (CATs) is promoted. As part of those CATs, templates for various purposes are defined, and having a CAT specifically designed for assessing modeling data could be added. The development of such modeling data-specific CAT was beyond the scope of the MAD group, but the principles outlined in this section could form the basis for future development.

4.3.1.1. Reliability

The reliability of a study relates to how well conducted and reported the study is. There are many aspects to this, and due to the wide range of study designs that can be needed to obtain the range of data required to develop, test, and use a model, it is not possible to be prescriptive about what constitutes a reliable study. The table below includes some guidance on the information that should be provided to justify the level of reliability of a study.

Table 4.3: Considerations for a study reliability.

Classification	Description
Fully reliable	• Scientifically sound study design with sufficient details provided to have confidence in reported results (e.g.): a. Methods fully explained b. Adequate replication c. Controls used when applicable d. Adequate data availability (raw data preferred, or if available on request; sufficient summary data to clearly show trends or differences between trials; provide information on variability...) e. Adequate statistical analysis f. Environmental conditions clear • Clear link provided between what was measured in the study and the information used in the model.
Reliable with restrictions	• Scientifically sound study design, there are limitations that prevent it being assessed as fully reliable. The restrictions to its reliability must be fully explained.
Unreliable	• Insufficient study details or poorly presented results such that no confidence in study conclusions (e.g.): a. Lack of information of test conditions b. Insufficient replication c. Lack of appropriate statistical methodology d. Deficiencies in reporting of study data

4.3.1.2. Relevance

The relevance of a study relates to how suitable it is for the specific use it is being employed for. A study can be fully reliable, but not relevant for the purpose it is used for. While the reliability of a study can be somewhat separated from the use in modeling, the relevance is fully linked. The more information about the relevance (and boundaries of the relevance) can be given, the more useful it will be for assessing the overall suitability of the regulatory model for addressing the regulatory question.

The relevance of the study cannot be defined by a simple yes / no classification, so a more extensive definition is required. When a study has been conducted for a different purpose than the modeling it is likely that parts of the study will be relevant, but other parts may not be.

The table below provides some guidance on the information that should be considered when justifying the level of relevance of a study and its limit of use.

Table 4.4: Considerations for a study relevance.

Classification	Aspects of the study	Limit of use
Relevant	Specify here what parts of the study are relevant to you model (for example, the study could cover two species, one relevant for your model, one not).	What are the limits on how this information can be used (for example it may only be suitable for models used in certain crops)?
Some relevance	Are there aspects of the study which inform the model, but are not directly relevant?	Where could this information be used?
No relevance	Parts of the study that are not relevant for the model can be specified here and no further reporting of these aspects is required (for the example of two species, only summarize the information on the species of concern).	N/A
Other comments		

When assessing the relevance of a study the regulatory risk assessor should consider and then clearly explain the relevance of the different aspects of the study and any limitations regarding how data from the model can be used in risk assessment. While some aspects of relevance and reliability can be judged on a study-by-study basis, ultimately it will also need to be determined and reported on an endpoint-by-endpoint basis, where information from the individual study evaluations will need to be sorted by endpoints, particularly as some values used in the model may come from more than one study.

4.3.2. How to summarize data including templates for study summaries

It is necessary to provide sufficient information in the summary of each study that is being relied on for transparency about what the study measured or tested, how this information has been used to obtain information for the model, and what possible restrictions on its use exist. The standard study summary templates used for regulatory assessments are more detailed than is likely to be required for many of the studies used, but flexibility is required. Key studies used to obtain critical information for a regulatory model require a detailed summary to allow appropriate evaluation. It is suggested that the model developer and/or user classifies studies according to whether they provide key information (for example data to define one or more influential parameters in the model) or provide less critical information (for example is one of several sources defining a value, or is used to define a less sensitive parameter). Such a summary must allow the reader to understand how a specific model value was derived. In particular, the reader should be able to reproduce the steps performed by the applicant

to derive the specific values used in the model. Uncertainties around these values used should also be clearly described in the summary and the selection of the parameter value justified.

It may be necessary to provide information on how the data were extracted from studies (e.g., obtained from authors, copied from tables or appendices, digitized from plots) and pre-processed (e.g., conversions of units, fitting of models or derivation of summary statistics), as well as if and why some data or part of the data were excluded to explain how the final dataset used in the model was obtained. The process used to select the final value is important, not just the final result. Additional information could be included, such as approaches applied to minimize bias, and strategies employed to set assumptions when information is lacking or unclear (Page et al., 2021).

A template for summarizing the data used is shown below. The model developer and/or user may use this template to provide information on the study and how data from the study are used in the risk assessment. The regulatory risk assessor can then review this information and use this template to add their consideration regarding the reliability and relevance of the study and/or data.

Table 4.5: A template for summarizing data used for modeling.

Applicant summary	
Reference	Author, year. Study title. Journal reference.
Which information is summarized? For what purpose?	Identify which aspects of the study are summarized and which parameter(s) in the model the study provides data on.
Materials and methods	Summarize the purpose of the study, how the study was conducted, and under what conditions.
Results and conclusion	Summarize the findings of the study, including any statistical analysis for relevant studies, and the conclusions of the study author(s).
Use in model	Set out how the specific information in the study has been used in the model. If any transformation of the information presented in the study was required, include this here.
Endpoints for modeling	List specific values that are taken forward to the model.
Evaluation by regulatory authority	
Reliability	Comment and conclude on reliability status based on information provided in part A and B.
Relevance	Comment and conclude on relevance status based on information provided in part A and B.

4.3.3. Examples summary

Disclaimer: The examples provided below do not reflect applicant or risk assessor official points of view regarding the studies referenced in these examples. They only provide examples of the level of detail that is expected for part B summaries. It should be noted that an evaluation of uncertainty stemming from literature data is important, but was not performed as a part of this evaluation exercise.

Table 4.6: An example of summarized data used for modeling (honey bee).

Applicant summary	
Reference	(Pernal & Currie, 2000) Pollen quality of fresh and 1-year-old single pollen diets for worker honey bees (Apis mellifera L.). Apidologie. 31: 387–409.
Which information is summarized? For what purpose?	Information is summarized on pollen consumption by worker bees. This is relevant for the daily pollen need per adult model parameter. In the model. the adults consume this amount of pollen from the pollen stores. This study also includes information on dietary protein content, bee mortality, protein content of the hypopharyngeal gland, and ovary development but these are not reported in this summary.
Materials and methods	In the study of Pernal and Currie (2000), newly emerged honey bees were fed sucrose syrup and seven different single-pollen diets in controlled laboratory conditions. In comparison to single pollen a commercial bee food BeePro were given. The investigated single pollen diets were: *Malus domestica* Borkh., *Brassica campestris* L., *Phacelia tanacetifolia* L., *Melilotus officinalis* (L.) Pall., *Helianthus annuus* L., *Pinus banksiana* (Lamb.). Pollen was collected from blooming trees and field crops using honey bee (Apis mellifera) colonies fit with OAC pollen traps in Manitoba, Canada. The pollen was given either fresh or after 1 year of storage in the freezer. The bees were tested in cages in cohorts of 150 newly emerged adult workers, with sucrose solution and pollen diet fed ad libitum. Diet was replenished on days 3 and 8. Test cages were incubated at 30 ± 1 °C and 70% relative humidity. The consumed diet per cage was measured by weighing the diet trays before and after the bees consumed the diet at day 3, 8 and 14. The weight was corrected for water loss from not consumed duplicates. The mean pollen consumption per bee per day was calculated. The Mean numbers of bees in treatments were calculated from the populations in cages at the midpoint of each experimental time interval, corrected for sampling loss and mortality, and weighted by the duration of the interval. Differences in diet consumption between treatments were examined using a split-plot design ANOVA with pollen age, replicate and diet as main factors, and time as a repeated measure.
Results and conclusion	Consumption rates varied between 0.2 and 6.2mg/bee/day for all different pollen and time points. Workers consumed significantly more fresh than old pollen. The consumed diet also varied significantly with type of pollen, with BeePro and Pinus being eaten less. This difference could not be explained by protein content. Maximum consumption occurred within the first 3 days and decreased significantly for the 8-day and 14-day measurement, except for Pinus pollen diet.
Use in model	For the modeling approach we selected the measurement after 14 days, because we are interested in the pollen requirement of mature worker bees. The stored pollen showed much higher variability than the fresh one and are not relevant for the environment so were excluded from the analysis. For the fresh pollen we excluded Pinus and BeePro because they seem to be suboptimal for the bees. For the other pollen types consumption rate varied between 1.2 mg/bee/day and 1.5 mg/bee/day. We decided to take the maximum pollen consumption of fresh pollen at day 14 as parameter into the model. This selection is assumed to be a realistic worst-case, because higher pollen consumption means higher doses per bee and higher demand of pollen for the hive.
Endpoints for modeling	DAILY_POLLEN_NEED_ADULT = 1.5 mg/bee/d

Evaluation by regulatory authority	
Reliability	The study is clearly reported and includes key information regarding the methodology and results. Six individual trials were performed, enabling all treatment combinations to be replicated three times. The full range of values is not presented but means and standard deviations are reported. Differences in pollen consumption between pollen types, pollen ages and over time have been analyzed statistically. The parameter value selected from this study is the worst-case mean 14-day pollen consumption per bee. How this value has been derived is clearly explained. The variability of this parameter within and between plant species is relatively low in this study, except for Jack pine (which is of limited relevance for the risk assessment). The link between what is measured in the study (pollen consumption per worker bee) and how it is used in the model (daily pollen requirement per adult) is clear. Classification: Fully reliable
Relevance	The study contains data on pollen consumption by worker bees, which is relevant for the daily pollen requirement per adult bee model parameter. The species studied, the honey bee (*Apis mellifera* L.).), is the species of interest for the risk assessment. The study was conducted under controlled conditions, so some consideration of extrapolation of the data to field conditions is required. Pollen consumption data are available for a range of crop and other plants species – apple, cabbage, Lacy phacelia, yellow sweet clover, common sunflower and Jack pine. This dataset includes a relevant plant species for the risk assessment for pome fruit (apple). Relevant – Results for pollen consumption per bee per day. Some relevance – Data on bee mortality input into the calculation of pollen consumption per bee. No relevance – Results for dietary protein content, bee mortality, protein content of the hypopharyngeal gland and ovary development are not used in the model and are not presented in this summary.
Comments	No additional comments

Table 4.7: An example of summarized data used for modeling (common shrew).

<table>
<tr><th colspan="2">Applicant summary</th></tr>
<tr><td>Reference</td><td>Schmidt, N.M, Olsen, H., and Leirs, H. (2009). Livestock grazing intensity affects abundance of Common shrews (Sorex araneus) in two meadows in Denmark. BMC Ecology. 9(2). (https://bmcecol.biomedcentral.com/articles/10.1186/1472-6785-9-2).</td></tr>
<tr><td>Which information is summarized? For what purpose?</td><td>Provides information relevant for the litter size parameter. Only results related to foetus numbers and date of capture have been considered.</td></tr>
<tr><td>Materials and methods</td><td>Common shrew population characteristics were investigated in semi-natural grasslands. To assess potential differences according to grazing treatment, 2 ungrazed control, 2 low-grazing intensity (sheep) and 2 high-grazing intensity (cattle) meadows were sampled. The plots were distributed across 2 sites in Western Denmark, separated by approximately 4 km. Meadow vegetation included Festuca rubra, Phleum pratense, Poa trivialis, Poa pratensis, Ranunculus repens and Bryophytes. Meadows were described as waterlogged during the study period. Environmental conditions during the study period are not reported in further detail.

Small mammal trapping was conducted from summer 1998 to spring 2000 on the 6 study plots. Common shrews were captured using Ugglan traps, with 36 traps used per plot. Traps were bedded with hay, and baited with rolled oats, apple, and in some sessions minced meat. Every 4 weeks trapping was conducted for periods of 3 consecutive days and nights (16200 trap nights in total). Live individuals were handled in the field, tagged for identification, and released at the point of capture. Common shrews found dead in traps were autopsied. The following parameters were measured– body mass, number of uterine foetuses and number of uterine scars. The sex ratio was also determined for each plot.</td></tr>
<tr><td>Results and conclusions</td><td>During the trapping period 570 individual Common shrews were captured and autopsies were performed on 465 individuals. The number of individual shrews caught in each trapping session showed large fluctuations between trapping sessions as well as between years. Results for the number of fetuses found in autopsied female common shrews are summarized in the following table.
<table>
<tr><th rowspan="2">Meadow</th><th rowspan="2">Grazing</th><th rowspan="2">Number trapped</th><th rowspan="2">Number autopsied</th><th colspan="2">Number of foetuses</th></tr>
<tr><th>Mean</th><th>Standard deviation</th></tr>
<tr><td>East</td><td>None</td><td>163</td><td>156</td><td>7.67</td><td>0.82</td></tr>
<tr><td></td><td>Low</td><td>186</td><td>144</td><td>5.50</td><td>2.38</td></tr>
<tr><td></td><td>High</td><td>71</td><td>56</td><td>6.33</td><td>1.54</td></tr>
<tr><td>West</td><td>None</td><td>81</td><td>49</td><td>8.00</td><td>Not reported</td></tr>
<tr><td></td><td>Low</td><td>65</td><td>60</td><td>6.5</td><td>0.71</td></tr>
<tr><td></td><td>High</td><td>4</td><td>0</td><td>-</td><td>-</td></tr>
</table>
There were no statistically significant differences in the number of fetuses between grazing treatments. No difference in the number of fetuses between sites is apparent, noting the smaller dataset available for the West meadow.</td></tr>
</table>

Use in model	The dataset on the number of fetuses recorded per autopsied female is used in the model in conjunction with data from several other studies for the litter size parameter. Data are available from 2 sites, each with 3 different grazing regimes, showing considerable variability (see also standard deviations). The information from this study can be used to estimate embryo numbers for comparison with other studies. However, the study authors reported only embryo numbers, not litter size. In the model, the latter is needed.
Endpoints for modeling	Data not directly used in the model but used in conjunction with other studies to derive the litter size parameter.
Evaluation by regulatory authority	
Reliability	The study is clearly reported and includes key information regarding the methodology and results. Two replicate sites with 3 meadows per site were included. The dataset is extensive and covers multiple years. The full range of values is not presented but means and standard deviations are available. Differences in measured parameters across sites and across management practices have been analyzed statistically. Basic information is provided on the study locations, but environmental conditions have not been reported. The link between what is measured in the study (number of fetuses per female) and how it is used in the model (litter size parameter) is clear. Classification: Fully reliable
Relevance	The study does not contain direct information on litter size but contains data on the number of fetuses per female, which is indirectly relevant for the litter size model parameter. The species studied, the common shrew, is the species of interest for the risk assessment. The study was conducted in meadows in Denmark and is therefore relevant for this northern zone grassland risk assessment. Extrapolation of the data to other habitats and to Southern EU Member States would need to be justified. Relevant – Results for common shrew fetuses per female are indirectly relevant for the litter size parameter. Some relevance – Date of capture potentially provides information that would allow consideration of how fetus number varies with season, however, the study report does not include litter size data for specific seasons. No relevance – Results for the number of individuals, body mass, number of uterine scars and sex ratio are not used in the model and are not presented in this summary.
Comments	No additional comments

4.4. Bibliography Chapter 4

Augusiak, J., Van den Brink, P. J., & Grimm, V. (2014). Merging validation and evaluation of ecological models to 'evaludation': A review of terminology and a practical approach. *Ecological Modelling*, *280*, 117-128. https://doi.org/10.1016/j.ecolmodel.2013.11.009

Cooper, C., Booth, A., Varley-Campbell, J., Britten, N., & Garside, R. (2018). Defining the process to literature searching in systematic reviews: A literature review of guidance and supporting studies. *Bmc Medical Research Methodology*, *18*, 14, Article 85. https://doi.org/10.1186/s12874-018-0545-3

EFSA (European Food Safety Authority). (2011). Submission of scientific peer-reviewed open literature for the approval of pesticide active substances under Regulation (EC) No 1107/2009. *EFSA Journal*, *9*(2), 2092. https://doi.org/10.2903/j.efsa.2011.2092

EFSA (European Food Safety Authority). (2015). Tools for critically appraising different study designs, systematic review and literature searches. EFSA supporting publication 2015:EN-836. https://efsa.onlinelibrary.wiley.com/doi/pdf/10.2903/sp.efsa.2015.EN-836

EFSA PPR (Panel on Plant Protection Products & their Residues). (2014). Scientific Opinion on good modelling practice in the context of mechanistic effect models for risk assessment of plant protection products. *EFSA Journal*, *12*(3), 3589. https://doi.org/10.2903/j.efsa.2014.3589

EFSA PPR (EFSA Panel on Plant Protection Products and their Residues). (2018a). Scientific opinion on the state of the art of toxicokinetic/toxicodynamic (TKTD) effect models for regulatory risk assessment of pesticides for aquatic organisms. *EFSA Journal*, *16*(8), 5377-5565. https://doi.org/10.2903/j.efsa.2018.5377

Grimm, V., Augusiak, J., Focks, A., Frank, B. M., Gabsi, F., Johnston, A. S. A., Liu, C., Martin, B. T., Meli, M., Radchuk, V., Thorbek, P., & Railsback, S. F. (2014). Towards better modelling and decision support: Documenting model development, testing, and analysis using TRACE. *Ecological Modelling*, *280*, 129-139. https://doi.org/10.1016/j.ecolmodel.2014.01.018

Klimisch, H. J., Andreae, M., & Tillmann, U. (1997). A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. *Regulatory Toxicology and Pharmacology*, *25*(1), 1-5. https://doi.org/10.1006/rtph.1996.1076

Lahr, J., Arts, G., Duquesne, S., Mazerolles, V., de Jong, F., Moermond, C., van der Steen, J., Alalouni, U., Baujard, E., van den Berg, S., Buddendorf, B., Faber, M., Mahieu, K., Montforts, M., Smit, E., van Spronsen, R., Swarowsky, K., Chaton, P.-F., Foldrin, J., Lambin, S., & Pieper, S. (2023). Proposal of critical appraisal tools for the evaluation of ecotoxicology studies. *EFSA supporting publication* 2023:EN-7787. https://efsa.onlinelibrary.wiley.com/doi/epdf/10.2903/sp.efsa.2023.EN-7787

Moermond, C. T. A., Kase, R., Korkaric, M., & Agerstrandk, M. (2016). CRED: Criteria for reporting and evaluating ecotoxicity data. *Environmental Toxicology and Chemistry*, *35*(5), 1297-1309. https://doi.org/10.1002/etc.3259

Munoz, C. C., Hendriks, A. J., Ragas, A. M. J., & Vermeiren, P. (2021). Internal and maternal distribution of persistent organic pollutants in sea turtle tissues: A meta-analysis. *Environmental Science & Technology*, *55*(14), 10012-10024. https://doi.org/10.1021/acs.est.1c02845

Munoz, C. C., & Vermeiren, P. (2020). Maternal transfer of persistent organic pollutants to sea turtle eggs: A meta-analysis addressing knowledge and data gaps toward an improved synthesis of research outputs. *Environmental Toxicology and Chemistry*, *39*(1), 9-29. https://doi.org/10.1002/etc.4585

O'Dea, R. E., Lagisz, M., Jennions, M. D., Koricheva, J., Noble, D. W. A., Parker, T. H., Gurevitch, J., Page, M. J., Stewart, G., Moher, D., & Nakagawa, S. (2021). Preferred reporting items for systematic reviews and meta-analyses in ecology and evolutionary biology: a PRISMA extension. *Biological Reviews*, *96*(5), 1695-1722. https://doi.org/10.1111/brv.12721

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Pernal, S. F., & Currie, R. W. (2000). Pollen quality of fresh and 1-year-old single pollen diets for worker honey bees (*Apis mellifera* L.). *Apidologie*, *31*(3), 387-409. https://doi.org/10.1051/apido:2000130

Schmidt, N. M., Olsen, H., & Leirs, H. (2009). Livestock grazing intensity affects abundance of Common shrews (*Sorex araneus*) in two meadows in Denmark. *BMC Ecology*, *9*(1), 2. https://doi.org/10.1186/1472-6785-9-2

Schmolke, A., Thorbek, P., DeAngelis, D. L., & Grimm, V. (2010). Ecological models supporting environmental decision making: a strategy for the future. *Trends in Ecology & Evolution*, *25*(8), 479-486. https://doi.org/10.1016/j.tree.2010.05.001

Skarén, U. (1973). Spring moult and onset of the breeding season of the common shrew (*Sorex araneus* L.) in Central Finland. *Acta Theriologica*, *18*(23), 443-458. http://rcin.org.pl/ibs/dlibra/publication/edition/10175

Stall, S., Yarmey, L., Cutcher-Gershenfeld, J., Hanson, B., Lehnert, K., Nosek, B., Parsons, M., Robinson, E., & Wyborn, L. (2019). Make all scientific data FAIR. *Nature*, *570*(7759), 27-29. https://doi.org/10.1038/d41586-019-01720-7

Yang, K.-C., Peng, Z.-W., Lin, C.-H., & Wu, M.-C. (2021). A new design of bee cage for laboratory experiments: nutritional assessment of supplemental diets in honey bees (*Apis mellifera*). *Apidologie*, *52*(2), 418-431. https://doi.org/10.1007/s13592-020-00832-8

5. Evaluation of modular modeling approaches

Tido Strauss, Fabienne Ericher, Benoit Goussen, Joachim Kleinmann, Josef Koch, Hanna Schuster, Thomas G. Preuss

5.1. Introduction

Mechanistic effect models are increasingly considered useful tools for ecological risk assessment (ERA) of chemical stressors because they can be used to predict species population dynamics under toxic stress for different environmental scenarios (Forbes et al., 2011; Hommen et al., 2016). In particular, toxicokinetic-toxicodynamic, population, and community models are important for extrapolating toxicological effects between different environmental conditions and across biological levels (e.g., from individuals to populations). For a more general introduction to the role of mechanistic effect models in ERA, see the previous chapters, specifically Chapter 1. The aim of this chapter is to introduce the concept of model modularization for the specific case of mechanistic effect models and to explain how it could be applied more effectively in the risk assessment context.

Typically, mechanistic effect models (MEMs) need a certain level of complexity to make precise predictions about real ecological systems (Evans et al., 2013). Individual-based models (IBM) consist of a high number of states, variables, parameters, model inputs, processes, and feedback loops. This is because they need to include all relevant ecological processes to capture the dynamics of the respective system according to the purpose of a MEM, as well as how these processes adapt to relevant environmental scenarios. Due to the often-observed complexity of MEMs such as IBMs, a transparent and detailed communication of the models in ERA is key. Accordingly standardized model documentation protocols such as ODD (Overview, Design concepts and Details; Grimm et al., 2006; Grimm et al., 2010; Grimm, 2020) and TRACE (TRAnsparent and Comprehensive model Evaluation; Grimm et al., 2014; Schmolke et al., 2010) have been developed and refined during the last two decades and are consequently cited in EFSA's guidance on good modeling practice (EFSA PPR, 2014).

Beyond model documentation standards, a number of systematic approaches were developed for the application of MEMs in regulatory risk assessment. These have been published in recent years and are based

on the definition of the fundamental processes and decision steps required to answer regulatory questions in ERA (Awkerman et al., 2020; EFSA PPR, 2014; Raimondo et al., 2021; Schmolke et al., 2017). So far, these approaches and protocols have largely ignored the modular structure of complex MEMs, resulting in high workloads during their evaluation. In this chapter, we will highlight, and further elaborate on, the advantages and benefits of explicitly considering the modular structure of complex models. We will discuss how the subdivision of models into meaningful and clearly defined modules allows for their individual documentation, testing, evaluation, and assessment. Ultimately, the consideration of MEMs as modules will enable a more appropriate communication and more efficient assessment of complex MEMs.

Due to the multitude of possible combinations of modules, with potentially different qualities concerning evaluation and assessment, it is not possible to develop simple evaluation schemes for complex, modular MEMs. However, consistent approaches or protocols for the evaluation of modular models will benefit developers, reviewers, and model users alike. It will allow definition and verification of expectations for model documentation. Those in charge of the authorization of chemicals would be able to formulate their assessments on the usability and reliability of the models for risk assessment purposes if differentiating a complex model into modules. Such approaches have successfully been introduced to document the structure of IBMs by developing the ODD protocol proposed by Grimm and co-workers (Grimm et al., 2006; Grimm et al., 2010; Grimm, 2020). Here, the documentation of submodels with respect to their implementation, parametrization, as well as a testing with independent, previously unused empirical data follows the same logic as for the overall models, emphasizing the possibility of a separate consideration of submodels (EFSA PPR, 2014; Grimm et al., 2010; Schmolke et al., 2010). Grimm et al. (2020) already recommend adding a subsection in the ODD to describe under which framework submodels were coupled, how and when they interact. These aspects will be also addressed in more detail throughout this chapter.

Considering modularized modeling approaches, more advanced questions quickly appear. For example, questions of transferability arise: Which modules are already sufficiently validated and can be further re-used in other models? Can previously documented and accepted modules be freely combined in new models? Can modules easily be substituted with simpler or more complex alternatives? These considerations highlight the importance of interfaces between modules (this will be elaborated on in section 5.4.3) and lead to further questions, for example: What is an appropriate assessment strategy for evaluating a modular model? What else needs to be evaluated if a model consists mainly of previously evaluated or accepted model approaches? In this chapter we will discuss that model structure, parametrization, and validation (including domain of applicability) should be considered separately per module. Criteria might be derived to determine which model parts can possibly be adopted unchanged and which might need to be re-evaluated for an application. This procedure can then be adopted to more complex, modularly organized models (e.g., coupled population or ecosystem models including toxicokinetics-toxicodynamics [TKTD] models). This would simplify communication about complex models and allow better distinction between robust and uncertain parts of the models.

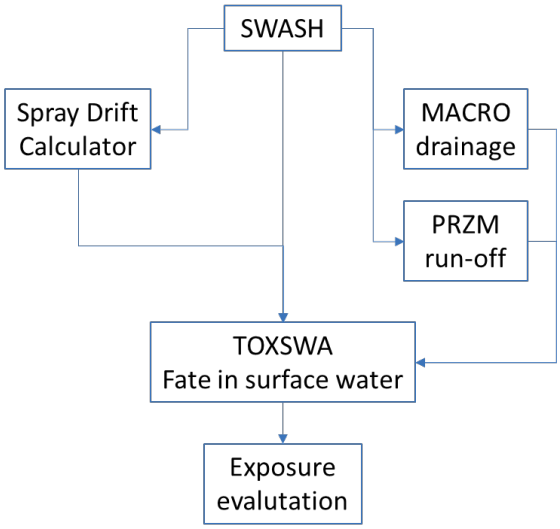


Figure 5.1: The modular structure of the FOCUS SWASH Suite to calculate surface water concentrations.

Furthermore, modularization can help to overcome challenges related to model over-parametrization by partial and independent parametrization and validation of modules, especially for models with high complexity.

Examples of three population models with different structures are given in Appendix 8.3 to illustrate how models might be described in more detail with respect to their modular structure, and also to their module-specific parametrization, validation, and range of applicability.

The terms *model* and *module* (or *submodel*) may be used in different contexts for the same model unit: Modules of a model can sometimes also be used as stand-alone model applications. Similarly, formerly independent models can become modules in the sense of sub-units (submodels) within an overarching model. The GUTS model, for example, can be used as a stand-alone model for individual survival of an organism or as one of multiple modules within a population model. Therefore, the same units in this text may sometimes be referred to as models or modules, depending on the context.

5.2. What is a modular model approach?

The modular model approach described in this chapter follows the concept of Reynolds & Acock (1997), who suggested that modular model structure should be (i) based on modules that relate directly to real world components or processes, (ii) have input and output variables that are measurable values, and (iii) communicate solely via these input and output variables. The authors state in addition that "a good modular design should decompose the problem into smaller, independent subproblems (modules), each of which can be understood and solved separately" (Reynolds & Acock, 1997).

A good example of a modular model approach is the FOCUS model suite[5]. Here, four models for water flow and pesticide fate and transport in soil (PRZM, PELMO, PEARL, MACRO) are used for groundwater risk assessment under different environmental scenarios.

However, two of these pesticide fate models (MACRO and PRZM) are also used in the surface water risk assessment as a module to simulate the input from the soil (from runoff and drainage) to surface water, and the other two (PEARL and PELMO) are used in addition for soil risk assessment. The advantage is that a specific model, in this instance a model for fate in soil (sorption, degradation and transport), needs to be evaluated and assessed only once for its use in regulatory risk assessment, and subsequently the standalone *model* can be implemented as a *module* in a more complex modeling framework, or to provide input for other modules, without the need for repeated evaluation. Another advantage is that modules can be replaced by simpler or more sophisticated versions, if necessary, without the need to change other modules, as long as the same interfaces are used. In the examples of the FOCUS suite, it is possible to conduct surface water risk assessment using either PRZM or MACRO as a soil model, and to combine either with TOXSWA or STEPS 1-2-3-4 as surface water models (Figure 5.1). This flexibility is possible due to the high level of abstraction of the FOCUS framework, with modules that are loosely coupled, enabling the design of adapters connecting the output of one module to the input of the next.

Another benefit of modularity is that it fosters interdisciplinary model development. Different modules can be developed and assessed by experts from distinct scientific areas (e.g., chemical exposure and fate, hydrology, soil science, ecology, and ecotoxicology), who apply their domain-specific expertise within the corresponding modules. These individual modules can be combined with each other into a single model with one graphical user interface, thereby integrating the complexity in the computer code (e.g., FOCUS model suite or PERSAM). In the EFSA scientific opinion on good modeling practice (EFSA PPR, 2014), a regulatory model is separated into a computer model and an environmental scenario with corresponding pesticide properties, where the computer model includes a fate (and exposure) and an effect model.

This basic concept can be expanded (Chapter 2, Figure 2.2) by modularizing relevant parts of the modeling framework for regulatory risk assessment. A special focus is given to the effect model in this section. An example application of this scheme for a population model for the midge *Chaoborus* is given in Appendix 8.3.1.

5 https://esdac.jrc.ec.europa.eu/projects/focus-dg-sante

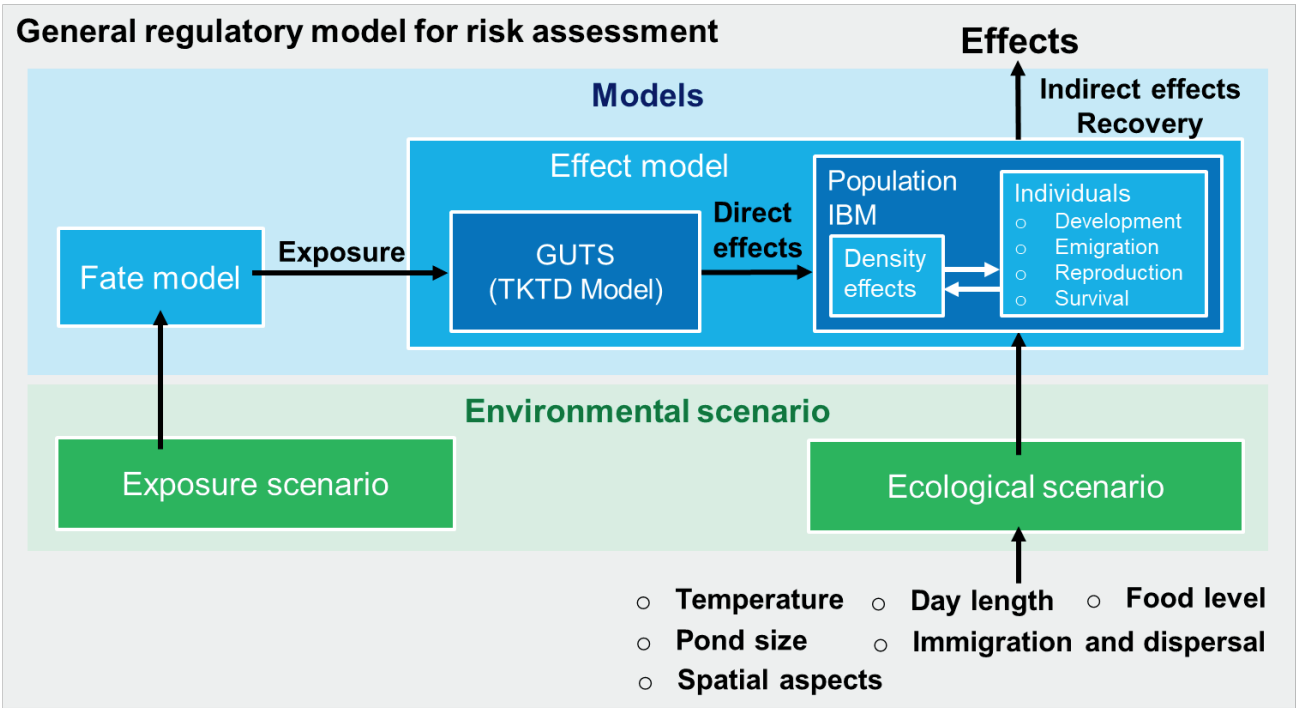


Figure 5.2: Modular representation of the *Chaoborus* population model within the effect model and its links to the fate model and the environmental scenario (details in Appendix 8.3.1.).

The modeling framework for the use of effect models in regulatory risk assessment combines a fate model that provides the exposure of a modeled individual to a toxic compound with an ecotoxicological effect model. The effect model consists of a toxicological module, for example, GUTS (General Unified Threshold model for Survival) or a dose response model, describing toxicokinetic and toxicodynamic processes to link exposure and effects, and an ecological module, typically representing a population or an ecosystem. Fate and effect models share a common environmental scenario but require different inputs. The environmental scenario and its subunits the exposure scenario and the ecological scenario, are explained in detail in Chapter 3.

The described fate and exposure-related modules form an overarching modeling framework. The ecotoxicological effect model can, vice versa, be composed of separate modules. In this context it appears useful to separate the toxicological (toxicokinetic and toxicodynamic) module, which is compound – and species-specific, from the ecological module, which is species – but not compound-specific (Figure 2.2). The complexity of the overall model can vary greatly and, depending on the respective question and the protection goal, can reflect direct and possibly indirect effects of stressors at the level of individuals, populations, or communities.

Environmental scenarios help to properly set-up interfaces between modules. For example, information on the physicochemical properties of the toxic compound is mainly needed in the fate model, whereas ecotoxicological information is required as input for the toxicological module. Defined scenarios

are very useful to distinguish which model parts or modules may need to be modified in their structure, or re-parametrized, when something in the specific modeling framework changes (e.g., the use of a different substance, changing application patterns, or new toxicological test data), and which modules can remain unchanged.

In most cases, no feedback between the fate model and the effect model will need to be considered (Figure 2.2). However, there may be processes that require feedback loops between the effect and the fate model. For example, the formation of tubes resulting from the burrowing activities of earthworms in soil can increase the vertical transport of substances; increased phytoplankton productivity can alter pH values and light intensities and potentially accelerate the degradation rate of a substance in water. In addition, the adsorption of a substance to algae or detritus can influence the bioavailability for organisms within food webs. Conceptually, these processes would still be a part of the fate module, although the technical implementation of such additional processes is often performed within the effect module.

The complexity of each module can vary, from a simple static parameter value that represents an ecological interaction, to complex dynamic models, which themselves might be built from smaller submodules. Schematic relationships between model components as shown in Figure 2.2 can be adapted to represent specific implementations of the generic modeling framework. Respective examples for three models can be found in Appendix 8.3.

The following sections further illustrate the modular approach by breaking down the toxicological and ecological modules into smaller modules and submodules.

5.2.1. Ecological module

The ecological module can simulate different biological levels (individual, population, community) depending on the risk assessment question. Currently, community models are often simulated as non-structured populations, representing a population by, for example, the specific biomass as compartments using differential equations (e.g., Aquatox [Park & Clough, 2018], CASM [Bartell et al., 2019]).

More recent individual-based population models (IBMs) follow a bottom-up approach starting at the individual level. A dynamic energy budget (DEB) model can be used as module to simulate growth, maturation and reproduction of individuals, in response to specific environmental conditions such as temperature and food availability (Gergs et al., 2014; Rakel et al., 2020). Additional submodules could be implemented to address behavior (e.g., moving) and ecological processes (e.g., overwintering strategies). However, simpler modules for individual growth can be used, for example, such as simple growth equations. Population dynamics can then evolve from interacting individuals, which is the basic principle of individual-based models (DeAngelis & Grimm, 2014; Grimm & Railsback, 2005). A detailed discussion about modular modeling approaches with special respect to soil risk assessment is given in Roeben et al. (2020).

This concept can be extended to communities, where IBMs simulate populations as part of multi-species communities, which has become more common with increasing computational power (e.g., IBC-grass; Reeg et al. [2017]). Here, each population is modeled using an individual-based module and is integrated within a community module (DeAngelis & Diaz, 2019). Communities would thus be built from the individual as the smallest unit upwards.

More details about modules that are relevant for this approach are given in section 5.4.2.

5.2.2. Toxicological module

The toxicological module translates the toxic exposure into direct effects in the ecological module using the ecotoxicological information. It can range from a simple threshold function to a dose-response function to a complex toxicokinetic-toxicodynamic (TKTD) model (EFSA PPR, 2018a). The toxicological module can either solely cover mortality or can include sublethal effects, for example, on growth or reproduction. The same threshold or dose-response function can be used to simulate lethal and sublethal effects. However, in a TKTD concept, different approaches are needed to simulate lethal and sublethal endpoints (Ashauer et al., 2011). The GUTS model is a TKTD framework explicitly designed to simulate lethal effects (Jager et al., 2011). The DEB-TKTD is a framework that also allows the user to address sublethal effects (Jager, 2020; Sherborne et al., 2020). Like the name already indicates, toxicokinetic-toxicodynamic modules can be further divided into two submodules.

First, a dose metric is calculated in the toxicokinetic (TK) module. The most useful dose metric would be the actual internal concentration at the target site of action, which has a direct link to the effect that is caused. However, this is mostly unknown and surrogate dose metrices are generally used. In standard dose-response curve analyses, this dose metric is the external concentration (e.g., aquatic or bee studies), and is not governed by a toxicokinetic module. In many other cases, the internal concentration is used, for example in the bird & mammal assessment or in critical body burden approaches (McElroy et al., 2011). The GUTS framework established three toxicokinetic dose metrices and one toxicodynamic dose metric (Jager et al., 2011). These metrices have an increasing need for data to predict more and more accurately the concentration at target site of action:

- External concentration
  a. Concentration in the media.
  b. No additional data requirement.
- Scaled internal concentration
  a. A simple dynamic toxicokinetic model (assuming uptake rate equals elimination rate), which can be calibrated solely on effect data fitting a dominant rate constant.
  b. No additional data requirement.

- Internal concentration
    a. Translating external concentrations to internal concentrations as a function of time. This internal concentration can be simulated, for example, by first order kinetics (Nyman et al., 2012), by a two-compartment (Gergs, Gabsi, et al., 2016), or by a comprehensive multi-compartment physiological-based kinetic (PBK) model (Grech et al., 2019).
    b. Requirement for additional measurements of internal concentrations over time.

- Scaled damage (toxicodynamic module)
    a. If measured body residues are available, and the time course of internal concentration does not explain the temporal pattern of effects, then scaled damage should be used as dose metrics. The scaled damage describes the damage accrual, damage recovery, and damage dilution, and is therefore strictly speaking not a part of the toxicokinetic but the toxicodynamic module.
    b. Additional measurements of internal concentrations over time required.

In the toxicodynamic module (TD part) any processes that describe what the chemical does to the body are included and the toxicokinetic dose-metric is linked to the effect using a stress function. Therefore, the scaled damage as dose metrics is part of the toxicodynamic module. In the GUTS framework, two different submodules are used for the TD part, namely *individual tolerance* and *stochastic death* or a combination of these (Jager et al., 2011). The options within each submodule can be chosen and combined, and are based on the problem formulation, its complexity, and available data for model calibration. A number of case studies use these different approaches; some examples include: *Daphnia* and triphenyltin (Gergs, Gabsi, et al., 2016), *Daphnia* and dispersogen A (Gabsi et al., 2014), *Gammarus pulex* and propiconazole (Nyman et al., 2012), *Americamysis bahia* and tembotrione (Gabsi et al., 2019), *Oncorhynchus mykiss* and copper and zinc (Janssen et al., 2021).

The DEB-TKTD model is also built as a modular approach to simulate sublethal effects on animals. Here, a dose metric is used, analogous to the TK part in GUTS, together with the physiological part of the species (TD part), consisting of a dynamic energy budget (DEB) model. The DEB model provides rules (based on first principle) for the assimilation and utilization of energy, which is used for the maintenance, growth, and reproduction of individual organisms. Based on the dose-metric, stress functions are applied to different processes within the DEB model depending on the physiological mode of action (Gergs et al., 2021; Jager et al., 2022), for example, interference with growth, reproduction, and maturation.

5.3. Model complexity, transferability, and testability

5.3.1. Why are MEMs often complex?

Because complex questions often evolve in risk assessment, MEMs must include the most relevant ecological processes in relation to the respective question and the corresponding environmental scenario. Consequently, MEMs are often complex to generate good predictions about real ecological systems (Evans et al., 2013). Complexity here means an increased number of processes that need to be parameterized. The complexity of mechanistic models often increases model generality, so that not just a single dataset can be precisely predicted, but a number of different datasets at different levels of biological organization. This is referred to as pattern-oriented modeling (Grimm & Railsback, 2005). This generality of complex models, not being specifically designed to fit one specific dataset, allows for transferability, which is a fundamental requirement that describes the ability to make accurate predictions with new data (Gray et al., 2009). Model complexity is driven by the research question(s) and relevant ecological scenarios, which in turn have an impact on the number of relevant processes and vice versa. The structure of a MEM determines the scope of environmental scenarios that can ultimately be simulated; for example, if temperature dependency is not included, the extrapolation of model simulations to different temperature scenarios is not immediately achievable.

5.3.2. Testing and transferability

The use of MEMs for extrapolation to different environmental conditions can increase confidence in the predictive power and generality of a model but requires more data and increased degrees of freedom in model behavior, thus increasing model complexity.

Model complexity supports the flexibility of models to represent scenarios under changing environmental conditions and thus allows transferability between different scenarios. Less complex models might represent a specific baseline scenario well but may fail to be accurate for different environmental conditions (Fulton et al., 2003). For example, limited spatial resolution, highly simplified food webs, or oversimplified implementation of mortality in populations can lead to poor model performance when predicting population behavior under changed environmental conditions (Fulton et al., 2003). The transferability of models or even of single modules can be tested by using data observed for environmental conditions different from those that have been used for parametrization (Gray et al., 2009).

5.3.3. Overparametrized models and modularization

Complex models usually require more parameters than can be derived from individual available datasets. With respect to the problem of developing and using overparametrized models, Omlin et al., (2001) stated, "The parsimony principle of system identification states that a model should not be more complicated than necessary for the description of the data. This implies that its parameters are identifiable from the available dataset. While this principle is very important for gaining information on poorly known systems, there are also good reasons for using overparametrized models for the description of environmental systems, especially in the context of prediction of expected changes under changing environmental conditions." While the term "overparameterization" implies that a simpler model might have been preferable, it can be required to use models with high degrees of freedom, to reach specific goals, including (i) combining the current state of knowledge from different scientific disciplines for a complex system or issue under consideration, and (ii) predictions for changing systems or environmental conditions (Omlin et al., 2001).

Complex models have the added advantage that modules can be switched off and tested independently to use the same model for different questions and scenarios. For example, in the PEARL model, processes were implemented that are in principle relevant but switched off in the current FOCUS version (e.g., snow cover, surface water runoff), because they were currently not considered to be of relevance for the regulatory question of groundwater leaching. However, due to these additional processes the models can be used for different regulatory questions, for example, the PEARL model can be used to assess leaching to groundwater (European Commission, 2014), and with a different parameterization and a different environmental scenario, the same model can be used for soil exposure assessment (EFSA, 2017). Another example of deliberately overparameterized models are complex whole body physiological-based kinetic (PBK) models (Baier et al., 2022; Larisch et al., 2017; Mavroudis et al., 2018). Whole body PBK models explicitly simulate all organs of the organism and their connection by blood flow and have a multitude of species-specific physiological and compound-specific physico-chemical parameters. Most questions for which these models are used, such as the elimination of a substance from a human, fish, or bird, can also be answered for a specific application by fitting a simple first-order kinetic to data from a single experiment. However, the elimination rate constant of this first-order kinetic may be driven by different processes such as metabolization, kidney excretion, or blood flow rate. This makes it impossible or uncertain to extrapolate from experimental data based on only one first-order kinetic rate constant to other scenarios. In contrast, these whole-body PBK models can be used for straightforward extrapolation from *in vitro* to *in vivo*, as they explicitly simulate these processes, with the relevant rate emerging from the compound-species combination (Brinkmann et al., 2014; Najjar et al., 2022; Stadnicka-Michalak et al., 2015).

This dilemma, that complex models often require more data than may be necessary to answer a question in a specific case, and that these can hardly be parameterized based on one single dataset, can be solved by following a modularized view of complex models: Each single module can be parameterized, tested, and used independently of the other modules. In this modular view, sub-models or modules are

adding new processes to the model, which ideally can be parametrized independently of the overall model (see section 5.4 for more details). If their parametrization is either generally valid and stays in their domain of applicability or is done for a special case with additional data, the number of remaining parameters that need to be (re)parametrized for the overall model is significantly reduced. One example for this modular feature includes the FOCUS surface water suite, which couples different models to calculate the input via drift, run-off, and drainage in specific modules, and a hydrological module that calculates the fate of the compound in the water and sediment. Another example is the hybrid model DaLaM, which couples an IBM for *Daphnia* and an ecosystem model, each parameterized on different datasets. The combination of these models can be used to predict the population dynamics of daphnids under semi-field conditions without the need for reparameterization (Strauss et al., 2017). Similarly, the standard DEB model with the AddMyPet parametrization is based on broad database and can be used to evaluate toxicity tests where not all parameters can be estimated (Jager et al., 2023).

For example, the GUTS model can be used as a generic module in a population model to describe mortality due to dynamic pesticide exposure over time. GUTS has more parameters than simpler toxicological approaches (e.g., dose-response curves, NOECs) and adds complexity to a population model. However, if the GUTS model used has already been calibrated and validated on additional, independent, species – and substance-specific datasets, it does not add additional uncalibrated parameters to the overall model. In other words, overparameterization of a model is of no concern if the additional processes are encapsulated in modules that are parameterized and tested independently, and the module remains in its domain of applicability. Therefore, modularization allows complex models to be parametrized and validated appropriately. The problem of over-parametrization is reduced by no longer having to rely on individual datasets to parametrize the entire model.

5.3.4. Definition of the domain of applicability for submodules

The precise definition of the domain of applicability of a MEM and its respective submodules is of utmost importance for consistent model use. The clearer the domain of applicability is defined in the model documentation, the easier it is for an evaluator to decide whether a model can be used with confidence for a particular scenario or risk assessment question. Definitions of the domain of applicability can help with answering questions such as: Can a DEB-TKTD model be validated for a midge be used to predict toxicity on mayflies? Can a GUTS model be used to simulate sublethal effects?

The domain of applicability of a model depends on the domain of applicability of its (sub-)modules. A a more specific definition states that the domain of applicability of a module indicates the purposes for which the module can be used. The domain of applicability of a model or module can be defined by answering several questions:

- What type of predictions can the module produce (e.g., survival pattern over time, growth and reproduction over time, population dynamic)?
- What are the limitations of the module in temporal and spatial scales?
- For which environmental conditions or external driving factors was the module calibrated or tested and proven to perform well (e.g., trophic state, climatic conditions, availability of food resources)?
- For which exposure conditions is the module calibrated or tested and applied (e.g., range of concentrations, time-variable exposure)?
- Can the module be used for interspecies extrapolations?
- Can the module be used for extrapolation to untested life-stages or are the relevant life-stages included in the model?
- Can the module be used for extrapolations to other chemicals?

For the combination of a number of modules, a relevant question is whether the domains of applicability of these modules fit together. For example, are these modules designed, parametrized, and tested for the same spatial and temporal scales, temperatures, or trophic conditions?

It should be noted, that a comprehensive list of what determines in general the applicability range of mechanistic models was beyond the scope of the MAD working group, although it was considered very relevant. This section 5.3.4 and the subsections outline some principles that may be useful for a future working group dealing specifically with this topic.

5.3.4.1. Important aspects to define the domain of applicability

Mechanistic models allow extrapolation outside of the data range for which they were calibrated, as long as they contain the relevant mechanisms. In contrast, statistical models can only interpolate within the range of the data used for calibration (Baker et al., 2018). This implies that the domain of applicability for mechanistic or statistical models are defined on a different basis. For statistical models, the domain of applicability is given by the range of the calibration dataset and extrapolation outside this parameter set is only possible with increased uncertainty.

There are more aspects relevant for the definition of the domain of applicability. The range of a dataset used for the calibration of a module, and hence, the module's domain of applicability must match the respective parameter range of the intended application. Even in the case of the toxicological, generic GUTS framework, the calibration dataset should be checked concerning the range of application. For example, if the concentration range of the ecotoxicological data used to calibrate the toxicological module is far outside the concentration range of the scenario to be simulated, or if the experimental duration in the laboratory

is significantly shorter than the intended duration of exposure in the simulations, the applicability of the TKTD parameter set should be verified experimentally if possible (e.g., with mesocosm or field data).

The range of applicability of a model might also depend on the risk assessment question(s). For example, the use of the model for the phantom midge *Chaoborus crystallinus* (Strauss et al., 2016) is only valid in fish-free waters, because no predators are considered in the model. The spatial migration and dispersal module for the adult midges has been tested for aquatic mesocosm scenarios (Dohmen et al., 2016), but the spatial behavior of *Chaoborus* in the field is poorly studied, leading to high parameter uncertainties for their dispersal in the landscape, which is typical for many organisms (although these scenarios can become very relevant for risk assessment). In contrast, the empirical module for winter dormancy induction and termination of larvae in the same model is dependent on environmental conditions such as temperature, photoperiod, and food (Strauss et al., 2016). The implemented dormancy induction, although conceptually plausible, is parametrized and tested only for central Europe. Application of the model in northern Europe (shorter summers) is more plausible than for southern Europe (higher temperatures and longer summers), but currently there is little robust field data for validation in either climate zone. Therefore, this model, in its current state, is only applicable to fish-free waterbodies in central Europe. If simulating the recovery of populations after disturbance, the spatial framework would have to be carefully adjusted to account not only for isolated ponds, but also for exchanges among several populations across the landscape.

In the case that individual modules are to be embedded into a broader model context, the need for modifications could arise. If, for example, a TKTD model such as GUTS is used independently, each species – and compound-specific TKTD parameter set represents the sensitivity of the life stage that was previously used for TKTD parametrization. For the use of TKTD models within a differentiated population structure (e.g., IBMs, matrix models), a stage – or size-dependent dynamic use of TKTD parameters would be conceptually useful to account for the potentially variable sensitivities of individuals in different life stages. If this is not considered, but instead a fixed TKTD parameter set is used for all life stages, this could over – or underestimate the overall effect of a compound on a population. If TKTD model calibration was performed with the most sensitive life stage, overestimation could occur, conversely it may underestimate the overall effect if less sensitive life stages are used. In that sense, it is in practice often challenging to precisely define the range of applicability, and model precision, and therefore, predictive power might decrease when the model is used for conditions at the edges or outside of the range domain of applicability. Regardless, using models in these cases can be useful as long as these uncertainties are clearly documented.

5.3.4.2. How to increase the domain of applicability

A module that is used outside its domain of applicability is not necessarily to be considered invalid. Outside this range, the module can still give meaningful, though more uncertain and probably less reliable, results. For example, without explicitly implemented temperature dependency, toxicological and physiological processes of an animal can be simulated only for the temperature of the experiment, which is usually in the

thermoneutral zone of the species. The use of a simple Arrhenius equation would give acceptable results for temperatures below its thermoneutral zone and could even be used as a first approximation for extrapolation to wider temperature ranges in the absence of experimental data. In these cases, uncertainty of model predictions can be expected to increase, and it depends on the question to be answered by the modeling as to how far such extrapolations might be acceptable. Where reliable temperature extrapolation is desired, additional data will need to be provided. In general, an analysis of the trade-off between increased uncertainty and the extension of the scope and scenarios could lead to a reasonable use of modules outside their experimentally validated domain of applicability. Simple modules with only a small number of equations and parameters can, thereby, greatly expand the possible domain of applicability of a model. For example, simple assumptions for predation such as seasonal, density-dependent predation rates, can allow to interpret the population dynamics of a single species in the context of food webs without detailed modeling of the predator's population dynamics, and, thus, account for toxic and natural stressors acting simultaneously on a simulated population (e.g., Gabsi et al. 2014).

In a regulatory sense, intraspecies extrapolation from one life-stage to another appears possible, if a TKTD model is calibrated and/or validated against data for the most sensitive life stage or for several life stages. In the first case, model predictions are likely to be more conservative for a less sensitive life-stage (EFSA PPR, 2018a). Such model use would not aim primarily at a precise description of natural processes, but would tolerate deviations in the prognosis, if they tend to overestimate the effects of a substance (conservative effect assessment in risk assessment). In the second case, the model explains or describes the differences in sensitivity between the life-stages and can be used for realistic predictions. In this case the sensitivity of the population depends on the population structure which might change over the year or between different scenarios (Gergs et al., 2013).

Interspecies extrapolation is more challenging than intraspecies extrapolation because the driver of the interspecies differences in sensitivity needs to be known. For example, whether the difference in TKTD dynamics is driven by toxicokinetics (e.g., uptake or metabolization) and/or by the toxicodynamics (sensitivity at target site). If the difference is driven by toxicokinetics, a mechanistic approach for uptake, together with physiologically based kinetic models (PBK-models) might address or predict these differences correctly if they are related to adsorption, distribution, or elimination (Brinkmann et al., 2016). Changes in species sensitivity can be caused by changes in metabolization, either the metabolism rate or the composition of metabolites. Here, specific *in vitro* assays measuring the metabolization in cells, liposomes, or enzymes might be helpful (e.g., OECD Test No 319A & 319B). Simpler approaches using DEB or GUTS models for species extrapolation are available in literature (Baas & Kooijman, 2015; Gergs et al., 2015; Gergs et al., 2019), currently, however, these approaches use mostly empirical relationships of parameters. Because such extrapolations are not based on a mechanistic understanding, it is necessary to ascertain whether the empirical relationships hold true for the mechanism of action of the compound-species combination of interest. Hypothesis testing for a mechanistic model will require much less data than for empirical relationships. Therefore, it might be appropriate for a Species-Sensitivity Distribution (SSD) of a TKTD model

(e.g., GUTS) to parametrize GUTS for several macroinvertebrate species (Singer et al., 2023). The model is now able to explain the response of the full set of species. So, the domain of applicability has changed from a single species to crustaceans, insects, or macroinvertebrates in general depending on the dataset used for model calibration. This also means that validation of the model to test whether the GUTS model can predict effects for time variable exposure only needs to be performed on a few selected species and not necessarily on all potentially relevant species.

5.4. Evaluation of modular model approaches

The EFSA scientific opinion on good modeling practice (EFSA PPR, 2014) does not specifically define modules or a modular model, but clearly acknowledges that many models are built on a modular basis. For such models, the EFSA opinion gives some recommendations regarding their documentation and verification. Section 10.2 ("Evaluation of the conceptual model") states that the conceptual model should be described "generally" and "qualitatively." Diagrams or text describing processes within the system to be modeled and their interactions should be provided in a way that sufficiently ensures an understanding of the structure of the modeling approach. If a modular approach is used, not only the modules should be described, but also the relationships between those modules. In addition, it should be ensured that spatial and temporal scales of all modules are consistent with each other. In a version control system, it should be clearly stated which modules and module versions constitute a certain model. Regarding model verification, the opinion states that both the respective implementations of the modules and the interactions between the modules are to be verified. This verification should be carried out by using a "series of hierarchically linked unit tests." These tests should be documented with explicit statements about the methods and the extent of the verification performed. After verifying each module separately, groups of modules should be verified as one unit at a higher hierarchical level. This process should be continued until the entire model has been tested. In that sense, the EFSA opinion in 2014 provided valuable suggestions as to how complex, modular models can be evaluated.

5.4.1. Classification of modular models and relevance for evaluation

The classification of models and/or modules can be done based on different categories; for example, by differentiating between dynamic and static models (e.g., TKTD vs. dose-response relationship). Another common differentiation can be made between mechanistic models that are based on first principles, and empirical (statistical) models that are data driven. Whereas the empirical models typically reproduce

calibration data, they lack a mechanistic principle and their transferability to untested cases is more uncertain. While the above-described classifications are in general helpful, it should be noted that models and modules are often located between these two extremes, for example, being strictly based on first principles versus purely data-driven. For example, the *Daphnia* DEB IBM (Gergs et al., 2014) is fully based on energy balance as a first principle. In contrast, the *Daphnia* model IDamP (Preuss et al. 2009) uses a mechanistic structure to describe an organism, but in parts, simulated processes are based on empirical observations. Both models simulate *Daphnia* populations in an agent-based approach, both were calibrated with the same data, and both produced a comparable output. Yet, the underlying principles are fundamentally different. Because the DEB IBM (Gergs et al., 2014) uses a DEB module for simulations of individual level processes and is based on first principles, it is straight forward to implement new influencing factors such as temperature dependency of growth. In contrast, in the IDamP model (Preuss et al., 2009) processes are empirically derived, so adding temperature dependency would require additional empirical data for each individual process at different temperatures to allow for a proper parametrization. Therefore, for a model application in the original domain of applicability, it may not make a difference whether it is based strictly on first principles (mechanistic) or empirical (statistic). However, if the model is to be used for another purpose, such differences become important (for more details on domains of applicability see section 6.2).

A crucial question in the practical work with modular models is whether models or respective modules need to be re-validated for each new application, or whether transferability of models or modules between applications is possible. In this context, a characterization of models with respect to *specificity* versus *generality* of their model structures, parametrization, and previous validation studies is considered helpful. For a meaningful evaluation of a modular model, it is essential to have a clear understanding of which modules are generally applicable, which are organism – or ecosystem specific, and which are limited to specific domains of applicability. In addition, it is important for the transferability of a model to know how generic respectively specific the submodules are.

Therefore, we suggest defining three different categories: *All-purpose*, *generic*, and *specific* models or modules. An *all-purpose* module would then be a module that is independent of a specific chemical, and valid for the (ranges of) environmental conditions relevant for the risk assessment question (e.g., a hydrological module for a lake). In contrast, a *generic* module might be used with the same model structure but with different parameters for a compound – or species-specific adaptation (e.g., DEB or GUTS models and/or modules). *Specific* modules are empirically based and specific for a process under the conditions for which they were designed (e.g., special behavioral models that apply to one species only).

As a rule of thumb, ecological modules describe the response of species at different environmental conditions without the impact of any toxicant. Ecological models and/or modules might therefore often be all-purpose models that can be used in different ways according to their case study-independent structure and parametrization. For example, an aquatic ecosystem model designed for standing waters

such as lakes and ponds can also be used for mesocosms by changing the ecological scenario, including input parameters. This means that no fundamental re-parametrization or recalibration of the model is required for a specific application, except for case-study-specific parameters, such as nutrient release rates in pond sediments or biological oxygen demand of the water column. Therefore, all-purpose modules are usually not changed for a specific application in risk assessment and can be applied for new questions if they have been validated and the environmental conditions for the risk assessment remain in the domain of applicability.

In contrast to ecological modules, a toxicological module is usually both species – and compound-specific. An important characteristic of a model or module is whether a species – and/or compound-specific adaptation includes only model parameters or the changes to the model structure, as well. Generic models offer a possibility to reuse a previously validated model structure and implementation, by simply using another set of (calibrated) parameters. A specific definition considers "a module to be generic if it can simulate several functionally equivalent systems just by using different values of the parameters" (Meyer [1990], cited in Reynolds & Acock [1997]). Generic modules, like GUTS or DEB are used with the same structure and equations but with different parameters for compound – and species-specific adaptation. Hence, the evaluation of an agreed generic toxicological module can focus on the parametrization as suggested for TKTD models (EFSA PPR, 2018a).

5.4.2. Model structure, processes, and submodules, and relevance for evaluation

The modeling framework for risk assessment introduced in Chapter 2 can be divided into the environmental scenario and the actual models (Figure 2.2). The models can be further separated into fate and effect modules, and the effect modules into toxicological and ecological modules. This structure shows the hierarchical principle that is often seen with complex modeling approaches being used in risk assessment. Nevertheless, despite differentiation at these levels, models or modules themselves may in turn contain a number of submodules that describe clearly defined, separate processes within MEMs.

Table 5.1 and Table 5.2 provide an overview of possible processes and respective submodules for the ecological and toxicological modules, and the associated potential evaluation question, with some examples. A preliminary classification of these processes and submodules is given in the table and considers whether these are generally expected to be species – or compound-specific. As mentioned before, such classification is relevant for a case-specific assessment of these modules with regard to structural adaptation, parametrization, and validation.

Populations can be used as a module in community or ecosystem models, where possible ecological interactions such as predation or competition with other species, and potentially underlying food web structures, are also defined. Equally, for population models used in isolation, questions about ecological interactions and food dependencies are relevant and need to be answered. Therefore, population models

can be extended by dynamic simulation of additional species that affect the modeled species or such influences might be described by a simple parameter. In the event that community models are to be used and evaluated for risk assessment, important aspects include the feeding interactions between trophic levels in food webs, such as species-specific prey selectivity and dynamic prey preferences in predator-prey interactions, can have a strong influence on the overall system behavior.

Population models can be age-structured (for example, matrix models, e.g., Ibrahim [2014]), or unstructured compartment models based on differential equations (Park et al., 2008). More detailed modeling approaches are individual-based (IBMs; also sometimes called agent-based population models [ABMs]; Railsback & Grimm [2019]), which were independently invented in the late 1970s by biologists in the US and Europe (Deangelis et al., 1980; Kaiser, 1979). In these approaches, the focus is shifted from the generic population level to the level of individuals, which is easier and can be defined more concretely. The population dynamics emerge in such IBM models from the interaction of the individuals with each other and with the environment, which often allows extrapolation to untested situations (Agatz et al., 2019; Gergs, Gabsi, et al., 2016; Strauss et al., 2017). These different modeling approaches use different process descriptions to simulate population dynamics. Whereas IBMs simulate individuals in detail and therefore need submodules that describe individual changes over time (e.g., growth in size; development from juvenile stages to adults), matrix models calculate changes in population structure, and compartment models merge these processes together into a few parameters, for example, population growth rates or capacities. Regardless of the model choice, basic processes for simulating individuals or populations need to be implemented. For processes such as growth, established modules are available for compartment, cohort, or individual-based models. Dynamic processes such as individual lifespan and starvation or the use of life-stage-specific mortality parameters are difficult to implement in unstructured population models.

Table 5.1 shows a list of processes that should or could be covered by population models. The implementation of biological processes at individual and population levels needs to be documented and might include interaction of individuals in population models. Submodules that account for basic processes might be used in different model implementations. The implementation of relevant abiotic factors such as temperature needs to be explained, including their integration for specific submodules.

Toxicology modules can vary from generic (e.g., dose-response or GUTS) to specific individual modules, which might be tailored precisely to a problem – and species-specific question. Yet, they all share common elements and processes. Table 5.2 provides a brief overview of relevant aspects of toxicological modules that may have a significant role for lethal and sublethal toxicological effects in the case of TBTK-based and DEB concepts.

Table 5.1: Overview of most important processes and corresponding submodules that should be addressed in the ecological part of a modeling framework for risk assessment , together with potential evaluation questions and examples.

Process or Submodule	Potential evaluation questions	Relevant aspects
Community level		
Population [S]	How do populations interact with other populations? How many trophic levels are present?	Predation based on food selection preferences; interspecific exploitative and interference competition; other interactions (e.g., kairomone-induced change of behavior)
Population level		
Density dependence and/or Individual interaction [S]	What kind of density dependencies are implemented in the model? What density dependencies are enforced or imposed in the model? What density dependencies emerge from interactions in the model?	Intraspecific exploitative competition; competition for space; movement and dispersal; cannibalism; chemical-induced crowding; sexual partner availability
Individual level		
Feeding [S]	Is food implicitly or explicitly modeled? How is the interaction with food implemented? Are different food types with food preferences considered? Is there any feedback loop on the diet? Constant or dynamic assimilation efficiency?	Size selective feeding; food item selectivity (static, dynamic); assimilation efficiency in function of food availability
Development [S]	How is development implemented? Does development need energy? Are all or only parts of life stages implemented?	DEB concept; implementation of pupation and emergence in insects (also connected to growth and reproduction below)
Growth [S]	How is growth implemented? Is growth emerging from available resources, or is it enforced in the model? Are limiting factors included such as food quality (e.g., by nutrient limitation)?	DEB concept; von Bertalanffy growth function
Reproduction [S]	How is reproduction implemented? Is reproduction emerging from the available resources or energy, or is it forced into the model? Is senescence considered? Do the species care about their newborns? Is mating implicitly or explicitly modeled? Continuous batch spawner versus seasonal reproduction? Environmental conditions triggering onset or offset of reproduction season.	DEB concept; litter per female; fixed reproduction intervals; resource-dependent timing of broods; insect emergence as reproduction event occurring only once; viviparous reproduction

Survival [S]	How is the lifespan of organisms modeled? What can reduce the lifespan of an organism? Does the lifespan emerge from the model or is it fixed? Rules for starvation, pathogens or predation?	Starvation; life-stage-dependent mortality; temperature dependent mortality; tolerance to other factors (e.g., lack of oxygen, flow, flooding, drought)
Autecology [S]	What special autecological behavior does the species have to address unfavorable situations? How is the start and end of the reproduction season defined?	Annual migration patterns; hibernation or dormancy and their triggers; seasonal patterns in reproduction
Movement behavior [S]	What movement model is used? What are the underlying assumptions? Is movement triggered by the scenario or by the intrinsic need of the species? Is immigration and/or emigration considered? Is there exchange with other populations of the same species (meta-population approach)? How is the attractiveness of habitats considered?	Directed or random movement; territoriality; fixed proportion of migrating individuals versus dynamically controlled migration; temporal pattern of vertically movement in water or soil; search for sexual partners
Toxicological module [C, S]	What are the primary effects of toxicants? What secondary (indirect) effects of the toxicant can emerge from the model? How far do the toxicological modules translate exposure into reduction of organism fitness or reduction of population growth rate, for example, what are the effects of body weight reduction or delayed maturity?	GUTS; DEB-TKTD; dose-response models

[S] Species specific, [C]compound specific.

Table 5.2: Possible submodules within the toxicological module, together with potential evaluation questions and examples.

Process or Submodule	Potential evaluation questions	Relevant aspects
Toxicokinetics [C, S]	Which dose metrics are used? Which uptake routes are implemented?	Uptake via body surface (bioconcentration) or via feeding (biomagnification); internal concentration or scaled internal concentration
Toxicodynamics [C, S]	Which modes of action are implemented? Do the dose metrics fit to the mode of action? Are several key events simulated or only the adverse effect itself?	Scaled damage used as additional dose metrics; lethal and/or sublethal mode of actions, for example, effects on growth, feeding, reproduction, etc.
Physiology[S]	What are the considered compartments relevant for toxicokinetics? Impact on toxicity by growth dilution, temperature dependency, size dependency, starvation.	One-compartment versus multi-compartment approaches; PBTK or DEB concepts (structure and reserve).

[S] Species specific, [C]compound specific.

5.4.3. Evaluation of interfaces between modules

The division of complex systems into biologically meaningful modules, which are in the best case independent of each other, is advantageous for testing and evaluating modular models (Reynolds & Acock, 1997). In consequence, an effect model can be built from modules of varying complexity, ranging from simple equations to modules that are in turn built from a number of submodules. Moreover, modules within effect models might be exchanged for different model applications, for example, a dose-response equation might be replaced by a TKTD model to account for dynamic exposure. For such an exchange of alternative modules of different complexity, which represent the same process or take the same function within an effect model, it is essential to maintain identical interfaces between modules.

For an evaluation of a model, the interaction and information exchanges at the interfaces between the modules need to be considered. Interfaces need to be designed to be transparent, flexible, and maintainable, and provide clearly documented communication processes. Moreover, interfaces need to be conceptually consistent, which means parameter values of specified types are exchanged and modules operate within compatible domains.

5.4.3.1. Design of modules based on software engineering principles

Software engineering is the term used for the formal design, creation, and testing of code. It takes a specific set of requirements as the starting point and verifies the code. When the software correctly implements the science, stakeholders can be confident that the output is based on science and not the result of inaccurately coded functionality. This trust is vital in the regulatory arena. Maintainability (i.e., the provision of possibilities for maintenance) and reusability (future compatibility) are key aspects to the design of complex software. These concepts can be adapted for the design of models for use in ERA because the stringent definition and implementation of modules is key. This is especially true for complex models.

Ideally, modules provide a high level of cohesion (e.g., all fate processes modeled inside the fate module) to allow for convenient maintenance, update, and/or replacement of the module. Modules are preferably loosely coupled (i.e., having limited and clearly specified points of interaction), so that connections between modules are simple to be established, tested, and maintained, even when modules are updated or exchanged. The implementation of established software engineering principles can improve the flexibility and maintainability of a model. As an example, the façade design pattern, which is commonly used in object-oriented programming, may be considered (Erich Gamma, 1994). The implementation of a façade design can help achieve loose coupling, clarifying input and output, and simplifying the exchange between modules (see Figure 5.3B). A façade masks the underlying structure and code (that can be simple or complex). This technique is especially useful for complex models, which can be difficult to understand due to the many interdependent processes included. It also provides a mechanism to control user input, therefore reducing the risk of errors being introduced to the code via the user interface.

The benefit of using a façade design becomes clear when we look at an example of a complex model with a variety of modules; for example, a model that aims to calculate the impacts of multiple stressors at landscape levels. If the example model was be coupled without such a system (Figure 5.3A), several connections between module 1 and module 2 would exist, implemented in different parts of the code. If module 2 needed to be replaced with an updated version, it would be complicated to restructure existing code, and there would be a risk of overlooking single, possibly hidden, connections. This risk is reduced using a façade design pattern, because clear interfaces are defined and implemented at the same location in the code.

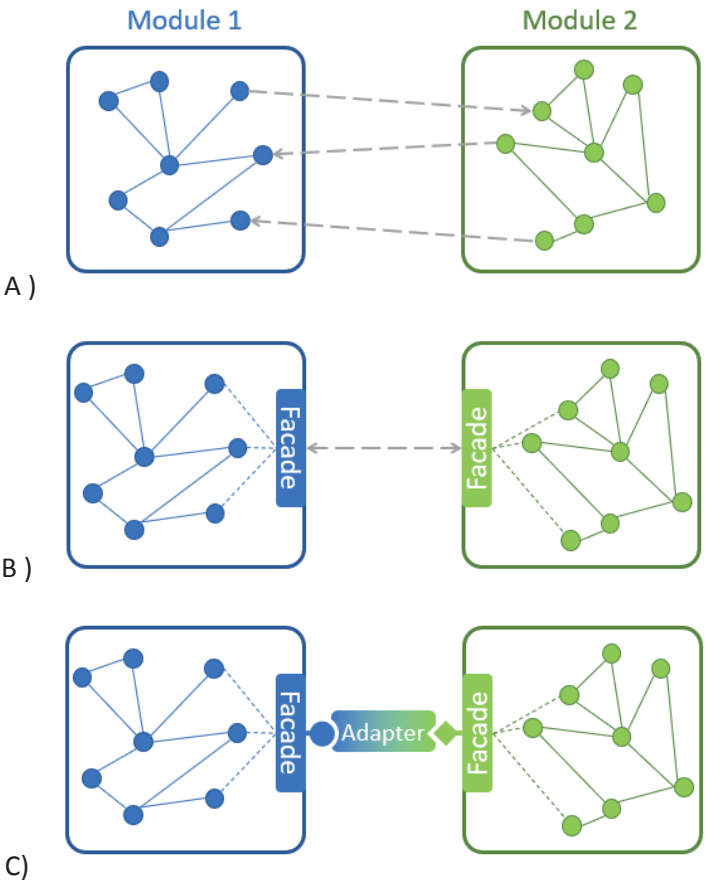


Figure 5.3: Conceptual representation of relationship between two modules and their submodules. A: a non-modular approach without a clear interface between the 2 modules; B: Conceptual representation of the implementation of a facade design pattern, providing a single point of entry for information flow; C: Conceptual representation of the implementation of an adapter design pattern to join modules, for example, using different languages, development environments or units.

It is important to be aware of abstraction, a fundamental concept in modular model design, which, in the context of software engineering, helps to describe the potential versatility of a system. To improve the reusability of modules, it is helpful to clearly separate the core underlying algorithms, which are as complex as required, from submodules that are used to handle input or to process output data and interact with the user.

For example, a console application that takes input and releases output in text format is typically easier to adapt and reuse (high level of abstraction) than a module that is reliant on manual data entry through a graphical user interface (GUI; low level of abstraction). In the first case, the creation of an adapter for integration into a larger automated framework will be more straightforward than compared with a module that is closely coupled to a user interface. From a software engineering perspective, data validation considerations need to be taken into account when designing an adapter, so one can understand whether it is easier for code maintenance or to handle data validation (e.g., whether an input variable is within the defined rage of calibration of the model) in the user interface or in the adapter.

The DaLaM Model (Strauss et al., 2017) is a good example of an appropriately abstracted system (Figure 5.3C). Here two independent models, the IDamP model (Preuss et al., 2009) and the StoLaM model (Strauß, 2009), were joined, using a common data file for data exchange, which was updated at each time step by both models. Within this data file, interactions within the two models are encapsulated. In such an approach, the replacement of the IDamP model by another MEM would be a straightforward task, and the risk of overlooking relevant model interactions is minimized.

The simplest way to join models is by data exchange. Fundamentally, exposure data can be directly used as model input via data files, if there is no feedback between the modules. If feedback is required between modules during runtime, linking distinct software modules into a single framework can be a complex task. In practice, implementation of a modular model approach can be achieved by coupling different models using a software (e.g., Knime or a bespoke script or application), while modules can be implemented in different languages. Alternatively, modules might be (re-)coded together into a single implementation. In any case, it is important to ensure that within the implementation, the modules remain separated, and the interfaces are clearly defined, conceptually and ideally, in the model architecture as well. Such separation can make it easier to maintain the model code for use in different operating systems (e.g., Windows or Linux) where file structure control codes may differ.

Regardless of the way communication between modules is implemented, it must be ensured and verified that parameter values are passed between the modules using correct time steps and units. If extensive conversions between modules are required, adapters can also be defined as independent modules.

5.4.3.2. Examination of the interface between modules

Analyzing the interface between modules provides a more detailed view on the model. Crucial questions in this context are: Are there feedback loops in the model? Do modules use the same temporal or spatial scales? Which parameters are exchanged, and is parameter transfer and use consistent? In this section, these questions will be considered in more detail.

Documentation of feedback loops

An important question for model evaluation is what links between modules have been implemented. Without feedback loops, a technical verification of the interface may be sufficient if both modules are already validated and stay within their domain of applicability. Coupling modules without feedback (also called loose coupling) may require justification (e.g., lack of scientific evidence for feedback; missing information for process formulation).

Feedback in usually loosely coupled modules, for example, between toxicological TKTD modules and ecological modules, may result from the inclusion of individual growth data in individual-based models and might then require a tight coupling, including feedback. Variable sizes of individuals affect the toxicological sensitivities of organisms (Gergs, Gabsi, et al., 2016), and in DEB-TKTD models, internal toxic concentrations can be reduced by dilution effects (Rosland et al., 2013). Likewise, the transfer of toxic substances during reproduction, via egg production for example (Awkerman et al., 2020), can lead to a reduction of the toxic body burden.

Organisms' behavior can also have an impact on their exposure and result in different toxicological output. Such feedback loops between an ecological module, exposure, and a toxicological module need to be considered as an essential property of interfaces. For example, in the POLARIS model, this is the case when the biomagnification of a toxic substance is determined by a mammal's food intake, and thus by the ecological module (see example in Appendix 8.3.3). Another example of a behavioral feedback loop is the active vertical migration of soil earthworms, which can lead to individual exposure to toxic substances that exhibit vertical gradients (Roeben et al., 2020). In such earthworm models, the individual-level submodule for behavior as part of the ecological module simulates vertical movement and feeding of earthworm species.

An important type of feedback loop is caused by density-dependent processes affecting individual physiological processes (e.g., in IBMs) and resource availabilities. Density dependence can manifest itself in direct interaction between individuals (interference competition) or indirect interaction for resources (exploitation competition). For the feedback loop, it is not relevant what caused the density change (e.g., toxicologically induced mortality, density regulation at the population level, or catastrophic events in the environment). What is important is an appropriate implementation of the feedback loop between organisms and their environment in the ecological module of the model in response to the density change.

The relevance of including feedback processes for specific questions needs to be decided on a case-by-case basis. If feedback loops are implemented, a joint validation of the involved modules might be required (see section 5.5).

Do modules use the same temporal or spatial scales?

Population models often use different temporal scales for simulating population-relevant processes, that can range from minutes, to hours, to days, or even longer periods. In contrast, fate modules increasingly provide concentration time series of chemicals on an hourly basis (e.g., FOCUS surface water). Fate and

TKTD modules commonly use differential equations and therefore calculate effects continuously or in short time intervals to avoid numerical artefacts. The output of such modules is normally on an hourly or daily basis, depending on the implementation and model settings. To maintain temporal consistency at different time scales between those modules, the temporal resolution of exposure data must be adjusted to the toxicological module, and in addition, temporal consistency needs to be carefully considered between the toxicological module and the connected ecological module. An example can be taken from the StoLaM lake model (Strauss et al., 2017). Here, different time steps in an ecosystem model and coupled IBMs may cause inconsistency. If algal growth occurs in smaller time steps than zooplankton grazing, optimally the processes related to algal feeding and excretion should be matched to the time step of algal dynamics. However, it is more often the case that modules compute processes at different temporal resolutions, often as a result of the module-specific trade-off between accuracy requirements and computation time. In such cases, pragmatic approaches in which the process with the smaller time step sends a time-integrated response to the coupled module become sufficient. Because these highly dynamic processes are considered, adaptation of the temporal resolution might result in changes in simulation results. Therefore, the chosen method for time alignment between modules should be explained and justified in detail. Modules can also use different spatial scales or resolution. For example, in lake models, nutrients or phytoplankton can be represented in a vertically differentiated way, whereas higher trophic levels such as zooplankton or fish are often programmed without spatial differentiation (e.g., StoLaM lake model, [Strauss et al., 2017]). For processes that link these levels, such as algal feeding or nutrient excretion of zooplankton, the levels must be conceptually aligned, for example, by also vertically differentiating the activity of the higher trophic levels. Also, when coupling spatially explicit exposure models and spatially explicit effect models, spatial scales and resolution consistency is crucial.

Semantic aspects for the consistent meaning of variables with the same name can be significant. For example, temperature can be measured in °C or °F, but it can also be measured in different intervals such as minutes, hours, or daily. Even more complex is that a daily temperature in °C can have different origins. It can be the daily mean, max, or minimum temperature for the day, or for the day and night. However, it could also be the temperature measured at 12:00. Clearly defining these parameters in the form of a software requirements design document can help to ensure that input variables are consistently handled throughout the code without errors occurring due to different units, especially when different teams are working on different parts of the code. NASA's Mars Climate Orbiter is a good example of where this went wrong.

Consistency between modules

The consistent use of parameters across modules can be analyzed from a structural point of view, or with regards to the consistency of parameter values or environmental variables.

From a structural point of view, it needs to be checked whether the relevant dependencies are consistent in all modules. If, for example, weather or climate data with varying temperatures are used in a complex MEM, this information can be passed from input data or a separate temperature module to other modules

(e.g., fate module, toxicological module, physiological or ecological module, etc.; e.g., POLARIS model; Kleinmann & Wang [2017]; see Appendix 8.3.3). A relevant question is whether all temperature-dependent processes are also implemented as temperature-dependent in each respective module. This is often overlooked, even in modules that are known to be clearly temperature dependent. For example, it is often missing in TKTD modules. Consideration of temperature dependencies in the current aquatic Tier 2c model for lethal effects of aquatic organisms (EFSA PPR, 2018a) can be used as an example for a structural consistency check. In its current form as proposed (EFSA PPR, 2018a), the model uses the FOCUS SW model as the exposure module and the GUTS model as the effect module. In the FOCUS SW model temperature has an influence on several important processes like degradation and triggers the exposure profile. On the other hand, in the standard GUTS model temperature influence on toxicokinetics and toxicodynamics is not implemented, even if it is known that temperature has an influence (Huang et al., 2023). Therefore, from a structural point of view there are clear inconsistencies in the approach with regard to handling temperature. For the regulatory question, it is then important to know if these inconsistencies are of relevance overall, or if they might prove conservative. For example, if tests are conducted at 20°C and surface water temperatures in Europe are on average lower this could be considered a conservative approach.

Another aspect of model consistency is the possible complexity differences of the coupled modules: Does the overall model provide the appropriate complexity for taking advantage of detailed modules, or is complexity lost during the transition between modules? For example, if a complex module is coupled with a simple module, the resulting overall model may not properly capture the complexity of the individual modules. As an example, an individual-based population model with a mixed age and/or size population structure might be used, but toxic effects are simulated with a GUTS model calibrated only for one specific age and/or size. In this case, the information about different sensitivities of sizes and/or ages cannot be adequately addressed, which might lead to a mismatch on population level (Gergs, Gabsi, et al., 2016). Thus, the performance of an overall model does not necessarily reflect the maximum capabilities of the individual modules. Attention must be paid to ensure that relevant features of individual modules are passed appropriately via the interfaces.

The consistency of parameter values or environmental variables that are shared between modules is another important question. Do the different modules use the same parameters or state variables for the same process? Even if a consistent temperature dependence of exposure, toxicological, and ecological modules is structurally considered in models, it is not always used in a consistent manner. For example, exposure from FOCUS scenarios is often used modeling as a fixed exposure pattern for aquatic effect modeling, but then combined with weather data from another ecological scenario that differs from the weather data originally underlying the FOCUS exposure scenarios.

Another example is the ecotoxicological relevant concentration as introduced in the aquatic risk assessment scheme (EFSA PPR, 2013): "Lack of a clear conceptual basis for the interface between the exposure and effect assessment may lead to a low overall scientific quality of the RA. This interface is defined by

EFSA (2005) and Boesten et al. (2007) as the concentration that gives an appropriate correlation to ecotoxicological effects, and is called the ecotoxicologically relevant type of concentration (ERC). In the RA, the ERC needs to be consistently applied so that field exposure estimates (PECs) and RACs can be compared as readily as possible."

Here, EFSA defines the risk assessment by an exposure module and an effect module, and it is important that the concentration in both modules have a consistent meaning.

5.5. Validation of modular models

The term *model output corroboration* was defined by Augusiak et al. (2014) as the "comparison of model predictions with independent data and patterns that were not used, and preferably not even known, while the model was developed, parametrized, and verified." It is good scientific practice to demand performance tests before a model is considered to deliver reliable results for decision making, and these performance tests are often called "validation." If a model can predict phenomena that were not even considered during model development, we have the strongest trust in its structural realism, because these phenomena emerge from the mechanistically correct implementation of the underlying processes. However, achieving this gold standard is rarely possible with ecological models, because empirical experiments with ecological systems as a whole are not often feasible. In contrast, models are often developed to address questions that arise precisely because empirical experiments are regularly not possible.

Although validation of complete models for compound-specific scenarios following the described gold-standard would be desirable, this is only achievable in exceptional cases. However, independent predictions of modules can be tested instead. Therefore, we suggest validating a complex MEM in a practical approach based on its modules, considering the ecology module and the ecotoxicological effect model separately.

It should be noted, that the formulation of clear guidelines for validating modular approaches was beyond the scope of the MAD working group, although very useful. The principles outlined in this section may be useful for a future working group dealing specifically with this topic.

5.5.1. Testing of modules

Using a modular model approach allows for model developers, users, and evaluators to tailor validation efforts to the most relevant areas in which uncertainty of the model performance may lie. It appears essential that models or modules can be tested experimentally, but testing experimentally might not always be possible for certain modules, nor for the overall model. This is especially true for vertebrates, including birds and mammals. Therefore, an important criterion for meaningful choice of modularity in a complex MEM is that each module can be parameterized and tested as independent from other modules as possible (Reynolds & Acock, 1997). In this context, the consideration of experimental possibilities for measuring input and output of a module can be helpful for modularization and increases its testability. This implies a different view on modularity: Which modules can be validated at all and independently of each other?

In principle, it makes sense to distinguish whether a model has (a) been tested for an application with regard to structure and parameters, (b) could, in principle, be used due to the structure (parametrization and/or testing possible but not yet carried out), or (c) cannot be tested even with a suitable structure due to methodological problems or missing parameters. The distinction between case study dependent and independent parameters in terms of their calibration as well as validation is an essential prerequisite for deciding whether parameters are transferable to other scenarios and are part of the model parametrization.

This is especially useful if the model in its full complexity cannot be validated. For example, it is impossible to validate a model that aims at simulating sublethal population-level effects on migrating birds. However, it is possible to break down validation into separated, simpler modules, focusing on the validation status of the different modules and the interfaces between them. In this example, module 1 could be identified as a DEB model to simulate growth, development, and reproduction of a bird, and module 2 could be a TKTD model approach to simulate the sublethal effects in this physiological model. In addition to these two modules, a behavior module; a feeding and a habitat distribution module could be identified. All these modules can be validated independently from each other, where the DEB module can be tested against appropriate observation data, ecological modules can be tested against ecological data from bird monitoring, and the emergent predictions of the model under a toxicant can be tested against real world data. For example, from an accident in which a large volume of a pesticide, or another chemical or similar acting stressors, were released. In that way, a modular approach helps by breaking down the complexity of an ERA into simpler tasks through the evaluation of modules, and the transparent evaluation of the interaction of these modules reduces structural uncertainty. If all these different modules are validated independently of each other it is important to evaluate the interaction of the different modules. Most importantly, are there feedback loops that are expected and implemented? In the case of existing feedback loops (e.g., DEB status feedback on bird behavior), it needs to be determined whether the emerging properties of these interactions are already covered by the validation or need to be assessed independently.

Validation of density-dependent modules as an important type of feedback loop can be done by either testing the resulting ecological model with the implemented density dependencies at the population level

(for example, carried out for *Chaoborus* [Strauss et al., 2016] and *Daphnia* [Strauss et al., 2017]), or by testing the response of the individual-related modules to the parameters influenced by variable population densities in the required range of values. For example, a reduction in population size may result in increased food availability to individuals. In this case, it would be necessary to check whether the physiological module (e.g., an IBM model) has already been tested for food levels in the relevant range of values (e.g., including food limitation to food excess). A model parameterization of an IBM for only one food level would not be sufficient in this case.

5.5.1.1. Disturbance: Substance-independent stress to test ecological modules

Models developed for use in risk assessment should be able to adequately represent effects of disturbances on biological or ecological systems, where both the derivation of effect thresholds and systems recovery depend on an accurate model formulation. Here, it is an important feature that the recovery of disturbed populations can be assumed to be independent from the disturbance itself (Gergs, Classen, et al., 2016), which is a definite way to understand the predicted response of an ecological model in a substance-independent way. When testing a MEM for its suitability for chemical risk assessment, a disturbance does not necessarily need to be substance-specific, but could also be effect-specific (e.g., effects of 70% mortality or reproductive inhibition). In that sense, ecological modules can be tested against real-world observations of recovery after disturbance without the need to collect substance-specific data (Gabsi et al., 2014).

This allows a generic validation of the system recovery independent of the specific PPP, which is very helpful because population experiments which include recovery are quite complex or impossible to conduct for single PPPs.

5.5.1.2. Compound-specific validation

In contrast to testing a model's performance in predicting recovery, which can be done independently of a specific pesticide, toxicological modules would need to be assessed on compound-specific data. Therefore, according to current guidelines (e.g., EFSA PPR [2018a]), a compound-specific calibration and validation is necessary. It is necessary that validation of the toxicological module would typically cover the extrapolation of effects from the calibration data specifically and could be achieved based on laboratory experiments. Typical types of extrapolations for toxicological modules can be different exposure patterns (Nyman et al., 2012), different temperatures (Gergs et al., 2019), different life-stages (Gergs, Gabsi, et al., 2016), or even different species (Gergs et al., 2015; Gergs et al., 2019). Such a compound-specific, validated module can then be used in a modular model approach in combination with an ecological module, which in turn was validated against recovery data, without the need to validate the full model against compound-specific data. It might be the case that mesocosm data was available to test a full model approach concerning maximum effect and time to recovery, as shown, for an IBM developed for the phantom midge *Chaoborus* (Dohmen et al., 2016).

5.5.2. Schematic decision tree for model validation to take a modular approach into account

In earlier examples, models were validated as a whole, following actual regulatory guidance (EFSA PPR, 2014). Examples are documented for the BEEHAVE model for honey bees (Agatz et al., 2019; Schmolke et al., 2020), the POLARIS model, or *Daphnia* models (Gergs et al., 2014; Preuss et al., 2009). However, in these examples, validation was performed by considering the ecological module separately, and in an independent, consecutive step the toxicological module (Agatz et al., 2019; Gabsi et al., 2014; Gergs, Gabsi, et al., 2016).

Here, we propose a decision tree describing how validation of a modular model approach can be tailored to specific risk assessment questions (Figure 5.4). The first question (**Q1**) is whether the model approach is modular, which means whether the model consists of, or can be subdivided into, separate modules that are standardized or previously validated. With a standardized module we define a model that is accepted in regulatory circles in its generic state, such as GUTS, the LEMNA model, or the FOCUS modeling suite. If this is not the case, the entire model needs to be validated, according to EFSA PPR (2014). If Q1 indicates a modular approach, question two (**Q2)** follows: Does the model consist of sufficiently validated, possibly standardized submodules (e.g., FOCUS PRZM for run-off), or were some modules developed specifically for the application in question and have not yet been validated? If case-specific modules are used for the given question, validation of these modules would need to be performed and documented either for each individual module and their interfaces, or for the overall model. If validated standardized modules are used, the next question verifies that each module is used within the tested domain of applicability (**Q3**, for details section 7). If the domain of applicability of the module was exceeded, this would need to be sufficiently justified. For non-standard modules, validation should be set up in a way that the domain of applicability for the risk assessment question is reached. If no feedback loops between the modules are implemented (**Q4**), the modules can be used serially, and if modules are validated and stay in their domain of applicability, the verification of the interface implementation is sufficient. For example, using GUTS as an effect module and FOCUS models as a fate module to calculate surface water exposure under agreed regulatory scenarios would not require a validation of the overall model approach. However, if feedback loops exist between modules, the joint validation of both modules will be required, because their combination might change their previously tested individual behavior. Therefore, a clear justification of the implemented feedback loops by the model developer and a rigorous evaluation of potential feedback loops by the model evaluator is essential. For validating coupled modules with feedback loops, cross-module tests using new, appropriate datasets can be carried out. One example for this approach is the coupling of an IBM for *Daphnia* (IDamP, Preuss et al. [2009]), which was validated for laboratory conditions, with the lake model StoLaM (Strauß, 2009), which was tested on deep and shallow lakes of variable trophic states, to form the hybrid model DaLaM (see Appendix 8.3.2). This module coupling could be jointly tested with a new long-term dataset for daphnids in aquatic outdoor mesocosms, integrating the feedback between the modules (Strauss et al., 2017).

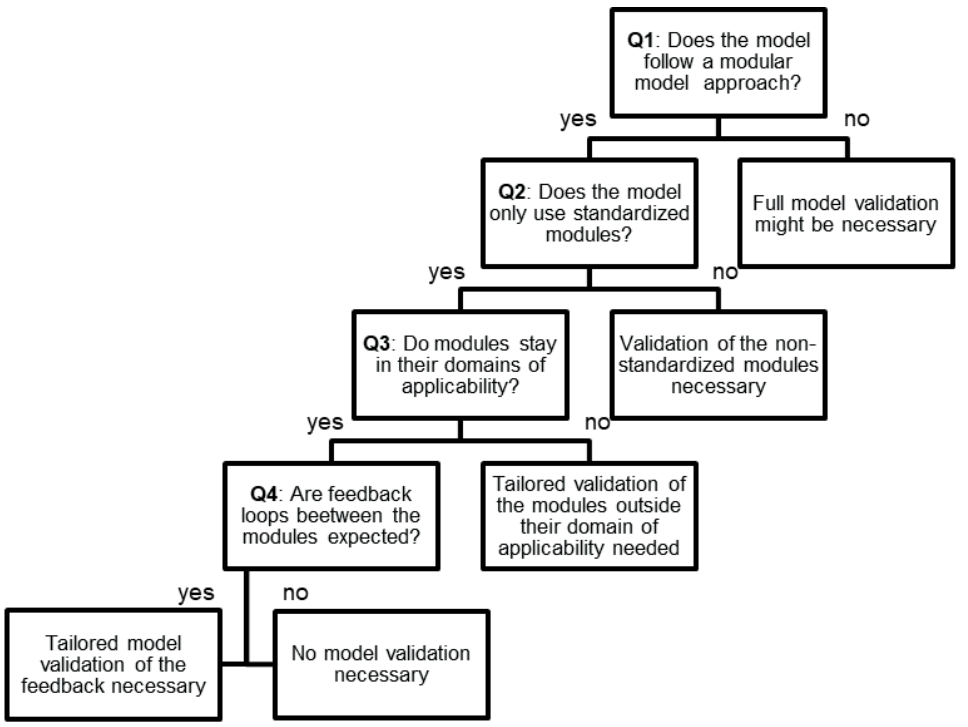


Figure 5.4: Decision tree to tailor validation actions for a modular model approach for a given risk assessment question.

All questions need to be addressed for an evaluation of the model. If the question is answered with "yes," the left box always applies; if it is answered with "no," the right box always applies.

5.5.2.1. Q1: Full model validation necessary

If the model does not follow a modular approach, a full model evaluation is necessary according to EFSA PPR (2014).

5.5.2.2. Q2: Validation of non-standardized modules

If the model follows a modular approach, but some modules have not been validated before, these modules need to be validated according to EFSA PPR (2014). After this validation step, the domain of applicability of other modules and interfaces need to be assessed as well. Here it should be noted that the ecological module can be tested and validated separately from the specific mode-of-action and the toxicology module.

5.5.2.3. Q3: Tailored validation of modules outside their domain of applicability

If the model follows a modular model approach and all modules are basically validated, but some modules are to be used outside of their domain of applicability, it should be evaluated whether the modules can still be used for the question at hand, and whether the overall model would still provide reliable simulation results. When a statistical module is used, a new parametrization might be required that covers the domain

of interest. For a mechanistic module, a clear hypothesis should be generated and tested in a tailored experimental approach. It is preferable to test simple hypotheses in tailored laboratory experiments, for example, the predictive ability of a module for individual growth and development under different temperature regimes, as this can be better tested under controlled conditions in the laboratory. In contrast, predictions on scenarios above a certain complexity or spatial extent can often only be tested in (semi-)field experiments, for example, the seasonal population dynamics of a widely dispersing population under variable environmental conditions.

The frequency of effect model validation for the many possible specific combinations of compounds and physiological endpoints is an important issue due to the high practical effort required for laboratory testing. If an effect model has been shown to be able to make predictions for a particular endpoint of a species (e.g., survival, number of offspring), it can be assumed that the model will be able to give reliable results that are also possible for other compounds acting on that endpoint. If effects on other endpoints are observed, the combination of toxicological and ecological models should be revalidated for at least one substance.

Generic models are a special case whose structure has proven to be fit for purpose (e.g., the GUTS module was approved to be ready for use in aquatic risk assessment [EFSA PPR, 2018a]), and therefore do not require further structural validation. This is also true, for example, for a GUTS module in a population model, which would not need to undergo repeated evaluations of its structural appropriateness for handling time-varying exposure. If, for example, a species-specific sub-model that claims generality and has neither compound – nor case-study-specific aspects has been validated previously, there may be no need to validate it again for a different application. A further example might be DEB-based IBMs that had been validated for many species (Gergs et al., 2014; Rakel et al., 2020). In those models, the physiology of individuals is represented by the DEB models to describe assimilation of resources and allocation to growth, reproduction, and maintenance, as well as survival, which are key elements at the individual level driving population dynamics. The DEB models are generic in a way that can be applied to a wide range of species, as long as they have similar life histories and only differ in parameter values while maintaining the same basic physiological structure. In these cases, IBM submodels such as DEB modules might not be re-evaluated each time the model is applied to a new species – given that parameter values are reliably estimated and/or parameter uncertainty is accounted for in the forward predictions, and sensitivity analyses have been performed.

5.5.2.4. Q4: Tailored validation of the interfaces

If the modules interact with each other and feedback loops exist, new properties can emerge from the coupled model. These new emerging properties need additional testing efforts, especially the propagation of effects between modules from physiology to population to ecosystem level. If modules are not connected in feedback loops within the model, it remains to be evaluated if technical implementation and data handling is consistent, for example, related to meaning and units of exchange variables. This could be achieved by model verification and would not need validation of the overall model against independent data.

Validation is not necessarily required if the modules (including their structure and case study-independent parametrization) have been sufficiently tested, are used within their domain of applicability, and no feedback loops are implemented.

Further details about the evaluation of interfaces between modules are given in section 5.4.3.

Application example for the schematic decision tree for model validation

It should be noted that the example provided does not reflect applicant or risk assessor official points of view regarding the validity of the IBM *Chaoborus* model referenced in these examples. They only provide examples of how to run through the decision tree.

The proposed decision tree for modulated model validation of population models in ERA is illustrated using the IBM *Chaoborus* model based on the case study on the pesticide "modelmethrin" (Dohmen et al., 2016). In this example, the GUTS model was parametrized for *Chaoborus crystallinus* after laboratory testing for modelmethrin, tested using field mesocosm studies, and applied to various Europe-wide scenarios.

The IBM Chaoborus model follows a modular approach (**Q1**); therefore, the entire model does not need to be validated. Standardized modules such as the toxicological model GUTS are used, alongside a non-standard ecological module (**Q2**).

However, submodules of this IBM for *Chaoborus* have already been tested in aquatic mesocosms for the Central European climate zone (Strauss et al., 2016). In addition, an eight-week semi-field mesocosm experiment conducted in the Netherlands with a 6-day in situ bioassay serves as a validation dataset for the IBM *Chaoborus* model. This allowed a joint validation of the toxicological (GUTS) and ecological modules for modelmethrin under field conditions (Dohmen et al., 2016), which would have not been needed applying strictly the modular approach as outlined in this chapter.

The applications for the IBM Chaoborus model shown in this case study (fishless ponds, but for different climates) were only partially within the previously tested application range (**Q3**). The application for the Central Zone has been validated, but has not for the Northern or Southern Zones, a tailored validation of the modules for these zones is still pending.

No feedback loops (**Q4**) exist between the main modules, neither between the fate module and the toxicological module (GUTS), nor between GUTS and the ecological module, as long as the size dependence of the TKTD module is turned off. Density-dependent feedback loops occur only within the ecological module, which have already been tested experimentally at the population level (Strauss et al., 2016). Therefore, due to the serial connection, no further model validations are required for the interaction of the main modules.

5.6. Conclusion

Environmental risk assessment is a complex task, and key ecological processes need to be included in a modeling approach, which is why the corresponding models are, in many cases, also complex (Evans et al., 2013). The subdivision of models into meaningful modules enables a more sophisticated testing, verification, and validation of complex models, particularly by distinguishing which model parts have been sufficiently proven and which may need to be further developed.

Using a modular model approach would enable the validation efforts to be tailored to the necessary areas in which uncertainty of the model performance exists. It is essential that models or modules can be tested experimentally, and the final test of a model is its ability to make predictions about observations unused in model design and parametrization (Evans et al., 2013). Therefore, an important criterion for meaningful modularity is that each module can be parametrized and tested as independently as possible from other modules, whereby a measurability of the input and output variables of a module increases its testability (Reynolds & Acock, 1997).

The precise definition of the domain of applicability of a MEM and respective submodules is important for consistent model use. The clearer the domain of applicability is defined in the model documentation, the easier it is for an evaluator to decide whether a model can be used with confidence for a particular scenario or risk assessment question. For the combination of a number of modules, a relevant question is whether the domains of applicability of these modules fit together. For example, are these modules designed, parametrized, and tested for the same spatial and temporal scales, for example, temperatures or trophic conditions? Mechanistic models allow extrapolation outside the data range they were calibrated to, as long as they contain the relevant mechanisms, in contrast to statistical models, which can only interpolate within the range of the data used for calibration (Baker et al., 2018). This implies that the domain of applicability for mechanistic or statistical models are defined on a different basis.

We propose a decision scheme for how validation of a modular model approach could be tailored to specific risk assessment questions (Figure 5.4). If the model consists of, or can be subdivided into, separate modules that are standardized or already validated, if no feedback loops between the modules are considered, and if the modules remain in their domain of applicability, the verification of the interface implementation appears sufficient as a model validation. Currently this case will rarely occur, although this decision scheme will promote tailored and efficient development, validation, and evaluation of modeling approaches using a modular model structure.

5.7. Outlook

So far, the evaluation of complex models for ERA purposes has commonly been undifferentiated with respect to assessment of strengths and weaknesses of individual submodels, modules, or specific model parts. Ultimately, a systematic approach to model modularization will reduce the workload in model development and evaluation. This is mainly due to the fact that generic modules (i.e., modules that can be used without structural adaptation in different overarching models) that have already been extensively validated do not need to be re-evaluated for each model application.

Within Table 5.1 and Table 5.2 an overview of possible processes and respective submodules for the ecological and the toxicological modules are provided. This can form the basis of the development and testing of generic modules for these processes, which clearly need to be addressed within an ERA. A possible first choice would be the GUTS model as a generic module to simulate survival or the DEB-TKTD model for sublethal effects. These models are generic in their structure and can be parameterized depending on the species and toxic compound, which has for example been demonstrated for the GUTS model using fish (Ashauer et al., 2013), several aquatic invertebrates (EFSA PPR, 2018a), and bees (Baas et al., 2022) as examples. For such generic modules, the evaluation will focus on model parameterization as is the case for exposure models.

Other generic modules to consider are individual-based population models (IBMs) to extrapolate from the individual to the population in a standardized approach, or DEB models to simulate the growth, development, and reproduction of an individual. Both approaches have already been used for a broad range of different species, but are often developed *ad hoc*, following no specific structure. For the DEB model the structural diversity is surprisingly low to explain the difference in life history strategies in the animal kingdom (AddmyPet database[6]). In contrast, the diversity of approaches for IBMs is currently high because no standard modules exist, and these models are often built empirically from scratch. However, first guidance for a more standardized development of population models including IBMs for ecological risk assessment is available (Forbes et al.; Raimondo et al., 2021).

There is currently a lack of generic modules for behavior, movement, and density-dependent interactions. Having generic and tested modules for these processes would enable developers to derive new effect models to answer regulatory questions based on standardized modules and would reduce validation needs to the verification of interfaces and the domain of applicability of the modules.

6 https://www.bio.vu.nl/thb/deb/deblab/add_my_pet/

5.8. Bibliography Chapter 5

Agatz, A., Kuhl, R., Miles, M., Schad, T., & Preuss, T. G. (2019). An evaluation of the BEEHAVE Model using honey bee field study data: Insights and recommendations. *Environmental Toxicology and Chemistry*, *38*(11), 2535-2545. https://doi.org/10.1002/etc.4547

Ashauer, R., Agatz, A., Albert, C., Ducrot, V., Galic, N., Hendriks, J., Jager, T., Kretschmann, A., O'Connor, I., Rubach, M. N., Nyman, A.-M., Schmitt, W., Stadnicka, J., van den Brink, P. J., & Preuss, T. G. (2011). Toxicokinetic-toxicodynamic modeling of quantal and graded sublethal endpoints: A brief discussion of concepts. *Environmental Toxicology and Chemistry*, *30*(11), 2519-2524. https://doi.org/10.1002/etc.639

Ashauer, R., Thorbek, P., Warinton, J. S., Wheeler, J. R., & Maund, S. (2013). A method to predict and understand fish survival under dynamic chemical stress using standard ecotoxicity data. *Environmental Toxicology and Chemistry*, *32*(4), 954-965. https://doi.org/10.1002/etc.2144

Augusiak, J., Van den Brink, P. J., & Grimm, V. (2014). Merging validation and evaluation of ecological models to 'evaludation': A review of terminology and a practical approach. *Ecological Modelling*, *280*, 117-128. https://doi.org/10.1016/j.ecolmodel.2013.11.009

Awkerman, J., Raimondo, S., Schmolke, A., Galic, N., Rueda-Cediel, P., Kapo, K., Accolla, C., Vaugeois, M., & Forbes, V. (2020). Guidance for developing amphibian population models for ecological risk assessment. *Integrated Environmental Assessment and Management*, *16*(2), 223-233. https://doi.org/10.1002/ieam.4215

Baas, J., Goussen, B., Miles, M., Preuss, T. G., & Roessink, I. (2022). BeeGUTS-A toxicokinetic-toxicodynamic model for the interpretation and integration of acute and chronic honey bee tests. *Environ Toxicol Chem*, *41*(9), 2193-2201. https://doi.org/10.1002/etc.5423

Baas, J., & Kooijman, S. (2015). Sensitivity of animals to chemical compounds links to metabolic rate. *Ecotoxicology*, *24*(3), 657-663. https://doi.org/10.1007/s10646-014-1413-5

Baier, V., Paini, A., Schaller, S., Scanes, C. G., Bone, A. J., Ebeling, M., Preuss, T. G., Witt, J., & Heckmann, D. (2022). A generic avian physiologically-based kinetic (PBK) model and its application in three bird species. *Environment International*, *169*, 107547. https://doi.org/10.1016/j.envint.2022.107547

Baker, R. E., Pena, J.-M., Jayamohan, J., & Jérusalem, A. (2018). Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biology letters*, *14*(5), 20170660.

Bartell, S. M., Schmolke, A., Green, N., Roy, C., Galic, N., Perkins, D., & Brain, R. (2019). A hybrid individual-based and food web-ecosystem modeling approach for assessing ecological risks to the Topeka shiner (*Notropis topeka*): A case study with Atrazine. *Environ Toxicol Chem*, *38*(10), 2243-2258. https://doi.org/10.1002/etc.4522

Boesten, J. J. T. I., Köpp, H., Adriaanse, P. I., Brock, T. C. M., & Forbes, V. E. (2007). Conceptual model for improving the link between exposure and effects in the aquatic risk assessment of pesticides. *Ecotoxicology and Environmental Safety*, *66*(3), 291-308. https://doi.org/10.1016/j.ecoenv.2006.10.002

Brinkmann, M., Eichbaum, K., Buchinger, S., Reifferscheid, G., Bui, T., Schaffer, A., Hollert, H., & Preuss, T. G. (2014). Understanding receptor-mediated effects in rainbow trout: In vitro-in vivo extrapolation using physiologically based toxicokinetic models. *Environmental Science & Technology*, *48*(6), 3303-3309. https://doi.org/10.1021/es4053208

Brinkmann, M., Schlechtriem, C., Reininghaus, M., Eichbaum, K., Buchinger, S., Reifferscheid, G., Hollert, H., & Preuss, T. G. (2016). Cross-species extrapolation of uptake and disposition of neutral organic chemicals in fish using a multispecies physiologically-based toxicokinetic model framework. *Environmental Science & Technology*, *50*(4), 1914-1923. https://doi.org/10.1021/acs.est.5b06158

Deangelis, D. L., Cox, D. K., & Coutant, C. C. (1980). Cannibalism and size dispersal in young-of-the-year largemouth bass: Experiment and model. *Ecological Modelling, 8*, 133-148. https://doi.org/10.1016/0304-3800(80)90033-2

DeAngelis, D. L., & Diaz, S. G. (2019). Decision-making in agent-based modeling: A current review and future prospectus. *Frontiers in Ecology and Evolution*, *6*, 15, Article 237. https://doi.org/10.3389/fevo.2018.00237

DeAngelis, D. L., & Grimm, V. (2014). Individual-based models in ecology after four decades. *F1000prime reports*, *6*.

Dohmen, G. P., Preuss, T. G., Hamer, M., Galic, N., Strauss, T., van den Brink, P. J., De Laender, F., & Bopp, S. (2016). Population-level effects and recovery of aquatic invertebrates after multiple applications of an insecticide. *Integrated Environmental Assessment and Management*, *12*(1), 67-81. https://doi.org/10.1002/ieam.1676

EFSA (European Food Safety Authority). (2005). Opinion of the PPR Panel on a request from EFSA related to the evaluation of dimoxystrobin. *EFSA Journal*, *3*(3), 178.

EFSA (European Food Safety Authority). (2017). EFSA Guidance Document for predicting environmental concentrations of active substances of plant protection products and transformation products of these active substances in soil. *EFSA Journal*, *15*(10), e04982. https://doi.org/10.2903/j.efsa.2017.4982

EFSA PPR (EFSA Panel on Plant Protection Products and their Residues). (2013). Guidance on tiered risk assessment for plant protection products for aquatic organisms in edge-of-field surface waters. *EFSA Journal*, *11*(7), 3290. https://doi.org/10.2903/j.efsa.2013.3290

EFSA PPR (Panel on Plant Protection Products & their Residues). (2014). Scientific opinion on good modelling practice in the context of mechanistic effect models for risk assessment of plant protection products. *EFSA Journal*, *12*(3), 3589. https://doi.org/10.2903/j.efsa.2014.3589

EFSA PPR (EFSA Panel on Plant Protection Products and their Residues). (2018a). Scientific opinion on the state of the art of toxicokinetic/toxicodynamic (TKTD) effect models for regulatory risk assessment of pesticides for aquatic organisms. *EFSA Journal*, *16*(8), 5377-5565. https://doi.org/10.2903/j.efsa.2018.5377

Erich Gamma, R. H., Ralph Johnson, John Vlissides. (1994). *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley Professional.

European Commission (European Commission, Directorate-General for Health and Consumers). (2014). Addressing the new challenges for risk assessment. https://doi.org/10.2772/37863

Evans, M. R., Grimm, V., Johst, K., Knuuttila, T., de Langhe, R., Lessells, C. M., Merz, M., O'Malley, M. A., Orzack, S. H., Weisberg, M., Wilkinson, D. J., Wolkenhauer, O., & Benton, T. G. (2013). Do simple models lead to generality in ecology? *Trends in Ecology & Evolution*, *28*(10), 578-583. https://doi.org/10.1016/j.tree.2013.05.022

Forbes, V. E., Accolla, C., Banitz, T., Crouse, K., Galic, N., Grimm, V., Raimondo, S., Schmolke, A., & Vaugeois, M. Mechanistic population models for ecological risk assessment and decision support: The importance of good conceptual model diagrams. *Integrated Environmental Assessment and Management*. https://doi.org/10.1002/ieam.4886

Forbes, V. E., Calow, P., Grimm, V., Hayashi, T. I., Jager, T., Katholm, A., Palmqvist, A., Pastorok, R., Salvito, D., Sibly, R., Spromberg, J., Stark, J., & Stillman, R. A. (2011). Adding value to ecological risk assessment with population modeling. *Human and Ecological Risk Assessment*, *17*(2), 287-299, Article Pii 936287334. https://doi.org/10.1080/10807039.2011.552391

Fulton, E. A., Smith, A. D. M., & Johnson, C. R. (2003). Effect of complexity on marine ecosystem models. *Marine Ecology Progress Series*, *253*, 1-16. https://doi.org/10.3354/meps253001

Gabsi, F., Hammers-Wirtz, M., Grimm, V., Schaffer, A., & Preuss, T. G. (2014). Coupling different mechanistic effect models for capturing individual – and population-level effects of chemicals: Lessons from a case where standard risk assessment failed. *Ecological Modelling*, *280*, 18-29. https://doi.org/10.1016/j.ecolmodel.2013.06.018

Gabsi, F., Solga, A., Bruns, E., Leake, C., & Preuss, T. G. (2019). Short-term to long-term extrapolation of lethal effects of an herbicide on the marine mysid shrimp Americamysis Bahia by use of the General Unified Threshold Model of Survival (GUTS). *Integrated Environmental Assessment and Management*, *15*(1), 29-39. https://doi.org/10.1002/ieam.4092

Gergs, A., Classen, S., Strauss, T., Ottermanns, R., Brock, T. C. M., Ratte, H. T., Hommen, U., & Preuss, T. G. (2016). Ecological Recovery Potential of Freshwater Organisms: Consequences for Environmental Risk Assessment of Chemicals. In P. DeVoogt (Ed.), *Reviews of Environmental Contamination and Toxicology, Vol 236* (Vol. 236, pp. 259-294). Springer. https://doi.org/10.1007/978-3-319-20013-2_5

Gergs, A., Gabsi, F., Zenker, A., & Preuss, T. G. (2016). Demographic toxicokinetic-toxicodynamic modeling of lethal effects. *Environmental Science & Technology*, *50*(11), 6017-6024. https://doi.org/10.1021/acs.est.6b01113

Gergs, A., Hager, J., Bruns, E., & Preuss, T. G. (2021). Disentangling mechanisms behind chronic lethality through toxicokinetic-toxicodynamic modeling. *Environmental Toxicology and Chemistry*, *40*(6), 1706-1712. https://doi.org/10.1002/etc.5027

Gergs, A., Kulkarni, D., & Preuss, T. G. (2015). Body size-dependent toxicokinetics and toxicodynamics could explain intra – and interspecies variability in sensitivity. *Environmental Pollution*, *206*, 449-455. https://doi.org/10.1016/j.envpol.2015.07.045

Gergs, A., Preuss, T. G., & Palmqvist, A. (2014). Double trouble at high density: Cross-level test of resource-related adaptive plasticity and crowding-related fitness. *Plos One*, *9*(3), 13, Article e91503. https://doi.org/10.1371/journal.pone.0091503

Gergs, A., Rakel, K. J., Liesy, D., Zenker, A., & Classen, S. (2019). Mechanistic effect modeling approach for the extrapolation of species sensitivity. *Environmental Science & Technology*, *53*(16), 9818-9825. https://doi.org/10.1021/acs.est.9b01690

Gergs, A., Zenker, A., Grimm, V., & Preuss, T. G. (2013). Chemical and natural stressors combined: from cryptic effects to population extinction. *Scientific Reports*, *3*, 8, Article 2036. https://doi.org/10.1038/srep02036

Gray, T. N. E., Borey, R., Hout, S. K., Chamnan, H., Nigel, Collar, J., & Dolman, P. M. (2009). Generality of models that predict the distribution of species: Conservation activity and reduction of model transferability for a threatened bustard. *Conservation Biology*, *23*(2), 433-439. https://doi.org/10.1111/j.1523-1739.2008.01112.x

Grech, A., Tebby, C., Brochot, C., Bois, F. Y., Bado-Nilles, A., Dorne, J. L., Quignot, N., & Beaudouin, R. (2019). Generic physiologically-based toxicokinetic modelling for fish: Integration of environmental factors and species variability. *Science of the Total Environment*, *651*, 516-531. https://doi.org/10.1016/j.scitotenv.2018.09.163

Grimm, V., Augusiak, J., Focks, A., Frank, B. M., Gabsi, F., Johnston, A. S. A., Liu, C., Martin, B. T., Meli, M., Radchuk, V., Thorbek, P., & Railsback, S. F. (2014). Towards better modelling and decision support: Documenting model development, testing, and analysis using TRACE. *Ecological Modelling*, *280*, 129-139. https://doi.org/10.1016/j.ecolmodel.2014.01.018

Grimm, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., Goss-Custard, J., Grand, T., Heinz, S. K., Huse, G., Huth, A., Jepsen, J. U., Jorgensen, C., Mooij, W. M., Muller, B., Pe'er, G., Piou, C., Railsback, S. F., Robbins, A. M., . . . DeAngelis, D. L. (2006). A standard protocol for describing individual-based and agent-based models. *Ecological Modelling*, *198*(1-2), 115-126. https://doi.org/10.1016/j.ecolmodel.2006.04.023

Grimm, V., Berger, U., DeAngelis, D. L., Polhill, J. G., Giske, J., & Railsback, S. F. (2010). The ODD protocol A review and first update. *Ecological Modelling*, *221*(23), 2760-2768. https://doi.org/10.1016/j.ecolmodel.2010.08.019

Grimm, V., & Railsback, S. F. (2005). *Individual-based Modeling and Ecology* (STU – Student edition ed.). Princeton University Press. http://www.jstor.org/stable/j.ctt5hhnk8

Grimm, V., Railsback, S. F., Vincenot, C. E., Berger, U., Gallagher, C., DeAngelis, D., L., Edmonds, B., Ge, J. Q., Giske, J., Groeneveld, J., Johnston, A. S. A., Milles, A., Nabe-Nielsen, J., Polhill, J. G., Radchuk, V., Rohwader, M. S., Stillman, R. A., Thiele, J. C., Ayllon, D. (2020). The ODD protocol for describing agent-based and other simulation models: A second update to improve clarity, replication, and structural realism. *JASSS – the Journal of Artificial Societies and Social Simulation*, *23*(2), 20, Article 7. https://doi.org/10.18564/jasss.4259

Hommen, U., Forbes, V., Grimm, V., Preuss, T. G., Thorbek, P., & Ducrot, V. (2016). How to use mechanistic effect models in environmental risk assessment of pesticides: Case studies and recommendations from the SETAC Workshop MODELINK. *Integrated Environmental Assessment and Management*, *12*(1), 21-31. https://doi.org/10.1002/ieam.1704

Huang, A., Mangold-Döring, A., Guan, H., Boerwinkel, M. C., Belgers, D., Focks, A., & Van den Brink, P. J. (2023). The effect of temperature on toxicokinetics and the chronic toxicity of insecticides towards Gammarus pulex. *Sci Total Environ*, *856*(Pt 2), 158886. https://doi.org/10.1016/j.scitotenv.2022.158886

brahim, L. P., T. G.; Schaeffer, A.; Hommen, U. (2014). A contribution to the identification of representative vulnerable fish species for pesticide risk assessment in Europe-A comparison of population resilience using matrix models. *Ecological Modelling*, *280*, 65-75. https://doi.org/10.1016/j.ecolmodel.2013.08.001

Jager, T. (2020). Revisiting simplified DEBtox models for analysing ecotoxicity data. *Ecological Modelling*, *416*, 108904. https://doi.org/10.1016/j.ecolmodel.2019.108904

Jager, T., Albert, C., Preuss, T. G., & Ashauer, R. (2011). General unified threshold model of survival – A toxicokinetic-toxicodynamic framework for ecotoxicology. *Environmental Science & Technology*, *45*(7), 2529-2540. https://doi.org/10.1021/es103092a

Jager, T., Goussen, B., & Gergs, A. (2023). Using the standard DEB animal model for toxicokinetic-toxicodynamic analysis. *Ecological Modelling*, *475*, 110187. https://doi.org/10.1016/j.ecolmodel.2022.110187

Jager, T., Trijau, M., Sherborne, N., Goussen, B., & Ashauer, R. (2022). Considerations for using reproduction data in toxicokinetic–toxicodynamic modeling. *Integrated Environmental Assessment and Management*, *18*(2), 479-487. https://doi.org/10.1002/ieam.4476

Janssen, S. D., Viaene, K. P. J., Van Sprang, P., & De Schamphelaere, K. A. C. (2021). Integrating bioavailability of metals in fish population models. *Environmental Toxicology and Chemistry*, *40*(10), 2764-2780. https://doi.org/10.1002/etc.5155

Kaiser, H. (1979). Dynamics of populations as result of the properties of individual animals. Fortschritte Der Zoologie, 25(2-3), 109-136.

Kleinmann, J. U., & Wang, M. (2017). Modeling individual movement decisions of brown hare (lepus europaeus) as a key concept for realistic spatial behavior and exposure: a population model for landscape-level risk assessment. *Environmental Toxicology and Chemistry*, *36*(9), 2299-2307. https://doi.org/10.1002/etc.3760

Larisch, W., Brown, T. N., & Goss, K. U. (2017). A toxicokinetic model for fish including multiphase sorption features. *Environ Toxicol Chem*, *36*(6), 1538-1546. https://doi.org/10.1002/etc.3677

Mavroudis, P. D., Hermes, H. E., Teutonico, D., Preuss, T. G., & Schneckener, S. (2018). Development and validation of a physiology-based model for the prediction of pharmacokinetics/toxicokinetics in rabbits. *Plos One*, *13*(3), e0194294. https://doi.org/10.1371/journal.pone.0194294

McElroy, A., Barron, M., Beckvar, N., Driscoll, S., Meador, J., Parkerton, T., Preuss, T., & Steevens, J. (2011). Use of the tissue residue approach for organic and organometallic compounds. *Integrated Environmental Assessment and Management*, *7*, 50-74. https://doi.org/10.1002/ieam.132

Meyer, B. (1990). *Object-Oriented Software Construction* (2nd ed.). Prentice-Hall.

Najjar, A., Punt, A., Wambaugh, J., Paini, A., Ellison, C., Fragki, S., Bianchi, E., Zhang, F., Westerhout, J., Mueller, D., Li, H., Shi, Q., Gant, T. W., Botham, P., Bars, R., Piersma, A., van Ravenzwaay, B., & Kramer, N. I. (2022). Towards best use and regulatory acceptance of generic physiologically based kinetic (PBK) models for in vitro-to-in vivo extrapolation (IVIVE) in chemical risk assessment. *Archives of Toxicology*, *96*(12), 3407-3419. https://doi.org/10.1007/s00204-022-03356-5

Nyman, A. M., Schirmer, K., & Ashauer, R. (2012). Toxicokinetic-toxicodynamic modelling of survival of Gammarus pulex in multiple pulse exposures to propiconazole: Model assumptions, calibration data requirements and predictive power. *Ecotoxicology*, *21*(7), 1828-1840. https://doi.org/10.1007/s10646-012-0917-0

Omlin, M., Brun, R., & Reichert, P. (2001). Biogeochemical model of Lake Zurich: Sensitivity, identifiability and uncertainty analysis. *Ecological Modelling*, *141*(1-3), 105-123. https://doi.org/10.1016/s0304-3800(01)00257-5

Park, R. A., & Clough, J. S. (2018). *AQUATOX (Release 3.2) – Modeling Environmental Fate and Ecological Effects in Aquatic Ecosystems. Volume 2: Technical Documentation*. https://www.epa.gov/ceam/aquatox-32-supporting-documents.

Park, R. A., Clough, J. S., & Wellman, M. C. (2008). AQUATOX: Modeling environmental fate and ecological effects in aquatic ecosystems. *Ecological Modelling*, *213*(1), 1-15. https://doi.org/10.1016/j.ecolmodel.2008.01.015

Preuss, T. G., Hammers-Wirtz, M., Hommen, U., Rubach, M. N., & Ratte, H. T. (2009). Development and validation of an individual based *Daphnia magna* population model: The influence of crowding on population dynamics. *Ecological Modelling*, *220*(3), 310-329. https://doi.org/10.1016/j.ecolmodel.2008.09.018

Railsback, S. F., & Grimm, V. (2019). *Agent-based and individual-based modeling: a practical introduction*. Princeton university press.

Raimondo, S., Schmolke, A., Pollesch, N., Accolla, C., Galic, N., Moore, A., Vaugeois, M., Rueda-Cediel, P., Kanarek, A., Awkerman, J., & Forbes, V. (2021). Pop-GUIDE: Population modeling guidance, use, interpretation, and development for ecological risk assessment. *Integrated Environmental Assessment and Management*, *17*(4), 767-784. https://doi.org/10.1002/ieam.4377

Rakel, K. J., Preuss, T. G., & Gergs, A. (2020). Individual-based dynamic energy budget modelling of earthworm life-histories in the context of competition. *Ecological Modelling*, *432*, 7, Article 109222. https://doi.org/10.1016/j.ecolmodel.2020.109222

Reeg, J., Schad, T., Preuss, T. G., Solga, A., Korner, K., Mihan, C., & Jeltsch, F. (2017). Modelling direct and indirect effects of herbicides on non-target grassland communities. *Ecological Modelling*, *348*, 44-55. https://doi.org/10.1016/j.ecolmodel.2017.01.010

Reynolds, J. F., & Acock, B. (1997). Modularity and genericness in plant and ecosystem models. *Ecological Modelling*, *94*(1), 7-16. https://doi.org/10.1016/s0304-3800(96)01924-2

Roeben, V., Oberdoerster, S., Rakel, K. J., Liesy, D., Capowiez, Y., Ernst, G., Preuss, T. G., Gergs, A., & Oberdoerster, C. (2020). Towards a spatiotemporally explicit toxicokinetic-toxicodynamic model for earthworm toxicity. *Science of the Total Environment*, *722*, 137673. https://doi.org/10.1016/j.scitotenv.2020.137673

Rosland, R., Alunno-Bruscia, M., Duinker, A., Strand, O., & Strohmeier, T. (2013). *A DEB based analysis of growth and toxin elimination processes in mussels (Mytilus edulis) exposed to Diarrhetic Shellfish Toxins (DST)* DEB 2013 – third DEB symposium, 23-26 April 2013, Texel, Netherlands. https://archimer.ifremer.fr/doc/00148/25936/

Schmolke, A., Abi-Akar, F., Roy, C., Galic, N., & Hinarejos, S. (2020). Simulating honey bee large-scale colony feeding studies using the BEEHAVE Model-Part I: Model validation. *Environmental Toxicology and Chemistry*, *39*(11), 2269-2285. https://doi.org/10.1002/etc.4839

Schmolke, A., Kapo, K. E., Rueda-Cediel, P., Thorbek, P., Brain, R., & Forbes, V. (2017). Developing population models: A systematic approach for pesticide risk assessment using herbaceous plants as an example. *Science of the Total Environment*, *599*, 1929-1938. https://doi.org/10.1016/j.scitotenv.2017.05.116

Schmolke, A., Thorbek, P., DeAngelis, D. L., & Grimm, V. (2010). Ecological models supporting environmental decision making: a strategy for the future. *Trends in Ecology & Evolution*, *25*(8), 479-486. https://doi.org/10.1016/j.tree.2010.05.001

Sherborne, N., Galic, N., & Ashauer, R. (2020). Sublethal effect modelling for environmental risk assessment of chemicals: Problem definition, model variants, application and challenges. *Science of the Total Environment*, *745*, 141027. https://doi.org/10.1016/j.scitotenv.2020.141027

Singer, A., Nickisch, D., & Gergs, A. (2023). Joint survival modelling for multiple species exposed to toxicants. *Science of the Total Environment*, *857*, 159266. https://doi.org/10.1016/j.scitotenv.2022.159266

Stadnicka-Michalak, J., Schirmer, K., & Ashauer, R. (2015). Toxicology across scales: Cell population growth in vitro predicts reduced fish growth. *Science Advances*, *1*(7), 8, Article e1500302. https://doi.org/10.1126/sciadv.1500302

Strauß, T. (2009). *Dynamische Simulation der Planktonentwicklung und interner Stoffflüsse in einem eutrophen Flachsee* (Vol. 44). Umwelt-Forum der RWTH Aachen.

Strauss, T., Gabsi, F., Hammer-Wirtz, M., Thorbek, P., & Preuss, T. G. (2017). The power of hybrid modelling: An example from aquatic ecosystems. *Ecological Modelling*, *364*, 77-88. https://doi.org/10.1016/j.ecolmodel.2017.09.019

Strauss, T., Kulkarni, D., Preuss, T. G., & Hammers-Wirtz, M. (2016). The secret lives of cannibals: Modelling density-dependent processes that regulate population dynamics in *Chaoborus crystallinus*. *Ecological Modelling*, *321*, 84-97. https://doi.org/10.1016/j.ecolmodel.2015.11.004

6. Evaluating the calibration and validation of mechanistic effect models

Simon Hansul, Peter Vermeiren*, Eugenia Chaideftou, Michail Gioutlakis, Udo Hommen, Mira Kattwinkel, Judith Klein, Stefan Reichenberger, Johannes Witt, Sandrine Charles*

** These authors contributed equally*

6.1. Introduction

Once a Mechanistic Effect Model (MEM) has been developed, but before it is applied as part of an environmental risk assessment (ERA), it is crucial that rigorous checks of model outputs are performed to evaluate the quality of the model and its usefulness for predictions. Such analyses of model outputs fall in two broad categories: (i) a comparison between model output and data as seen in calibration and validation, and (ii) assessment of model sensitivity and uncertainty. In this chapter, methods for comparing model outputs with data are described; in Chapter 7, concepts and methods for uncertainty and sensitivity analyses are described. Note that in addition, internal checks such as evaluating parameter correlations and distribution shapes, and the process of obtaining and collating data provide insights into the quality of a model's output, but these will not be covered in this chapter. While the aims of calibration and validation differ, the methods used to compare model output to data are in parts the same or similar. Calibration aims to tailor the parameter values in reducing the discrepancy between data and model outputs, whereas validation aims to check how well the previously calibrated model's output match independent data. In this chapter, data is only loosely defined and might for instance include detailed datasets from comprehensive field studies (e.g., on population size and age structure over time), data from laboratory studies, and observations in natura about the species in question (e.g., about movement patterns, lifespan, or maximum size). Normally more than one method should be employed to ensure a robust assessment of the quality of model outputs in reference to data (i.e., a multi-criteria assessment). Methods range from visual assessments over simple statistics to sophisticated quantitative methods. This chapter aims to provide an overview of approaches used for calibration and validation, with the objective of helping modelers choose the right methods for

the different tasks, supporting model assessors to follow the reasoning behind method selection and assess the respective results and so the model quality.

As will be discussed in this chapter, different methods may be more appropriate than others, depending on the model complexity and on the step in the modeling cycle (Figure 6.1). For example, for simpler models (e.g., General Unified Threshold model of Survival, GUTS), outputs can likely directly be compared to one observed dynamical pattern (survival over time; for definition of "pattern" see section 6.2). For models of medium complexity (e.g., Dynamic Energy Budget coupled Toxicokinetic Toxicodynamic models, DEB-TKTD), multi-pattern comparisons may be useful if applicable (e.g., patterns of body weight as well as reproduction over time). For more complex models such as DEB implementations within individual-based models (DEB-IBM) or ecosystem models, typically several different outputs (e.g., population dynamics, body size, location) will be compared to several observed patterns, not only in a one-on-one comparison, but also across available patterns as a whole pattern matching (see section 6.2).

During calibration and validation steps, model outputs are compared to experimentally collected data or observed patterns, for example from literature. Approaches for evaluating the quality of model outputs can broadly be divided in qualitative (section 6.3.1) and quantitative approaches (section 6.3.2). This division is similar to the concepts of exploratory versus confirmatory data analysis in statistics (Gelman, 2004). Exploratory data analysis focuses on detecting unanticipated areas of model fit, while confirmatory data analysis is focused on checking if discrepancies could have occurred by chance. In statistics, qualitative methods are often favored during exploratory analysis, while quantitative methods are usually preferred for confirmatory data analysis, although often both types of approaches need to be used in a complementary manner (Gelman, 2004). Similarly, in mechanistic modeling qualitative methods are often used to understand a model (e.g., to diagnose causes of poor fit) or to compare general alignment with several expected output patterns whereas quantitative methods are used to compare quantitative model outputs to data.

In the remainder of this chapter, we make the distinction between quantitative and qualitative methods to roughly categorize the different evaluation approaches we present, noticing that it is not always possible nor useful to make a strict distinction. There is a large body of literature on model output evaluation for statistical models, but less for mechanistic models, especially complex models. It is a complex issue, and terminology is not always harmonized (Mentré & Escolano, 2006). The aim of model output evaluation is not to determine whether a model is "correct" or "wrong," but whether "the model's deficiencies have a noticeable effect on the substantive inferences" (Gelman et al., 1995). In line with this thinking, this chapter does not aim at presenting specific cut-off values for specific indices, but rather presents guidance on which topics are of concern and which methods can be used to evaluate them. It should be kept in mind that all models are a simplification of reality. Hence, assessing a model's calibration and validation outcomes using a variety of methods should always consider the model's aims, context, and its application. Indeed, a key question, particularly in a risk assessment context, is whether a model is fit-for-purpose, which includes an assessment of whether it is reliable and can obtain an adequate level of certainty within a specific application

context (Hamilton et al., 2022). The standardized evaluation of calibration and validation against a strict, predetermined set of assessment methods (and linked cutoff values) can create a sense of pseudo-rigorousness when it is interpreted as a mere "ticking the boxes" exercise. Rather, the methods presented in this chapter should be discussed and interpreted within the context of model development and application, and not as a hunt for hitting as many method thresholds as possible.

Some recommendations on model output evaluation of MEMs have already been given in the European Food Safety Authority (EFSA) scientific opinions (SO) on good modeling practice (EFSA PPR, 2014) and the SO on TKTD modeling (EFSA PPR, 2018a). While the EFSA SO on TKTD modeling focuses on three specific types of models (DEB-based toxicity models, GUTS, and TKTD models for primary producers), the recommendations given in the present document are intended to be applicable to MEMs of any type (e.g., also population or ecosystem models). In addition to the recommendations given in the EFSA scientific opinion on good modeling practice, this chapter aims to give an overview and more detailed advice on the specific approaches and tools used for model output evaluation throughout the modeling cycle. First, Pattern-Oriented Modeling (POM, section 6.2) is presented as an approach for model output evaluation, then, visual (section 6.3.1) and quantitative (section 6.3.2) assessment of the correspondence between model outputs and observations are discussed. Evaluation approaches are illustrated with minimal examples chosen specifically for this purpose.

6.1.1. Model evaluation and the modeling cycle

Before coming to more specific aspects of model output evaluation, the position within the modeling cycle is outlined to give some orientation to the reader. Qualitative and quantitative criteria for model output evaluation are used at different points throughout the modeling process (or cycle). Figure 6.1 shows the modeling scheme from Chapter 1, expanded by an indication of elements that require model output evaluation. In correspondence with schemes that are reported in other documents (EFSA PPR, 2014; Schmolke et al., 2017), the same scheme can also be shown in cyclic form.

The modeling process starts with "Problem formulation" (step 1), "Model formulation" (step 2), and "Model formalization" (step 3), all of which are based on a conceptual and formal level of the model. Because there is no model output yet, neither qualitative nor quantitative criteria apply to these steps.

Model output is first produced during "Model implementation" (step 4). At this step, it must be checked if the model implementation is correct (i.e., the results are not influenced by programming "bugs") and produces plausible outputs, namely, reasonable magnitude of outputs, plausible dynamics over time, and reasonable reactions to changes in model inputs. The assessment of model output in step 4 is thus mostly qualitative. During "Model setup" (step 5), parameter values are set, some of which may not be directly measurable, and must be inferred by calibration. Calibration can in certain cases be conducted and evaluated based on qualitative (namely, visual) criteria, for instance for movement rules in spatially explicit

IBMs. However, calibration often requires that the discrepancy between model output and observed data is expressed quantitatively (section 6.3.2). In any case, a qualitative evaluation, and the quantitative metrics are complementary, because they can give information about different aspects of model performance and therefore using both gives a more robust assessment of model output quality. Calibration is of particular importance for parameters that are too abstract to be directly measurable or otherwise hard to measure directly, but still can be inferred from data, for example in generic TKTD models.

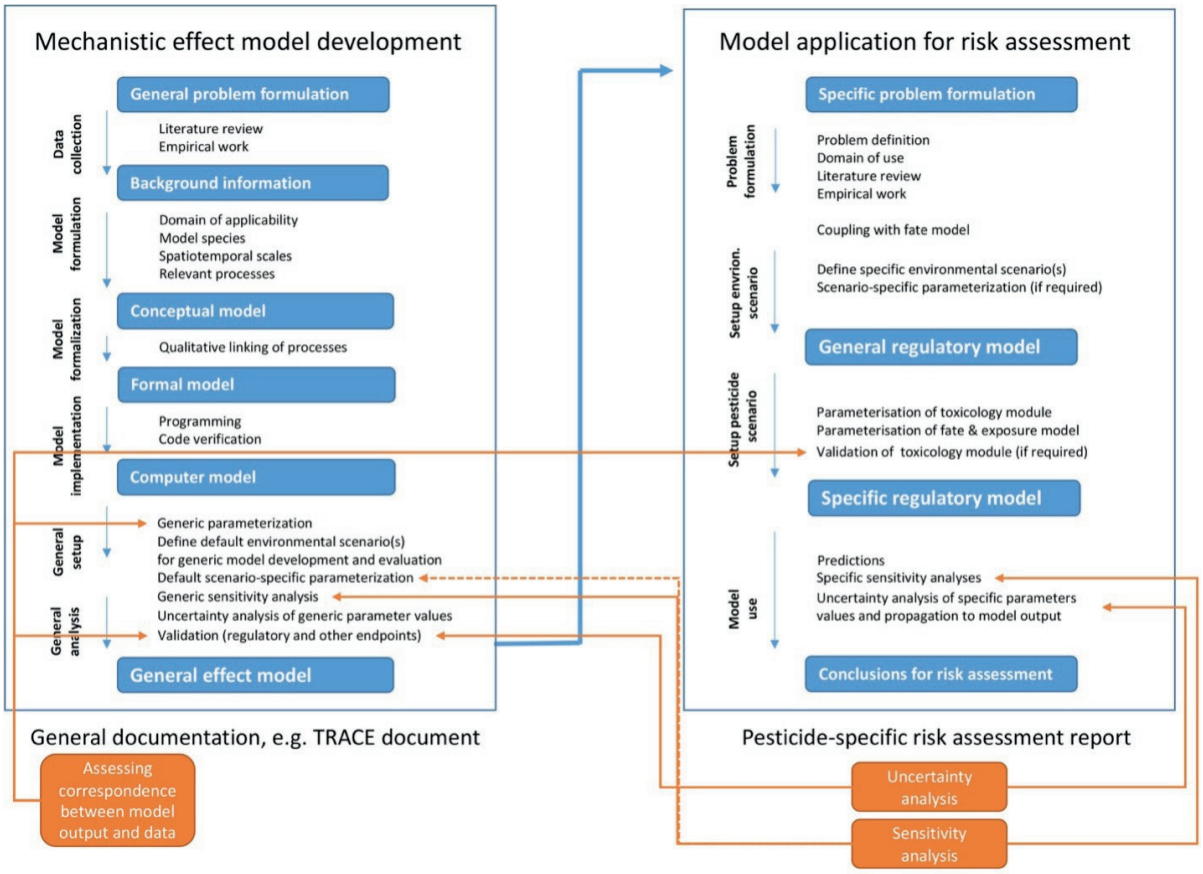


Figure 6.1: The model evaluation scheme (Figure 2.3: Proposed concept for the evaluation of mechanistic effect models and their application for regulatory risk assessment.) with additional boxes indicating elements discussed in this chapter.

Once a model has been developed, calibrated, and validated, and the scenarios for a specific risk assessment (i.e., environmental, exposure, and ecological scenarios) have been set up, the model is finally applied to the risk assessment, that is to say, it might, for instance, be used to predict the effects of the toxicant in question to individual organisms, a population, or a community under untested (field) conditions. Typically, these model predictions are done for situations where observations are not available, and sometimes it is not possible to obtain such data for practical or ethical reasons. Nevertheless, before the results are used in the risk assessment, the quality of model application outputs should be evaluated based on plausibility, which may include studying extreme cases to demonstrate situations when clear effects are expected (e.g., a "toxic standard").

6.2. Pattern-oriented modeling

The aim of MEMs is to reproduce certain patterns observed in nature, based on a mathematical formulation of underlying processes. A key feature of MEMs is that, generally, the model outputs integrate multiple patterns. A "pattern" is a characteristic, clearly identifiable structure in nature itself or in the data extracted from nature (Grimm et al., 1996). Examples of patterns that were so far used in ecological modeling can found in Gallagher et al. (2021). Patterns can be distinguished along a continuum from "strong" to "weak." A pattern is "strong" when it describes outputs that are easily quantifiable and often closely influenced by the model's structure, parameters, and their values. For example, the presence and/or absence of effects on a given endpoint, a change in population size, population age structure, or composition of species assemblages are patterns for which quantitative datasets maybe available when developing a population or community model. Meanwhile, "weak" patterns (also called "vague" patterns; Wiegand et al., 2003) are more qualitative, descriptive outputs that are often less strongly tied to the model's design and parameterization. In the context of MEMs, a weak pattern might be the increase or decrease in effect strength during a time period without exact numeric specification, or the timeframe over which oscillations in population sizes occur. Individual weak patterns only allow for weak (i.e., highly uncertain) inference of model structure and parameters, but combinations of weak patterns may allow for strong inference, possibly more so than a single strong pattern.

Around this idea of designing and evaluating models explicitly in relation to multiple patterns (rather than in reference to one outcome or pattern), the framework of Pattern-Oriented Modeling (POM) has been developed. The initiators of the framework define POM as "the multi-criteria design, selection, and calibration of models of complex systems" (Grimm & Railsback, 2012). POM is also useful for validation: Complex models can be easily calibrated to match a single pattern, but the pattern may not be matched for the right reasons. In contrast, it is hard to match several patterns simultaneously unless the model captures the functioning of the system (Grimm & Railsback, 2012). While not all MEMs are necessarily associated with complex systems, POM is highly suitable for the design and evaluation of MEMs and provides a framework for the integrated use of quantitative and qualitative criteria during model output evaluation (Schmolke et al., 2010). Particularly, POM provides a structured model evaluation framework that is particularly useful in the following four situations, the first three of which were also suggested by (Grimm & Railsback, 2012):

1. Determine which scales, entities, variables, and processes the model needs (setting the model domain, and focusing on which patterns to evaluate);
2. Test and select (sub)models to represent key processes (model evaluation);
3. Estimate useful parameter values during calibration (iterative model parameterization);
4. Judge if predictions are plausible (i.e., match with known ecological behavior).

Which patterns the model should reproduce depends on the model purpose, data availability, and the processes the model intends to capture to fulfil that purpose. The identification of relevant patterns to be reproduced is often clear on an intuitive level. Nevertheless, defining these patterns a priori, based on theoretical and empirical understanding of the studied system allows for increased objectivity in the evaluation of model performance, the selection of appropriate quantitative and qualitative methods, and anchoring visual assessments around key features of the pattern to be replicated. Additionally, it must be clearly reported which patterns are considered relevant (and to which degree the model should be able to replicate them), during which stages of the modeling cycle, and why. The relevance of patterns might change across steps of the modeling cycle, for example if these steps correspond to different levels of organization. A number of selection criteria, discussed below (section 6.2.1), can help to identify and describe patterns that are to be considered during POM. Once the patterns of interest have been identified, they help to determine the temporal and spatial scales (Grimm et al., 1996), and the levels of biological organization on which the model operates. This, then, provides a frame for modelers to define the domain of applicability of the model, and for regulators to assess the applicability of the model to specific situations.

The use of multiple patterns is a key concept of POM (Grimm & Railsback, 2012), which makes POM particularly relevant for the evaluation of MEMs. The reason is that a specific pattern of a model output can often be generated by multiple generating mechanisms, formulated as different model structures or parameterizations of the underlying processes. A combination of multiple patterns, however, is much harder to replicate. Consequently, the more patterns are identified and evaluated, the higher the confidence that the model correctly integrates the causative mechanisms, that is to say, the mechanisms that control the system dynamics (Railsback & Johnson, 2011). This property of POM provides a natural basis for model validation that is particularly valuable because it increases our understanding of underlying mechanisms. Additionally, POM also provides a basis for calibration as it allows different parameterization options (i.e., model calibrations) to be compared. An example of POM applied to calibration is demonstrated in a study on Lesser Spotted Woodpecker (*Picoides minor*), where pre-breeding survival rates were missing for the study sites (Rossmanith et al., 2007). Using POM, the authors were able to calibrate parameters for pre-breeding survival rates against patterns regarding population structure (sex ratio and pair ratio), nesting success, breeding success, and territory occupancy at the population level (Rossmanith et al., 2007). In this context, it should be clear which datasets are used to inform calibration, and which are used for validation. While calibration and validation are often an iterative process and data limitations are frequently a restricting factor, it is advisable to have some patterns or data designated only for the final model validation. In summary, both model developers and evaluators benefit from POM as a clear, a priori, listing of the (multiple) relevant patterns (and associated data) to be replicated by the model and a subsequent framework against which to conduct model output evaluation.

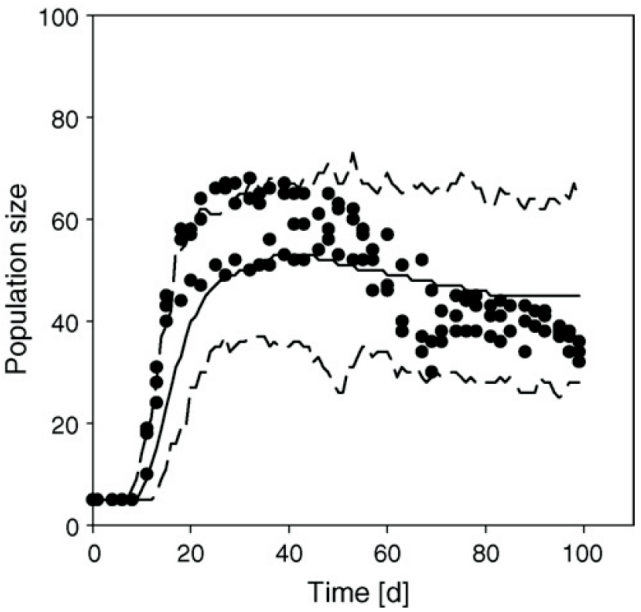


Figure 6.2: Logistic growth of D. magna populations in a flow-through system as the result of crowding, resource competition, and starvation. Lines show the minimum, maximum, and mean of 1000 simulations. Individual points indicate experimental observations (n = 3) of population size over time. Reprinted with permission from Preuss et al. (2009).

6.2.1. Criteria to identify patterns of interest

In this section, we will discuss three key criteria that can be used to identify patterns of interest when using POM as a model output evaluation framework, specifically, the criteria of "emergence," "patterns across multiple levels of organization," and "relating to scientific hypotheses or specific problem formulations."

6.2.1.1. Criterium of emergent properties

A core concept in relation to POM are emergent properties (also referred to as emergent behavior or emergent dynamics). Emergent properties are patterns which result from the interaction between model elements, without being explicitly part of the model formulation, or directly deducible from observation of subsystems (Breckling et al., 2005). Note that we are using a relatively broad definition of emergent properties, which does not require adaptive behavior of model components. If, for example, an IBM predicts a population to grow logistically as a result of resource competition between individuals, the logistic growth pattern can be considered an emergent property: The logistic growth equation appears nowhere in the IBM, but the pattern results from individuals being limited by a shared resource (See Figure 6.2). Table 6.1 gives more literature examples of emergent properties. Emergent properties can be the result of complex interactions and are one of the main reasons that mechanistic models can be used to extrapolate to untested conditions at all. Both quantitative metrics and qualitative visualizations should be used to analyze those emergent properties.

6.2.1.2. Criterium of patterns across multiple organizational levels

MEMs can produce outputs at different levels of organization, such as model state variables that are related to certain levels of biological organization, such as individual-level length and weight, or population-level age structure, as well as derived higher-level summary statistics, such as predicted ECx estimates. As an example, consider a DEB-IBM implementation where effects of a toxicant on a species are extrapolated from the individual – to the population-level (e.g., Vlaeminck et al., 2021). In the population context, with resource competition as a relevant process, the maximum length that individuals reach can be considered an emergent property, because it is the result of interaction between individuals and the (heterogeneous) environment, as well as indirect interactions between individuals via resource competition. This is an example of an emergent property at the level of the state variable. Population density, meanwhile, is a (simple) derived higher-level output, because state variables have to be aggregated to quantify it. Likewise, in this model, a summary statistic such as population growth rate, or an EC10 or EC50 on population growth rate, is a higher-level model output that has to be derived from population densities. We previously noted that POM aids in identifying generative mechanisms, essentially the formulation of model equations and structures based on underlying biological processes, with the aim of generating output patterns emerging from these processes (Grimm et al., 2012). In this context, "emergence" and "patterns across levels of organization" are relevant criteria in model output evaluation as they allow for evaluation and development of MEMs based on mechanistic understanding.

Table 6.1: Examples for emergent properties in ecological and ecotoxicological models.

Model	Underlying mechanisms	Emergent properties	Study
Matrix population model	Age and/or cohort-specific survival and fecundity rates	Long-term population growth rate, stable age structure	(Klanjscek et al., 2006)
D. magna individual-based model	Starvation, Crowding, Resource competition	Population dynamics, Life-stage structure	(Preuss et al., 2009)
Individual-based model of *D. galeata x hyalina* vertical migration and life-history	Visual, size-selective predation Light – and predator-induced movement Light extinction Food search	Diel vertical migration, Population growth rate, Life-stage structure	(Rinke & Petzoldt, 2003)
BEEHAVE model of *Apis mellifera* colonies and effects of parasites	Complex interaction between and within submodels (landscape, mite, colony)	For example, egg laying rate, use of flower patches, Number of brood cells, Number of workers	(Becher et al., 2014)
Dynamic Energy Budget Individual-based Model of bioaccumulation of PCBs in *Solea solea*	Inter-individual variability in feeding	Population-level variation in contamination levels	(Mounier et al., 2020)

Individual-based exposure model for pronghorn (*Antilocapra americana)*	Individual movement following habitat preferences	Population-level distribution of daily exposure doses	(Purucker et al., 2007)
Modular model	Altered foraging behavior in response to the exposure histories of individuals to noise pollution during 3D movements, coupled with a bioenergetics model. Impacts of oil exposure on demographic transitions	Population dynamics	(Farmer et al., 2018)
Food web models	Predator prey relations	Bottom-up, top-down effects	(Lombardo et al., 2015)

An additional advantage of MEMs that produce outputs at different levels of organization is that, often, the data available for calibration and validation might be limited to a specific level of organization (e.g., individual-level measurements of toxic effects) while the interest of the model is at a different level (e.g., population-level resilience measures). Hence, we can indirectly calibrate and validate a model with data from levels of organization that are not the key focus of model's predictions.

One will often focus an initial model evaluation on investigating derived (i.e., higher level) results with relevance to the risk assessment question that the model is aimed at answering, for example, population size or time to recovery. At some point, however, the connection to lower levels of organization, down to the model's state variables, should ideally be made if a deeper mechanistic understanding is desired, or if the model's mechanistic basis is to be evaluated. Examples of this include age of first foraging for honey bees (Becher et al., 2014) or specific seasonal growth rates in population models. Such coupling (referred to as "assessment of emergence") should be part of model analysis after a first investigation (Grimm & Railsback, 2012). Nevertheless, for complex models, emergence assessment is not a trivial task; indeed, one of the reasons that mechanistic models are needed is that humans are not able to integrate all relevant relationships by common sense reasoning. To move toward better understanding of the link between state variables and derived outputs at different levels of organization within complex models such as MEMs, a stepwise approach in model development could be followed. This may start with observation of high-level (often emergent) patterns of model output, followed by formulation of a candidate mechanism (i.e., a rationale explaining how model entities and state variables interact to produce the higher-level pattern). By defining alternative representations of certain mechanisms in the model structure or its parameter values (Railsback & Harvey, 2002), simulation experiments can be used to identify mechanisms that are necessary or sufficient to produce a certain pattern. Alternatively, the inverse approach (starting with lower-level patterns) can be used. In either case, the use of patterns across multiple levels of model output increases understanding of model behavior and the identification of generative mechanisms.

There is an inherent difference in the selection of relevant patterns between the calibration, validation, and prediction steps. Patterns for validation and prediction tend to be given by the problem formulation. These might be certain aspects of population or community dynamics that are directly linked to the risk assessment question (e.g., time until recovery after application of a pesticide), with the possible addition of patterns that are not directly related to the risk assessment question but help to assess the plausibility of model output (e.g., age and/or size structure, reproduction, and mortality rates). In contrast, the patterns that are selected for model calibration imply a scientific hypothesis about the processes that drive the validation patterns (e.g., individual life-history traits, which are believed to drive the rate of population recovery). A failure to match patterns during validation can have multiple reasons, but one possibility is a mismatch between calibration and validation patterns (i.e., the patterns which were selected to evaluate model output during calibration are not truly the most relevant for matching validation patterns). This implies that the choice of calibration patterns might have to be re-evaluated if a developed model is applied to new problems.

6.2.2. Weighing multiple patterns

POM evaluates a model (during calibration, validation, or prediction) against multiple patterns. Data regarding some patterns, however, might be easier to obtain than others. Additionally, different levels of detail can be derived for different patterns, that is to say, some patterns are "strong" and others "weak" (Grimm et al., 2020). Observations on internal concentrations of pollutants, for example, can be limited in replication due to the cost of the analysis and ethical restrictions in destructive sampling. Additionally, instrument detection limits and accuracy can limit the information retained within the data. Such experimental limitations might limit the ability to use patterns in such data. Alternatively, some patterns are more directly influenced by the parameters, and the specific parameter values, in a model. Such strong patterns are highly useful in evaluating models. (Wiegand et al., 2003) Nevertheless, weak patterns are often well-known features of system behavior, and a set of multiple weak patterns can be very useful for model calibration and validation. Reproducing multiple weak patterns, each reflecting different aspects of the system, puts stringent constraints on the possible model structure and parameter values that can create these patterns (Grimm & Railsback, 2005; Rossmanith et al., 2007). Thus, several aspects of the pattern(s) should be considered, including how different patterns complement and/or overlap with each other.

A single pattern might outweigh multiple others if it is of particular relevance for the risk assessment question. If, for example, the timing of the application of a plant protection product is part of the risk assessment question, correctly predicting population structure and spatial distributions may be outweighed by a failure to predict the timing in oscillations of population dynamics. Consequently, some level of expert judgement will be necessary to weigh patterns on a case-by-case basis.

6.3. Assessing correspondence between model outputs and observations

Correspondence between model output and observations is an essential part of model output evaluation. A combination of different assessments will lead to more robust conclusions about model output reliability than any single assessment on its own. This section outlines visual and quantitative approaches for assessment of model output correspondence with the aim to give an overview of commonly used metrics. It should be noted that a detailed comparison between these approaches, their origins, and their current use is beyond the scope of the MAD working group.

6.3.1. Visual predictive check

The method of visually comparing model outputs with observations can be summarized under the term "Visual Predictive Check" (VPC), introduced in the context of pharmacokinetic-pharmacodynamic models (Mentré & Escolano, 2006). However, VPC is not restricted to predictions in the sense of Figure 6.2, but can be applied in all modeling steps that require assessment of correspondence between model outputs and observations.

In this section, we aim to provide some recommendations in relation to VPC of MEMs. This includes two main issues: Firstly, how does one evaluate whether a visual evaluation has been designed carefully and fits the purpose? Secondly, how can we interpret a visually detected mismatch between predictions and observations, informing further steps to reach a more objective conclusion?

The latter issue is related to the obvious disadvantage of VPC that it is to some degree subjective and can be affected by how the data are presented (e.g., scaling of axes). VPC is nevertheless indispensable: It allows for a relatively quick evaluation of model outputs, including non-linear patterns that might be easily obscured in scalar evaluation metrics. Importantly, often the discrepancies identified during VPC can help identifying which model process(es) needs to be improved, whereas a simple scalar evaluation metric will simply tell that the fit is poor, but not why it is poor.

6.3.1.1. Evaluating the design of VPC

The method of visualizing and interpreting predicted versus observed data points as done during VPC is in principle simple and intuitive. However, VPC is sensitive to confirmation bias, for example, the enthusiastic modeler may see a good fit, whereas a model skeptic may see a poor fit in the same plots. We therefore recommend some steps to make VPCs more robust and transparent; the points mentioned below are not exhaustive and we do not imply that deviations render the VPC invalid. However, the mentioned points will aid evaluators to judge whether VPC-type visualizations were designed carefully and fit for purpose.

6.3.1.2. Use of multiple visualizations

Just as individual quantitative measures have weaknesses and should be used in combination with each other, each kind of visual examination can only show one angle of the problem. Part of a rigorous qualitative model evaluation should therefore be to effectively show the model outputs from multiple angles and at different scales and dimensions. Table 6.2 shows an overview of plot types that are frequently used, and the context within which they may be useful; other plot types may also be helpful for some model types, for example, spatial plots of movement patterns in landscapes as predicted by models and observed with radio tracking. An example for how VPC might be done is given for a simple case (calibration and validation of GUTS model for *Gammarus pulex* exposed to dimethoate) shown in Figure 6.3.

6.3.1.3. Match between patterns of interest and visualization

We assume that a set of patterns of interest has been identified a priori and communicated explicitly. The visualizations chosen for VPC should show these patterns as directly as possible. Problems with this may arise when model output is aggregated over multiple simulation runs, as this might obscure patterns (e.g., due to oscillatory dynamics slightly shifted in time between simulation runs).

Table 6.2: Non-exhaustive overview of common plot types for Visual Predictive Check, used to convey different kinds of information and complemented by different metrics.

Type of plot	Type of information	Suggested complementary metrics (examples)	Example
Predicted versus observed scatterplot	Error, bias, deviation, accuracy	Posterior Predictive Check (PPC) Nash-Sutcliffe Efficiency (NSE) Mean Absolute Error (MAE) Mean Relative Error (MRE), (Normalized) Root Mean Squared Error (NRMSE)	Figure 6.3
Residuals per treatment over time	Time-point and treatment-specific error and bias over time	NSE, MAE, MRE	For example, (FOCUS, 2006), Figure 6.2
Predicted and observed values per treatment over time	Specific error over time, biological plausibility, emerging patterns	NSE, MAE, MRE per treatment and/or observed time-point	For example, (Hansul et al., 2021)
Predicted and observed concentration / dose-response	Approximation of relative effect	Fold-difference in ECx of interest Absolute difference in ECx of interest Standardized Prediction Error (SPE)	For example, Figures 2-12 Annex G (EFSA, 2023c)

Note that the recommendations given below in principle also apply to the interpretation of quantitative criteria, but the rationale is often easier to apply with VPC as this more easily detects mismatches that only occur in specific treatments (or at specific time-points, etc.).

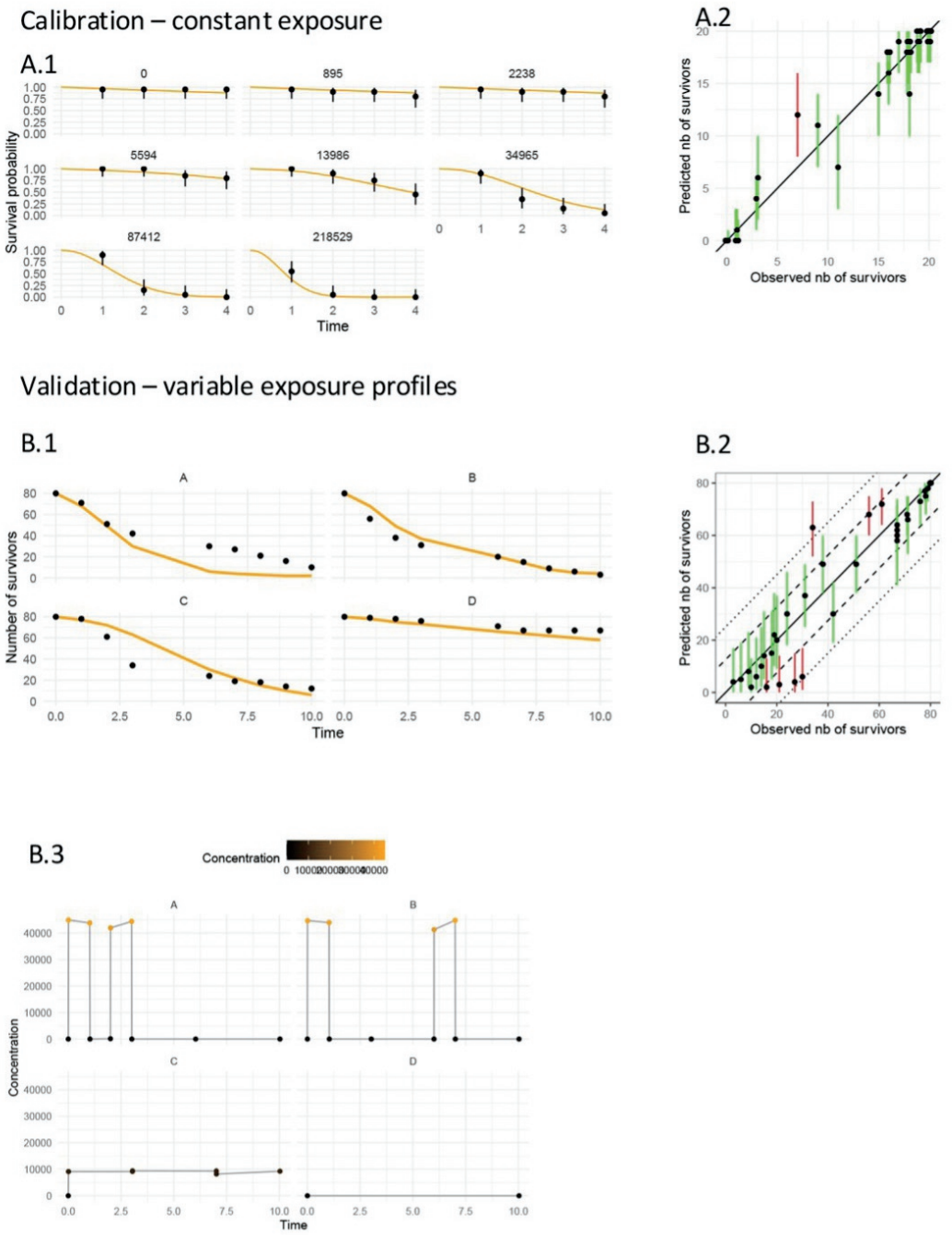


Figure 6.3: Calibration and validation steps on survival data obtained for *Gammarus pulex* exposed to dimethoate (Ashauer et al., 2020). (A.1) Fit values of the GUTS-RED-SD model superimposed on observations collected according to a standard protocol under an increasing range of constant exposure concentrations; (A.2) the corresponding Posterior Predictive Check (PPC) where each black dot corresponds to one observation from the dataset, the corresponding observed value being read on the x-axis, while the y-axis reports the value predicted by the fit model, as well as the 95% uncertainty interval. The interval is depicted in green if it contains the observed value, in red otherwise; (B.1) simulations from the previously fit GUTS-RED-SD model with the propagated uncertainty, superimposed to new observations collected under time-variable exposure profiles (B.3); (B.2) the corresponding PPC of this validation step providing 80.5% of the data within the uncertainty band. Concentrations are given in nmol/L.

The problem is exacerbated when values are averaged over different models and/or parameterizations, for example, when parameter uncertainty is propagated. At the same time, the practice of averaging over simulation runs is useful to display overall trends and should not be dismissed entirely. Using multiple visualizations (e.g., time-series in combination with concentration-response curves) as mentioned above can help to minimize the chance that relevant patterns produced by the model are overlooked due to the choice of visualization.

6.3.1.4. Source of prediction intervals

If an interval for predicted values is shown, it should be clearly stated how the interval was obtained. A prediction interval might indicate variability in model output due to stochasticity (including individual variability), be the result of uncertainty propagation, or both. Depending on the source(s), the interval should be interpreted differently: If variability in model output stems from stochasticity, it is inherent to the system and cannot be reduced. By contrast, if it stems from uncertainty, the prediction interval could possibly be reduced by gathering more (different) data for parameterization or choosing a more appropriate model formulation.

6.3.1.5. Nature of the prediction

When predicted and observed values are compared, it has to be made clear in every case whether the prediction is showing the result of a calibration (that is, data was used to fit the model), cross-validation (data that was not used to fit the model, but was a subpart of the calibration dataset, also referred to as internal validation) or an independent validation (such as population-level effects predicted from individual-level toxicity; for validation independent data that has not been used for calibration has to be used, also termed external validation). In each case, the predicted and observed values have to be compared differently. If a model shows an insufficient goodness of fit (GOF) during calibration, this can have many reasons, such as inappropriate model formulation, calibration algorithm, error function, etc. If a model fulfils most of the goodness-of-fit criteria but a relatively bad cross-validation output, this generally indicates overfitting. If a model exhibits both good calibration and cross-validation, but inaccurate independent validation, this likely points at problems with respect to model formulation or limits to the model's domain of applicability. The different visual interpretation also feeds back on interpretation of the corresponding quantitative criteria and can help to diagnose the underlying causes of discrepancies.

6.3.1.6. Multi-dimensionality

Different subsets of the data (e.g., treatments) are often indicated by color or shape encoding. Whether this is effective depends on the number of subsets, whether they lie on a categorical or at least an ordinal scale, and how many dimensions are shown. For high-dimensional data, small multiples (panels) are often more effective than complex color or shape encodings, as the latter might obscure relevant patterns in the data.

6.3.1.7. Interpreting mismatches between model outputs and observations

Whatever the modeling step (here mainly, calibration, validation, or prediction), the most common plot types for VPC show predicted (i.e., the model outputs) and observed values as function of an independent variable (e.g., time, space), or display the match and/or mismatch as a scatterplot of predicted versus observed values. There is no general rule for deciding which deviations between observation and model output should lead one to conclude that the model performance is insufficient. However, some recommendations can be provided as to which actions should follow from a visually detected mismatch during calibration or validation. These recommendations are most relevant for borderline cases where it is not obvious how to judge the extent of the mismatch.

6.3.1.8. Mismatches during calibration

If a mismatch occurs during calibration, it is desirable to first identify the processes and/or parameters that cause the mismatch. For simple models, this may be done based on analytical knowledge of the system. Even for complex models, looking at where and when the mismatches occur is an important diagnostic tool that can point to the processes that should be improved. For more complex models, one might also consult the results of a previously conducted sensitivity analysis (SA) or conduct an additional parameter sweep (i.e., running the model iteratively with a range of alternative parameter values).

When several patterns are used, the match of a model's outputs to the different patterns have to be weighed against each other (in case some patterns are of higher importance and/or relevance to the application of the model) to assess whether the overall fit is acceptable in the context of the model's aim and application. In some cases, the conclusion is that the mismatch can only be resolved at the expense of correct prediction of another pattern of interest, or by increasing model flexibility and therefore complexity. If that is the case, it is worthwhile to get an idea of the consequence of the mismatch in the validation scenario. If the extent of the mismatch has no unacceptable consequences for the patterns of interest defined for the validation scenario, it may not justify increased model complexity (see also Chapter 7 on sensitivity and uncertainty analysis). However, it should be clear that the mismatch does not imply a biologically implausible model parameterization, and that the mismatch is without unacceptable consequences in all validation and/or prediction scenarios to which the model is developed to be applied.

6.3.1.9. Mismatches during validation

If a mismatch that is considered relevant for the problem formulation occurs during validation, a first step is – as for calibration – to seek to explain what causes it. If the mismatch can be linked to lower-level processes (e.g., patterns of interest), the principles for interpreting mismatches during calibration apply. In addition, uncertainty analysis (UA) may help to understand whether the mismatch only occurs for the point estimates of parameters or will be mitigated for other probable parameter values (possibly at the expense of predicting other patterns and/or parts of the data well).

If the mismatch cannot successfully be linked to lower-level processes, this may prompt a refinement of the assessment of emergence. In any case, mismatches that occur during validation can only partially be interpreted from a purely scientific perspective, as the conclusion whether the mismatch is within an acceptable range or not also depends on the application of the model, and whether the level of inaccuracy with which the final risk assessment question will be answered is acceptable from a decision-making perspective.

6.3.2. Quantitative measures

This section describes measures and considerations for quantifying the correspondence between model outputs and observations. This is followed by a section on model selection criteria. Finally, a brief outlook is given on how to derive acceptability criteria for specific models.

6.3.2.1. Quantifications of the correspondence between model outputs and observation

Quantitative measures on the deviations between model outputs and observations are needed in three steps during the modeling cycle:

1. As an optimization criterion within a calibration algorithm;
2. To check the goodness of fit of a calibration step, for example, to select between different models or to decide whether a model has been calibrated sufficiently well before continuing in the modeling cycle;
3. To describe the predictive accuracy of the model (e.g., Reichenberger et al. [2019]), that is, how well the model can predict observations not used for calibration (validation, output corroboration).

Not all deviation measures are equally suitable for all three steps. Also, for the same measure, potential cut-off or threshold values should be chosen differently for a calibration dataset and a validation dataset. Further, the measures to be used depend on the response types (i.e., whether the variable to be predicted is continuous, binary or quantal, counts [discrete] or time-to-event values). Finally, the suitability of a measure may also depend on the type of model; For example, different measures might be suitable for data with a monotonous decline, an exponential increase, or with seasonal variability. Therefore, a level of expert judgement is necessary when choosing appropriate measures for the task at hand, nevertheless, clear justification should be given to help non-experts assess whether chosen measures are adequate.

During model calibration, a single scalar value is usually needed to describe the deviation between model output and observed data for every set of parameter values that is evaluated. Nevertheless, it is crucial to keep in mind that, in the end, several evaluation criteria may be considered together and

comparatively before making a final decision on model calibration. The choice of a quantitative deviation criterion should not be biased toward larger numbers (as, e.g., Euclidean distance would be) and should be symmetrical (i.e., negative and positive values must not cancel each other out, as would be the case for the mean error). If only a single observed variable is used to calibrate the model, a simple deviation function such as the squared sum of errors will most likely be suitable. If multiple observed variables are used, this introduces two problems: Firstly, the variables might have different dimensions and secondly, the datasets might contain different numbers of observations (data points). The problem of different dimensions is solved by scaling each dataset, that is to say, by dividing all values in each dataset by a summary statistic that has the same unit as the observed variable, such as standard deviation (van der Vaart et al., 2015) or maximum of all observed values. The problem of different numbers of data points can be solved by weighing each dataset inversely proportional to its number of data points. By scaling and weighting variables, the correspondence between model output and complex datasets can be appropriately expressed in a single summary quantity. If a calibration that involves multiple observed variables does not result in a reasonable fit, it is worthwhile to check whether the optimization criteria, when calculated for individual variables separately, are within similar orders of magnitude. If this is not the case, a problem related to scaling or weighting variables may be the cause.

In some cases, the ideal optimization criterion used for calibration might not be intuitively interpretable, making it more difficult to judge the acceptability of the calibration outcome. For example, if variables were scaled by standard deviations and if fitting errors for different datasets were summed up, the resulting optimization function might be well-suited to identify the best fit parameters but will be hard to make sense of. Different approaches might be taken to deal with that: an intuitively interpretable deviation measure can be calculated, in addition to the optimization criterion, for the sole purpose of a more accessible interpretation.

Specific model assumptions can be checked and the deviation from data can be evaluated by variance-based methods, such as the coefficient of determination used under a frequentist framework. Null-Hypothesis Significance Testing (NHST) can also be used to check for specific trends. Such an approach is not suitable under a Bayesian framework where comparisons can be based directly on posterior probability distributions of model outputs. Generally, different measures are needed in combination given that there is no single indicator which can be considered as the best (EFSA PPR, 2018a; FOCUS, 2006; Reichenberger et al., 2019).

Often, the same deviation measures used to check the goodness of fit achieved by calibration can be applied equally as a measure for accuracy of a prediction during model validation. For example, this is the case of the PPC and its associated percentage of data falling within the prediction uncertainty interval. However, a lower deviation will be expected for calibration data to which the model is optimized than for validation data, which are compared to "blind" model simulations. Therefore, acceptability criteria for validation will generally be less strict than for calibration. A non-exhaustive overview of commonly used quantitative deviation measures is given in Table 6.3, which mainly focuses on global comparisons

between model outputs and data. Also, Table 6.3 focuses on deviation measures for regression models. Measures related to classification are presented, for example, in Lee & Lim (2019). Note that most measures are suitable to highlight a specific aspect of disagreement between model output and data, like bias or linearity, but none of them excels at capturing all aspects.

One frequent question when using measures, such as those listed above, to compare model and observations is about acceptability criteria: Which values of the specific measures can be regarded as acceptable in regulatory risk assessment, and where is the cut? Unfortunately, answering this question in general is impossible, and we think that it should probably be explicitly avoided, because what is acceptable depends on the quality of the observations, the specific model purpose, and the area of application. For instance, for an application in regulatory risk assessment, specific protection goals may need to be taken into account, the criteria for model acceptability may be related to acceptance criteria for study types used in standard risk assessments; for instance, if data from a certain study type is known to be highly variable, a model may not be able to match all studies well.

Similarly, more stringent criteria may be needed for endangered species than for common species or for vertebrates compared with invertebrates.

Nevertheless, a generic way how appropriate acceptability criteria for a specific context could be derived is briefly described in the outlook (section 6.3.3). In general, quantitative measures should not be considered in isolation, but ideally in combination, and together with different types of visualization. For instance, the quantitative measure may indicate a good fit, but a visual inspection may reveal that the calibration has been driven by the highest values and the fit of more relevant scenarios for risk assessment is poor. Visual assessment may help to better assess, interpret, and understand the quantitative measures; for example, a high root mean squared error may be caused by a large scatter in the data, by a single outlier, or by systematic over – or underestimation of the data. These different explanations may have different consequences in regulatory risk assessments. In turn, quantitative approaches provide an objective measure to support the inherently subjective visual impression, and they help in highlighting features that may have been overlooked by a more visual assessment. As mentioned before, the combination of quantitative metrics with more qualitative evaluation methods (cf, section 6.2) is crucial, because MEMs can not typically provide the level of numerical validity that is suggested when model output evaluation is done based on purely quantitative criteria. For the use of models in a risk assessment context, accuracy or precision may also be balanced with protectiveness, though the decision to favor protectiveness over accuracy lies ultimately in the hands of risk assessors and/or risk managers. If a model provides inaccurate predictions, but shows a bias to conservative predictions, it can still be useful for risk assessment.

Table 6.3: Non-exhaustive overview of commonly used measures to compare observations and model results for regression models.

Name	Calculation	Suggested by / used for
Adequacy (A)	$\frac{\min(UL(P), UL(O)) - \max(LL(P), LL(O))}{UL(O) - LL(O)}$	High value if the range of observations is within the range of predictions; (Schmolke et al., 2020; Scholten & Van Der Tol, 1994)
Coefficient of determination (R2); Nash-Sutcliffe coefficient of Efficiency (NSE); Model efficiency (ME)	$1 - \frac{\sum(O_i - P_i)^2}{\sum(O_i - \bar{O})^2}$	R^2 as GOF measure, NSE or ME for comparison of observed and predicted values; conventional R2 derived from linear regression[7] (Reichenberger et al., 2019)
Mean (absolute) Percent Error (MAE)	$\frac{100}{\mathrm{n}} \cdot \sum abs(\frac{P_i - O_i}{O_i})$	Ignores the direction of the deviation; does not work for zero values of observations
Mean Percent Error (MPE)	$100 \cdot \frac{\sum \frac{(P_i - O_i)}{O_i}}{n}$	Indicates trend of over – or underestimation; does not work for zero values of observations
PBIAS	$\frac{\sum O_i - P_i}{\sum O_i} \cdot 100\%$	GOF and predictive accuracy of regression model (Reichenberger et al., 2019)
Posterior predictive check (PPC)	percentage of predictions falling within the uncertainty band	For assessing what percentage of the observed data is matched by the model predictions (including model uncertainty); (EFSA PPR, 2018a);
Predictive squared correlation coefficient (Q2)	$1 - \frac{\left[\sum_{i=1}^{n_{val}}(O_i - P_i)^2\right] / n_{val}}{\left[\sum_{i=1}^{n_{cal}}(O_i - P_i)^2\right] / n_{cal}}$	For cross-validation of a regression model (Reichenberger et al., 2019)
Reliability (R)	$\frac{\min(UL(P), UL(O)) - \max(LL(P), LL(O))}{UL(P) - LL(P)}$	High value if the range of predictions is within the range of observations; (Scholten & Van Der Tol, 1994)
RMSE-standard deviation ratio (RSR)	$\sqrt{\frac{\sum(O_i - P_i)^2}{\sum(O_i - \bar{O})^2}}$	Indicates how well the model explains the variance in the observations; sensitive to outliers; (Bennett et al., 2013; Schmolke et al., 2020)

7 Modifications for nonlinear models exist (pseudo-R2), e.g., McFaddens's R2, Nagelkerke's R2. Pseudo R2 cannot be interpreted independently or compared across datasets, they are valid and useful in evaluating multiple models predicting the same outcome on the same dataset.

Scaled / Normalized Root Mean Squared Error	$\frac{1}{\bar{O}}\sqrt{\frac{\sum(P_i - O_i)^2}{n}}$	For validation of TKTD models (EFSA PPR, 2018a); For GOF and predictive accuracy (Reichenberger et al., 2019)
Scaled Total Error (STE)	$\frac{\sum abs(P_i - O_i)}{\sum O_i}$	GOF and predictive accuracy
Squared Pearson correlation coefficient (r^2)	$\left(\frac{\sum(O_i - \bar{O})(P_i - \bar{P})}{\sqrt{\sum(O_i - \bar{O})^2}\sqrt{\sum(P_i - \bar{P})^2}}\right)^2$	GOF and predictive accuracy (Reichenberger et al., 2019)
Survival Probability Prediction Error (SPPE)	$\left(\frac{O_{end}}{O_{init}} - \frac{P_{end}}{P_{init}}\right) \cdot 100$	For validation of GUTS models; evaluates model overestimation or underestimation only at the end of the tested exposure profile (EFSA PPR, 2018a)

Abbreviations: Pi: prediction of calculation i; Oi: observation i; end: time of the last data point; init: time of the first data point; n: number of data points; nval: number of data points for validation; ncal: number of data points for calibration; :mean of observations; : mean of predictions; UL and LL: upper limit and lower limit of the observations or predictions.

6.3.2.2. Model selection

In most projects, modelers can select between multiple alternative model formulations. Typical examples are the distinction between stochastic death and individual tolerance mechanisms in GUTS or the identification of Physiological Modes of Action (PMoA) in DEB-TKTD models. In these cases, the decision of using either one of the model variants can usually not be made during earlier steps of the modeling cycle because this decision depends on the outcome of the calibration and possibly of the validation. In other words, the decision for a specific model formulation relates to how well the model formulation performs for a specific species and chemical compound under consideration, in addition to biological considerations pointing toward one formulation rather than another one.

When a biologically based decision is not appropriate, quantitative criteria can be used. Evaluation criteria used for model calibration might seem a good start to compare model fits. However, important aspects of model selection can be missed if only such quantitative criteria are used to compare models. Most importantly, one may want to apply a penalty for models with a larger number of parameters. Doing so is an application of the principle of parsimony, also known as Occam's razor: If two models explain the data equally well, the simpler one should by default be preferred. For this reason, common quantitative model selection criteria, like the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) depend linearly on the number of parameters in the model. In addition, BIC also considers the number of data points used in the calibration step.

Both AIC and BIC are based on the likelihood, that is the probability of the data given that a model and set of parameters represent the "true" process to be modeled. For complex models, the likelihood may be intractable, or too resource intensive (i.e., time consuming) to compute. In recent years, the development of Bayesian likelihood-free methods, summarized under the term Approximate Bayesian Computation (ABC) methods, progressed rapidly; see Sunnåker et al. (2013) for a review. While these rather computationally expensive methods are mostly used for calibration of complex models, they also offer an effective method of model selection, for example when combined with a Sequential Monte Carlo algorithm (SMC-ABC); (Toni et al., 2008). In this approach, model calibration and model selection are carried out simultaneously, and the posterior probabilities of alternative models are derived from the frequency at which they repeatedly satisfy a common acceptance criterion, which is in turn defined based on the optimization criterion. The algorithm proposed by Toni et al. (2008) includes an inherent penalty against models with many parameters. However, as in all Bayesian model selection approaches, a penalty for the number of parameters can in addition be included in the prior probabilities of the model. A posterior comparison of pairs of models can be made with the Bayes factor, which can be interpreted as the degree to which one model is favorable over the other.

For nested models, the F-test can be used in a frequentist framework to compare models based on their goodness of fit regarding the same dataset. The F-test assumes that model residuals are normally distributed, and that data points are independent. Under a Bayesian framework, either the Deviance Information Criteria (DIC) or the Widely Applicable Information Criterion (WAIC) is used to perform model comparisons. These criteria apply whether or not the models are nested given they have been fit to an identical dataset.

There is no general consensus among statisticians on how to approach model selection. Also, multi-model inference can be thought of from many different perspectives, including model averaging. Thus, "model selection" can be used somewhat generically, including model comparison and multi-model inference. Indeed, many of the methods inherently consider multiple models (sometimes infinitely many) but are not dedicated to model averaging in the conventional sense (Hooten & Hobbs, 2015). A non-exhaustive overview of commonly used model selection criteria is given in Table 6.4.

Similar to the evaluation of calibration outcome, we recommend that multiple measures are used, and that quantitative measures are combined with plausibility checks and visual methods. If the differences between models are seemingly small, one should consider performing predictions using model averaging method, because models that generate similar results during calibration might generate very different results when used to extrapolate to a different scenario.

Table 6.4: Non-exhaustive overview of commonly used model selection criteria, used to compare models fit on the same dataset.

Name	Calculation	Suggested by / used for
Akaike information criterion (AIC)	$2k - 2\ln(L)$	A within-sample non-Bayesian score for prediction also used for comparisons of nested models with a decreasing number of parameters; related: corrected AIC (for small sample sizes).
Bayesian Information Criterion (BIC, Schwartz, 1978)	$k \ln(n) - 2\ln(L)$	A misleading name, as it is a within-sample non-Bayesian score including a measure of model complexity as a penalty. Good for small sets of well-justified models. Can also be used for model averaging if parameters can easily be counted, priors are vague, and posterior modes are used as point estimates for parameters (Hooten & Hobbs, 2015).
Deviance Information Criterion (DIC)	$D(\bar{\theta}) + 2p_D$ where $D(\theta) = -2\log p(y\|\theta)$ and $p_D = \overline{D(\theta)} - D(\bar{\theta})$	A within-sample quasi-Bayesian score for prediction. Widely used for Bayesian model comparison. It approximates a penalized loss function based on the deviance, with a penalty derived from a cross-validation argument. Valid only when the effective number of parameters in the model is much smaller than the number of independent observations (Plummer, 2008).
Leave-one-out cross-validation LOO, Gelfand (1996)	$\sum_{i=1}^{n} \log p(y_i\|y_{-i})$ Importance sampling LOO criteria formula: $CV_2 \cong -\frac{1}{n}\sum_{i=1}^{n} log \frac{\mathbb{E}_w\left[p(X_i\|w)p(X_i\|w)^{-\beta}\right]}{\mathbb{E}_w[p(X_i\|w)^{-\beta}]}$ n is the posterior sample size; Ew[ ] stands for the posterior average.	A method to estimate the generalization performance of a model, done by training the model on all observations except one then predicting the hold-out observation, and repeating n times. Asymptotically equivalent to AIC and WAIC when maximizing the likelihood. See (Watanabe, 2010) for details.

Widely Applicable Information Criterion (aka. Watanabe AIC; WAIC); (Watanabe 2010)	$$\text{WAIC}(n) \equiv B_t L(n) + \frac{\beta}{n} V(n)$$ with $$V(n) = \sum_{i=1}^{n} \left\{ \mathbb{E}_w \left[\left(log\, p(X_i \mid w) \right)^2 \right] - \mathbb{E}_w [log\, p(X_i \mid w)]^2 \right\}$$ and $$B_t L(n) = -\frac{1}{n} \sum_{i=1}^{n} \log p^*(X_i)$$	Method for estimating pointwise out-of-sample prediction accuracy from a fit Bayesian model using the log-likelihood evaluated at the parameter estimates. Vehtari et al. (2021) for details, Vehtari et al., (2017) for implementation in the R software, and Baudrot & Charles (2018) for use in ERA). WAIC is fully Bayesian as it is using the entire joint posterior distribution. Asymptotically equal to the Bayesian cross-validation. Unlike DIC, WAIC is invariant to parametrization and also works for singular models. Also, asymptotically equal to LOO. Note that (Watanabe, 2013) also mention a widely applicable Bayesian information criterion (WBIC) version of BIC.

Abbreviations: k: number of parameters estimated by the model; L: maximum value of the likelihood function; n: number of observations; q: estimated model parameters D(q): deviance; $\overline{\mathrm{D}(\theta)}$: posterior expectation of the deviance; $\mathrm{D}(\bar{\theta})$: deviance evaluated at the posterior mean of the parameters; p(y|q): likelihood function; yi: data point i, y-i: the data without data point i.

6.3.3. Outlook: Deriving acceptability criteria for specific models

While we suggest that proposing general and universal acceptability criteria and thresholds for MEMs may not be possible, it appears more feasible to select generic acceptability criteria for situations where the context can be more clearly defined. For example, a primary producer TKTD model used at Tier 2C of environmental risk assessment, with laboratory data for calibration and validation, has a well-defined modeling purpose and an area of application, with observations of the same type, following typical patterns. In such setting, a systematic approach to investigate potential acceptability criteria for a specific model type might be as follows:

- Collect a number of example datasets for different substances, such that enough data are available to ensure proper calibration and validation for each substance. Within these datasets, different assignments of the data to calibration or validation could be tested, resulting in a number of different calibrations and validations for each substance. For the sake of this exercise, it may not be most important to get reasonable assignments of the data to calibration or validation, nor to obtain good model predictions, but to yield a considerable number of example datasets with good and bad model calibrations and validations.

- Various acceptability criteria could then be applied to these different sets. The criteria can be compared to the visual assessment of several experts (loosely, or in a more formalized Expert Knowledge Elicitation procedure), to identify the criterion (or, more likely, the combination of criteria) that seems the best in line with the expert assessment. Note that even acceptability criteria derived in this way should not be considered in isolation, but in combination with a visual assessment.

6.4. Conclusions

To assess correspondence between model output and observations, a wide variety of visual, qualitative and quantitative measures are applicable. This diversity of approaches provides the opportunity to select the most appropriate measures for the research and/or regulatory question to be addressed and the MEM developed and/or applied to answer it. Additionally, the diversity of approaches also provides an excellent foundation to conduct a multi-angled model evaluation where multiple visual checks and quantitative measures are employed complementary to each other. The results of visual and quantitative measures should not be interpreted only in isolation or in reference to complementary measures, but also in context of the problem formulation, model use, and the biological system being modeled.

Setting definitive acceptability criteria for MEMs in general is probably not realistic or even convenient given the variety of models and their applications as well as different types and tiers of risk assessments. Nevertheless, in the future, acceptability criteria could be set for specific models and applications such as in regulatory settings, where more stringent regulatory guidance can assist in increasing familiarity with MEMs and their evaluation and can help build confidence among regulators evaluating these models. Nevertheless, in some model applications expert judgement is still likely to play a role in assessment of the quality how well the model outputs match data, and to aid such expert judgement, the present chapter gives an overview of the methods available and how they can be used. In conclusion, our considerations highlight that model calibration and validation are not tasks with binary outcomes (e.g., the model is valid vs. invalid). We can only estimate degrees of validity, based on a combination of assessment tools. This is inherent to biologically based models, which have to deal with the natural variability and the inherent uncertainty associated with both observable and unobservable quantities.

6.5. Bibliography Chapter 6

Ashauer, R., Kuhl, R., Zimmer, E., & Junghans, M. (2020). Effect modeling quantifies the difference between the toxicity of average pesticide concentrations and time-variable exposures from water quality monitoring. *Environmental Toxicology and Chemistry*, *39*(11), 2158-2168. https://doi.org/10.1002/etc.4838

Baudrot, V., & Charles, S. (2018). Recommendations to address uncertainties in environmental risk assessment using toxicokinetics-toxicodynamics models. *bioRxiv*, 356469. https://doi.org/10.1101/356469

Becher, M. A., Grimm, V., Thorbek, P., Horn, J., Kennedy, P. J., & Osborne, J. L. (2014). BEEHAVE: a systems model of honey bee colony dynamics and foraging to explore multifactorial causes of colony failure. *Journal of Applied Ecology*, *51*(2), 470-482. https://doi.org/10.1111/1365-2664.12222

Bennett, N. D., Croke, B. F. W., Guariso, G., Guillaume, J. H. A., Hamilton, S. H., Jakeman, A. J., Marsili-Libelli, S., Newham, L. T. H., Norton, J. P., Perrin, C., Pierce, S. A., Robson, B., Seppelt, R., Voinov, A. A., Fath, B. D., & Andreassian, V. (2013). Characterising performance of environmental models. *Environmental Modelling & Software*, *40*, 1-20. https://doi.org/10.1016/j.envsoft.2012.09.011

Breckling, B., Müller, F., Reuter, H., Hölker, F., & Fränzle, O. (2005). Emergent properties in individual-based ecological models—introducing case studies in an ecosystem research context. *Ecological Modelling*, *186*(4), 376-388. https://doi.org/10.1016/j.ecolmodel.2005.02.008

EFSA (European Food Safety Authority). (2023c). Supplementary information to the revised guidance on the risk assessment of plant protection products on bees (*Apis mellifera*, Bombus spp. and solitary bees). *EFSA Supporting Publications*, *20*(5), 7982E. https://doi.org/10.2903/sp.efsa.2023.EN-7982

EFSA PPR (Panel on Plant Protection Products & their Residues). (2014). Scientific opinion on good modelling practice in the context of mechanistic effect models for risk assessment of plant protection products. *EFSA Journal*, *12*(3), 3589. https://doi.org/10.2903/j.efsa.2014.3589

EFSA PPR (EFSA Panel on Plant Protection Products and their Residues). (2018a). Scientific opinion on the state of the art of toxicokinetic/toxicodynamic (TKTD) effect models for regulatory risk assessment of pesticides for aquatic organisms. *EFSA Journal*, *16*(8), 5377-5565. https://doi.org/10.2903/j.efsa.2018.5377

Farmer, N. A., Baker, K., Zeddies, D. G., Denes, S. L., Noren, D. P., Garrison, L. P., Machernis, A., Fougères, E. M., & Zykov, M. (2018). Population consequences of disturbance by offshore oil and gas activity for endangered sperm whales (*Physeter macrocephalus*). *Biological Conservation*, *227*, 189-204. https://doi.org/10.1016/j.biocon.2018.09.006

FOCUS (2006). Guidance Document on Estimating Persistence and Degradation Kinetics from Environmental Fate Studies on Pesticides in EU Registration. *EC Document Reference Sanco/10058/2005 version 2.0*, 434. https://esdac.jrc.ec.europa.eu/public_path/projects_data/focus/dk/docs/finalreportFOCDegKinetics.pdf

Gallagher, C. A., Chudzinska, M., Larsen-Gray, A., Pollock, C. J., Sells, S. N., White, P. J. C., & Berger, U. (2021). From theory to practice in pattern-oriented modelling: identifying and using empirical patterns in predictive models. *Biological Reviews*, *96*(5), 1868-1888. https://doi.org/10.1111/brv.12729

Gelman, A. (2004). Exploratory data analysis for complex models. *Journal of Computational and Graphical Statistics*, *13*(4), 755-779. https://doi.org/10.1198/106186004X11435

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian Data Analysis* (1st ed.). Chapman and Hall. https://doi.org/10.1201/9780429258411

Grimm, V., Frank, K., Jeltsch, F., Brandl, R., Uchmanski, J., & Wissel, C. (1996). Pattern-oriented modelling in population ecology. *Science of the Total Environment*, *183*(1-2), 151-166. https://doi.org/10.1016/0048-9697(95)04966-5

Grimm, V., Johnston, A. S. A., Thulke, H. H., Forbes, V. E., & Thorbek, P. (2020). Three questions to ask before using model outputs for decision support. *Nature Communications*, *11*(1), 3, Article 4959. https://doi.org/10.1038/s41467-020-17785-2

Grimm, V., & Railsback, S. F. (2005). *Individual-based Modeling and Ecology* (STU – Student edition ed.). Princeton University Press. http://www.jstor.org/stable/j.ctt5hhnk8

Grimm, V., & Railsback, S. F. (2012). Pattern-oriented modelling: a 'multi-scope' for predictive systems ecology. *Philos Trans R Soc Lond B Biol Sci*, *367*(1586), 298-310. https://doi.org/10.1098/rstb.2011.0180

Hamilton, S. H., Pollino, C. A., Stratford, D. S., Fu, B., & Jakeman, A. J. (2022). Fit-for-purpose environmental modeling: Targeting the intersection of usability, reliability and feasibility. *Environmental Modelling & Software*, *148*, 105278. https://doi.org/10.1016/j.envsoft.2021.105278

Hansul, S., Fettweis, A., Smolders, E., & De Schamphelaere, K. (2021). Interactive metal mixture toxicity to *Daphnia magna* populations as an emergent property in a dynamic energy budget individual-based model. *Environ Toxicol Chem*, *40*(11), 3034-3048. https://doi.org/10.1002/etc.5176

Hooten, M. B., & Hobbs, N. T. (2015). A guide to Bayesian model selection for ecologists. *Ecological Monographs*, *85*(1), 3-28. https://doi.org/10.1890/14-0661.1

Klanjscek, T., Caswell, H., Neubert, M. G., & Nisbet, R. M. (2006). Integrating dynamic energy budgets into matrix population models. *Ecological Modelling*, *196*(3), 407-420. https://doi.org/10.1016/j.ecolmodel.2006.02.023

Lee, S., & Lim, H. (2019). Review of statistical methods for survival analysis using genomic data. *Genomics Inform*, *17*(4), e41. https://doi.org/10.5808/GI.2019.17.4.e41

Lombardo, A., Franco, A., Pivato, A., & Barausse, A. (2015). Food web modeling of a river ecosystem for risk assessment of down-the-drain chemicals: A case study with AQUATOX. *Science of the Total Environment*, *508*, 214-227. https://doi.org/10.1016/j.scitotenv.2014.11.038

Mentré, F., & Escolano, S. (2006). Prediction discrepancies for the evaluation of nonlinear mixed-effects models. *Journal of Pharmacokinetics and Pharmacodynamics*, *33*(3), 345-367. https://doi.org/10.1007/s10928-005-0016-4

Mounier, F., Pecquerie, L., Lobry, J., Sardi, A. E., Labadie, P., Budzinski, H., & Loizeau, V. (2020). Dietary bioaccumulation of persistent organic pollutants in the common sole Solea solea in the context of global change. Part 1: Revisiting parameterisation and calibration of a DEB model to consider inter-individual variability in experimental and natural conditions. *Ecological Modelling*, *433*, 109224. https://doi.org/10.1016/j.ecolmodel.2020.109224

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics*, *9*(3), 523-539. https://doi.org/10.1093/biostatistics/kxm049

Preuss, T. G., Hammers-Wirtz, M., Hommen, U., Rubach, M. N., & Ratte, H. T. (2009). Development and validation of an individual based *Daphnia magna* population model: The influence of crowding on population dynamics. *Ecological Modelling*, *220*(3), 310-329. https://doi.org/10.1016/j.ecolmodel.2008.09.018

Purucker, S. T., Welsh, C. J. E., Stewart, R. N., & Starzec, P. (2007). Use of habitat-contamination spatial correlation to determine when to perform a spatially explicit ecological risk assessment. *Ecological Modelling*, *204*(1), 180-192. https://doi.org/10.1016/j.ecolmodel.2006.12.032

Railsback, S. F., & Harvey, B. C. (2002). Analysis of habitat-selection rules using an individual-based model. *Ecology*, *83*(7), 1817-1830. https://doi.org/10.1890/0012-9658(2002)083[1817:AOHSRU]2.0.CO;2

Railsback, S. F., & Johnson, M. D. (2011). Pattern-oriented modeling of bird foraging and pest control in coffee farms. *Ecological Modelling*, *222*(18), 3305-3319. https://doi.org/10.1016/j.ecolmodel.2011.07.009

Reichenberger, S., Sur, R., Kley, C., Sittig, S., & Multsch, S. (2019). Recalibration and cross-validation of pesticide trapping equations for vegetative filter strips (VFS) using additional experimental data. *Sci Total Environ*, *647*, 534-550. https://doi.org/10.1016/j.scitotenv.2018.07.429

Rinke, K., & Petzoldt, T. (2003). Modelling the effects of temperature and food on individual growth and reproduction of Daphnia and their consequences on the population level. *Limnologica*, *33*(4), 293-304. https://doi.org/10.1016/S0075-9511(03)80024-5

Rossmanith, E., Höntsch, K., Blaum, N., & Jeltsch, F. (2007). Reproductive success and nestling diet in the Lesser Spotted Woodpecker (Picoides minor): the early bird gets the caterpillar. *Journal of Ornithology*, *148*(3), 323-332. https://doi.org/10.1007/s10336-007-0134-4

Schmolke, A., Abi-Akar, F., Roy, C., Galic, N., & Hinarejos, S. (2020). Simulating honey bee large-scale colony feeding studies using the BEEHAVE Model-Part I: Model validation. *Environmental Toxicology and Chemistry*, *39*(11), 2269-2285. https://doi.org/10.1002/etc.4839

Schmolke, A., Kapo, K. E., Rueda-Cediel, P., Thorbek, P., Brain, R., & Forbes, V. (2017). Developing population models: A systematic approach for pesticide risk assessment using herbaceous plants as an example. *Science of the Total Environment*, *599*, 1929-1938. https://doi.org/10.1016/j.scitotenv.2017.05.116

Schmolke, A., Thorbek, P., DeAngelis, D. L., & Grimm, V. (2010). Ecological models supporting environmental decision making: A strategy for the future. *Trends in Ecology & Evolution*, *25*(8), 479-486. https://doi.org/10.1016/j.tree.2010.05.001

Scholten, H., & Van Der Tol, M. W. M. (1994). Towards a Metrics for Simulation Model Validation. In J. Grasman & G. van Straten (Eds.), *Predictability and Nonlinear Modelling in Natural Sciences and Economics* (pp. 398-410). Springer Netherlands. https://doi.org/10.1007/978-94-011-0962-8_33

Sunnåker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., & Dessimoz, C. (2013). Approximate Bayesian Computation. *PLOS Computational Biology*, *9*(1), e1002803. https://doi.org/10.1371/journal.pcbi.1002803

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., & Stumpf, M. P. H. (2008). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface*, *6*(31), 187-202. https://doi.org/10.1098/rsif.2008.0172

van der Vaart, E., Beaumont, M. A., Johnston, A. S. A., & Sibly, R. M. (2015). Calibration and evaluation of individual-based models using Approximate Bayesian Computation. *Ecological Modelling*, *312*, 182-190. https://doi.org/10.1016/j.ecolmodel.2015.05.020

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413-1432. https://doi.org/10.1007/s11222-016-9696-4

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved Rˆ for assessing convergence of MCMC (with Discussion). *Bayesian Analysis*, *16*(2). https://doi.org/10.1214/20-ba1221

Watanabe, S. (2010). Asymptotic equivalence of Bayes Cross Validation and widely applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research, 11*, 3571-3594

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, *14*(Mar), 867—897.

Wiegand, T., Jeltsch, F., Hanski, I., & Grimm, V. (2003). Using pattern-oriented modeling for revealing hidden information: a key for reconciling ecological theory and application. *Oikos*, *100*(2), 209-222. https://doi.org/10.1034/j.1600-0706.2003.12027.x

7. Sensitivity and uncertainty of mechanistic effect models

Simon Hansul, Peter Vermeiren*, Eugenia Chaideftou, Michail Gioutlakis, Udo Hommen, Mira Kattwinkel, Judith Klein, Stefan Reichenberger, Johannes Witt, Sandrine Charles*

** These authors contributed equally*

7.1. Introduction and definitions

In the context of Mechanistic Effect Models (MEMs), uncertainty analysis (UA) investigates the uncertainty in model outputs (including sources, magnitude, and direction), particularly those outputs that are subsequently used in risk assessment decision making. One important aspect of such uncertainty analysis is covered by sensitivity analysis (SA), which focuses on evaluating the impact of uncertainties in model assumptions and parameter values on the outcome. Hence, SA assists in identifying sources of uncertainty. Sensitivity analysis is in a certain sense a backward-oriented method in the step of model analyses, as it traces the uncertainty back to the source in the model and helps to rank model input (parameters, initial conditions, driving data, etc.) with respect to their influence on the model output. Thereby, it aims at identifying non-influential model inputs that can be fixed when calibrating the model or making predictions.

Uncertainty should be considered for all scientific assessments done under the EFSA remit, and EFSA has outlined corresponding guidance for the requirements in different assessments (EFSA SC, 2018a). Practical definitions regarding UA and SA are given in this guidance (EFSA SC, 2018a), specifically: "uncertainty is […] referring to all types of limitations in available knowledge that affect the range and probability of possible answers to an assessment question." Uncertainty analysis is defined as characterizing "the uncertainty in model prediction, without identifying which assumptions [i.e., input factors, sources of uncertainty] are primarily responsible." Sensitivity analysis is defined as "the study of how the uncertainty in the output of a model (numerical or otherwise) can be apportioned to different sources of uncertainty in the model input" (Saltelli, 2002). Sensitivity analysis is also defined as "a tool to understand how model outputs respond to changes in parameter values" and other assumptions (EFSA PPR, 2018a).

Sensitivity analysis, and UA as a whole, are often focused on model parameterization (i.e., the values assigned to model parameters). For example, the EFSA guidance (EFSA SC, 2018a) outlines methods for uncertainty analysis ranging from qualitative to quantitative and how to report them, yet takes a predominantly statistical approach to UA mainly assuming the model formulation is correct and the main concern is parameter uncertainty. While parameterization certainly is an important aspect to consider in UA, many other aspects can influence model uncertainty. Examples of aspects influencing model uncertainty include the initial conditions, environmental scenarios and driving data (e.g., quality of temperature records), and model assumptions and structure (switch on/off processes, different equations describing a process). Therefore, Saltelli and colleagues use the more general term "input factors" for the items that are varied in a UA and SA (Saltelli et al., 2019). In line with this broader view, we argue for a multifaceted approach to SA and UA that considers all inputs to a model (including inputs into the formulation of a model's concept and structure) as potential factors contributing to uncertainty.

A second major question to consider in UA is which model outputs should be the focus of the UA. Within a risk assessment context, model outputs that relate to assessment endpoints are the main interest, as these are the model outputs on which decisions are made. Hence, uncertainty is ideally minimized in model outputs that relate to these endpoints. Nevertheless, refinements to a model following UA should consider the overall coherence and realism of the model. Similar to the evaluation of the calibration and validation of MEMs discussed in Chapter 6, we promote a multi-angled, multi-faceted approach to UA and SA that considers patterns in range of model outputs and state variables, and utilizes a range of complementary quantitative to qualitative methodologies to evaluate sources, as well as the magnitude and direction of uncertainty.

There are two main approaches to UA in the context of MEMs: (i) inverse uncertainty quantification, which is akin to pattern-oriented validation (and indeed one of the strongest arguments for pattern-oriented modeling is that by using several patterns simultaneously, structural model uncertainty is reduced). In this approach, uncertainty is estimated by comparing model predictions to data. However, a more common approach is (ii) forward uncertainty propagation, which is also the approach EFSA has focused on (EFSA SC, 2018a). In forward uncertainty propagation, a Monte Carlo style approach is generally taken to see how uncertainty in the model propagates to the results. Here, it is worthwhile to reflect on what contributes to model uncertainty, specifically the contribution of natural variability versus model uncertainty. In EFSA's uncertainty guidance document, they suggest treating natural variability and uncertainty together, as both contribute to the total observed uncertainty in model outputs. However, it should be noted that while natural variability cannot be reduced, uncertainty can be reduced with the addition of more knowledge and data. Therefore, if an uncertainty analysis of a model indicates an expected high level of uncertainty, it may be worth exploring whether this is driven by natural variability or model uncertainty. One way to separate the two is via an SA where only the model itself (i.e., parameter values, model structure) is varied. For example, natural variability is separated from uncertainty in Monte Carlo approaches. When uncertainty stems from the model, it should be considered whether more data, information, or studies can be obtained to reduce this uncertainty. In this context, it should be noted that natural variability is not necessarily identical to experimental variability. The

variability observed in experiments under controlled conditions does not necessarily reflect the natural variability and can be reduced to a certain extent.

Considering that both natural variability and model uncertainty contributes to the total observed uncertainty in model outputs, modeling results and the derived risk assessments cannot take deterministic model predictions alone, as there is always uncertainty stemming from different sources that needs to be considered in the decision making. It is important to note that uncertainty is not only associated with mechanistic effect modeling but is an unavoidable property of all methods for risk assessment (EFSA SC, 2018b) and life in general. Nevertheless, MEMs offer a method to explicitly quantify uncertainty and variability, beyond often qualitative assessments and expert opinion, which should be a main impetus to increase the uptake of MEMs in regulatory risk assessment.

A final consideration in UA, next to the scope of the analysis and the endpoints to focus on, as discussed in previous paragraphs, relates to the magnitude of the uncertainty and how this relates to the model's behavior. An input factor (e.g., parameter value) is influential if small variations of it significantly affect the model outputs (Ciric et al., 2012). An important model input factor is both influential and uncertain: If it is influential but well known (for instance, density of water, or the application rate of a pesticide), it will not cause substantial uncertainty in the model output. For this reason, it is important to use realistic probability distributions for model inputs included in an uncertainty analysis. If the input factor is uncertain but not influential (e.g., diffusion coefficient of a substance in water when simulating a fast-flowing waterbody), it will not contribute substantially to uncertainty either and can be fixed to a default value (within a physically reasonable range) for model calibration and predictive simulations.

In practice, both UA and SA can be quite demanding in terms of both information requirements and computational effort. For complex models with many parameters and input variables, it may be particularly hard to describe reasonable probability distributions for all the parameters involved. Moreover, many simulations with large samples from the possible distributions of parameters and input variables are needed to quantify the resulting uncertainty in the model outcomes. This can become difficult to conduct for models with long running times such as some IBM implementations and might lead to the impression that an elaborate UA and SA is not possible for more complex models. However, how can we trust modeling results if we (i) would not be able to quantify the uncertainty in parameters and inputs, and (ii) would refrain from analyzing the impact of our imperfect knowledge on the outcome? For complex models in particular, and situations or questions in regulatory ERA, it is hardly possible to estimate the (joint) impact of the different sources of uncertainty just by expert judgment. Moreover, a rigorous UA and SA may lead to the subjective conclusion that the results of MEMs are too uncertain and too sensitive to be used in risk assessment. However, they only make the sources of uncertainty, their impact on the outcomes and the influential input factors explicit; a step that is often just ignored in other methods for risk assessment, although required for all EFSA assessments (EFSA SC, 2018a).

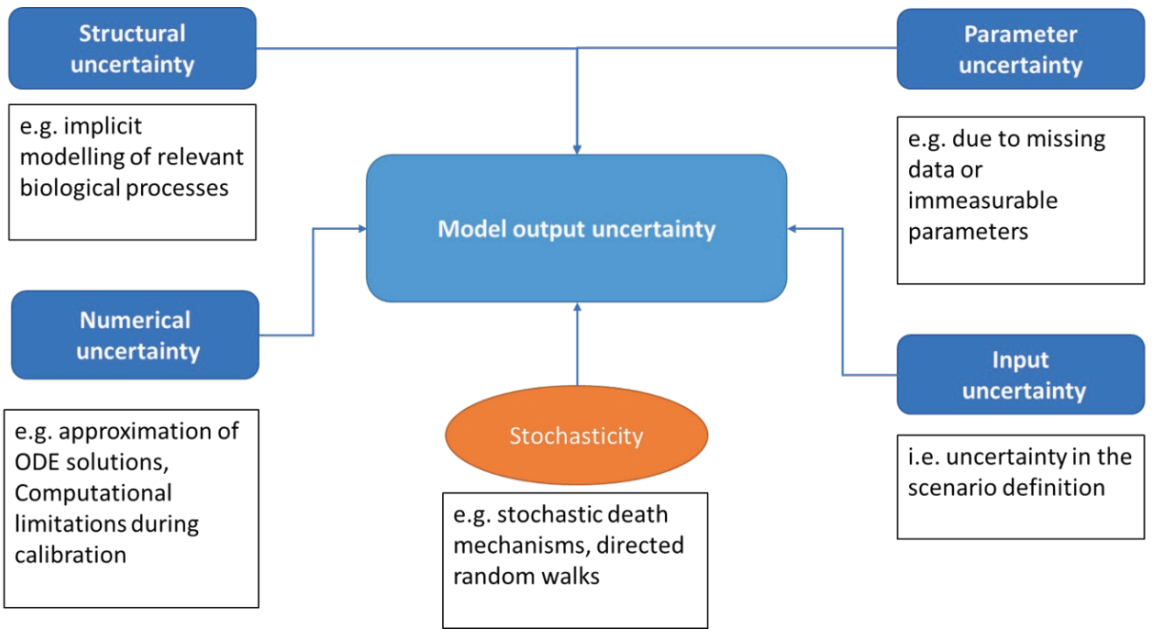


Figure 7.1: Sources of model output uncertainty and examples of their causes. Stochasticity is conceptually different from uncertainty as it may be part of an accurate representation of the modeled system but contributes to model output uncertainty.

7.2. Sources of uncertainty in model outputs

Several factors affect the uncertainty in the model output (Figure 7.1) and this output may be more or less sensitive toward a certain factor. Before we can assess the effect of the different sources on the model output, we need to clearly define the sources and specify a reasonable probability distribution for each input source. Usually, uncertainty is separated from variability, but here natural variability and uncertainty are dealt with together, because for the purpose of regulatory risk assessment the main question is how accurate the model predictions are for a given application.

- **Structural or model uncertainty:** The model structure, for example what level of biological organization is accounted for, or what processes are being used and how they are included, depends on the question to be answered by the model, but also on the available knowledge, data, and other sources of information. There is usually more than one way to describe a system and the associated processes, and it may be impossible to state *a priori* which is preferrable. These different models unavoidably result in different model outputs, and we cannot tell a priori which one is the most "correct" one regarding the asked question. Structural uncertainty is often a large source of uncertainty and hard to quantify.

- **Parameter uncertainty:** This is the source of uncertainty that probably first comes to our mind when thinking about uncertainty and sensitivity in modeling. No model parameters can be known with absolute certainty. This applies for both calibrated parameters and those taken from literature. The uncertainty of calibrated parameters can be assessed during the calibration process, for example, by confidence or credible intervals. The uncertainty of values taken from literature is often harder to quantitatively assess, particularly when numerical values are transferred from one species to another or were determined under different experimental or environmental conditions. To counterbalance this, it may be possible to use various sources or given ranges of measured parameter values as a measure of uncertainty in the parameter values.
- **Input uncertainty:** Most models depend on certain input variables such as temperature, precipitation, or soil conditions. There are two types of uncertainty related to input variables. First, if the model is applied to measured conditions there may be measurement errors, for example, stemming from the measurement devices or the temporal or spatial resolution of the measurements. This can be accounted for by using error distributions or ranges for the input. Second, if the model is extrapolated to future conditions, it is uncertain how the input will be in new situations. Scenario analysis is a way to deal with this. Because there is an unlimited number of possible scenarios one must make sure to cover the relevant range of inputs.
- **Numerical uncertainty:** Numerical solutions, for example, of differential equations, inherently contain small errors. Likewise, numbers are rounded during computer simulation. Numerical uncertainty is usually very small compared with uncertainty from other sources. However, there are some exceptions when numerical issues might become important. First, certain techniques for parameters estimation depend on the convergence of a certain quantity (e.g., optimization algorithms). Hence, it should be checked that convergence of the Monte Carlo Markov Chains (MCMC) has indeed been reached for example with the Gelman-Rubin diagnostic expressed through the potential scale reduction factor (PSRF; Vehtari et al., 2021). Furthermore, strategies for hyperparameter optimization can be explored (i.e., optimization of the parameters of calibration algorithms). Application of such strategies provide advantages with respect to the replicability of modeling processes and harmonization across use cases. It should be acknowledged that computational limitations may stand in the way of a rigorous hyperparameter optimization. As a minimum requirement, we suggest documenting and justifying how hyperparameters are set. This includes defining prior distributions for Bayesian approaches and setting initial values and parameter limits for frequentist approaches (e.g., Nelder-Mead).

If there is no strategy for how hyperparameters are set or calibration is diagnosed, the calibrated values are associated with additional uncertainty of unknown magnitude. Second, stochastic models, like IBMs, try to capture the stochasticity of natural systems and therefore result in different model outputs for each replicate run. Several runs of the model with identical settings need to be simulated and somehow

aggregated as model results. If the number of replicates is too small, the results become unstable and therefore unreliable. Check to ensure the number of replicate runs is large enough; if new (aggregated) simulation results with identical settings provide a different output or a different answer to the question at hand. The numerical approximation of solutions for differential equations can also result in some degree of model output uncertainty. This is especially true if input factors take on extreme values (e.g., during calibration). The exact model output may vary based on the chosen meta-parameters of ordinary differential equation (ODE) solvers (solving algorithm, tolerances). This is sometimes exacerbated for models with a low degree of continuity (e.g., hard switches in model behavior occurring at life stage transitions).

- **Uncertainty due to inherent variability in natural systems (stochasticity):** Variability is an intrinsic characteristic of biological systems and results in randomness in the systems investigated in risk assessment. Uncertainty will result in the model outcomes if the true variability is not described in the model by every aggregation necessary (e.g., when modeling population size instead of single individuals, or when modeling the macroscopic behavior of an individual and not the biochemical processes resulting in this behavior). Some model types (e.g., individual-based models) try to capture this inherent variability of natural systems by introducing a stochastic representation of the processes modeled. Natural variability also manifests in variability among different organisms of the same species or spatial variability of soil characteristics, for example. Typically, variability results in uncertainty in parameter values or initial conditions of a system, which can be quantified during the calibration process (see above).

The first four sources of uncertainty are also called epistemic (reducible) uncertainties, caused by a lack of knowledge. "This refers to uncertainty which is due to lack of information and of understanding of the natural and human systems that encompass us. Theoretically, epistemic uncertainty is reducible with further investigations and empirical research" (Kirchner et al., 2021). The latter one (variability) is also referred to as stochastic (irreducible) uncertainty, or, "uncertainty due to the inherent variability of natural and human systems (also labelled natural, ontological, phenomenological, or aleatory uncertainty). This refers to uncertainty that is in practice irreducible, such as natural variability (e.g., flood events, precipitation events) or chaotic systems (e.g., cloud behavior, social dynamics); (Kirchner et al., 2021). In theory, the uncertainty may reducible, but it may not be practically possible or advantageous to explicitly include the corresponding detail of processes. Other examples related to MEMs in risk assessment are individual variability between organisms (or replicates) or the exact initial state in a population model.

Because of the difficulty of accounting for uncertainty and/or variability, they are often ignored in modeling approaches, which risks providing irrelevant and unreliable model outputs. However, the same applies also for other approaches in risk assessment.

7.3. When to perform uncertainty and sensitivity analyses?

In most cases, both UA and SA should be performed. Once an analyst has performed an uncertainty analysis and is informed of the robustness of the inference, it would appear natural to ascertain with an SA where the uncertainty is coming from (Saltelli et al., 2019). Usually, a sensitivity analysis without uncertainty analysis is not meaningful, because the relative importance of a factor on the model output has a different relevance depending on whether the output has a small or large variance. However, there are cases where a pure SA may be sufficient: For instance, studies to identify the dominant effects on the output for a subsequent model reduction or calibration analysis (Saltelli et al., 2019).

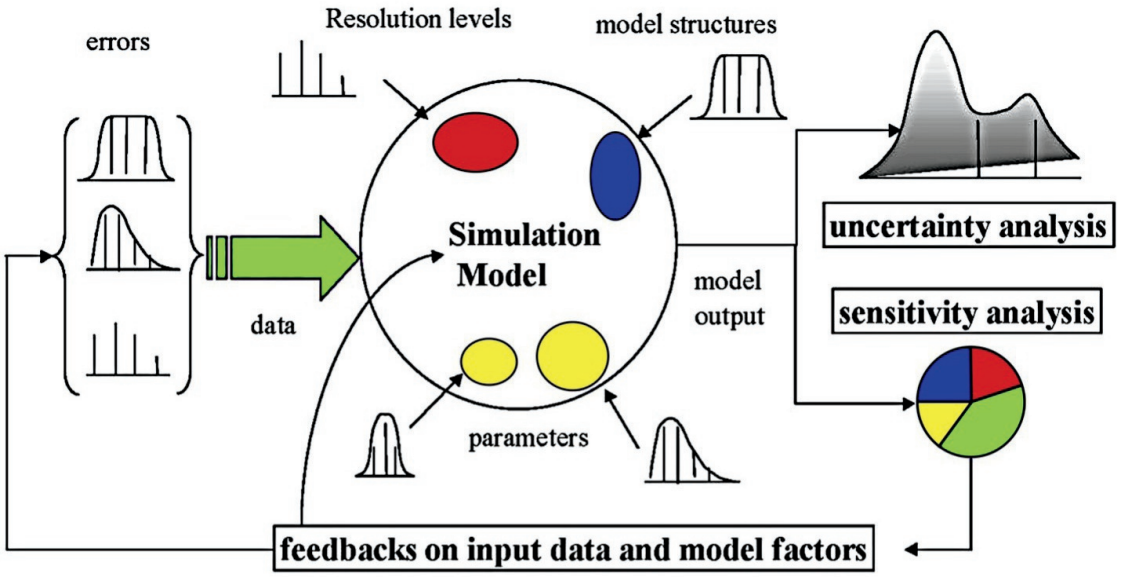


Figure 7.2: Idealized uncertainty and sensitivity analysis. Reproduced from Saltelli et al. (2019), licensed under Creative Commons.

Uncertainty coming from heterogeneous sources is propagated through the model to generate an empirical distribution of the output of interest (gray curve). The uncertainty in the model output, captured, for example, by its variance, is then decomposed according to source, thus producing a sensitivity analysis (Saltelli et al., 2019).

According to (Saltelli et al., 2019), UA should ideally precede SA (Figure 7.2). Before uncertainty can be apportioned it needs to be estimated. They also point out that it is essential to focus the sensitivity analysis on the question addressed by the model rather than more generally on the model. This implies that the probability distributions of the input factors should correspond to the problem at hand. For instance,

exploring the full realistic parameter space of TKTD parameters for a range of species (e.g., a family or order) requires much wider distributions than assessing the effect of the variability of these parameters for a single species. Consequently, varying all input factors in the same way (e.g., by a number of fixed percentages or factors) can lead to the wrong conclusions.

In the modeling cycle (e.g., EFSA PPR [2014]), there are typically three different stages when UA and SA can be applied. The results from the first two stages may be reported in a TRACE document (Grimm et al., 2014), while the third step refers to a specific use case.

1. During model setup: Assume one has set up a complex model with many parameters that are not well known. An "a priori sensitivity analysis" (Beck, 1983) can help the model developer decide which parameters are non-influential and can be fixed during calibration and prediction, and which should be parameterized using additional experimental data or calibration.
2. During model analysis: This analysis still refers to a generic model, without application to a specific chemical. For example, for a fully parameterized and validated population model the uncertainty in the model outcomes (e.g., for the outputs like abundance or biomass) and the sensitivity of these model outcomes toward various ecological parameters and scenario settings is assessed.
3. During model application (model use): For a parameterized model used for predictive purposes in a risk assessment context (for a given scenario, application scheme, organism, and substance), both UA and SA are performed to apportion the uncertainty in the model output to the model input factors.

It is preferable that at least the UA and SA is performed for all model parameters, but it may not be feasible to include all input variables (EFSA PPR, 2014). An initial screening can be used to identify particularly influential parameters and processes in the model before embarking on a full uncertainty and sensitivity analysis. Clearly discuss which input factors were excluded for what reason and include the initial conditions in the SA (EFSA PPR, 2014). Additionally, if there is indication that one of the other sources of uncertainty may be important (e.g., competing theories for model structure) they should also be assessed.

When applying a previously used model to a new problem definition, re-use of the UA and SA results for the same species and compound may be appropriate, given that the probability distributions of all input factors remained valid.

7.3.1. Performing and interpreting uncertainty analysis

The goal of uncertainty analysis is to quantify the uncertainty of model predictions caused by the different sources. There are two steps necessary to analyze model uncertainty. The first one is to describe and quantify the uncertainty related to the different sources. The second step is to propagate the uncertainty from the different sources to the model outputs and assess how this affects the answer to the risk assessment question.

Hence, in the first step we should try to quantify our lack of knowledge related to the different sources of uncertainty. For parameter uncertainty this means defining probability distributions for the parameter values. For calibrated parameters, these can be quantified from confidence or credible intervals. If there is any indication that estimates for parameters are correlated with each other, it is essential to take this correlation into account. Values taken from the literature or derived from expert knowledge can be harder to estimate. Since we cannot assume that these values are certain in the context of the model at hand or in general, it is necessary to construct ranges or distributions of reasonable parameter values, for example by basing them on different (published) experiments or by expert judgment. Note that giving a range for parameter values corresponds to a uniform probability distribution. The model input uncertainty stemming from measurement errors or variability (e.g., weather conditions) can be described similarly by ranges or probability distributions for the values used. This might become demanding for time series data, like temperature or precipitation, for example. To develop model inputs that predict future events, it is necessary to build reasonable scenarios and to try to assess the range of possible values, as well.

Structural uncertainty is difficult to describe and quantify, as there is virtually an unlimited number of possible model structures, often equivalent and/or complementary so that one or the other cannot be preferred. Hence, even though this source of uncertainty is often responsible for a large part of model uncertainty (O'Hagan, 2012), we can only concentrate on some aspects, such as known competing hypotheses or biologically based reasoning. For example, in TKTD modeling this could be stochastic death versus individual tolerance, two versions of the GUTS modeling approach that are often similarly appropriate (Baudrot & Charles, 2018; Gabsi et al., 2019). As mentioned above, numerical uncertainty usually plays only a minor role and is not further discussed here.

In the second step, which consists in propagating the uncertainty toward model outputs, both parameter and input uncertainty can be accounted for similarly. For simple models, it may be possible to assess the effect on the model outcome of these two sources of uncertainty with classical error propagation. For more complex models, the derived probability distributions of parameter values and input variables can be used to run the model with many samples from the possible values and investigate potential effects on model outputs. Thereby, the uncertainty is propagated to the model outcomes. An example using the GUTS-RED-SD model with data on carbendazim toxicity to *Gammarus pulex* illustrates such computational uncertainty propagation procedure (section 8.4). In this example, the joint posterior distribution is approximated by a set of accepted parameter combinations. By repeatedly sampling from this set and evaluating the model for each set, the parameter uncertainty can be mapped to model output uncertainty, with account for correlation between parameters.

Outcomes of different model structures can be compared to assess uncertainty in the second step. Another possibility it to use model averaging or include different (nested) models in an inference process, so that when calibrating model parameters, an additional parameter can be a switch between model structures. Either way, it is difficult to describe and quantify structural uncertainty, which makes it necessary to explain and justify the chosen model structure with enough details in the model description.

In the instance that an uncertainty analysis does show that virtually every outcome is possible with the same probability, we should indeed not fully trust the modeling results and try to improve our knowledge. An honest uncertainty analysis can improve our understanding of the model and the system it describes, as well as build trust and improve acceptance in the model if interpreted correctly.

A crucial point in uncertainty analysis of effect models for regulatory risk assessment of pesticides is to understand *model outputs* in relation to the risk assessment question and the specific protection goals at hand. Therefore, looking at uncertainty bands of raw predictions might not be enough, but we need to evaluate the uncertainty in the outcome of the assessment. For example, often treatment model runs are compared with control runs to decide whether there is only an acceptably small effect present or if recovery is observed after a given time (depending on the question). Here, the same parameter and input uncertainty applies to both treatment and control runs. Even though raw predictions might be very uncertain, one of the predicted effects based on paired runs may not. Thus, these effects should be the presented and interpreted *model output* for the uncertainty analysis.

Regarding the evaluation of an uncertainty analysis, both the modeler and the risk assessor should keep in mind that we will never be absolutely certain about any risk assessment results no matter if derived by modeling, laboratory, or field methods. The question is whether and how the uncertainty is quantified, and how much uncertainty we are willing to accept. For example, an uncertainty analysis may show that in 90% of all samples from uncertain parameter and input values the modeled treatment population recovers within the requested time after the treatment. We might decide to accept this uncertainty or not, but if we do, use the modeling results as evidence in the risk assessment. Another option may be to use the average of all model predictions and apply the assessment factor of the corresponding experimental approach. We should also be aware that with modeling we can evaluate the effect of variability and uncertainty while by experimental approaches, especially of field tests are often limited to single case studies for a specific location and time and can only roughly cover the uncertainty by use of assessment factors.

7.3.1.1. Visualizing distributions

Visualizing distributions is an inherent part of presenting uncertainty. However, it is required in different contexts of model output evaluation. Some examples include:

- A distribution of parameters, such as prior and posterior distributions, that can be visually compared to check if there was a gain of information from fitting the model to data and compared against prior information;
- A distribution of individual attributes, for example, when evaluating population structures;
- A spatial distribution, including the visualization of discrete home ranges from a distribution of individuals.

Most distribution visualizations require some sort of transformation of the data, such as binning (histograms) or even fitting of a distribution (e.g., Kernel Density Estimates, KDE). While most visualization tools provide defaults for performing these transformations, we suggest that these choices should be made consciously by modelers and ideally reported together with convincing arguments. Some technical considerations for the visualization of distributions are given in section 8.5.1.

7.3.2. Performing and interpreting sensitivity analysis

Methods for sensitivity analysis can be broadly divided into local and global methods. Depending on the method, different sensitivity measures are used Table 7.1 and sigma (measure for interaction with other input factors; Table 7.1

7.3.2.1. Local sensitivity analysis

This is a sensitivity analysis around a fixed base scenario (i.e., parameter set and initial conditions). A local SA consists in varying only one input parameter value from one simulation to the next (One At a Time = OAT), and then returning to the base scenario. Hence, all simulations (parameter sets) differ from the base scenario only by one parameter value. Sensitivity is then usually computed as an approximated first partial derivative, or change of the target model output variable divided by change of the value of the varied parameter (ratio of variation [ROV]; Table 7.1). Changes can be absolute or relative to the values of the base scenario; however, in the case of absolute changes the obtained sensitivity coefficients are not dimensionless and cannot be compared between parameters. The term "local" refers to the fact that all derivatives are taken at a single point, known as "baseline" or "nominal value" point, in the hyperspace of the input factors (Saltelli & Annoni, 2010). Examples of a local SA are given in Dubus et al. (2003) or EFSA PPR (2018a).

In the EFSA opinion on GMP (EFSA PPR 2014), OAT is indicated as the simplest sensitivity analysis and aims to identify the most influential parameters. The EFSA GMP opinion mentions two (relatively similar) alternatives for an OAT design (EFSA PPR, 2014; p. 52). The first option is:

1. Set all parameters to reference values (what we called "base scenario" above).
2. Vary one parameter at a time by ±10%.
3. Calculate the sensitivity coefficient (relative change in model output divided by relative change in parameter value, which is to say the first partial derivative); average over the two values (+10% and – 10% variation). Note that averaging might lead to a loss of information as the impact may not be symmetric.
4. Report as table.

The second option is:

1. Set all parameters to reference values.
2. Vary one parameter at a time over a reasonable range of values (e.g., ±2%, ±5%, ±10%, ±20%, ±50%, ±100%, ± 200%).
3. Report as plot of model output versus change in parameter value (e.g., EFSA PPR [2014], Figure 12; EFSA PPR [2018a], Figure 11).

These two options of local SA suggested in the EFSA opinion on GMP (EFSA PPR, 2014) can be used in the model setup phase even when information on input parameter distributions is not available yet.

In practice, however, some parameters will have a narrow probability distribution, and some will have a wide one (due to uncertainty and/or variability). If one varies all parameters in the same way (i.e., by fixed percentages or factors), one may get wrong ideas about the actual importance of each parameter.

Moreover, there are several methodical problems with local SA:

1. The sensitivity measures used in local SA are inadequate for non-linear and non-additive models (Gardner et al., 1981; Saltelli & Annoni, 2010), unless only one specific base parameter set is of interest. Saltelli and Annoni (2010) state: "In principle, local analyses cannot be used for the robustness of model-based inference unless the model is proven to be linear (for the case of first order derivatives) or at least additive (for the case of higher and cross order derivatives). In other words, derivatives are informative at the base point where they are computed, but do not provide for an exploration of the rest of the space of the input factors unless some conditions (such as linearity or additivity) are met in the form of the mathematical model being represented." In other words, the computed sensitivity measures (derivatives) are valid only for the pre-defined base scenario, unless the model is linear. If a model is nonlinear or non-additive, which is probably the majority of MEMs, the sensitivity of a given parameter determined with a one-at-a-time (OAT) analysis depends on the base value of this parameter.
2. Local OAT designs cannot effectively explore a multidimensional space (Saltelli et al., 2019). The more parameters there are, the worse the problem gets ("curse of dimensionality").
3. Effect of parameter interactions cannot be explored, because this would require moving more than one factor from the base scenario. In the simplest case, the value of one parameter affects the sensitivity of another parameter. (Saltelli & Annoni, 2010).

In summary, the use of local SA can only be justified for models that are linear, additive, and have few parameters, or if only the base scenario is of interest.

7.3.2.2. Global Sensitivity Analysis

Global Sensitivity Analysis (GSA) can be defined as follows: "The 'global' term denotes that GSA studies output variability when all input factors vary globally, within their validity domain defined by probability distribution functions (PDFs), as opposed to locally, (one at a time), i.e., around an arbitrary range from a base value. GSA allows for simultaneous estimation of the factors' individual importance (first order effects) and interactions (higher order effects)." (Lauvernet & Munoz-Carpena, 2018).

In a GSA, the input factors (parameters, initial conditions, model switches, etc.) are usually sampled from probability distributions, using different sampling strategies. The GSA allows taking parameter interactions and non-linearities into account (EFSA PPR, 2014). EFSA PPR (2018a; p. 39) recommends a global sensitivity analysis to "quantify the sensitivity of the model for changes in one parameter also in interaction with the other parameters."

The simplest evaluation of global SA is to create partial scatter plots of model output over the individual parameters. Such plots already provide information on the relevance of the single parameters (and their uncertainty) to model output and the shape of the relation (see e.g., EFSA PPR [2014]; Figure 12).

Examples for GSA methods are:

- Monte Carlo simulation with subsequent multiple linear regression (MLR); (Saltelli et al., 2008)
- Extended Fourier Amplitude Sensitivity Testing (eFAST); (Munoz-Carpena et al., 2010; Saltelli et al., 1999)
- Sobol' method (Gatel et al., 2020)
- Morris or Elementary Effects method (Campolongo et al., 2007; Ciric et al., 2012; Morris, 1991; Munoz-Carpena et al., 2010)

Monte Carlo simulations with subsequent MLR can help detect non-linearity and non-additivity by means of the coefficient of determination (R2). If R2 is low, the model is non-linear and/or non-additive. However, in this case the obtained standardized regression coefficients are not useful for ranking input factors (Saltelli & Annoni, 2010).

Both the Sobol' method and eFAST are variance-based GSA methods that use sophisticated quasi-random sampling strategies. The total variance of the target model output variable is decomposed in parts attributed to each input factor's direct effects or to factor interactions (Lauvernet & Munoz-Carpena, 2018), and Sobol' sensitivity indices (Si and STi); (Sobol, 1993) are calculated. The first-order sensitivity index (Si) for each input factor Xi is defined as the fraction of the output variance associated with the direct effect of that factor (Table 7.1). In contrast, the total sensitivity index (STi) is calculated as the fraction of output variance associated with factor Xi *and* its interactions with other input factors. In case of a purely additive model – a model without interactions between inputs – both the sum of Si and the sum of STi are equal to 1. Otherwise, the sum of Si is <1 and the sum of STi is > 1. Variance-based measures of sensitivity

are attractive because they measure sensitivity across the whole input space, they can deal with nonlinear responses, and they can measure the effect of interactions in non-additive systems (Saltelli & Annoni, 2010). The downside of Sobol' sensitivity indices is that they are difficult to interpret when there is dependence between input factors (Broto et al., 2020). To overcome this issue, an alternative sensitivity index called Shapley effect has recently been proposed (Broto et al., 2020; Owen, 2014).

Results of a variance-based GSA are usually presented as column charts of the sensitivity indices Si and STi (Gatel et al., 2020).

Due to the dense sampling scheme used, eFAST and Sobol' also allow us to compute probability distribution functions of the target output variables (uncertainty analysis); (Ciric et al., 2012; Lauvernet & Munoz-Carpena, 2018; Muñoz-Carpena et al., 2010). However, it is precisely the sampling density required by true variance-based methods such as eFAST or Sobol' that can be prohibitive for computationally expensive models with a large number of model parameters. For such models, Saltelli and Campolongo propose the use of the Morris Method (or Elementary Effects Method) as an effective sensitivity screening method (Campolongo et al., 2007; Saltelli et al., 2005). The computational cost of a Morris sensitivity screening is not higher than the cost of a local SA (Saltelli & Annoni, 2010). The Morris method can also do a one-at-a-time (OAT) variation but without a base scenario. Several trajectories with the length (number of parameters + 1) are laid across the parameter space. This way the whole parameter space can be explored with little computational cost. Several algorithms for trajectory sampling exist to optimize the exploration of the parameter space (Khare et al., 2015). The sensitivity measures μ^* (measure for total sensitivity; Table 7.1) and sigma (measure for interaction with other input factors; Table 7.1) computed by the Morris methods were found to be a good approximation to variance-based sensitivity measures (Campolongo et al., 2007; Lauvernet & Munoz-Carpena, 2018). A Morris sensitivity screening can also be performed for coupled models.

Table 7.1: Non-exhaustive overview of commonly used measures for sensitivity analyses.

Abbreviations: y = value of target output variable Y; x = value of input factor X; subscript i = index of varied input factor; subscript 0 refers to base scenario.

Method	Name of measure	Calculation	Suggested by / used by	Explanation / advantages / disadvantages
Local SA	Absolute ratio of variation	$absROV = \frac{\Delta y}{\Delta x_i}$	(Brylinsky, 1972)	Advantage: when computed for different values of Δxi ROV can at least detect non-linearity; disadvantages: only valid for base scenario; works mainly for linear and additive models; unit depends on input factor
Local SA	Relative ratio of variation	$relROV = \frac{\Delta y}{\Delta x_i} * \frac{x_{i,0}}{y_0}$	(Brylinsky, 1972; Dubus et al., 2003)	Advantage: when computed for different values of Δxi ROV can at least detect non-linearity; disadvantages: only valid for base scenario; works mainly for linear and additive models
Local SA	Absolute first partial derivative	$\frac{\partial y}{\partial x_i}$	(Brylinsky, 1972; Gardner et al., 1981; Turányi, 1990)	Disadvantages: only valid for base scenario; only works for linear and additive models); unit depends on input factor
Local SA	Sigma-normalized first partial derivative	$\frac{\partial y}{\partial x_i} * \frac{s_{x_i}}{s_y}$ with sxi = standard deviation of xi sy = standard deviation of y	(IPCC, 1999, 2000; Saltelli et al., 2008)	Disadvantages: only valid for base scenario; only works for linear and additive models

Monte Carlo + subsequent MLR	Standardized regression coefficient β*	$\beta^* = \beta \frac{s_{x_i}}{s_y}$ with β = regression coefficient sxi = standard deviation of xi sy = standard deviation of y	(Saltelli et al., 2008)	Advantage: method gives an indication of the degree of non-linearity or non-additivity; disadvantage: method not useful for ranking input factors if model is non-linear or non-additive
Morris	Mean of elementary effects (μ)	$\mu_i = \frac{1}{r}\sum_{j=1}^{r} EE_i^j$ with EEi = Elementary Effect $EE_i = \frac{y(x_1, \ldots, x_{i-1}, x_i + \Delta, x_{i+1}, \ldots, x_k) - y(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_k)}{\Delta}$ Δ = jump length in the quantile space of the input factor; usually set to = p/(2(p-1)) for even p. p = number of levels in the quantile space from which samples are drawn (often 4, 6 or 8). r = number of trajectories through the transformed parameter space; usually r >= 10.	(Morris, 1991)	μ is a measure for the total sensitivity of an input factor; advantage: explores whole parameter space with low computational costs; disadvantage: positive and negative elementary effects cancel each other out
Morris	Mean of absolute elementary effects (μ*)	$\mu_i^* = \frac{1}{r}\sum_{j=1}^{r} \lvert EE_i^j \rvert$	(Campolongo et al., 2007)	μ* is a measure for the total sensitivity of an input factor; advantage: good approximation of STi, low computational costs; disadvantage: not fully quantitative

Morris	Standard deviation of elementary effects (σ)	$\sigma_i = \sqrt{\frac{1}{r}\sum_{j=1}^{r}\left(EE_i^j - \mu_i\right)^2}$	(Morris, 1991)	σ is a measure for the interaction of factor i with other input factors (higher-order interactions; advantage: good approximation of (STi – Si); low computational costs; disadvantage: not fully quantitative
Variance-based GSA	First-order sensitivity index Si	$S_i = \frac{V[E(Y\|X_i)]}{V(Y)}$ with V(Y) = unconditional variance of Y, obtained when all factors X_i are allowed to vary. E(Y|Xi) = mean of model output variable Y when factor Xi is fixed. V[E(Y|Xi)] = expected reduction in variance that would be obtained if Xi were fixed.	(Pearson, 1905; Sobol, 1993)	Si gives the effect of factor Xi by itself (without interactions); advantage: suitable measure to identify direct sensitivity; disadvantage: needs a lot of samples to compute it
Variance-based GSA	Total sensitivity index STi	$S_{Ti} = \frac{E[V(Y\|X \sim i)]}{V(Y)}$ with E[V(Y|X ~i)] = expected variance that would be left if all factors but Xi were fixed. V(Y) = unconditional variance of Y, obtained when all factors X_i are allowed to vary.	(Homma & Saltelli, 1996)	STi gives the total effect of a factor, including all its interactions with other factors; advantage: suitable measure to identify non-influential factors; disadvantage: needs a lot of samples to compute it

For instance, Bach and colleagues performed a Morris sensitivity screening for the FOCUSsw Step 3 package. The results of a Morris sensitivity screening are usually presented as scatterplots of the total effect measure μ* versus the interaction measure sigma (Bach et al., 2017).

For models of intermediate complexity and computational cost it may be advantageous to perform a sensitivity analysis in two steps: In the first step, perform a Morris sensitivity screening with all model parameters to filter out non-influential input factors. In the second step, a true variance-based GSA is conducted for the input factors that have not been filtered out.

7.3.2.3. Other aspects of SA

The results of a sensitivity analysis always depend on the chosen scenario as well as selected endpoints (i.e., target output variables). Thus, selecting a different endpoint can (and usually will) lead to a different sensitivity of the model to a given parameter. Hence, it is important to focus on the assessment endpoint, or, what model output is driving the risk assessment. Depending on the model and assessment, it might be useful to conduct UA and SA for different endpoints.

Sensitivity analysis can improve the understanding of the model structure, and help identify critical flaws. For instance, if one parameter is excessively influential or has very strong interactions with other parameters, these are warning signs that the model structure may be flawed.

Sensitivity analysis can also help identify areas of the parameter space where the model is not stable and crashes or yields nonsensical results.

7.3.2.4. Outlook

Global Sensitivity Analysis is more complex but considerably more useful than local SA. However, while GSA methods are common in exposure modeling, they have not been sufficiently adopted in effect and ecological modeling. One reason might be that exposure models are often simpler and mostly deterministic and effect models usually stochastic and more complex. This increases the challenge of conducting (runtime) and analyzing a GSA: for complex population models, computation time will often be an issue. However, in this case a Morris sensitivity screening could be performed, which is not more expensive than OAT. For local SA, modelers can refer to the examples in EFSA PPR Panel (EFSA PPR, 2014, 2018a), but for GSA there are only scatter plots in the EFSA PPR (2014). Consequently, more reference publications of GSA, such as Ciric et al. (2012), are needed in effect and ecological modeling to spread the use of GSA methods in these fields. Regarding the use of SA, some questions have been compiled which could help regulatory authorities during evaluation (section 7.3.4).

7.3.3. A practical example

A useful description of an uncertainty analysis for TKTD models can be found in the scientific opinion on the state of the art of toxicokinetic/toxicodynamic (TKTD) effect models for regulatory risk assessment of pesticides for aquatic organisms (EFSA PPR, 2018a). In Appendix 8.5.2, we exemplify the procedure of uncertainty and sensitivity analysis using a simple toy model of population dynamics. It is crucial to note that, for demonstration purposes, the process is simplified here by assuming uniform probability distributions of parameters and no correlations between parameters.

7.3.4. A set of questions regarding sensitivity analysis for MEMs

For future assessments of MEMs, the models might be evaluated at EU level, and national authorities might then evaluate the use of the models in specific cases. In both situations, evaluators will be confronted with sensitivity and uncertainty analyses. The following set of questions intend to support regulatory authorities during model assessments.

- Is there a way to decide whether a sensitivity analysis is needed?
  - → In theory it is always needed. The question is, to what extent? Should it be done for the whole model and all parameters for each application to learn something about unforeseen effects of combinations of scenario settings and parameters? Or, is it ok to do it once for the ecological model and afterwards only for new parameter values in each application? This is a question that is likely subjective and influenced by experience and perhaps the state of mind of the assessor and/or modeler in question. In the absence of current, concrete guidance or best practice consensus within the community, we would recommend looking at other comparable models, and discussing options with an experienced ecological modeler. Developing such consensus and guidance would be an important task to further the uptake and evaluation of MEMs within risk assessment.
- Is there a way to determine if the sensitivity analysis provided has been conducted with the most suitable methodology given the model that is being used?
  - → A global analysis would be preferred over a local one; the latter can only give information about the model behavior at one single point of the parameter space. For complex population models that are often computationally expensive, a Morris sensitivity screening could be performed, which is not more expensive than OAT.
- What are the expected results of sensitivity analysis, and what is the best way that they are presented so that they are clear and comprehensive?
  - → The SA tells us, to which parameters (and inputs) the model output is most sensitive, hence which parameters have the strongest influence on model output variation. If the total uncertainty of model predictions is high, these are the parameters for which more knowledge should be acquired. The presentation of the SA should be given in a way that the results are easy to understand.
- How do the results of sensitivity analysis give us any further information for the evaluation of the use of a specific model?
  - → The results of sensitivity analysis could give us information about what parameters need additional data. In that way, we can aim to reduce the uncertainty around the parameters with the largest influence on model outputs.

7.4. Bibliography Chapter 7

Bach, M., Guerniche, D., Thomas, K., Trapp, M., Kubiak, R., Hommen, U., Klein, M., Reichenberger, S., Pires, J., & Preuß, T. (2017). *Bewertung des Eintrags von Pflanzenschutzmitteln in Oberflächengewässer – Runoff, Erosion und Drainage, GERDA – GEobased Runoff, erosion and Drainage risk Assessment for Germany* (Text Issue).

Baudrot, V., & Charles, S. (2018). Recommendations to address uncertainties in environmental risk assessment using toxicokinetics-toxicodynamics models. *bioRxiv*, 356469. https://doi.org/10.1101/356469

Beck, M. (1983). Sensitivity analysis, calibration, and validation. In G. Orlob (Ed.), *Mathematical modelling of water quality: streams, lakes and reservoirs* (12 ed., pp. 425-467). Wiley.

Broto, B., Bachoc, F., & Depecker, M. (2020). Variance reduction for estimation of shapley effects and adaptation to unknown input distribution. *SIAM/ASA Journal on Uncertainty Quantification*, *8*(2), 693-716. https://doi.org/10.1137/18M1234631

Brylinsky, M. (1972). Steady-State Sensitivity Analysis of Energy Flow in a Marine Ecosystem. In B. Patten (Ed.), *Mathematical modelling of water quality: streams, lakes and reservoirs* (Vol. Volume II, pp. 591). Academic Press.

Campolongo, F., Cariboni, J., & Saltelli, A. (2007). An effective screening design for sensitivity analysis of large models. *Environmental Modelling & Software*, *22*(10), 1509-1518. https://doi.org/10.1016/j.envsoft.2006.10.004

Ciric, C., Ciffroy, P., & Charles, S. (2012). Use of sensitivity analysis to identify influential and non-influential parameters within an aquatic ecosystem model. *Ecological Modelling*, *246*, 119-130. https://doi.org/10.1016/j.ecolmodel.2012.06.024

Dubus, I. G., Brown, C. D., & Beulke, S. (2003). Sensitivity analyses for four pesticide leaching models. *Pest Manag Sci*, *59*(9), 962-982. https://doi.org/10.1002/ps.723

EFSA PPR (Panel on Plant Protection Products & their Residues). (2014). Scientific opinion on good modelling practice in the context of mechanistic effect models for risk assessment of plant protection products. *EFSA Journal*, *12*(3), 3589. https://doi.org/10.2903/j.efsa.2014.3589

EFSA PPR (EFSA Panel on Plant Protection Products and their Residues). (2018a). Scientific opinion on the state of the art of toxicokinetic/toxicodynamic (TKTD) effect models for regulatory risk assessment of pesticides for aquatic organisms. *EFSA Journal*, *16*(8), 5377-5565. https://doi.org/10.2903/j.efsa.2018.5377

EFSA SC (2018a). Guidance on uncertainty analysis in scientific assessments. *EFSA Journal*, *16*(1). https://doi.org/10.2903/j.efsa.2018.5123

EFSA SC (2018b). The principles and methods behind EFSA's Guidance on Uncertainty Analysis in Scientific Assessment. *EFSA Journal*, *16*(1), e05122. https://doi.org/10.2903/j.efsa.2018.5122

Gabsi, F., Solga, A., Bruns, E., Leake, C., & Preuss, T. G. (2019). Short-term to long-term extrapolation of lethal effects of an herbicide on the marine mysid shrimp Americamysis Bahia by use of the General Unified Threshold Model of Survival (GUTS). *Integrated Environmental Assessment and Management*, *15*(1), 29-39. https://doi.org/10.1002/ieam.4092

Gardner, R. H., Oneill, R. V., Mankin, J. B., & Carney, J. H. (1981). A comparison of sensitivity analysis and error analysis based on a stream ecosystem model. *Ecological Modelling*, *12*(3), 173-190. https://doi.org/10.1016/0304-3800(81)90056-9

Gatel, L., Lauvernet, C., Carluer, N., Weill, S., & Paniconi, C. (2020). Sobol global sensitivity analysis of a coupled surface/subsurface water flow and reactive solute transfer model on a real hillslope. *Water*, *12*(1), 121. https://www.mdpi.com/2073-4441/12/1/121

Grimm, V., Augusiak, J., Focks, A., Frank, B. M., Gabsi, F., Johnston, A. S. A., Liu, C., Martin, B. T., Meli, M., Radchuk, V., Thorbek, P., & Railsback, S. F. (2014). Towards better modelling and decision support: Documenting model development, testing, and analysis using TRACE. *Ecological Modelling*, *280*, 129-139. https://doi.org/10.1016/j.ecolmodel.2014.01.018

Homma, T., & Saltelli, A. (1996). Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering & System Safety*, *52*(1), 1-17. https://doi.org/10.1016/0951-8320(96)00002-6

IPCC. (1999). *IPCC expert meetings on good practice guidance and uncertainty management in national greenhouse gas inventories* (Background papers, Issue). http://www.ipcc-nggip.iges.or.jp/public/gp/gpg-bgp.htm

IPCC. (2000). *Good Practice Guidance and Uncertainty Management in National Greenhouse Gas Inventories*. https://www.ipcc.ch/publication/good-practice-guidance-and-uncertainty-management-in-national-greenhouse-gas-inventories/

Khare, Y., Muñoz-Carpena, R., Rooney, R. W., & Martinez, C. (2015). A multi-criteria trajectory-based parameter sampling strategy for the screening method of elementary effects. *Environmental Modelling & Software*, *64*. https://doi.org/10.1016/j.envsoft.2014.11.013

Kirchner, M., Mitter, H., Schneider, U. A., Sommer, M., Falkner, K., & Schmid, E. (2021). Uncertainty concepts for integrated modeling – Review and application for identifying uncertainties and uncertainty propagation pathways. *Environmental Modelling & Software*, *135*, 104905. https://doi.org/10.1016/j.envsoft.2020.104905

Lauvernet, C., & Munoz-Carpena, R. (2018). Shallow water table effects on water, sediment, and pesticide transport in vegetative filter strips – Part 2: Model coupling, application, factor importance, and uncertainty [Article]. *Hydrology and Earth System Sciences*, *22*(1), 71-87. https://doi.org/10.5194/hess-22-71-2018

Morris, M. D. (1991). Factorial sampling plans for preliminary computational experiments. *Technometrics*, *33*(2), 161-174. https://doi.org/10.1080/00401706.1991.10484804

Muñoz-Carpena, R., Fox, G., & Sabbagh, G. (2010). Parameter importance and uncertainty in predicting runoff pesticide reduction with filter strips. *Journal of environmental quality*, *39*, 630-641. https://doi.org/10.2134/jeq2009.0300

O'Hagan, A. (2012). Probabilistic uncertainty specification: Overview, elaboration techniques and their application to a mechanistic model of carbon flux. *Environmental Modelling & Software*, *36*, 35-48. https://doi.org/10.1016/j.envsoft.2011.03.003

Owen, A. B. (2014). Sobol' indices and shapley value. *SIAM/ASA Journal on Uncertainty Quantification*, *2*(1), 245-251. https://doi.org/10.1137/130936233

Pearson, K. (1905). *On the general theory of skew correlation and non-linear regression*. Dulau and co. https://openlibrary.org/books/OL6555066M/On_the_general_theory_of_skew_correlation_and_non-linear_regression.

Saltelli, A. (2002). Sensitivity analysis for importance assessment. *Risk Analysis*, *22*(3), 579-590. https://doi.org/10.1111/0272-4332.00040

Saltelli, A., Aleksankina, K., Becker, W., Fennell, P., Ferretti, F., Holst, N., Li, S. S., & Wu, Q. L. (2019). Why so many published sensitivity analyses are false: A systematic review of sensitivity analysis practices. *Environmental Modelling & Software*, *114*, 29-39. https://doi.org/10.1016/j.envsoft.2019.01.012

Saltelli, A., & Annoni, P. (2010). How to avoid a perfunctory sensitivity analysis. *Environmental Modelling & Software*, *25*(12), 1508-1517. https://doi.org/10.1016/j.envsoft.2010.04.012

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., & Tarantola, S. (2008). *Global Sensitivity Analysis: The Primer*. Wiley. http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470059974.html

Saltelli, A., Ratto, M., Tarantola, S., & Campolongo, F. (2005). Sensitivity analysis for chemical models. *Chemical Reviews*, *105*(7), 2811-2828. https://doi.org/10.1021/cr040659d

Saltelli, A., Tarantola, S., & Chan, K. P. S. (1999). A quantitative model-independent method for global sensitivity analysis of model output. *Technometrics*, *41*(1), 39-56. https://doi.org/10.2307/1270993

Sobol, I. M. (1993). Sensitivity Estimates for Nonlinear Mathematical Models. *Mathematical Modelling and Computational Experiments, 4*, 407-414.

Turányi, T. (1990). Sensitivity analysis of complex kinetic systems. Tools and applications. *Journal of Mathematical Chemistry*, *5*(3), 203-248. https://doi.org/10.1007/BF01166355

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved Rˆ for assessing convergence of MCMC (with Discussion). *Bayesian Analysis*, *16*(2). https://doi.org/10.1214/20-ba1221

8. Appendix

8.1. Appendices Chapter 3: Practical examples for environmental scenario documentation

8.1.1. Example for voles

This example illustrated the environmental scenario development for a hypothetical case study investigating the application of a selective herbicide applied once a year in a grassland habitat.

Table 8.1: An example of the environmental scenario for a hypothetical case study (voles).

	1. Risk assessment problem definition	
a	What is the regulatory context in which the model will be used?	The model's purpose is to simulate field population dynamics of the common vole (*Microtus arvalis*) for the specific purpose of identifying population-level responses to impacts of plant protection products (PPP). The model is used as a tool which can complement to the environmental risk assessment (ERA) of PPP following EU Regulation (EC) No. 1107/2009. Specifically, the model is used in a weight of evidence approach to refine the small herbivorous mammal "vole" scenario (EFSA PPR, 2009; Scenario 74 in Appendix A) that failed in the Tier-1 ERA. GAP: Selective herbicide applied once a year in grassland; application timing: March – September; application rate: 80g a.s./ha.
b	Which species or species group is going to be modeled, and why?	The modeled species is the common vole (*M. arvalis*), which is a common herbivore living in most parts of central Europe. It is one of the best studied small mammals in Europe. The species is common in agricultural land, particularly in grassland habitats. In the risk assessment of PPP, the common vole is included in the EFSA guidance document on birds and mammals risk assessment (EFSA PPR, 2009) as a representative species for the "small herbivorous mammal" generic focal species.

c	What is the protection goal (SPG attributes: ecological entity, attribute, magnitude, temporal and spatial scale)?	The model directly addresses the protection goal of no long-term repercussions on population abundance and diversity (also after year-on-year applications of the PPP). This actual protection goal is defined in the EFSA guidance document for the risk assessment for birds and mammals (EFSA PPR, 2009). Entity: Population. Attribute: Abundance or population size, reproduction. Magnitude: No long-term repercussions on population abundance (if population level effects are observed in the modeling, it is evaluated if recovery occurs within one year). Temporal scale: Long-term (the model investigates a 10-year treatment period of continuous year-on-year application. Spatial scale: "Field to landscape" scale.
d	Which model outputs are required to answer the risk assessment question(s), including the required performance criteria?	The model provides output that allows linking organism-level effects as measured in standard toxicity tests to the target organizational level of populations in the ERA (EFSA, 2009). The output required to address the specific risk assessment question is population abundance. In this case study, this is expressed in terms of population density (number of individuals per hectare), plus the analysis of exposure modification factors as optional information. This example handles two endpoints regarding the long-term risk for vole populations, specifically (1) the treatment-related relevant effects on population density due to the dietary exposure (i.e., due to the consumption of contaminated food) of vole individuals to the PPP and (2) the potential and time until complete recovery of the populations, in case affected, after the simulated year-on-year PPP applications cease. The model output in this example is evaluated by comparing the daily mean vole population densities in treatment simulations (where a PPP is applied) to those in control simulations (where no PPP application takes place).
e	How has the appropriate model been selected?	The mechanistic individual-based model (IBM), *eVole* (RIFCON, 2022)[8], was chosen for the population modeling because it considers ecological processes such as reproduction, survival, and spatial behavior, which are relevant for common vole population dynamics. *eVole* is also a spatially explicit model, which considers the dynamic spatial arrangement of vole home ranges in different habitats (which may or may not be contaminated) and the underlying dietary exposure when PPP residues are present on the vole food, to predict the resulting effects on each of those vole individuals. In short, *eVole* is an adequately complex tool and provides all relevant outputs for the assessment of potential adverse effects of PPP on vole population abundance and dynamics.
f	What is the domain of applicability of the model?	The standard formal model and its parametrization is applicable to central European conditions.

8 RIFCON 2022. TRACE for *eVole* (version 3.1). RIFCON GmbH, unpublished report.

g	How will exposure and effects be integrated?	The model includes an ecotoxicological module to simulate exposure and effects. The exposure of individuals is determined by their behavior (i.e., their dynamic daily home ranges). A PPP is applied on certain parts of the landscape. The dietary exposure profile for each individual vole is driven by substance residues on vegetation, or vole food, which each vole individual obtains from its dynamic daily home range. Long-term exposure is calculated based on the suggested good agricultural practice and using the EFSA parameters (EFSA PPR, 2009) for vole dietary uptake of PPPs, while assuming a 100% monocotyledon diet as a worst-case vole diet in this example. The individual-level effects that result from long-term (TWA21) exposure to the PPP are calculated for each vole individual using the dose-response relationships derived from toxicological studies. The toxicological effects interfere with the model processes which constitute the life-history of a vole (i.e., survival, reproduction, development).
	2. The environmental scenario layout	
a	Is it necessary to define crop-specific environmental scenarios?	Yes, the landscape spatial structure (shape, size, types and distribution of habitats), the vegetation characteristics of its constituent habitats (e.g., specific crop types), as well as the spatial distribution of PPP residues on vegetation (e.g., in-crop and off-crop, heterogeneous contamination in-crop) are decisive for the vole population dynamics and exposure and effect in *eVole*. In fact, vegetation characteristics in addition to social criteria decide the location and size of voles' daily home ranges. These daily home ranges, where the voles feed, influence exposure and effect dynamics. For example, when PPP residues on sprayed vegetation are heterogeneously distributed in the model landscape, the crop type(s) and their spatio-temporal dynamics are of major importance.
b	Is it necessary to define country – or zone-specific environmental scenarios?	Because the model and its parametrization are applicable (mainly) to central European conditions, it allows a direct application for the central zone. If relevant, country specific conditions (e.g., landscape structure, specific agricultural practice) might either be covered by the conservative (realistic worst-case) definition of the ES or addressed in additional uncertainty simulations.

c	Which biotic factors are potentially relevant for the environmental scenarios used in the risk assessment (e.g., interspecific variability in intrinsic sensitivity, food availability, food quality, dispersal, reproduction, predation, competition, habitat quality)?	Landscape-specific information on food and shelter availability as well as the spatial distribution of habitat types are of major importance. They are the main drivers of the voles' home range dynamics, which have a large impact on the voles' ecology. In the model this is implemented as a gridded model landscape that can comprise different habitat types, each defined by dynamics of vegetation height and cover. The vegetation dynamics can be realistic or generic worst-cases depending on the envisaged environmental (landscape) scenario. The characteristics of each landscape cell serve as proxies for food resource availability (vegetation height) and shelter (vegetation cover), which in turn influence the vole individual's habitat preference and the associated home range dynamics in the model. Other biological or ecological factors that are important for representing the voles' behavior are considered as standard ecological parameters in the model; thus, these factors are not included in the application specific environmental scenario. For example, the parameterization of the ecological model regarding reproduction and natural survival uses values derived from published literature with preference for field studies in the central European zone. Those parameters are standard and therefore not part of the application specific environmental scenario. Similarly, vole movement and behavior, including dispersal, is not provided explicitly as external input from the ecological scenario. It arises in the model from the interaction of each vole individual with its environment on daily basis.
d	Which abiotic factors are potentially relevant for the environmental scenarios used in the risk assessment (e.g., weather, temperature, pH, oxygen, hydrology, light, habitat)?	No abiotic factors need to be explicitly included in the environmental scenario. Abiotic factors such as weather, temperature, or hydrology are implicitly included in the defined vegetation time series. Accordingly, different abiotic conditions could be represented by altering dynamics in vegetation height or cover. Similarly, the impact of such abiotic factors on vole individuals is implicitly included in the standard parametrization of the ecological model for the CZ, as explained for the biotic factors. They are not provided as direct input in the context of the ES.
e	Which spatial scale and structure is chosen, and is this relevant to the problem definition?	The ES represents a conservative (realistic worst-case) landscape of 25 ha. It consists of two grassland fields where the PPP is applied, surrounded by an off-field grassland margin that is not treated. The in-field grassland covers 90% of the landscape and the untreated field margin 10%. This in-field/off-field ratio represents a realistic worst-case situation for central Europe, including a large proportion of treated area. The spatial extent (total area of the modeling landscape) is chosen to be the area in which voles can establish a viable population size (while still have a reasonable computational demand). Hence, the spatial scale is adequate to investigate the specific protection goal, i.e., potential long-term repercussions on population abundance after applications of PPPs.

f	Are there biotic or abiotic components that need to be standardized between the exposure and the ecological scenario? If so, are the exposure and the ecological scenario consistent with each other (for example, do the data used to define the exposure scenario and the ecological scenario originate from the same country or zone)? If the exposure and ecological scenario are not consistent, has it been demonstrated that the environmental scenario is conservative?	There are no components that need to be synchronized between the exposure and the ecological scenario. The vegetation time series is representative for grassland in the CZ. Accordingly, the exposure resulting from the application window specified in the GAP is consistent with the environmental scenario and a representative selection of application dates was chosen. The exposure and effect calculations are both built-in and are synchronized by concept. In treatment simulations, the model receives as input the PPP application pattern (rate and frequency) in space and time. The model then calculates for each vole a long-term dietary dose (21-day TWA). Using the calculated exposure, each vole will face an effect according to a dose-response relationship, which is also input to the model for each relevant effect. Furthermore, additional simulations (with changes in the vegetation characteristics and the proportion of treated or untreated area) were conducted to investigate whether the ES is conservative (see below).
g	How are the agronomic practices considered for the risk assessment? How has the timing of the application within the GAP been chosen (e.g., was the worst-case application day chosen or was a number of time points for application screened)? What are other agronomic practices that should be taken into account? How are agronomic practices and applications spread into the landscape (e.g., do all fields are treated identically or are there difference in agronomic practice included)?	The agronomic practices are represented by the definition of the landscape settings (i.e., spatial composition of the landscape, and the temporal vegetation dynamics of the different habitat types). Particularly agronomic interventions (such as mowing, harvesting, plowing, fertilizing, or wilting due to herbicide use) are directly represented by the dynamics of vegetation height and cover. Additionally, PPP applications as application rates in space and time are input to the model. Several application time points were chosen, and each time point was simulated in a separate exposure scenario. Overall, the simulated exposure scenarios covered the entire possible application time window (March–September) specified in the GAP. Altogether eight application dates were defined (at the first of each of the relevant months including October). This also covers the main annual breeding period of the vole. There are no further agronomic practices that have to be taken into account. Both habitat types (i.e., in-field and off-field grassland) were treated identically except of the pesticide application event that occurs only in the in-field grassland. The pesticide is applied once a year for a 10-year treatment period. This treatment period is preceded by a 10-year warm-up period (pre-treatment) and followed by a 10-year post-treatment period (where no PPP is further applied).

	3. Assessment and analysis of the most relevant environmental scenario components	
a	Has the influence and relevance (combination of influence and/or impact and variability or uncertainty) of each environmental scenario component (5.2) been assessed (either by expert-opinion, literature research, model simulation, or a combination of methods)?	Yes. A realistic worst-case environmental scenario has been chosen based on expert-opinion. The conservativeness of this environmental scenario and the uncertainty associated with relevant scenario components has been investigated by additional simulations altering single components of the chosen environmental scenario, including vegetation characteristics, landscape settings, application dates. Different **application timings** were investigated to identify the most adverse application time points. Less favorable **vegetation conditions** (in terms of reduced vegetation height) were analyzed to investigate whether this changes the vulnerability of the vole populations. For the specified environmental scenario of voles in grassland the **vegetation height** (proxy for food availability) will have an impact on the overall vole density. Because it cannot be ruled out that this might have an impact on the resilience of a vole population after a PPP application, the uncertainty related to this parameter was investigated by additional simulations. In contrast, the **vegetation cover** (proxy for shelter) in (managed) grasslands can assumed to be high (usually between 80%–100%). Realistic values of cover are always far beyond the threshold where voles would abandon a landscape cell in the model. Hence, particular consideration is not given to vegetation cover in the conducted environmental scenario uncertainty analyses. Regarding the **landscape structure**, usually a larger proportion of treated in-field habitat is assumed to be more conservative as more voles will be exposed over the entire landscape. However, a landscape that consists of 100% treated habitat would be worst-case but unrealistic. Accordingly, in the baseline environmental scenario, a conservatively low but still realistic percentage of 10% untreated off-field habitat is assumed. Nevertheless, the unrealistic but worst-case scenario of 100% treated area has been exemplified by an additional investigatory simulation. Several application time points were chosen to cover the entire possible application time window (March– September) specified in the GAP. This also covers the main annual breeding period of the vole. Moreover, in addition to the nominal application rates, increased rates using exposure multiplication factors (EMFs) of two to five were included as is recommended in the EFSA opinion on good modeling practice in the context of PPP risk assessment (EFSA PPR, 2014). The increased rates can be used to illustrate a safety margin for the proposed use pattern. In case obvious adverse effects arise at the higher EMFs, which in contrast did not show in the nominal application rate simulations (EMF 1). The higher EMFs could also serve in this case as a positive control to gain „confidence that the model used is able to demonstrate negative effects should they occur“ (EFSA PPR, 2014).
b	Has the natural range of each environmental scenario component relevant for the risk assessment been considered?	A realistic worst-case scenario was chosen as baseline environmental scenario based on expert judgement and literature. Different scenario components were varied in additional simulations to capture their variability.

	4. Selection of components for the environmental scenarios, and a check for representativeness and consistency	
a	Has a set of environmental scenarios been defined?	A realistic worst-case scenario has been defined as a baseline environmental scenario. Based on this baseline environmental scenario, additional simulations were performed to assess the uncertainty related with the different scenario components. The baseline scenario together with the additional simulations serve as the set of environmental scenarios.
b	Do the defined environmental scenarios provide a range of environmental conditions from favorable to unfavorable concerning the resilience of the modeled system?	The defined environmental scenario provides a realistic worst-case environmental scenario. In additional simulations for the environmental scenario uncertainty analysis, the vegetation time series has been changed to simulate less favorable food conditions. Also in an additional simulation, an unrealistically large proportion of 100% treated in-field area was simulated to exemplify the extreme assumption of an entirely exposed vole population.
c	Have most uncertain and important environmental scenario components been covered either in their natural or reasonable ranges or have conservative values been chosen?	The baseline environmental scenario is outlined as a realistic worst-case scenario. In additional scenario uncertainty analysis simulations, more conservative specifications of relevant scenario components (e.g., the vegetation height time series, in-field/off-field ratio) were analyzed. For the timing of applications, a regular grid of application dates is analyzed considering the entire application period of the PPP. The PPP application rate is varied in a range from the nominal to a five-fold increased application rate (EMFs) to investigate the margin of safety.
d	Have most relevant exposure conditions and exposure pathways been assessed and considered in the environmental scenario development?	Yes, the considered route of exposure of common voles to the PPP is dietary, which is to say via the ingestion of contaminated plant material from treated areas. This follows the standard risk assessment approach defined in EFSA PPR (2009). From that guidance document, equations for calculations of theoretical dietary exposure of common voles were also obtained and implemented in the model, as well as default parameter values for daily energy requirements (*DEE*), food energy content (*FE*), the residue unit dose (*RUD*). The proportion of food taken from treated fields (*PT*) emerges for the vole home-range dynamics in the model landscape. Mean body weights for adult vole males, adult vole females, and vole juveniles were defined using literature, for use in the exposure calculations in the model as the single default value given in the guidance document does not differentiate between sexes or life stages. To estimate the residues on vegetation each day that follows the day of application, first-order kinetics was considered. Therefore, the average disappearance time for 50% of the considered active substance's foliar residues on treated vegetation (DT50) was used. In this example, as a conservative assumption, a default DT50 of 10 was used. For each day and for each vole, the model estimates the day's exposure (daily dietary dose, DDD) in addition to the 21-day time-weighted average exposure (TWA21) of the DDDt. Those TWA21 doses as such estimate the average daily long-term exposure to the PPP (in line with EFSA PPR, 2009). This estimation of long-term doses is considered relevant for the chronic effects that are measured in the considered toxicity tests and implemented in the simulations.

e	For spatially explicit models: Do the chosen environmental scenarios result in realistic (spatial) behavior?	Initial control simulations with the ecological model (i.e., no exposure or effect included) and the environmental scenario landscape were conducted to demonstrate a realistic vole behavior in the model and ensure that they allow the establishment of a long-term stable vole population.
f	Has an estimation of conservatism of the environmental scenarios in ecological and exposure dimensions been provided?	Yes. In terms of exposure, we assumed a conservative diet of 100% monocots and a rather large proportion of treated area (90%) in the landscape. (In addition exemplarily a more conservative but unrealistic proportion of 100% treated area [i.e., no off-crop] was analyzed in additional investigatory simulations). For the vegetation characteristics, an average time series for food availability was chosen to provide habitat for a reasonably large vole population. Such an average population size facilitates detecting the significant deviation of an endpoint from the normal operating range (NOR) due to a PPP application. In contrast, a low population size is usually characterized by a larger variability (due to demographic stochasticity), which would result in a wider NOR that in turn would make it more difficult to detect an PPP effect. Nevertheless, additional simulations were conducted to demonstrate the result of more adverse food conditions. Moreover, a range of relevant application dates and application rates (exposure multiplication factors) were analyzed to find the most adverse application timings and to explore the margin of safety.
g	Are the environmental scenarios representative for the risk assessment under consideration (e.g., are the spatio-temporal scales of the environmental scenarios in line with the problem definition; does the conservatism of the environmental scenarios match with the conservatism of the problem definition)?	Yes, the baseline environmental scenario represents a realistic worst-case situation for the vole population considering the GAP and the SPG. The spatial and temporal scale allows to investigate the effect of year-on-year applications of the PPP on a (under untreated conditions) stable vole population. The setting of the model analysis enables the investigation of potentially adverse effects of GAP suggested application patterns on the population level, whether these would occur and if so, how long do they persist.

8.1.2. Example for honey bees

This example gives a hypothetical environmental scenario as input for BEEHAVE. The context is a regulatory risk assessment for the use of a PPP in apple cultivation in France.

Table 8.2: An example of an environmental scenario for a hypothetical case study (BEEHAVE).

	1. Risk assessment problem definition	
a	What is the regulatory context in which the model will be used?	Risk Assessment of a Plant Protection Product use in apple cultivation in a European Member State (France).
b	Which species or species group is going to be modeled, and why?	Honey bee. The honey bee is a focal species in regulatory risk assessment of pesticides. Beyond this, landscape scenarios are key for beekeeping management purposes (EFSA, 2023a; EFSA PPR, 2013; EFSA SC, 2021)
c	What is the protection goal (ecological entity, attribute, magnitude, temporal and spatial scale)?	Negligible effects on colony strength for honey bees, corresponding to a value of 10% as the maximum permitted level of honey bee colony size reduction (EFSA, 2023a) Entity: Honey bee colony Attribute: Colony strength (honey bee population) Spatial Scale: Primarily n.a., as referring to individual hive (corresponding to “field edge”) or hive population in the landscape (corresponding to “region”) Temporal Scale: Any time (in assessment period) Magnitude: Maximum of 10% reduction
d	Which model outputs are required to answer the risk assessment question(s), including the required performance criteria?	Primary model output: Bee colony population dynamics. Modeling study output: Assessment Endpoints addressing colony strength as protection goal (above). Colony strength is the sum of all bees and their brood (eggs, larvae, and pupae) in the colony. This, for a representative (set of) environmental scenarios, e.g., including different exposure scenarios (as well as “no exposure”as baseline), different beehive locations in cultivation regions. Supplemental risk assessment information: Safety factors as resulting from exposure (application rate) multiplication factors.
e	How has the appropriate model been selected?	The BEEHAVE model is capable to simulate colony dynamics and is freely available. It has been used as an appropriate model for analyzing honey bee colony dynamics in the context of pesticide risk assessment (EFSA, 2021).

f	What is the domain of applicability of the model?	The BEEHAVE model was designed to model honey bee colony (population) development for artificial and real-world bee forage provision (Becher 20143). Core unit is the honey bee hive. The model includes processes for varroa and virus mortality, as well as beekeeping activities (varroa treatment, feeding, swarming presentation). A foraging module simulates spatiotemporal bee foraging in the landscape. A toxicological module was added (Preuss et al., 2022) that allows the prediction of effects from a given exposure via nectar and pollen using the standard datasets for honey bees. Basically, BEEHAVE can be applied in any region where beekeeping happens as long as data on landscape-scale nectar and pollen provision can be provided, and the validity of the imposed egg-laying rate in BEEHAVE has been confirmed.
g	How will exposure and effects be integrated?	All processes and data are consistently integrated regarding to information (data) representation, their spatial and temporal scales, with respect to the resulting assessment endpoints and their certainty levels. This integration comprises agricultural activities (cropping, PPP use), exposure events and substance fate, as well as bee activity, toxicological and population effects. Consistent landscape development and environmental data (e.g., weather) are used throughout the processes. Currently, the example focuses on spatiotemporally explicit modeling of nectar and pollen provision. First exposure modules on spray drift and on residues in nectar from seed treatment are available. In this modular approach modules for all relevant exposure routes can be added at appropriate spatial and temporal scale that seamlessly fits to the requirements of the toxicological effect module(s).
	2. Definition of the environmental scenario layout	
a	Is it necessary to define crop-specific environmental scenarios?	Basically yes, depending on regional variability of landscape (land use or cover, environmental), crop cultivation and beekeeping conditions. For example, scenarios for apple cultivations will be driven by apple cultivation hot-spots and their climate conditions, whereas scenarios of maize cultivation might be located in regions of high maize growing densities considering crop rotations of possibly high combined pesticide use. However, the answer depends on the level of realism combined with a level of conservatism of the entire risk assessment approach: Whereas worst-case environmental scenarios might not need to be crop specific, more realistic scenarios might need to be, given the real-world variability of driving processes.

b	Is it necessary to define country – or zone-specific environmental scenarios?	Basically yes, depending on regional variability of landscape (land use or cover, environmental), crop cultivation and beekeeping conditions. A similarity analysis on driving factors can be used to clarify this question, in relation to the risk assessment level of reality (below). However, the answer depends on the level of realism combined with a level of conservatism of the entire risk assessment approach: Whereas worst-case scenario might not need to be country – or zone-specific, more realistic scenarios might need to be, given the real-world variability of driving processes. For example, a worst-case, lower-tier scenario for a rape indication might represent a generic (geographically non-specific) situation of an intense rape cultivation world with a minimum on additional bee forage combined with bad bee weather, just for the colony to not die of starvation. For more realistic scenario, higher-tier country and zone-specific rape scenarios need to be derived that reflect more realistic combinations of real-world land use (e.g., rape density and crop rotation), land cover (e.g., hedges, woodland, grassland), landscape structure (i.e., the structural design of land use or cover) and climate conditions.
c	Which biotic factors are relevant for the risk assessment?	Bee forage occurrence in space and time (nectar, pollen, honeydew). Varroa (and virus) infestation can be parameterized in different strengths.
d	Which abiotic factors are relevant for the risk assessment (e.g., weather, temperature, pH, oxygen, hydrology, light, habitat)?	Weather data (foraging hours in daily resolution, for the beehive location). Spatial and temporal explicit nectar and pollen availability.
e	Which spatial scale and structure is chosen, and how is this relevant to the problem definition?	One spatial scale of the scenarios is the “region”: Environmental conditions are assumed homogeneous for the region in which an individual beehive is located. This applies for example for daily temperature and sunshine hours. Another spatial scale is the “patch (field)”: Land use or cover, vegetation types and hence, bee forage occurrence is provided by patches (e.g., apple field, wood or field margin, grassland).
f	Are the exposure and the ecological scenarios consistent with each other? What biotic or abiotic components need to be standardized between the exposure and the ecological scenarios? If the exposure and ecological scenarios are not consistent, has it been demonstrated that the scenarios are conservative?	Yes. This data and information consistency is a key paradigm of our modular landscape modeling approach. Basically, all environmental, landscape management and biotic conditions are consistent that affect the behavior of the different model components (e.g., exposure, effect). For example, agricultural activities, weather, PPP use, bee forage occurrence and beekeeping are consistently modeled. Although base conditions should be consistent and integrated, spatial and temporal resolutions of data and information should fit to the individual requirements: For example, temperature is provided in daily resolution, whereas temperature-dependent bee forage occurrence is provided in monthly resolution.

g	How are the agronomic practices considered for the risk assessment? How has the timing of the application within the GAP been chosen? What are other agronomic practices that should be taken into account? How are agronomic practices and applications spread into the landscape?	An application window of 10 days (11–20. April) was defined according to product label and BBCH stage of apple cultivations in the scenario regions. With this approach, timing of application is independent of weather and crop development. All apple orchards in the landscape are sprayed in randomly in this time window. Beekeeping activities include feeding, varroa treatment, swarm prevention.
	3. Assessment and analysis of the most relevant scenario components	
a	Has the influence and relevance (combination of influence and/ or impact and variability or uncertainty) of each scenario component (5.2) been assessed (either by expert-opinion, literature research, model simulation, or a combination of methods)?	Yes. The definition, scales and resolutions of individual factors and their consistency throughout the modeling approach were evaluated and documented. At each individual development step, the level of certainty was considered with an overall aim to achieve a balanced certainty level across the individual scenario elements. This consideration reflected the definition of the SPGs, hence, the design elements of model assessment endpoints (entity, attribute, spatial – and temporal scales, spatial and temporal resolutions, effect magnitude, representativeness, etc.). Evaluation of certainty aspects contains quantitative estimations, qualitative evaluations, and expert judgement. Certainty aspects are inherent part of basically all scenario development process steps (concepts, approach, data, sub-models, assumptions), of which some are explicitly discussed, and others are communicated by transparent documentation and full access. For example: • Land use or cover: Spatial and temporal representation of apple orchards, grasslands, wood (margins), road (margins) can be considered >90% correct. • Weather data: Uncertainties in daily parameters of agri4cast data (JRC) are considered not relevant for overall BEEHAVE model outputs, hence, study assessment endpoints. • Bee forage thematic representation and phenology: At the scale of land use or cover patches bee forage was represented by a four-step category approach in monthly resolution. This approach reflects bee forage literature and is considered adequate with respect to BEEHAVE model processes and outcome at given scales. At landscape-scales currently no better information is available. Further aspects regarding certainty are given in the study documentation.
b	Has the natural range of each scenario component relevant for the risk assessment been considered?	Yes, with respect to the defined spatial and temporal scales and resolutions of the different landscape entities. Annual variation of cropping and weather can be covered by multi-year simulations. Remaining variability in ultimate scenario definition is explicitly addressed and is part of the analysis process with respect to SPGs: For example, instead of trying to identify a worst-case application timing, we screen the GAP-realistic variability of application timing using a Monte Carlo approach.

	4. Selection of components for the environmental scenarios, and a check for representativeness and consistency	
a	Has a set of scenarios been defined?	Yes. In the given example three scenarios are defined for three major apple cultivation regions covering three different climatic regions in France. However, the scenario "sampling" process was designed to be extendable, which is to say in case representativeness of scenarios is considered insufficient it can be extended.
b	Do the scenarios provide a range of environmental conditions from favorable to unfavorable concerning the resilience of the modeled population?	Such range is given implicitly and with limitations. A key element affecting honey bee colony resilience is in their management activities by the beekeeper, another one in the environmental conditions. The latter obviously affect honey bee forage behavior, which is, however, again linked to beekeepers' practice. Moreover, there is a range of landscape factors that affect the spatiotemporal variability of forage occurrence affecting honey bee vulnerability and resilience. In the example, these factors have not been quantitatively investigated across the focal geographic space that is intended to be represented by the scenarios (France), but are implicitly, to some extent, part of the scenario development approach of the example: The scenario locations, hence, environmental conditions, were intentionally chosen conservative with respect to the purpose of risk assessment for pesticide use. Scenario ranking was done using a vulnerability approach at the entire "statistical population" of potential scenarios across France. Honey bees are a managed species. Thus, their occurrence in the landscape depends on beekeepers' decisions. Here, we assume these decisions to be dominated by regional bee forage occurrence (density) plus a locally appropriate hive place by beekeepers' choices (accessible with car, protected from sight). Vulnerability with respect to colony stress is highly linked to beekeeper management (location, feeding, parasite control). Thus, it is assumed that vulnerability (risk) is mainly driven by exposure due to the target PPP. Because the PPP is used in apples and at the same time apples provide a mass bee forage, we can assume that vulnerability (risk) is related to apple cultivation density. Apple cultivations are well represented in available geodata with high quality and certainty. Thus, scenario site selection, or ranking of regions, is driven by apple cultivation density. It is well known that bee colony development depends on weather conditions (besides from forage, parasites, etc.). Because there is currently no explicit bee colony development model that can be applied at the regulatory scale of an entire country (here, France) to quantify vulnerability explicitly, we address vulnerability dependence from weather conditions by selecting three apple cultivation regions that occur in different climatic regions across the country. This way weather pattern variability is covered and so, potential vulnerability uncertainties to such site conditions are accounted for.

c	Have most uncertain and sensitive scenario components been covered either in their natural or reasonable ranges or have conservative values been chosen?	Like real bee colonies, BEEHAVE is sensitive to weather conditions (temperature, sunshine hours). Weather data were taken from JRC agri4cast database and are considered adequately certain for the given purpose. Honey bees are managed species. Bee health is the key interest of the beekeeper to produce honey. Thus, typical beekeeping activities were considered in the model simulation scenario approach. BEEHAVE is sensitive to bee forage provision (literature, expert judgement). By expert judgement we assume that all relevant bee forage provision units in the scenario landscape have been identified and considered. Their representation at patch-scale with monthly phenology in five categories (0-4, 4=mass forage) is considered adequately certain at landscape-scale. Nectar and pollen provision data is taken from literature. All processes and data are transparently documented. The large size of the individual scenarios of 9 km contributes to the representation of cultivation and natural representation of bee forage in a landscape.
d	Have most relevant exposure conditions and exposure pathways been assessed and considered in the environmental scenario development?	At the current state, the example focuses on spatiotemporally explicit modeling of nectar and pollen provision. First exposure modules on spray drift and on residues in nectar from seed treatment are available. In this modular approach modules for all relevant exposure routes can be added at appropriate spatial and temporal scale that seamlessly fits to the requirements of the toxicological effect module(s) (e.g., guttation, paddles). Available exposure modules will be published together with the entire approach.
e	For spatially explicit models: Do the chosen scenarios result in realistic (dispersal) behavior?	Dispersal behavior is not relevant for the managed species of honey bees (actually, dispersal is suppressed by the beekeeper). However, in case the question of dispersal includes the “dispersal,” which is distribution of foragers during foraging the scenarios provide all data (information) that is currently consumed by BEEHAVE to model foraging. Therefore, real-world landscape composition and structure allows to assess honey bee realistic behavior. Even surface water is included in case the bee foraging scouting model is set to prevent bee to cross large water areas. Because the scenario development approach follows a tiered framework of stepwise increase of the level of realism, future scenarios can represent more details of landscape composition and structure up to 3D representation (e.g., structure that affect bees‘ movement behavior, like woodland).

f	Has an estimation of conservatism of the scenario in ecological and exposure dimensions been provided?	The vulnerability approach for scenario site selection with respect to abiotic factors and exposure conditions is described above. Natural ecological factors play a less important role for honey bee colony development as honey bees are managed species. Presumably, biotic ecological factors like predators play a minor role for honey bee colony dynamics (even if climate change leads to population increase of predators like *Merops apiaster*). Also, competition does not occur for nesting, yet, possibly for forage, given the presence of further honey bee colonies and wild pollinators. For the density of honey bee colonies, beekeepers will likely take care that this is in good balance to local bee forage conditions. Food competition with wild pollinators is not included in the current example but might be considered in more realistic scenario development. However, such rather uncertain scenario factors are suggested to be treated with special care not to overparameterize a scenario and modeling approach. Parasites represent another ecological factor potentially affecting populations. Again, in real-world honey beekeeping, this factor is subject to beekeepers‘ management practice, hence, is realistically not an element in a scenario development.
g	Are the scenarios representative for the risk assessment under consideration (e.g., are the spatio-temporal scales of the scenarios in line with the problem definition; does the conservatism of the scenarios match with the conservatism of the problem definition)?	All aspects of scenario development were designed for the given purpose of honey bee risk assessment using the BEEHAVE model and in the framework of SPGs (and their dimensions and scales).

8.1.3. Example for macrophytes

This example is based on a case study used in the SETAC Modelink workshop (Hommen et al., 2016).

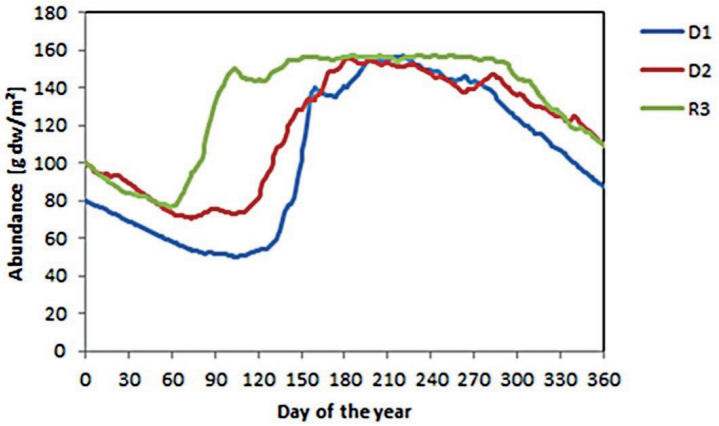
Table 8.3: An example of an environmental scenario for a hypothetical case study (macrophytes).

	1. Risk assessment problem definition	
a	What is the regulatory context in which the model will be used?	Risk assessment for the effects of an herbicide on macrophytes in edge-of-field surface waters in the EU.
b	Which species or species group is going to be modeled, and why?	*Lemna* sp. as Tier 1 macrophyte test species. Focal macrophytes species have not been defined yet. For Tier 2C (refined exposure tests and modeling), usually the most sensitive standard test species is used. Here, we assume that *Lemna* is the most sensitive macrophyte.

c	What is the protection goal (ecological entity, attribute, magnitude, temporal and spatial scale)?	(EFSA PPR, 2013): Ecological threshold option (ETO): Negligible effects on survival, growth, abundance, and/or biomass of populations Ecological recovery option (ERO): Small effects over months or medium effects over weeks on survival, growth, abundance, and/or biomass Spatial scale: Edge of field waterbody Laboratory tests address only the ETO, and the regulatory acceptable concentration in Tier 1 is derived by dividing the ErC50 by an assessment factor of 10. For evaluation of population model results these SPG must be made more specific, for example: ETO: Abundance or biomass should not be reduced by > 10% compared with control ERO: No deviations of abundance or biomass from control > 70% (large effects), 30%–70% no longer than 3 weeks and 10 – 30% no longer than 3 months
d	Which model outputs are required to answer the risk assessment question(s), including the required performance criteria?	Biomass or abundance over time
e	How has the appropriate model been selected?	A TKTD population model for *Lemna* is available (Schmitt et al., 2013) and considered acceptable in the EFSA TKTD opinion (EFSA PPR, panel 2018). The TKTD module can describe effect of time variable exposure on growth rate. The population model describes population growth under dynamic environmental conditions. The model has been evaluated in the TKTD opinion of the EFSA PPR panel (2018) as ready for use.
f	What is the domain of applicability of the model?	In strict sense, the TKTD module is parameterized and validated for lab test conditions. The population model is applicable to a broad range of environmental conditions. However, validation by means of field data is limited to a dataset for ditches in the Netherlands.
g	How will exposure and effects be integrated?	The exposure (PECsw) will be given by FOCUSsw Step 3 or 4 modeling (scenario approach). Effects of the plants on fate are not considered. Thus, PEC time series are used as model inputs defined in the exposure scenarios.
	2. The scenario layout	
a	Is it necessary to define crop-specific scenarios?	Use in different crops will results in different exposure patterns and thus, crop and GAP are part of the exposure scenario. The ecological scenario is considered not affected by the crops.
b	Is it necessary to define country – or zone-specific scenarios?	Driven by the exposure scenario, which also defines temperature and light as part of the ecological scenario. Nutrient levels are defined by the exposure scenarios but needed in the ecological scenario. There is no information on country – or zone-specific nutrient levels. Thus, different nutrient levels could be assessed in each climate scenario.

c	Which biotic factors are relevant for the risk assessment?	None. Predation and competition are not explicitly considered in the model, neither is habitat quality. A general carrying capacity is set assuming that there is a spatial limit for frond layers at the water surface. Resource availability or competition pressure can be varied by using different values for the carrying capacity K. However, K is a model parameter, and thus, constant over time. Similarly, predation pressure can only be changed via the reference loss rate. Differentiation between different loss rates is not included in the model.
d	Which abiotic factors are relevant for the risk assessment (e.g., weather, temperature, pH, oxygen, hydrology, light, habitat)?	Temperature, global radiation, nitrogen and phosphorus concentrations are forcing functions of the model. Temperature and global radiation are taken from the FOCUSsw Step 3 scenarios. The FOCUS scenarios do not include information on nutrient levels. It was assumed that edge-of-field surface water will receive nutrients input from the agricultural areas and that *Lemna* is typical for nutrient-rich water bodies. Thus, nutrients levels were set to not limiting values. The effect model does not consider habitat morphology, but the exposure scenario differentiates streams, ditches, and ponds. Note that *Lemna* might not live in streams, but it is modeled as surrogate for similar sensitive species that might live in streams.
e	Which spatial scale and structure is chosen, and is this relevant to the problem definition (e.g., the landscape structure influences the outcome of the risk assessment by the presence or absence of off-crop elements, connectivity, landscape management)?	The *Lemna* model is not spatially explicit. It is assumed that the toxicant and *Lemna* are homogenously distributed in space. From the exposure site, the PECs of the last segment (segment 20) of a ditch or stream along the treated field are used.
f	Are the exposure and ecological scenario consistent with each other? For example, is use of an exposure scenario for a certain country or zone and an ecological scenario based on data series from another country or zone avoided? What biotic or abiotic components need to be standardized between the exposure and the ecological scenario? If the exposure and ecological scenarios are not consistent, has it been demonstrated that the scenarios are conservative?	Yes, the weather data (temperature and light) of the exposure scenario are also used for the effect model. The type of the waterbody (stream, ditch, pond) is not considered in the effect model. Other factors (except the PEC) are not defined in the exposure scenario.

g	How are the agronomic practices considered for the risk assessment? How has the timing of the application within the GAP been chosen (e.g., was the worst-case application day chosen or were a number of time points for the application screened)? What are other agronomic practices that should be taken into account? How are agronomic practices and applications spread into the landscape (e.g., are all fields treated identically or are there differences in agronomic practice included)?	The agronomic practices are considered in the exposure assessment resulting in the PECs for the different FOCUSsw (surface water) Step 3 or 4 scenarios. The weather data and the PEC time series are used directly as forcing functions of the effect model.
	3. Assessment and analysis of the most relevant scenario components	
a	Has the influence and relevance (combination of influence or impact and variability or uncertainty) of each scenario component (5.2) been assessed (either by expert-opinion, literature research, model simulation, or a combination of methods)?	No, because temperature and radiation are defined by the FOCUSsw scenarios. Thus, the relevance of weather conditions is represented by the different FOCUS scenarios simulated. It was assumed that nutrients in edge-of-field water bodies in agricultural landscape are not limiting. Thus, the nutrient concentration was fixed to values not limiting photosynthesis. However, the effect of lower nutrient levels on the effects could be checked, e.g., by setting nutrient levels to half of the saturation constant. However, it should be considered that for situations where the nutrient level only allows very low abundance of *Lemna*, the relevance of growth inhibition of *Lemna* for ecosystem services is low.
b	Has the natural range of each scenario component relevant for the risk assessment been considered?	See 3a) The FOCUS scenarios are considered to represent the relevant range of conditions in the EU with respect to weather and exposure conditions. For the Modelink case study no review of nitrogen or phosphorus concentration in typical edge-of-field water bodies in the EU was conducted. The assumption that the input of fertilizers into water bodies is accompanied by the input of pesticides and thus creates good conditions for the growth of macrophytes seems to be justified in such a situation. This is supported by the fact that ditches are often dredged to reduce macrophytes.

	4. Selection of components for the environmental scenarios, and a check for representativeness and consistency	
a	Has a set of scenarios been defined? A short overview about reasons and choices should be given here.	Yes, the usual FOCUSsw scenarios including PEC, daily air temperature, and daily global radiation were used and no nutrient limitation was assumed. In the case study (Hommen et al., 2016), three FOCUS scenarios were analyzed. The FOCUS weather data resulted in different seasonal dynamics of unexposed *Lemna* populations as shown in the following figure. In applications for submission, all exposure scenarios without acceptable risk at Tier 1, Tier 2A or 2B based on constant exposure would be simulated.  Example of the population dynamics of unexposed *Lemna* populations resulting from temperature and radiation data in three FOCUSsw step 3 scenarios. Nutrients concentrations were set not limiting values. For details see Hommen et al. (2016).
b	Do the defined scenarios provide a range of environmental conditions from favorable to unfavorable concerning the resilience of the modeled system?	The weather conditions over at least one year cover different conditions for growth and the different scenarios describe conditions at different locations and in different years in the EU. The scenarios do not cover situations where *Lemna* does not grow in spring, but such situations are not relevant for the risk assessment of macrophytes.
c	Have most uncertain and important scenario components been covered in their natural or reasonable ranges?	With respect to T and radiation, yes. With respect to nutrient levels, no. However, different levels of nutrient concentrations can be simulated.
d	Have most relevant exposure conditions and exposure pathways been assessed and considered in the scenario development?	Yes. *Lemna* is exposed to toxicant solved in the water, described by the PEC time series of the FOCUS scenarios.
e	For spatially explicit models: Do the chosen scenarios result in realistic (dispersal) behavior?	Not applicable.

f	Has an estimation of conservatism of the scenario in ecological and exposure dimensions been provided?	No. As usual for aquatic risk assessment it has been assumed that the set of the FOCUS exposure scenarios represents a realistic worst case. For nutrient level, an analysis of realistic levels and their impact on the assessment is missing. Generically, a range of nutrient levels ranging from low to no limitation could be simulated to assess the effect of, for example, the maximum effect. However, if very low nutrient level result in stronger % effects, its relevance should be discussed – 20% reduction at a very low level might be less relevant than 10% reduction at a high level.
g	Are the scenarios representative for the risk assessment under consideration (e.g., are the spatio-temporal scales of the scenarios in line with the problem definition; does the conservatism of the scenarios match with the conservatism of the problem definition)?	Yes, because the ecological scenario is closely linked to the exposure scenarios based on current practice. However, the relevance and the impact of nutrient limitation is not checked yet.

8.1.4. Bibliography Appendices Chapter 3

EFSA (European Food Safety Authority). (2009). Risk assessment for birds and mammals. *EFSA Journal*, *7*(12), 1438. https://doi.org/10.2903/j.efsa.2009.1438

EFSA (European Food Safety Authority, Alessio Ippolito, Andreas Focks, Maj Rundlöf, Andres Arce, Marco Marchesi, Franco Maria Neri, Agnès Rortais, Csaba Szentes, Domenica Auteri). (2021). Analysis of background variability of honey bee colony size. *EFSA Supporting Publications*, *18*(3), 6518E. https://doi.org/10.2903/sp.efsa.2021.EN-6518

EFSA (European Food Safety Authority). (2023a). Revised guidance on the risk assessment of plant protection products on bees (*Apis mellifera*, Bombus spp. and solitary bees). *EFSA Journal*, *21*(5), e07989. https://doi.org/10.2903/j.efsa.2023.7989

EFSA PPR (EFSA Panel on Plant Protection Products and their Residues). (2013). Guidance on tiered risk assessment for plant protection products for aquatic organisms in edge-of-field surface waters. *EFSA Journal*, *11*(7), 3290. https://doi.org/10.2903/j.efsa.2013.3290

EFSA PPR (Panel on Plant Protection Products & their Residues). (2014). Scientific opinion on good modeling practice in the context of mechanistic effect models for risk assessment of plant protection products. *EFSA Journal*, *12*(3), 3589. https://doi.org/10.2903/j.efsa.2014.3589

EFSA SC (EFSA Scientific Committee, Simon More, Vasileios Bampidis, Diane Benford, Claude Bragard, Thorhallur Halldorsson, Antonio Hernández-Jerez, Susanne Hougaard Bennekou, Kostas Koutsoumanis, Kyriaki Machera, Hanspeter Naegeli, Søren Saxmose Nielsen, Josef Schlatter, Dieter Schrenk, Vittorio Silano, Dominique Turck, Maged Younes, Gerard Arnold, Jean-Lou Dorne, Angelo Maggiore, Stephen Pagani, Csaba Szentes, Simon Terry, Simone Tosi, Domagoj Vrbos, Giorgia Zamariola, Agnes Rortais). (2021). A systems-based approach to the environmental risk assessment of multiple stressors in honey bees. *EFSA Journal*, *19*(5), e06607. https://doi.org/10.2903/j.efsa.2021.6607

Hommen, U., Forbes, V., Grimm, V., Preuss, T. G., Thorbek, P., & Ducrot, V. (2016). How to use mechanistic effect models in environmental risk assessment of pesticides: Case studies and recommendations from the SETAC Workshop MODELINK. *Integrated Environmental Assessment and Management*, *12*(1), 21-31. https://doi.org/10.1002/ieam.1704

Preuss, T. G., Agatz, A., Goussen, B., Roeben, V., Rumkee, J., Zakharova, L., & Thorbek, P. (2022). The BEEHAVEecotox Model—Integrating a mechanistic effect module into the honey bee colony model. *Environmental Toxicology and Chemistry*, *41*(11), 2870-2882. https://doi.org/10.1002/etc.5467

8.2. Appendix Chapter 4: Example of a literature search about pollen consumption of adult honey bees

Date span of the search: 1900-01-01 – 2021-04-06.

Database Name	Date of last Database Update
Agricola	2021-03-08
Biosis	2021-03-31
CABA	2021-03-30
Chemical Abstracts	2021-04-05
Derwent Drug File (DRUGU)	2021-03-29
EMBASE	2021-04-05
Esbiobase	2021-03-31
IPA	2021-03-12
Medline	2021-04-05
PQSciTech	2021-03-25
Scisearch	2021-04-05
Toxcenter	2021-04-05
FSTA	2021-04-01

Search query:

(BEE OR APIS OR A MELLIFERA OR HONEY BEE)

AND

(WORKER OR IN HIVE OR FORAGER OR ADULT)

AND

(POLLEN(2A) (COMPOSITION OR DIET))

NOT P/DT

Search results:

- References found in total: 242
- References not dealing with Apis mellifera: 86
- References about Apis mellifera, but not relevant for daily pollen consumption of adult bees by the title: 132

We identified 24 articles to be relevant according to the title (Table 8.1).

After reading the abstracts, nine relevant articles for the daily pollen consumption of adult bees were identified (Table 8.2). For the 15 studies considered non-relevant according to the abstract, articles that have investigated colonies and not only adult bees as well as articles with artificial food have been excluded. The nine relevant articles will subsequently need to be assessed for reliability.

Table 8.4: Articles found relevant for pollen consumption of adult honey bees based on their title.

ID	Publication Year	Title	Author	Source	Reliability Score
3	2021	A new design of bee cage for laboratory experiments: nutritional assessment of supplemental diets in honey bees (*Apis mellifera*)	Yang, Kai-Chieh; Peng, Zhi-Wei; Lin, Chun-Hsien; Wu, Ming-Cheng	Apidologie (2021) Ahead of Print	
45	1967	Comparison of the protein quality of pollens for growth stimulation of the hypopharyngeal glands and longevity of honey bees, *Apis mellifera*	Standifer, Lonnie N.	Insectes Sociaux (1967), 14(4), 415-25 From: Apicult. Abstr. 1968, 19(3), 102	
46	2008	Comparisons of pollen substitute diets for honey bees: Consumption rates by colonies and effects on brood and adult populations	Degrandi-Hoffman, Gloria; Wardell, Gordon; Ahumada-Segura, Fabiana; Rinderer, Thomas; Danka, Robert; Pettis, Jeff	Journal of Apicultural Research, (2008) Vol. 47, No. 4, pp. 265-270. http://www.tandfonline.com/toc/tjar20/current.	
51	2013	Conversion of high and low pollen protein diets into protein in worker honey bees (Hymenoptera: Apidae)	Basualdo, M.; Barragan, S.; Vanagas, L.; Garcia, C.; Solana, H.; Rodriguez, E.; Bedascarrasbure, E.	Journal of Economic Entomology (2013), 106(4), 1553-1558	
65	2018	Digestibility and nutritional value of fresh and stored pollen for honey bees (*Apis mellifera* scutellata)	Nicolson, Susan W.; Da Silva Das Neves, Susana; Human, Hannelie; Pirk, Christian W. W.	Journal of Insect Physiology (2018), 107, 302-308	

66	1983	Disappearing disease – II. Effects of certain protein sources on brood rearing and length of life in the honey bee under laboratory conditions	Kulincevic, Jm; Rothenbuhler, Wc; Rinderer, Te	American Bee Journal [AM. BEE J.]. Vol. 123, no. 1, pp. 50-53. 1983. ISSN: 0002-7626	
72	2017	Ecological stoichiometry of the honeybee: Pollen diversity and adequate species composition are needed to mitigate limitations imposed on the growth and development of bees by pollen quality	Filipiak, Michal; Kuszewska, Karolina; Asselman, Michel; Denisow, Bozena; Stawiarz, Ernest; Woyciechowski, Michal; Weiner, January Filipiak, Michal (Reprint); Weiner, January Denisow, Bozena	PLOS ONE, (22 2017 AUG 2017) Vol. 12, No. 8. ISSN: 1932-6203.	
78	2005	Effect of some supplementary feeding on physiological characters of honeybee workers	Abdilla, F. S.	Assiut Journal of Agricultural Sciences (2005), Volume 36, Number 1, pp. 97-108, 20 refs. ISSN: 1110-0486 Published by: Faculty of Agriculture, Assiut University, Assiut	
79	1984	Effects of a single-cell protein and/or pollen diet fed to honeybee colonies. The hypopharyngeal gland	Omar, M.	Apicultura in Romania (1984), Volume 59, Number 10, pp. 5-7, B	
80	2021	Effects of artificial diets on growth and development and physiological metabolism of young worker bees	Zhuang Mingliang; Li Jianfei; Li Zhiyong; Wang Zhi; Niu Qingsheng; Chen Donghai; GE Peng; Zhang Fa; Zhuang, M. L.; Li, J. F.; Li, Z. Y.; Wang, Z.; Niu, Q. S.; Chen, D. H.; GE, P.; Zhang, F.	Chinese Journal of Animal Nutrition (2021), Volume 33, Number 2, pp. 1070-1080, 23 refs. ISSN: 1006-267X DOI: 10.3969/j.issn.1006-267x.2021.02.047 Published by: Chinese Association of Animal Science and Veterinary Medicine, Beijing URL (Availability): htt	
93	1984	Feeding preferences of *Apis mellifera* hymenoptera Apidae individual vs. mixed pollen species	Schmidt, J. O.	Journal of the Kansas Entomological Society, (1984) Vol. 57, No. 2, pp. 323-327.	

122	1988	Influence of nutritional stress and the age of adults on the morphometrics of honey bees *Apis mellifera* L.	Herbert, E. W. Jr; Sylvester, H. A.; Vandenberg, J. D.; Shimanuki, H.	Apidologie, (1988) Vol. 19, No. 3, pp. 221-230.	
128	2013	Influence of pollen nutrition on honey bee health: Do pollen quality and diversity matter?	Di Pasquale, Garance; Salignon, Marion; Le Conte, Yves; Belzunces, Luc P.; Decourtye, Axel; Kretzschmar, Andre; Suchail, Severine; Brunet, Jean-Luc; Alaux, Cedric Di Pasquale, Garance (Reprint) Di Pasquale, Garance (Reprint); Alaux, Cedric Di Pasquale, Ga	PLOS ONE, (5 2013 AUG 2013) Vol. 8, No. 8. ISSN: 1932-6203.	
135	2006	Influence of some protein diets on the longevity and some physiological conditions of honeybee *Apis mellifera* L. workers	Alqarni, Abdulaziz S.	Journal of Biological Sciences (Faisalabad, Pakistan) (2006), 6(4), 734-737	
143	2008	Laboratory studies of some diets as pollen substitutes on some biological activities and morphological measurements of caged worker bees at different ages	Ashour, A. T.; Hammad, H. M.; Nour, M. E.; Zakaria, M. E.	Bulletin of Faculty of Agriculture, Cairo University (2008), Volume 59, Number 2, pp. 103-110, 28 refs. ISSN: 0526-8613 Published by: Faculty of Agriculture, Cairo University, Giza URL (Availability): http://www.cu.edu.eg	
160	2014	New formula of pollen supplemental diets to study honey bee (*Apis mellifera* carnica) attractiveness	Aly, M. Z.; Osman, K. S.; Karem, M. M.; Elsayeh, W. A.	Egyptian Academic Journal of Biological Sciences: Entomology (2014), Volume 7, Number 2, pp. 47-55, 35 refs. ISSN: 1687-8809 Published by: Ain Shams University, Cairo URL (Availability): http://entomology.eajbs.eg.net/pdf/vol7-num2/6.pdf	

167	2016	On the effects of artificial feeding on bee colony dynamics: A mathematical model	Lisboa Mohallem Paiva, Juliana Pereira; Paiva, Henrique Mohallem (Reprint); Esposito, Elisa; Morais, Michelle Manfrini Lisboa Mohallem Paiva, Juliana Pereira; Esposito, Elisa Paiva, Henrique Mohallem (Reprint) Morais, Michelle Manfrini	PLOS ONE, (22 NOV 2016) Vol. 11, No. 11. ISSN: 1932-6203.	
176	2019	Physiological effects of some pollen substitutes diets on caged honey bee workers *Apis mellifera* (Hymenoptera: Apidae)	Abdulraouf, M. A.; Mohamed, S. Y.; Ayman, A. G.	Egyptian Journal of Plant Protection Research Institute (2019), Volume 2, Number 4, pp. 741-750, 33 refs. ISSN: 2356-9832 Published by: Plant Protection Research Institute, Giza URL (Availability): http://www.ejppri.eg.net/pdf/v2n4/27.pdf	
186	2016	Pollen nutrition in honey bees (*Apis mellifera*): impact on adult health	Frias, Bruna Estefania Diniz; Barbosa, Cosme Damiao; Lourenco, Anete Pedro	Apidologie (2016), 47(1), 15-25	
187	2000	Pollen quality of fresh and 1-year-old single pollen diets for worker honey bees (*Apis mellifera* L.)	Pernal, Stephen F.; Currie, Robert W.	Apidologie, (May-June, 2000) Vol. 31, No. 3, pp. 387-409. print.	
195	2013	Protein levels and colony development of Africanized and European honey bees fed natural and artificial diets	Morais, M. M.; Turcatto, A. P.; Pereira, R. A.; Francoy, T. M.; Guidugli-Lazzarini, K. R.; Goncalves, L. S.; De Almeida, J. M. V.; Ellis, J. D.; De Jong, D.	GMR, Genetics and Molecular Research (2013), 12(4), 6915-6922, 8 pp.	
197	1998	Quantification of hemolymph proteins as a fast method for testing protein diets for honey bees (Hymenoptera: Apidae)	Cremonez, Tania M.; De Jong, David; Bitondi, Marcia M. G.	Journal of Economic Entomology (1998), 91(6), 1284-1289	
218	1989	The nutritional value of typhalatifolia pollen for bees	Schmidt, J. O.; Buchmann, S. L.; Glaiim, M.	Journal of Apicultural Research, (1989) Vol. 28, No. 3, pp. 155-165.	

228	2017	The impact of pollen consumption on honey bee (*Apis mellifera*) digestive physiology and carbohydrate metabolism	Ricigliano, Vincent A.; Fitz, William; Copeland, Duan C.; Mott, Brendon M.; Maes, Patrick; Floyd, Amy S.; Dockstader, Arnold; Anderson, Kirk E.	Archives of Insect Biochemistry and Physiology (2017), 96(2), n/a	

Table 8.5: Articles found relevant for pollen consumption of adult honey bees based on abstract.

ID	Publication Year	Title	Author	Source
3	2021	A new design of bee cage for laboratory experiments: nutritional assessment of supplemental diets in honey bees (*Apis mellifera*)	Yang, Kai-Chieh; Peng, Zhi-Wei; Lin, Chun-Hsien; Wu, Ming-Cheng	Apidologie (2021) 52, pages 418–431
45	1967	Comparison of the protein quality of pollens for growth stimulation of the hypopharyngeal glands and longevity of honey bees, *Apis mellifera*	Standifer, Lonnie N.	Insectes Sociaux (1967), 14(4), 415-25 From: Apicult. Abstr. 1968, 19(3), 102
51	2013	Conversion of high and low pollen protein diets into protein in worker honey bees (Hymenoptera: Apidae)	Basualdo, M.; Barragan, S.; Vanagas, L.; Garcia, C.; Solana, H.; Rodriguez, E.; Bedascarrasbure, E.	Journal of Economic Entomology (2013), 106(4), 1553-1558
65	2018	Digestibility and nutritional value of fresh and stored pollen for honey bees (*Apis mellifera* scutellata)	Nicolson, Susan W.; Da Silva Das Neves, Susana; Human, Hannelie; Pirk, Christian W. W.	Journal of Insect Physiology (2018), 107, 302-308
78	2005	Effect of some supplementary feeding on physiological characters of honeybee workers	Abdilla, F. S.	Assiut Journal of Agricultural Sciences (2005), Volume 36, Number 1, pp. 97-108, 20 refs. ISSN: 1110-0486 Published by: Faculty of Agriculture, Assiut University, Assiut
135	2006	Influence of some protein diets on the longevity and some physiological conditions of honeybee *Apis mellifera* L. workers	Alqarni, Abdulaziz S.	Journal of Biological Sciences (Faisalabad, Pakistan) (2006), 6(4), 734-737

176	2019	Physiological effects of some pollen substitutes diets on caged honey bee workers *Apis mellifera* (Hymenoptera: Apidae)	Abdulraouf, M. A.; Mohamed, S. Y.; Ayman, A. G.	Egyptian Journal of Plant Protection Research Institute (2019), Volume 2, Number 4, pp. 741-750, 33 refs. ISSN: 2356-9832 Published by: Plant Protection Research Institute, Giza URL (Availability): http://www.ejppri.eg.net/pdf/v2n4/27.pdf
186	2016	Pollen nutrition in honey bees (*Apis mellifera*): impact on adult health	Frias, Bruna Estefania Diniz; Barbosa, Cosme Damiao; Lourenco, Anete Pedro	Apidologie (2016), 47(1), 15-25
187	2000	Pollen quality of fresh and 1-year-old single pollen diets for worker honey bees (*Apis mellifera* L.).	Pernal, Stephen F.; Currie, Robert W.	Apidologie, (May-June, 2000) Vol. 31, No. 3, pp. 387-409. print.

References found as part of the example literature search are not included in a separate bibliography.

8.3. Appendices Chapter 5: Examples for modular aspects in model description and assessment

In this appendix, model examples are given as case studies for a stronger accentuation of the modular aspects in complex model descriptions. Besides the always needed fundamental model overview, the emphasis is on the additional representation of the most important modules regarding their structure, parametrization, validation, and description of the main module interfaces, and their domain of applicability. With respect to key aspects of modularity such as module qualities, validity ranges, and testability of individual model components, these exemplary model documentations and their interactions are intended to enable a more differentiated evaluation and communication of complex models used in the ERA.

8.3.1. Example 1: Aquatic population-level model (Chaoborus)

8.3.1.1. Short overall model description

The IBM Chaoborus model (Strauss et al., 2016, Dohmen et al., 2016; current model version: 4.2) is designed to simulate the population dynamics of the aquatic phantom midge *Chaoborus crystallinus*, a common pelagic invertebrate predator in fish-free freshwater ponds (see Figure 8.1). This individual-based model is

based on the physiology and ecology of this species, and formulates the main processes (e.g., development, reproduction) using empirical, species-specific equations. This model is specific to *Chaoborus* only and can be used across studies as a general module for this species to simulate population dynamics based on a specific ecological scenario and the effects of a specific exposure scenario. The model calculates the abundance of larval stages and emergence of the adults over time with daily resolution. Density-dependent processes such as cannibalism as well as emigration rates are important factors for the regulation of the population density. With this model, it is also possible to simulate two neighboring populations, of which only one is exposed to the toxicant. The population density is influenced by the toxicant directly for the treated population and indirectly for the untreated population by migrating adults due to their spatial connectivity in the field (metapopulation approach, Dohmen et al. 2016).

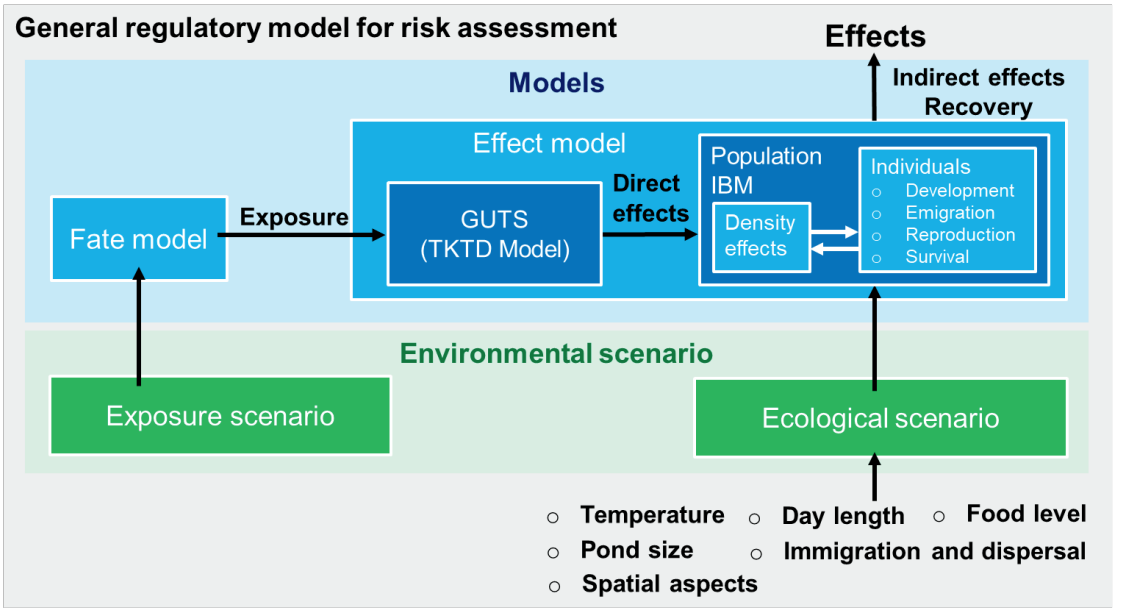


Figure 8.1: Modular representation of the IBM Chaoborus population model within the effect model and its links to the fate model and the environmental scenario.

Temporal scales

The simulation time step of the toxicological module is one day or one hour, depending on the resolution of exposure data. The ecological module is updated in daily steps. Mortality by toxic substances takes place at the end of each day. The effect model is designed for simulation periods ranging from weeks to several years.

Spatial scales

Scenarios of individual populations have been simulated so far for volumes in the range of 100 L to 10 m3. If a mesocosm scenario or meta-population approach is to be simulated instead of a single pond, different populations with and without toxic treatments can be considered assuming a maximum active migration distance up to 1 km, depending on the spatial scenario.

8.3.1.2. Modular model structure

Exposure

Default model input from the fate module of daily or hourly exposure concentrations of the toxic compound in the water column (e.g., from FOCUS[9] exposure scenarios. As an alternative to this input of an external fate module, up to three application peaks can be simulated with the possibility to consider the temperature dependence of the degradation rate [DT50, implemented in the effect model]).

Toxicological module

As toxicological modules for lethal effects, the generic TKTD model framework GUTS is used by default. If experimental data are available, the dependence of TKTD parameters on temperature as well as larval size can be also considered in the toxicological module. For very short exposure events and lack of data to calibrate TKTD models, a simple dose-response model is implemented as an alternative approach.

Ecological module

Input data

Daily values of water temperature and day length are required. The food level is assumed to be constant throughout each simulation run.

Individual level submodules:

Growth and reproduction

Food level and temperature influences the rate of individual development. A fixed clutch size per hatched female is used in the current model version.

Survival

Background mortality and density dependent cannibalism. No explicit consideration of starvation.

Overwintering of larvae

Induction and termination of larval dormancy during winter depending on temperature and day length. The parametrization of the timing of dormancy is based on laboratory data.

Movement

Only flight-capable adults show dispersal. The loss of a population by emigration is given as a fixed proportion of hatched adults, although little experimental data are available for this. An equal distribution of egg-laying females over closely neighboring ponds is integrated (mesocosm scenario). A hypothetical immigration from a distant uncontaminated pond can be switched on.

---

9 https://esdac.jrc.ec.europa.eu/projects/focus-dg-sante

Population level

Density dependence

Mortality and food availability are implemented density-dependently. Cannibalism between *Chaoborus* larvae is the biggest determination of the capacity of the population.

Food

Different food levels must be specified for small and large larvae for each simulation run, reflecting different food item availabilities. Currently, no feedback on food availability by foraging *Chaoborus* larvae is implemented. This very simple food module can be replaced by dynamic food resource modules in the future if needed.

Community level interaction

None.

Indirect effects

Depending on the landscape and exposure scenario, a toxic impact on a treated population may also have a significant impact on another population that is not itself contaminated, through a net loss of migrating adults (sink-source dynamics in metapopulations). Another indirect effect is that reduced larval survival due to the toxic stressor reduces population density, which in turn reduces density-dependent cannibalism. This mechanism may compensate for some of the toxin-induced mortality.

Feedback loops between main modules

There is feedback between the toxicological module and the ecological module through the dynamic consideration of larval size in both the individual-based population model and the GUTS parameters once the size dependence of the GUTS parameters is turned on.

No feedback occurs between the fate module and the toxicological or the ecological module. Therefore, the exposure can be independently calculated and read in as a file into the model.

The implemented pathway from toxic exposure in the water phase to the toxicological module is only bioconcentration in case of GUTS or a direct dose-response to the ambient concentration, which is not influenced by any behavior of the animals.

Consistency between modules for important variables

The daily water temperature is an important input parameter for the effect model and influences the physiological processes of the individuals in the ecological *Chaoborus* module. In the best case (though not always fulfilled in individual studies), the external fate module uses the same water temperatures for the calculation of, for example, the degradation rate of the compound (DT50) as the effect module. In the toxicological module of the *Chaoborus* model, an already implemented temperature dependence of the GUTS parameters can optionally be switched on, but this is often not used due to lack of data.

Model parametrization

The ecological components of the IBM Chaoborus model were parametrized based on species-specific laboratory data at the individual level (all physiological dependencies) and semi-field tests (clutch number per female, estimation of migration), and are therefore considered generally applicable to new case studies for this species. Only a few case-study-dependent parameters, such as the net emigration rate, must be recalibrated or specified in the case of each model application. Another exception in the present model version is the physiological dependence of winter dormancy of L4 larvae, because this are parametrized only for Central Europe and thus can be transferred to other regions only with high uncertainty.

Model testing and range domain of applicability

The IBM Chaoborus model has been tested so far at the population level for aquatic field mesocosms at different test sites in Germany, for different pesticides as well as without toxicants under different feeding conditions. However, an increased uncertainty remains for the migration rate of adults, for which assumptions still must be made based on limited data. The module for larval dormancy and overwintering has not yet been verified for northern or southern European regions with field data. Thus, the model is applicable under Central European weather conditions as long as the food conditions prevent complete starvation of the individuals. Transferability to northern or southern Europe is possible in principle but involves increased uncertainty due to a lack of field data on the timing of larval dormancy and therefore a lack of testing for these climatic regions.

8.3.2. Example 2: Aquatic model at ecosystem level (StoLaM+ model)

8.3.2.1. Short overall model description

A lake model-population IBM-hybrid model is presented as an example of an ecosystem model (based on the DaLaM model, Strauss et al. 2017). The lake model StoLaM is a general all-purpose mechanistic ecosystem model, which integrates physico-chemical dynamics of standing waters, phytoplankton and zooplankton communities responding to weather data with a high temporal resolution (Strauss 2009). The use of an hourly resolution of weather data ensures realistic integration of dynamic climate data. Therefore, a mechanistic, dynamic hydrodynamic module (HyLaM) is linked to the StoLaM to provide the physical environment in terms of water temperature, underwater light conditions, vertical turbulence and thermal stratification. StoLaM contains generic and easily parametrizable modules for phytoplankton and zooplankton species. These consist of unstructured compartment models with differential equations and case study independent parametrization (see phytoplankton module below for an example). Selected zooplankton species can be replaced by individual-based models (IBMs), such as the compartment model for daphnids by the IDamP (Preuss et al. 2009, Strauss et al. 2017). Figure 8.2 gives a schematic overview of the overall model.

Temporal scales

The simulation time step of the toxicological module (GUTS) is one day or one hour, depending on the resolution of exposure data. The ecological lake model StoLaM calculates in minute steps, while the time steps of the coupled IBMs can vary between minutes and daily steps (e.g., IDamP for *Daphnia magna*). The weather data driving the hydrodynamic module must have a temporal resolution of 10 minutes or one hour. The model is designed for simulation periods ranging from weeks to several years.

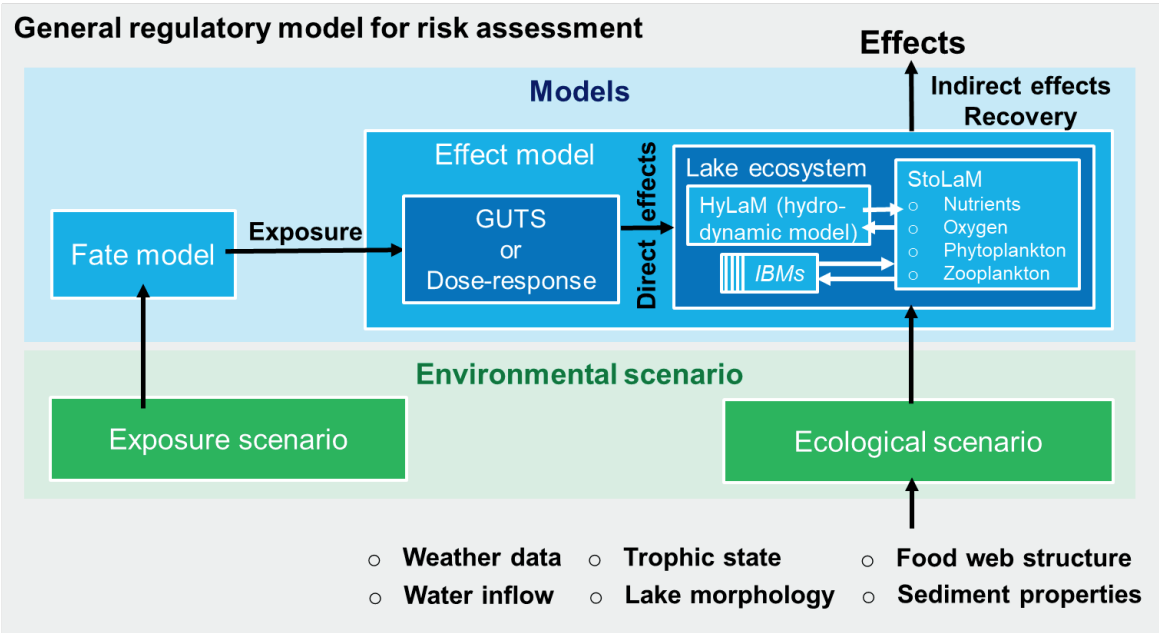


Figure 8.2: Modular representation of the ecosystem lake model and its submodels within the effect model and its links to the fate model and the environmental scenario.

Spatial scales

The StoLaM model is designed for water bodies starting from small ponds to large lakes and allows the consideration of vertical gradients (1D model).

8.3.2.2. Modular model structure

Exposure

A time-varying model input of daily or hourly substance concentration is used as exposure (assumed to be homogeneously distributed in the waterbody), no in-lake fate of the toxic compound is implemented yet.

Toxicological module

For the zooplanktonic IBMs, toxicological modules can be added either TKTD-based (GUTS, DEB-TKTD) or empirical dose-response models (e.g., ECx approaches for sublethal effects on filtration and growth). For

the phytoplankton module, the toxicological module is a dose-response model (ECx approach) on growth rate *(TD model for algae)*.

Ecological module

Input data: The required input data also determines the ecological scenario. In particular, the food web settings and the available data on nutrient inputs, for example, from sediment or tributaries, determine the performance of the simulation.

Individual-level submodules (IBMs)

The individual-based models describe the life cycle of individual species and are linked to the lake model via water temperature, food web interactions, and nutrient-related processes. The individuals respond to the toxicological modules. Such IBMs should have been previously parametrized and tested independently before.

Population level

The populations are either simulated directly as compartmental models in StoLaM (specific modules for phytoplankton and zooplankton taxa), or in the case of IBMs, are composed of the sum of their simulated individuals. Implemented populations interact through predator-prey relationships as well as competition for resources.

Community-level interaction

In this ecosystem model, many community-level interactions exist, such as through competition for nutrients (phytoplankton), competition for food (zooplankton), and predator-prey interactions. Through the chemical stoichiometry implemented in StoLaM, nutrient effects in phytoplankton can also cause feedback effects on zooplankton growth.

Indirect effects

Due to the high model complexity, several indirect effects are possible. These include shifts in dominance in plankton communities, increase in primary producers (e.g., algal blooms at reduced grazing), accelerated or slowed recovery of planktonic organisms as a function of nutrient turnover and availability.

Feedback loops between main modules

Within the lake model, there are interactions between the HyLaM hydrodynamic module and the StoLaM that influence temperature dynamics (Strauss et al. 2017). The individual-based zooplankton models interact with the StoLaM analogously to the zooplankton modules implemented as compartments via food intake and excretions. In the case of an appropriate research question and data availability, the toxicological model GUTS can account for the size of individuals from the IBM model as well as the temperature from the hydrodynamic module for adjusting the TKTD parameters. So far, no feedback of the effect model on exposure has been implemented.

The implemented pathway from toxic exposure in the water phase to the toxicological module is only bioconcentration in case of GUTS or a direct dose-response to the ambient concentration, both are not affected by any behavior of the organisms. Biomagnification, for example, via contaminated food, can be implemented in this model, but would lead to feedback from the feeding activity of organisms in the ecological module to the toxicokinetic part of the toxicological module.

Consistency between modules for important variables

The timing of interactions with different time steps between modules (here, ecosystem model and coupled IBMs) is solved as follows (Strauss et al. 2017): If algal growth occurs in steps ≤ 1h and *Daphnia* population dynamics are updated daily, at least algal grazing and possible nutrient excretion by daphnids should also occur in the same or similarly short time step (1h). The food ingested by the daphnids during this process is summed over one day for later calculation of growth at the end of the day. Mortality of daphnids by toxic substances in IDamP also occurs at the end of each day to date. Although the toxicological lethal effects module, GUTS, calculates in smaller time steps such as days or hours, in IDamP the damage is accumulated for each individual. The entire effect model is implemented as temperature-dependent for the most important processes. In standard applications for risk assessment, however, temperature is rarely applied to the parameters of the toxicological TKTD model. However, this function can be switched on in the present model if sufficient data are available. The lake model StoLaM considers the ecological stoichiometry especially of nitrogen and phosphorus in all essential nutrient-related processes. Therefore, at the interface to the IBM model IDamP, attention must also be paid to the utilization and excretion of food-bound nutrients as well as to a possible limitation of daphnids due to nutrient deficiencies in the diet.

Model parametrization

The ecological parts of the lake model StoLaM were mainly parametrized using literature and some laboratory data and are thus generally valid independently of new case studies. Only few waterbody-specific dependent parameters (e.g., nutrient release from sediment, turbidity of the water column) must be recalibrated or specified in each case of model application.

At the community level, the composition of the food web and their interactions (e.g., prey selectivity of consumers) will massively influence population dynamics, and therefore must be set and justified with special care for the scenarios being simulated.

The hydrodynamic module HyLaM has been fully parametrized and validated beforehand without the need for further case study-specific adaptation.

The IBMs that can be coupled in this modeling approach should have been previously parametrized using independent laboratory data and tested case study independently before. Therefore, these IBM would be usable in new case studies without further adaptation.

Model testing and domain of applicability

The lake model StoLaM (including the hydrodynamic module HyLaM) was tested as a water quality model under different trophic conditions (oligotrophic to hypertrophic lakes), different lake sizes (from shallow ponds and lakes to deep lakes with water depth > 20 m), and different weather conditions (data from years between 1995–2019, model applications for standing waters in Germany and China).

In principle, the ecological lake model StoLaM is generally applicable, because both the implemented hydrodynamic and biological processes are based on decades of experience of the scientific community and have been parametrized independently of case studies. The model is designed for the main processes of the planktonic community in lakes, embedded in detailed implemented nutrient cycles and hydrodynamic processes. Nevertheless, for individual model applications, the food web composition and few individual waterbody-specific parameters (e.g., nutrient release from sediment, turbidity of the water column) must always be re-parametrized.

Due to the lack of horizontal differentiation in StoLaM, horizontal gradients cannot be represented, and the use, for example, for reservoirs is not recommended.

The IBM *Daphnia* model IDamP has been tested many times under laboratory conditions in combination with the GUTS module, and the ecological module of IDamP exemplarily under perennial field conditions.

8.3.3. Example 3: Terrestrial spatially explicit population–level model (POLARIS)

8.3.3.1. Short overall model description

The POLARIS model is a spatially explicit and individual-based model framework for terrestrial small mammals that has been developed to simulate population-level effects in the context of risk assessment of plant protection products according to EFSA (2009);(Kleinmann and Wang, 2017; Wang and Luttik, 2013; Wang et al., 2022).

The implemented relevant European focus types include:

- Common vole (*Microtus arvalis*)
- Field vole (*Microtus agrestis*)
- Common shrew (*Sorex araneus*)
- Wood mouse (*Apodemus sylvaticus*)
- Brown hare (*Lepus europaeus*)
- European rabbit (*Oryctolagus cuniculus*)

POLARIS is modularly structured into four main models for landscape, fate, and the effect model, which is subdivided into the toxicological module and the ecological module, the latter consisting mainly of generic IBM population models (see Figure 8.3). Exposure depends on the exposure scenario (application), the landscape model including vegetation, the fate model modifying the toxic load of the plants, and animal behavior. The landscape and toxicology models are structurally generic and must be re-parametrized specifically for each risk assessment. A customized landscape can be used for each study (depending on the GAP of the substance and the species), and the toxicology module must also be selected for each risk assessment and parametrized for the specific combination of a toxicant and a species.

The population model including the IBM submodules is independent of the specific risk assessment in most cases. However sometimes specific model parameters need to be adapted, for example, when the model is used in a specific climatic region. For each species a fully parametrised and validated individual-based model is available. This consist of a set of generically designed submodules which were adapted and parametrised specifically for individual species.

In principle, the model is applicable over the entire range of natural occurrence of European focal species, as far as climate zone-specific parametrizations for all modules are available. However, the fully parametrized overall model, with its specific combination of landscape and toxicity, is only applicable for a specific risk assessment. This modular approach offers a high level of flexibility to simulate and analyze landscape-based scenarios, which can be tailor made for the specific requirements of a plant protection product.

Temporal scales

The effect model is designed for simulation periods ranging from weeks to several years (e.g., for validation studies depending on the dataset). For risk assessment applications, a duration of 20 years is often chosen. The simulation time step in the different modules is usually one day. However, for species such as brown hare and European rabbit, a daily walk is simulated. This means that the position of each individual is simulated in a smaller time step. The exact time step per day for these individuals is variable because it depends on the individual distance walked during the day. Within the toxicological module, daily or hourly steps are possible for the TKTD approach, whereby in both cases the effects only affect the individuals at the end of the day. The dose-response approach is only processed in daily time steps.

Spatial scales

The maximum area of the simulated landscape depends on both the acceptable simulation time and the computing power. The total area should not fall below a minimum value to ensure that the population does not go extinct by chance. Thus, this minimum area depends on the population density of the simulated species in untreated control scenarios. For example, while 1 ha is usually sufficient for the water vole, more than 100 ha are required for the brown hare. For the explicitly spatial subdivision of the total area into grid cells, a spatial resolution of 5x5 m per single cell is selected by default.

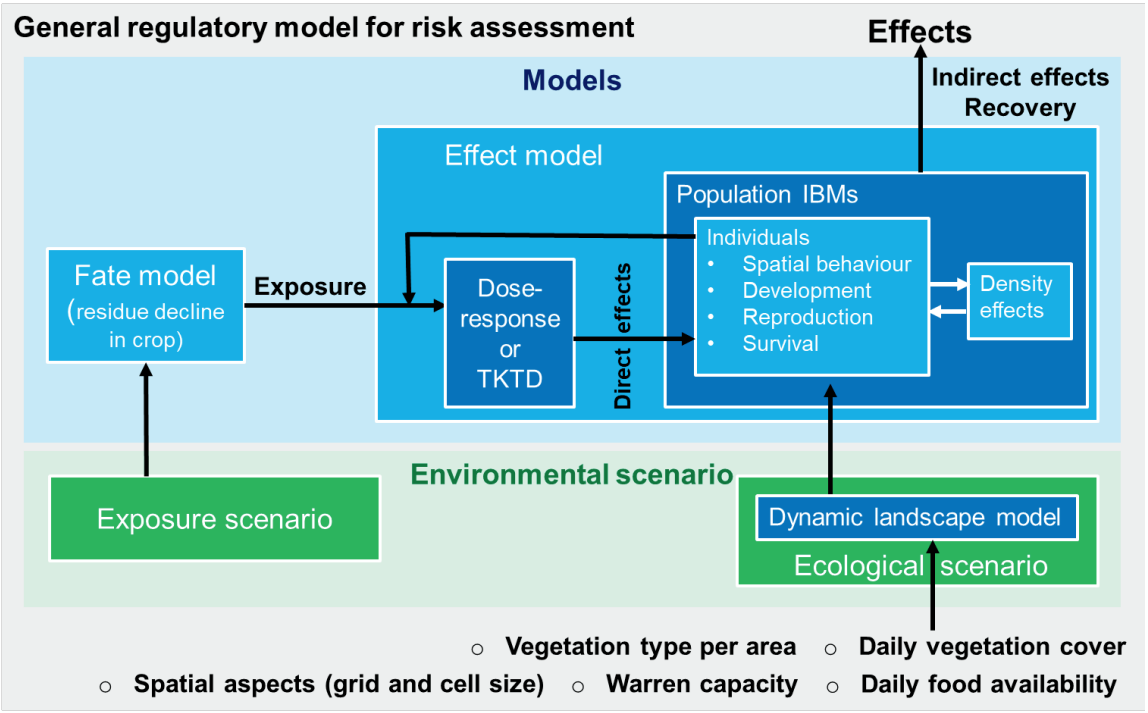


Figure 8.3: Modular representation of the spatially explicit POLARIS model and its submodels within the effect model, and its links to the fate model and the environmental scenario.

8.3.3.2. Modular model structure

Dynamic landscape model (as part of the ecological scenario)

The landscape is implemented as species – and compound-independent module, which makes it possible to use landscape scenarios for different species and toxicants. However, the minimum size of the landscape depends on the species to be simulated (see spatial scales). Depending on the crop type and spatial pre-setting of the landscape, a variety of different application scenarios are possible.

Currently, a static input of spatially structured time series of the vegetation composition and plant growth is still required. These scenarios already include the impact by agricultural practice and consist of in-field and off-field parts. What is needed for spatially input per grid cell is the vegetation type, the daily vegetation cover [%], the daily available food [mm], and the warren capacity [1/ha].

In the future, a dynamization of vegetation growth in the landscape model is planned. In principle, toxicological effects of herbicides on vegetation growth could also be simulated, and indirect effects on animals could be addressed. For this purpose, the ecological model could remain unchanged.

Fate and exposure

The exposure scenario provides spatially explicit application patterns of a toxicant with different application rates for defined vegetation types, considering the specific spatial settings in the landscape model. The concentration of the toxicant on the specific plants is subject to simple degradation kinetics with a given

DT50 to model the residue decline in crop, implemented in the fate module. The exposure pathway for individuals in this model is biomagnification. To this end, the toxicological submodule of the effect model requires the daily toxic dose of the individuals as input variable. This results from the interaction of the dynamic concentration of the substance in the respective vegetation types (output of the fate module) and the spatio-temporal behavior of the individuals (feedback from the ecological module).

Toxicological module

Lethal and several sublethal effects on the individual level can be selected and parametrized as dose-response relationships. If enough data is available, a TKTD model can also be parametrized and used. Both alternative toxicological submodules consider biomagnification as the exposure path using the dietary dose per time (daily or hourly), albeit in different ways.

The size of individuals, their food and habitat preferences determine their daily food intake, which, considering the toxic load of their plant food from the fate module, determines the toxicant dose per day. This daily dietary dose is determined by the portion of time consuming different toxically loaded food per day. In the TKTD approach, a weight-based internal toxicant concentration is calculated in the toxicokinetic submodule as a function of the daily toxicant dose, which determines the effect size in the toxicodynamic submodule. Because no internal toxicity is calculated in the dose-response submodule, a time-weighted average (TWA) approach over the last days is used to calculate the daily dose to be applied. This "moving window" approach is time adjustable, often the TWA is used for the daily dietary dose of the last 21 days and converted into a toxic effect on the individual by means of a dose-effect relationship.

For both types of toxicological submodules, the calculation of toxicological effects also considers the actual body weight of the individuals.

Ecological module

Input data

Depending on the species, daily values of vegetation cover, food availability, and the soil capacity for warrens are required and are provided by the landscape model. Likewise, the spatial resolution and the size of the total area are specified there on a species-specific basis.

Individual level submodules

Separate submodules of the IBMs are used for each individual for its spatial behavior, development, reproduction, and survival.

Growth and reproduction

Individuals grow according to a growth model. Start and end of the breeding season and litter size is parametrized by probability distributions. Female individuals need a male in their home range to get pregnant.

Survival

Depending on the species, mortality is simulated with background mortality, density-dependent mortality, disperser mortality, and habitat-specific mortality. Food shortages up to starvation are not explicitly implemented but are covered by reasonably high mortality rates throughout the year.

Spatial behavior

Depending on the species, for each individual a spatially explicit home range or a daily walk of individuals is simulated. Both depend on the needs of the corresponding species, which may be food availability, shelter, and proximity to the nest or warren. This spatio-temporal behavior also essentially determines the exposure and consequently the uptake of toxic substances.

Population level

Species-specific carrying capacity of the landscape is emerging from the simulations, as dynamic population size results from density dependent mortality (e.g., by disease), territorial behavior, and the ability to build warrens.

Community level interaction

None.

Indirect effects

A reduction of population density by toxic stressors may in turn reduce species inherent density dependent processes (e.g., transmission of disease, territorial fights). This may compensate for some toxin induced reductions of density.

Feedback loops between main modules

The weight-dependent biomagnification causes feedback of the ecological module to the toxicological module: The spatio-temporal behavior of the individuals (in combination with the current toxic load of the ingested vegetation) determines the uptake of the toxic substance. Furthermore, the variable body weight of individuals affects the weight-related concentration of the toxicant in the toxicological module, both in the dose-response approach and in the toxicokinetic submodule of the TKTD approach. Therefore, individual growth or loss of body weight changes the simulated dose in the toxicological module. No feedback of animals on vegetation is implemented in the landscape module, because only in rare exceptional cases significant feedback, for example, by resource use of animals on vegetation, can it be detected at all.

Consistency between modules for important variables

In the case of hourly calculation steps of the TKTD submodule, the calculation of the daily dietary dose in the ecological module must be adapted to the reduced time step. For this purpose, the movement steps within each day, from which the residence time of the individuals in differently exposed areas is derived, are converted into hourly sections to represent the hourly dietary dose. Individual body weight is used in the ecological as well as in the toxicological module and therefore should be used in the same unit. However,

care should be taken to ensure that all parameters that depend on climatic conditions (e.g., start and end of breeding season, vegetation growth) are from the same climatic region. Similarly, the choice of a DT50 value in the fate module that is appropriate for the chosen climatic conditions is important, because the disappearance of the toxic compound is not explicitly modeled as a function of temperature.

Model parametrization

Landscape parameters need to be adapted for each case study or scenario. The parameters of the toxicological module are always species and substance specific and are usually parametrized with data derived from laboratory tests. If toxicological laboratory data of the focal species to be simulated are not available, experiments on closely related animals are used, which are often rat or rabbit studies. Independent from the toxicological module, the ecological parameters of the species-specific IBMs are usually parametrized with species-specific laboratory or region-specific semi-field and field studies from regions for which many data are available, for example, the hare with data from UK, rabbit for Spanish regions. The parametrization therefore depends on the selected region, which then also specifies the climate. Thus, also the DT50 value in the fate module must be adjusted to these climatic conditions.

Model testing and domain of applicability

All species-specific individual-based population sub-models of POLARIS have been tested with many field data published in scientific literature. These data come from all fields of ecology (reproduction, survival, spatial behavior, population dynamics). Various patterns of species biology emerging from these submodules (e.g., the proportion of pregnant females emerging from the reproductive and spatial behavior modules) were used to validate the overall ecological module. Population densities or spatial patterns simulated for individual species using IBMs in combination with the underlying landscape model could be tested using field data (e.g., by telemetry studies) in the relevant climatic-geographic context.

Independent validation of the toxicology module would require further independent studies on the same substance, which are usually not performed for animal welfare reasons.

The overall effect model (including the parametrization of the toxicological module) of a particular toxicant is also hardly testable on the population level under field conditions, because usually no field studies with toxic effects are available for the substance in question. For this very reason, the landscape-based population model is used to fill this data gap. Therefore, the only remaining approach is to test the individual modules and their sub-modules as well as possible and to check the plausibility of the overall impact model.

An uncertainty that cannot be avoided comes from the use of closely related animals for toxicological laboratory experiments and the transfer of the parameters thus obtained to the focal species.

The lack of a food web for the simulated herbivore species is not considered a deficiency because an impact of predators is covered by a density-dependent mortality.

In the POLARIS model, climatic conditions are not explicitly simulated. However, the geographical and climatic range of applicability has already been defined by the selection of the focal species and their parametrization using regional datasets, as well as by the selected values for plant growth in the landscape model.

8.3.4. Bibliography examples Chapter 5

Dohmen, G. P., Preuss, T. G., Hamer, M., Galic, N., Strauss, T., van den Brink, P. J., De Laender, F., & Bopp, S. (2016). Population-level effects and recovery of aquatic invertebrates after multiple applications of an insecticide. Integrated *Environmental Assessment and Management, 12*, 67-81.

Kleinmann, J., & Wang, M. (2017). Modeling individual movement decisions of brown hare (*Lepus europaeus*) as a key concept for realistic spatial behavior and exposure: A population model for landscape-level risk assessment. *Environ Toxicol Chem* 36: 2299–2307.

Strauss, T. (2009). Dynamische Simulation der Planktonentwicklung und interner Stoffflüsse in einem eutrophen Flachsee. Ph.D. Thesis, RWTH Aachen University, Aachen, Germany, 2009.

Strauss, T., Kulkarni, D., Preuss, T. G., & Hammers-Wirtz, M. (2016). The secret lives of cannibals: Modeling density-dependent processes that regulate population dynamics in *Chaoborus crystallinus*. *Ecological Modeling, 321*, 84-97.

Strauss, T., Gabsi, F., Hammer-Wirtz, M., Thorbek, P., & Preuss, T. G. (2017). The power of hybrid modeling: An example from aquatic ecosystems. *Ecological Modeling, 364*, 77-88.

Preuss, T. G., Hammers-Wirtz, M., Hommen, U., Rubach, M. N., & Ratte, H. T. (2009). Development and validation of an individual based *Daphnia magna* population model: The influence of crowding on population dynamics. *Ecological Modeling, 220*, 310-329.

Wang M. and Luttik R. (2013). Development of Landscape Scenarios for Population-Level Risk Assessment. Poster presented at the SETAC Europe 23rd Annual Meeting, Glasgow.

Wang M., Park S.-Y., Dietrich C. and Kleinmann J. (2022). Selection of scenarios for landscape-level risk assessment of chemicals: Case studies for mammals. *Env Sci Europe, 34*, 35.

8.4. Appendices Chapter 6: Calibration and validation of GUTS-RED-SD –Carbendazim toxicity to *Gammarus pulex*

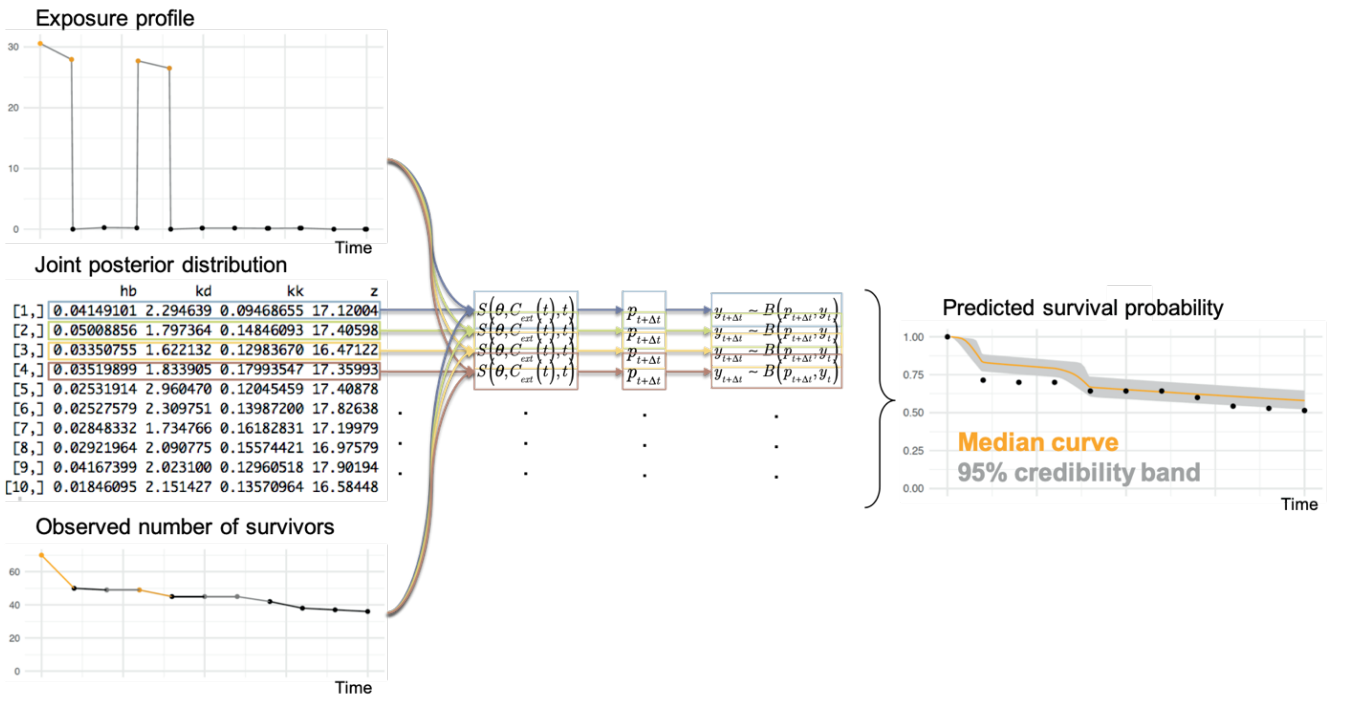


Figure 8.4: Illustration of a computational uncertainty propagation procedure.

Here, the joint posterior distribution is approximated by a set of accepted parameter combinations. By repeatedly sampling from this set and evaluating the model for each set, the parameter uncertainty can be mapped to model output uncertainty, with account for correlation between parameters.

8.5 Appendices Chapter 7

8.5.1. Visualization of distributions – Some basic principles of histograms

8.5.1.1. Histograms

Histograms are constructed by binning the data and counting the number of observations in each bin. The objective is usually to visualize the shape of the distribution. The number of bins needs to be large enough to reveal interesting features and small enough to not indicate statistical noise as trends in the data. A very small bin width can be used to look for rounding or heaping (for example in zero-inflated datasets). The effect of changing bin widths is illustrated in Figure 8.5. Common choices for the vertical scale are either bin counts (i.e., frequencies), or counts per unit (i.e., densities). The count scale is more interpretable for lay viewers. The density scale is more suited for comparison to mathematical density models. Finally, constructing histograms with unequal bin widths is possible but rarely a good idea.

8.5.1.2. Kernel density estimates

Kernel density estimates (KDEs) are a way to derive a smooth distribution from a discrete number of samples without assuming a certain parametric distribution globally. For KDEs, the type of kernel should be chosen and ideally reported. Most visualization packages will choose a Gaussian kernel by default. Other common options for KDE kernels are Tophat, Epanechnikov, Exponential, and Cosine. The choice of kernel affects the shape of the KDE estimate and might therefore affect the visual impression. Visualization of distributions by KDEs also requires the choice of a bandwidth, which determines the width of data intervals of values to which the kernel is fit. By default, most visualization software will use a heuristic such as Silverman's rule of thumb to derive the bandwidth. One should be aware that the choice of bandwidth can substantially affect the perceived width of the distribution. Also, large bandwidths might obscure multimodality, while small bandwidths might display statistical noise as multimodality (Figure 8.6). Additional care is required when KDEs are used to visualize distributions of parameters with definitive boundaries (e.g., parameters that are fractions), because the KDE estimate might very well extend beyond these boundaries.

8.5.1.3. Empirical cumulative distribution function

The empirical cumulative distribution function (eCDF) transforms a population of samples to a step-function that approximates the cumulative distribution function (CDF). The eCDF does not directly show some of the qualitative aspects of distributions (e.g., modality, kurtosis, skewness). However, when visualized with a step-line, the probability of any value being equal or smaller than a given value in the sample population can be directly read from the figure (Figure 8.7), which is particularly useful during uncertainty analysis.

8.5.1.4. Boxplots

Boxplots are extremely common to visualize empirical data, but might also be used to visualize model output, for example, to summarize distributions over a multitude of scenarios, discrete time-points, etc. However, the summary statistics that are used as defaults to draw boxes and whiskers may not always be the most appropriate to visualize model output. In particular, the whiskers usually show the 1.5-fold interquartile range (IQR). The 1.5-fold IQR is sometimes used to identify outliers, but for model output, there is no reason to assume that the 1.5-fold IQR is of any particular significance. Changing this to the minima and maxima, or a pair of relevant percentiles, is probably most appropriate. In either case, the identity of the estimate (middle line), boxes and whiskers should be reported explicitly.

Each subplot is generated with the same set of samples. The lowest bandwidth displays the most statistical noise. The largest bandwidth hides the bimodality of the distribution.

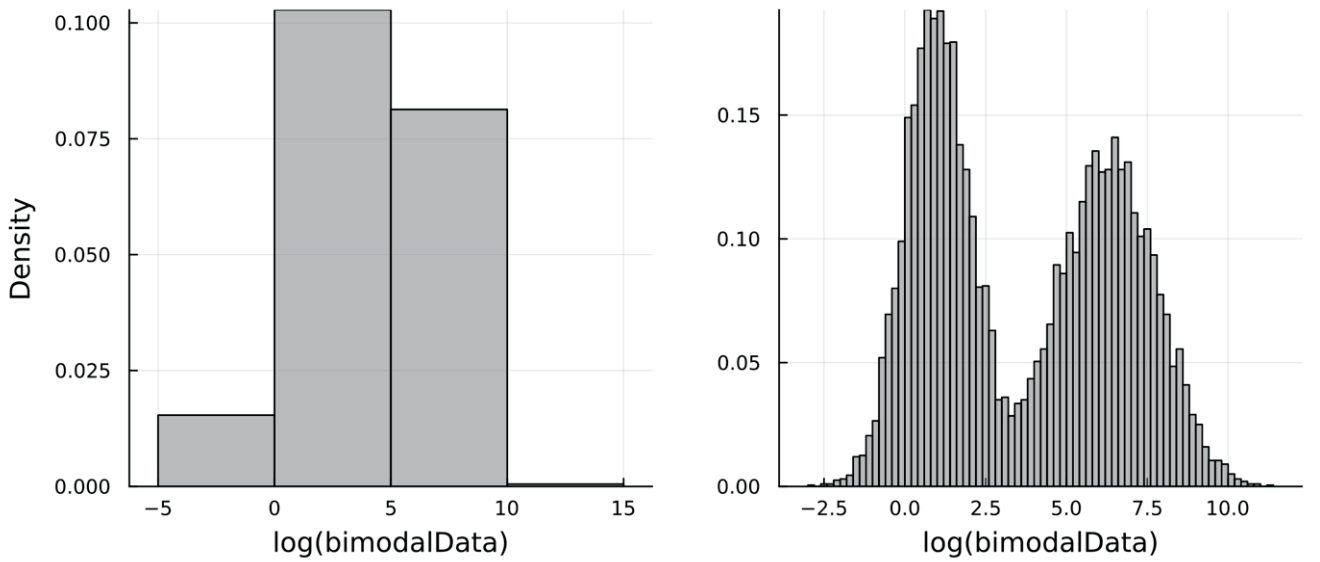


Figure 8.5: Bin widths of histograms can fundamentally alter the visualized distribution. Both plots show histograms of the same 10,000 samples from a bimodal distribution using either 6 bins (left) or 100 bins (right) to construct the histogram.

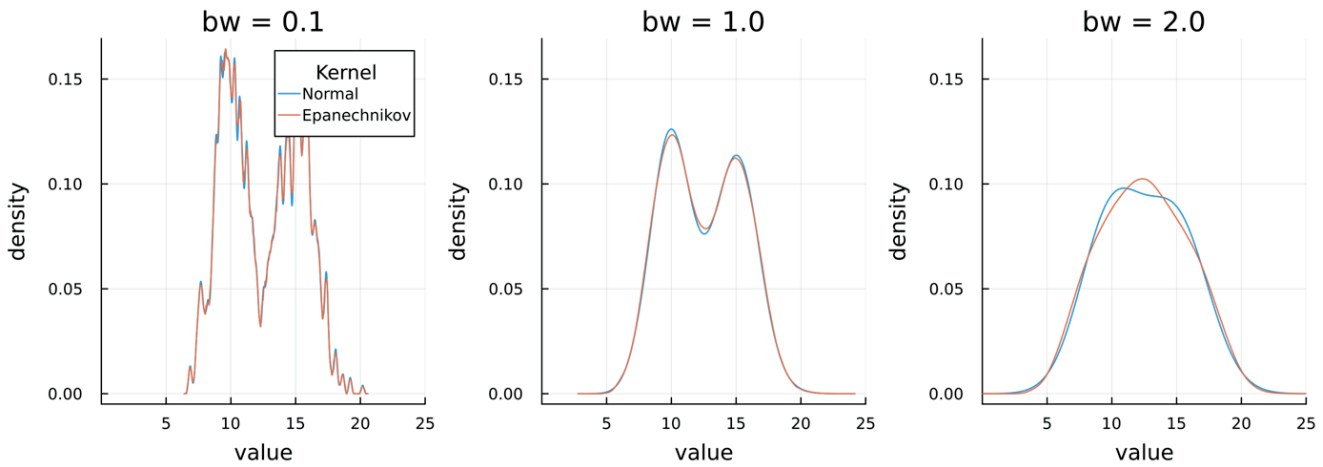


Figure 8.6: Visualizations of 10,000 samples from a bimodal distribution, using KDEs with different bandwidths (bw) and kernels (Normal, Epanechnikow).

Independent of which plot type (histogram, KDE, eCDF, boxplot) is used to visualize a distribution, it is often a good idea to either combine two different visualization types (e.g., histogram with KDE) and to add the raw data-points (e.g., histogram with rug plot, KDE with rug plot), allowing for a more flexible interpretation of the figure and to spot possible artifacts caused by the choice of data transformation. A further aspect that is independent of the plot type is the choice of the axis scale. Specifically, a distribution of log-normally distributed values will in many cases not lead to a useful visualization when displayed on a linear scale.

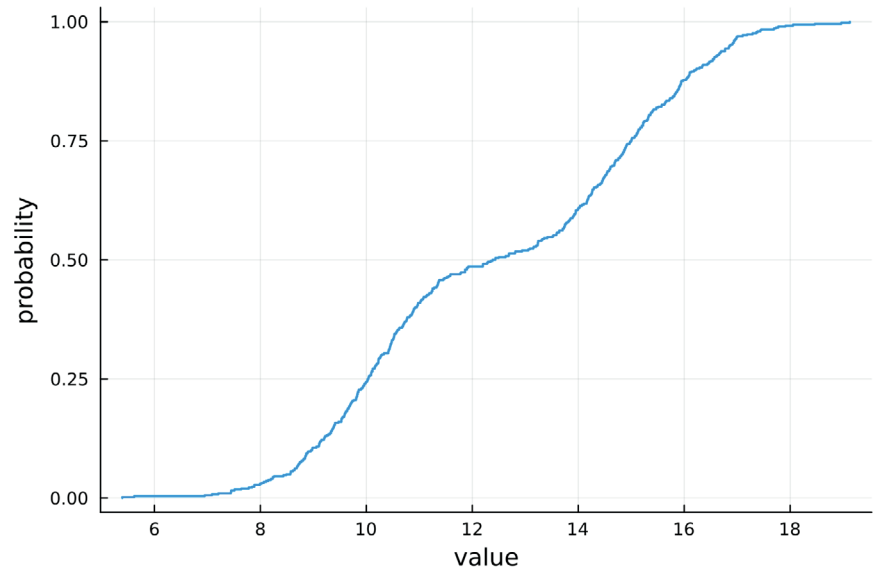


Figure 8.7: Visualization of a distribution as eCDF, using 200 samples from a bimodal distribution. The two modes, located at value = 10 and value = 15, can be identified from a steeper slope of the curve around these points.

8.5.1.5. Spatial distributions

Spatially explicit MEMs require that the distribution of organisms over space is visualized, usually on an appropriate grid. The grid resolution may be given by the spatial resolution of the model if this happens to be appropriate for visualization or chosen for the sole purpose of visualization. While individuals in an IBM might be displayed as individual points, this can easily lead to overplotting and make it hard to extract relevant information. In one – and two-dimensional continuous spaces, the same principles for visualization as previously discussed apply. A color map might be used to indicate density, in which case additional care should be taken in the choice of the color scale. It should be perceptually uniform (changes in values correspond to perceived changes in color) as well as robust to grayscale conversion and color vision deficiency (CVD). The "jet" color map is an example for a relatively widespread color map that fulfils neither of these criteria and should generally be avoided. A better choice are "viridis" and CVD-optimized versions thereof (Nuñez et al., 2018).

8.5.2. Uncertainty and Sensitivity Analysis: A practical example

Mira Kattwinkel, Stefan Reichenberger, Michail Gioutlakis

8.5.2.1. The model

Population dynamics

In this example, we use a simple deterministic population model with discrete time steps of a generic species. The model simulates logistic growth of a population with an additional mortality rate depending on temperature. Temperature is described by a sinus function with a period of year.

$$N_t = N_{t-1}\left(1 + b\left(1 - \frac{N_{t-1}}{K}\right)\right)\left(1 - d\left(\sin\left(2t\frac{\pi}{365}\right) + 1\right)\right)$$

N_t is the population size at time step t, depending on the population size of the previous time step ($N_{(t-1)}$), birth rate b, carrying capacity K, and the temperature dependent mortality rate d. This is the implementation of the model:

```
pop_model <- function(N0, K, b, d, ts){
  N <- vector(length = ts)
  N[1] <- N0
  for(i in 2:ts){
    N[i] <- N[i-1] * (1 + b  * (1 - N[i-1]/K)) *
     (1 - d * (sin(i * 2*pi/365) + 1))
  }
  return(N)
}
```

N_0 is the starting population size, and ts is the number of time steps to simulate. Hence, without any toxicant effect, this is a model of four parameters (N_0, K, b, d). The dynamic might look like this:

```
x <- 1:(3*365)
K_val <- 100
b_val <- 0.03
d_val <- 0.005
N0_val <- 50

N <- pop_model(N0_val, K_val, b_val, d_val, length(x))
par(cex = 0.8, mar = c(4,4,2,1))
plot(x, N, ylim = c(0,100),
     xlab = "time [days]", ylab = "population size",
     type = "l", lwd = 2, las = 1
     )
```

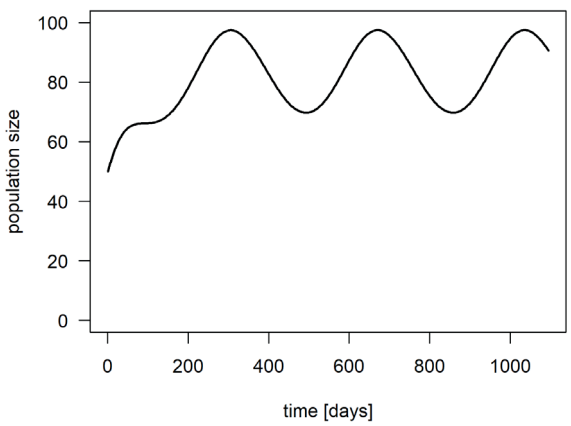


Figure 8.8: Population dynamic of the example model.

Toxicant effect

We assume that a toxicant is applied at a certain time that has an acute effect on survival, hence, reducing by a given fraction (*etox_mort*) after birth and death processes have taken place:

$$N_t^* = N_t \cdot (1 - e_{tox\_mort}) \quad \text{if} \quad t = t_{tox}$$

Additionally, we assume an effect on reproduction that reduces the birth rate by a given fraction (*etox_rep*). For simplicity, this effect lasts with a constant strength for a given number of time steps (*tstox_rep*).

$$b^* = b \cdot (1 - e_{tox\_rep}) \quad \text{if} \quad t \in [t_{tox}, t_{tox} + ts_{tox\_rep}]$$

Hence, with these four parameters for the toxic effect, this becomes a model with eight parameters (N_0, K, b, d, *ttox*, *etox_mort*, *etox_rep*, *tstox_rep*):

```
pop_model_tox <- function(N0, K, b, d, ts,
                          t_tox, e_tox_mort, e_tox_rep, ts_tox_rep){
  N <- vector(length = ts)
  N[1] <- N0
  etox <- 1
  for(i in 2:ts){
    # acute mortality
    if(i == t_tox+1){
      N[i-1] <- N[i-1] * (1 - e_tox_mort)
    }
    # effect on reproduction
    if(i %in% seq(t_tox, t_tox + ts_tox_rep, 1)){
      etox <- 1 - e_tox_rep
    } else {
      etox <- 1
    }
    N[i] <- N[i-1] * (1 + b * etox  * (1 - N[i-1]/K)) *
     (1 - d * (sin(i * 2*pi/365) + 1))
  }
  return(N)
}
```

Comparing the model with and without the toxicant effect looks like this:

```
x <- 1:(3*365)
K_val <- 100
b_val <- 0.03
d_val <- 0.005
N0_val <- 50

t_tox_val <- 400
e_tox_mort_val <- 0.4
e_tox_rep_val <- 0.3
ts_tox_rep_val <- 50

N <- pop_model(N0_val, K_val, b_val, d_val, length(x))
N_tox <- pop_model_tox(N0_val, K_val, b_val, d_val, length(x),
                       t_tox_val, e_tox_mort_val, e_tox_rep_val, ts_tox_rep
_val)

plot(x, N, ylim = c(0,100), xlab = "time [days]",
     ylab = "population size", type = "l", lwd = 2, las = 1)
lines(x, N_tox, lwd = 2, col = "red")
legend("bottomright", col = c("black", "red"), lty = 1, lwd = 2, bty = "n",
       legend = c("without toxicant effect", "with toxicant effect"))
```

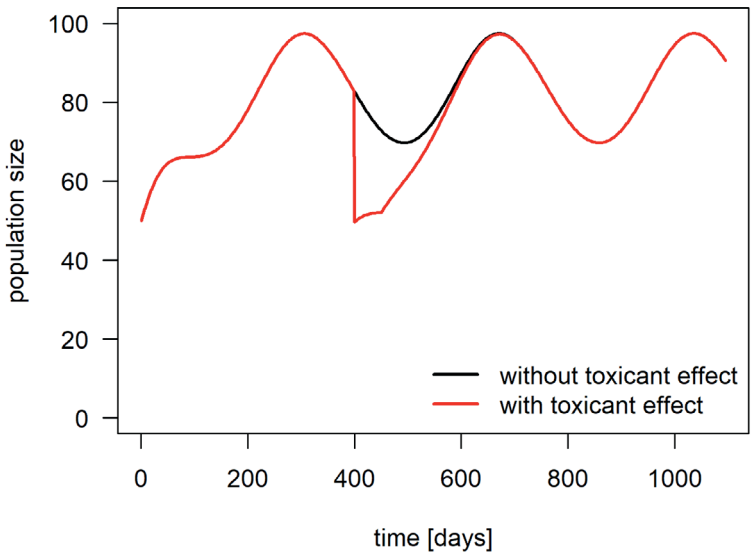


Figure 8.9: Population dynamics with and without toxicant effect.

8.5.2.2. Risk assessment

The question what effects and recovery time (if any) is acceptable in the risk assessment at hand has to be specified elsewhere, for example, at the definition of the protection goal of this specific assessment. Guidance can be found in the respective ERA guidance documents. In this simple example we assume that we accept a population reduction in the treatment down to at most 50% of that of the control population with recovery back to at least 90% of the control population size within 150 days after the treatment. In the given example, this would be the case:

```
# plot the ratio of treatment and control population size
plot(x, N_tox/N, ylim = c(0,1), xlab = "time [days]",
     ylab = "population size treatment / population size control",
     type = "l", lwd = 2, las = 1)
abline(h = 0.9, lty = 2, col = "darkorange")
a <- which(c(N_tox/N)[t_tox_val:length(x)] >= 0.9)[1]
b <- min(N_tox/N)
text(x = 520, y = 0.65, adj = c(0,0),
     labels = paste("90% of control reached\n", a, "days after treatment"))
arrows(x0 = 550, y0 = 0.8, x1 = a + t_tox_val-1, y1 = 0.9,
       length = 0.1, lwd = 2)
text(x = 150, y = 0.35,  adj = c(0,0), labels = paste("minimum relative
                              \ntreatment population size:", round(b,1)))
arrows(x0 = 370, y0 = 0.5, x1 = t_tox_val, y1 = b, length = 0.1, lwd = 2)
legend("bottomleft", bty = "n", col = c("black", "darkorange"), lty = c(1,2
), lwd = c(2,1), legend = c("relative treatment population size",
           "treatment population size = 90% of control population size"))
```

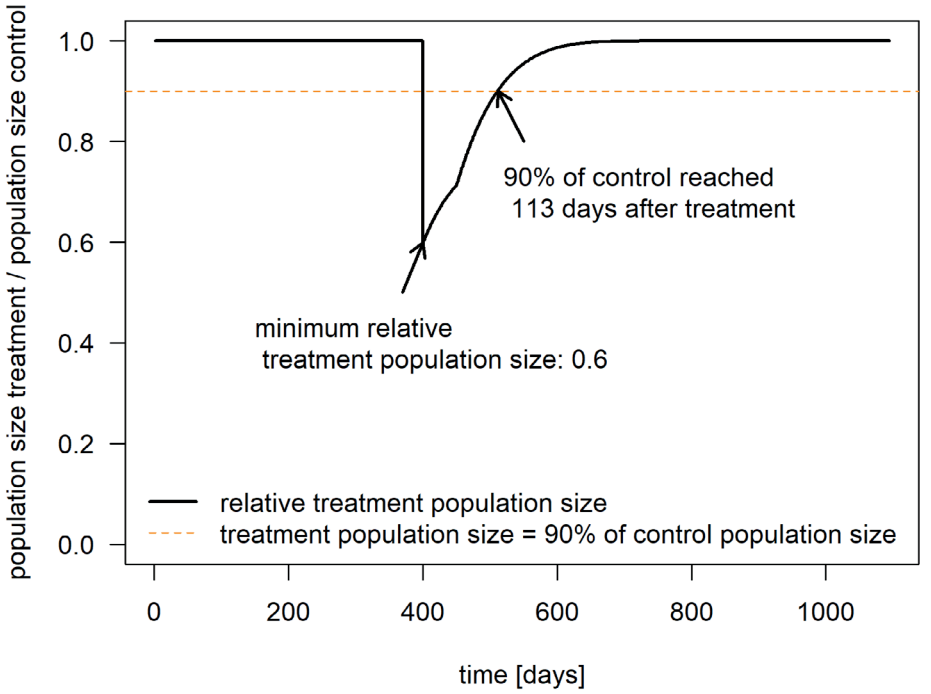


Figure 8.10: Relative population dynamic (treatment/control population size).

8.5.2.3. Uncertainty analysis

No toxicant effect, parameter uncertainty only, uniform distributions

First, we must specify the uncertainty of all parameters to be analyzed. The specification might be based on calibration, literature, or expert knowledge. A simple choice are ranges of possible parameter values. Note that such ranges are just uniform probability distributions, hence every parameter value within the ranges is equally likely while outside the range the probability is zero. In this example, we just make up the ranges. It is important to note that the process is simplified here by assuming no correlation between the parameters.

```
param_ranges <- NULL
# minimum and maximum values for each parameter
param_ranges$N0 <- c(1,100)
param_ranges$K <- c(70, 130)
param_ranges$b <- c(0.01, 0.15)
param_ranges$d <- c(0.001, 0.01)
```

Next, we take Latin hypercube samples from these ranges and let the model run many times. Note that there are more sophisticated options for sampling and that it must be ensured to take a large enough number samples. Here, we take 2000 samples.

```
# a function for the quantile functions for various distributions
# see e.g., ?qnorm
q_dist <- function(x, n, dist = "uniform", par1, par2){
  if(dist == "uniform"){
    return(qunif(x, min = par1, max = par2))
  } else {
    if(dist == "normal"){
      return(qnorm(x, mean = par1, sd = par2))
    } else {
      if(dist == "lognormal"){
        ## from ecosim::randnorm
        if (par1 <= 0) {
          warning("for 'lognormal' mean must be positive")
          return(rep(NA, n))
        }
        meanlog <- log(par1/(sqrt(1 + par2 * par2/(par1 * par1))))
        sdlog <- sqrt(log(1 + par2 * par2/(par1 * par1)))
        return(qlnorm(x, mean = meanlog, sd = sdlog))
      } else {
        warning("unknown distribution")
        return(rep(NA, n))      }
    }
  }
}

library(lhs) # for latin hypercube sampling
set.seed(13) # set seed for reproducibility
nsamp <- 2000    # number of samples
# create the samples
para <- randomLHS(nsamp, length(param_ranges))
for(i in 1:length(param_ranges)){
  para[,i] <- q_dist(para[,i], nsamp, dist = "uniform",
                     par1 = param_ranges[[i]][1],
                     par2 = param_ranges[[i]][2])
}
# run the model with all parameter combinations
N_para <- matrix(ncol = nsamp, nrow = length(x))
for(i in 1:nsamp){
  N_para[,i] <- pop_model(para[i,1],para[i,2],para[i,3],para[i,4],length(x)
)
}
```

We can plot the resulting population sizes and compare it to the outcome with the default parameter values:

```
matplot(x = x, y = N_para, type = "l", col = rgb(0,0,0,alpha = 0.03),
        lty = 1, las = 1, xlab = "time [days]", ylab = "population size")
lines(x,N, col = "red", lwd = 2)
quant <- t(apply(N_para, 1, quantile, probs = c(0.1,0.9)))
lines(x, quant[,1] , col = "royalblue1", lty = 2, lwd = 2)
lines(x, quant[,2] , col = "royalblue1", lty = 2, lwd = 2)
legend("bottomright", col = c("red", "grey30", "royalblue1"),
       lty = c(1,1,2), lwd = c(2, 1, 2), bty = "n",
       legend = c("default parameter values",
                  paste(nsamp, "sample runs"), "10th and 90th quantile"))
```

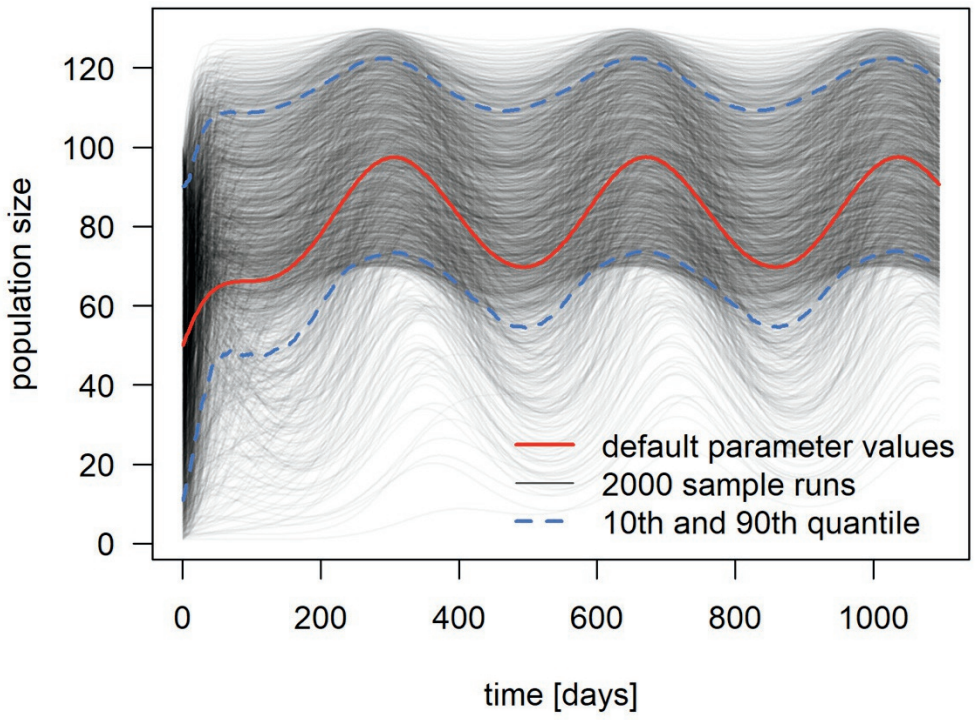


Figure 8.11: Population dynamic for 2000 sample parameters sets from uniform distributions.

Some model runs result in a larger population size compared with the default parameter values, some in smaller. Additionally, the seasonality is damped for some of the runs, resulting in oscillations with smaller amplitude.

No toxicant effect, parameter uncertainty only, different distributions

Usually, we have a bit more knowledge on the parameter values, that is to say, not all values are equally likely, but we can give distributions for the values even if they are specified with a very wide spread. We can assume that our default value is the most likely one, but other values are also possible with decreasing probability the farther away from the default value. For parameters that can only take positive values the log-normal distributions might be a good choice (but many other possibilities exist). Note that if estimates for parameter values come from Bayesian inference the joint posterior distributions (i.e., samples from the chains) would be a natural choice for these distributions. Likewise, it may be important to consider parameter covariance meaning that only certain combinations of parameter values are meaningful or are more likely than others.

In the definition below gives the distribution and and give parameters for this distribution. For this is the minimum and maximum value, and for the mean and standard deviation (on the normal scale, not on the log scale, see definition of *qdist* above). Usually, we would assume for the log-normal distributions that the default parameters values from above are also the most likely ones:

```
param_dist <- data.frame(
  parameter = c("N0", "K", "b", "d"),
  dist = c("uniform", "lognormal","lognormal","lognormal"),
  par1 = c(1, K_val, b_val, d_val),
  par2 = c(100, 20, 0.004, 0.001)
)# sample from the parameter distributions
para <- randomLHS(nsamp, nrow(param_dist))
for(i in 1:nrow(param_dist)){
  para[,i] <- q_dist(para[,i], dist = param_dist$dist[i],
                     par1 = param_dist[i, "par1"],
                     par2 = param_dist[i, "par2"])
}
# run the model
N_para_dist <- matrix(ncol = nsamp, nrow = length(x))
for(i in 1:nsamp){
  N_para_dist[,i] <- pop_model(para[i,1],para[i,2],para[i,3],para[i,4],leng
th(x))
}
```

The simulation results with log-normal probability distributions for all but *N*0 look like this:

```
matplot(x = x, y = N_para_dist, type = "l", col = rgb(0,0,0,alpha = 0.03),
        lty = 1, xlab = "time [days]", ylab = "population size", las = 1)
lines(x,N, col = "red", lwd = 2)
quant <- t(apply(N_para_dist, 1, quantile, probs = c(0.1,0.9)))
lines(x, quant[,1] , col = "royalblue1", lty = 2, lwd = 2)
lines(x, quant[,2] , col = "royalblue1", lty = 2, lwd = 2)
legend("topleft", col = c("red", "grey30", "royalblue1"), lty = c(1,1,2),
       lwd = c(2, 1, 2), bty = "n", legend = c("default parameter values",
       paste(nsamp, "sample runs"), "10th and 90th quantile"))
```

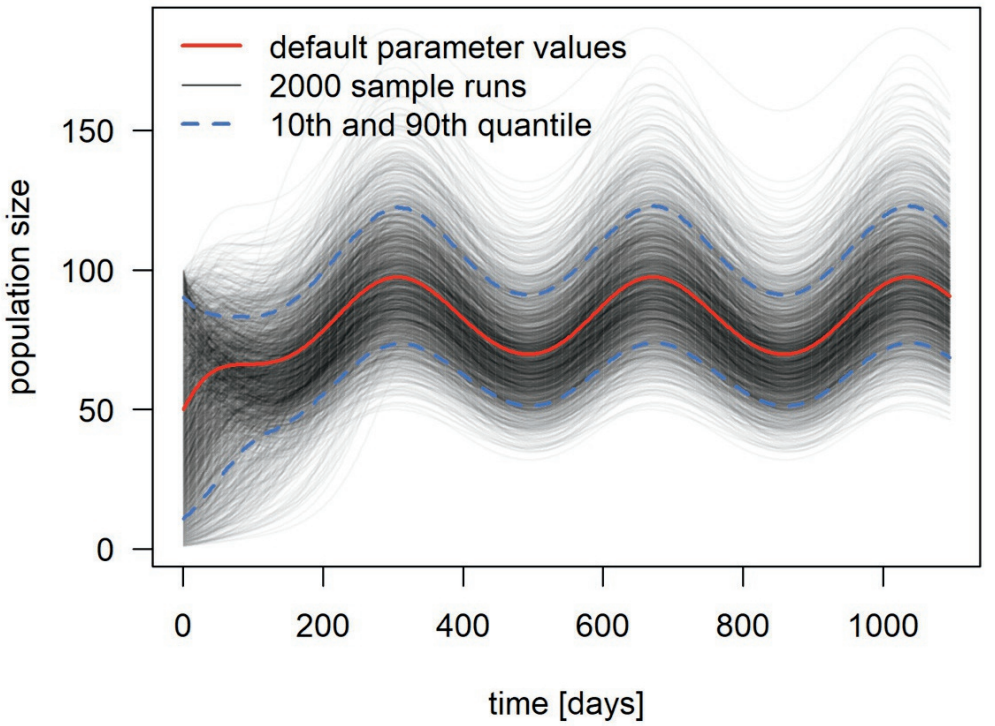


Figure 8.12: Population dynamic for 2000 sample parameters sets from different distributions.

Now the model outputs derived with 2000 parameter samples from the given distributions follow more closely that one of the default parameters because the sampled parameter values are closer to the default ones.

Compare treatment and control runs

Usually, it is not the pure population sizes that are interesting, but the comparison of treatment and control runs. For this we must make sure that the same parameter values are applied for the paired runs:

```
pop_model_tox_control <- function(N0, K, b, d, ts, t_tox,
                                  e_tox_mort, e_tox_rep, ts_tox_rep){
  N_c <- N_t <- vector(length = ts)
  N_c[1] <- N_t[1] <- N0
  etox <- 1
  for(i in 2:ts){
    if(i == t_tox+1){
      N_t[i-1] <- N_t[i-1] * (1 - e_tox_mort)
    }
    if(i %in% seq(t_tox, t_tox + ts_tox_rep, 1)){
      etox <- 1 - e_tox_rep
    } else {
      etox <- 1
    }
    N_t[i] <- N_t[i-1] * (1 + b * etox  * (1 - N_t[i-1]/K)) *
              (1 - d * (sin(i * 2*pi/365) + 1))
    N_c[i] <- N_c[i-1] * (1 + b  * (1 - N_c[i-1]/K)) *
              (1 - d * (sin(i * 2*pi/365) + 1))
  }
```

```
  return(cbind(N_c, N_t))
}

# create the samples
para <- randomLHS(nsamp, nrow(param_dist))
for(i in 1:nrow(param_dist)){
  para[,i] <- q_dist(para[,i], dist = param_dist$dist[i],
                     par1 = param_dist[i, "par1"],
                     par2 = param_dist[i, "par2"])
}

# run the model with all parameter combinations
N_para_c <- matrix(ncol = nsamp, nrow = length(x))
N_para_t <- matrix(ncol = nsamp, nrow = length(x))
for(i in 1:nsamp){
  Ni <-  pop_model_tox_control(para[i,1],para[i,2],para[i,3],para[i,4],leng
th(x),
                               t_tox_val, e_tox_mort_val, e_tox_rep_val, ts
_tox_rep_val)
  N_para_c[,i] <-Ni[,1]
  N_para_t[,i] <-Ni[,2]
}

N_para <- N_para_t/N_para_c
matplot(x = x, y = N_para, type = "l", lty = 1, col = rgb(0,0,0,alpha = 0.0
3),
        las = 1, ylim = c(0.2,1), xlab = "time [days]",
        ylab = "treatment population size / control population size")
lines(x,N_tox/N, col = "red", lwd = 2)
quant <- t(apply(N_para, 1, quantile, probs = c(0.1,0.9)))
lines(x, quant[,1] , col = "royalblue1", lty = 2, lwd = 2)
lines(x, quant[,2] , col = "royalblue1", lty = 2, lwd = 2)
abline(h = 0.5, col = "magenta", lty = 2)
abline(h = 0.9, col = "darkorange", lty = 2)
abline(v = t_tox_val + 150, col = "darkorange", lty = 2)
```

```
legend("bottomright", col = c("red", "grey30", "royalblue1"), lty = c(1,1,2
), lwd = c(2, 1, 2),
       bty = "n", legend = c("default parameter values",
                             paste(nsamp, "sample runs"), "10th and 90th qu
antile"))
legend("bottomleft", col = c("darkorange", "magenta"), lty = 2, lwd = 1,
       bty = "n", legend = c("treatment population size = 90%\nof control p
opulation size,\n150 days after treatment",
                             "maximum effect = 50%"))
```

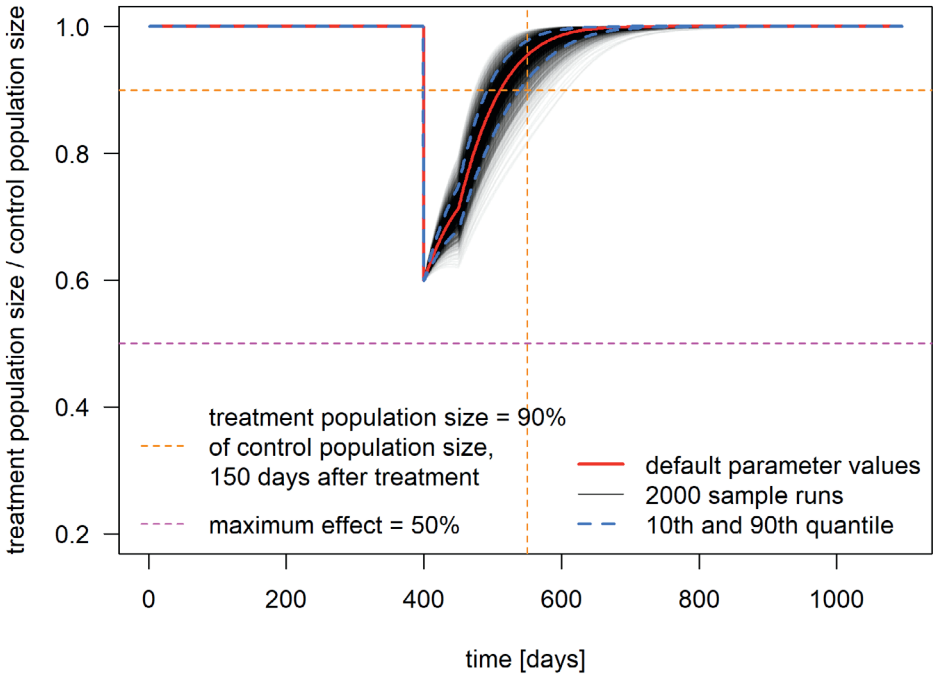


Figure 8.13: Relative population size (= treatment / control population size) for 2000 sample parameter sets (black) and the default parameter set (red).

Before, the treatment ($t < 400$) treatment and control runs were identical, hence the relative treatment population size was 1 (keep in mind that this is a deterministic model). After the treatment, some model runs show a faster recovery compared with the default parameter values, some a slower one. The lower 10th quantile (lower blue dashed line) recovers within the set time of 150 days after the treatment. Likewise, all runs result in a population reduction in the treatment of less than 50%. Hence, we may conclude that we accept the uncertainty and take the modeling results as evidence in the risk assessment. Note that we keep the toxicant effect and its timing fixed here.

Compare treatment and control, parameter uncertainty including toxicant effect parameters

Now, we also define probability distributions for the parameters describing the toxicant effect (, , ,):

```
param_dist <- data.frame(
  parameter = c("N0", "K", "b", "d",
                "t_tox", "e_tox_mort", "e_tox_rep", "ts_tox_rep"),
  dist = c("uniform", "lognormal","lognormal","lognormal",
           "uniform", "uniform", "uniform", "lognormal"),
  par1 = c(1, 100, 0.03, 0.005, 385, 0.15, 0.2, 50),
  par2 = c(100, 20, 0.004, 0.001, 415, 0.5, 0.5, 0.2)
)
```

Note that we increase the number of samples to 5000 as we deal with more parameters now.

```r
# create the samples
nsamp <- 5000
para <- randomLHS(nsamp, nrow(param_dist))
for(i in 1:nrow(param_dist)){
  para[,i] <- q_dist(para[,i], dist = param_dist$dist[i],
                     par1 = param_dist[i, "par1"],
                     par2 = param_dist[i, "par2"])
}
# round to integer for t_tox
para[,5] <- round(para[,5],0)

# run the model with all parameter combinations
N_para_c <- matrix(ncol = nsamp, nrow = length(x))
N_para_t <- matrix(ncol = nsamp, nrow = length(x))
for(i in 1:nsamp){
  Ni <-  pop_model_tox_control(para[i,1], para[i,2], para[i,3], para[i,4],
                    length(x), para[i,5], para[i,6], para[i,7], para[i,8])
  N_para_c[,i] <-Ni[,1]
  N_para_t[,i] <-Ni[,2]
}

N_para <- N_para_t/N_para_c
x_para <- matrix(ncol = nsamp, nrow = length(x), byrow=F, x)
t_diff <- para[,5] - t_tox_val
x_para <- sweep(x_para, 2, t_diff)
matplot(x = x_para, y = N_para, type = "l", lty = 1, col = rgb(0,0,0,alpha
= 0.03),
        xlab = "time [days]", ylab = "treatment population size / control p
opulation size",
        las = 1, ylim = c(0.2,1))
lines(x, N_tox/N, col = "red", lwd = 2)
quant <- t(apply(N_para, 1, quantile, probs = c(0.1,0.9)))
lines(x, quant[,1] , col = "royalblue1", lty = 2, lwd = 2)
lines(x, quant[,2] , col = "royalblue1", lty = 2, lwd = 2)
abline(h = 0.5, col = "magenta", lty = 2)
abline(h = 0.9, col = "darkorange", lty = 2)
abline(v = t_tox_val + 150, col = "darkorange", lty = 2)

legend("bottomright", col = c("red", "grey30", "royalblue1"),
       lty = c(1,1,2), lwd = c(2, 1, 2), bty = "n",
       legend = c("default parameter values",
                  paste(nsamp, "sample runs"), "10th and 90th quantile"))
legend("bottomleft", col = c("darkorange", "magenta"), lty = 2, lwd = 1,
       bty = "n", legend = c("treatment population size = 90%\nof control
     population size,\n150 days after treatment", "maximum effect = 50%"))
```

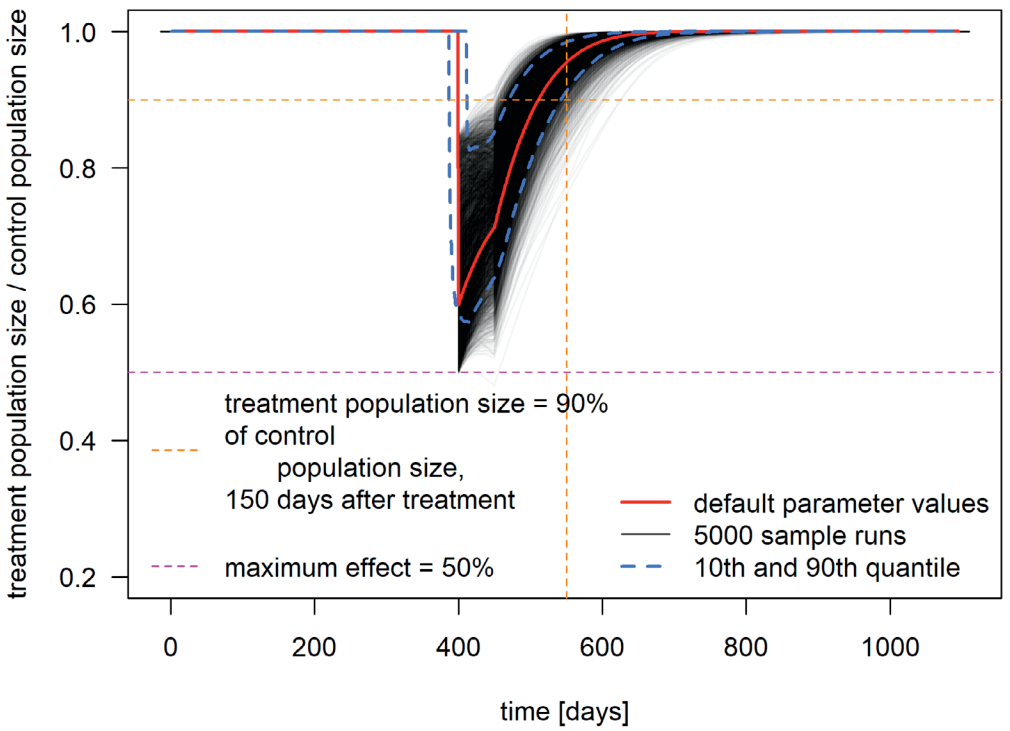


Figure 8.14: Relative population size for 5000 sample parameter sets including uncertainty in the parameters defining the toxicant effect.

Again, most of the model runs result in relative treatment populations sizes within the chosen ranges (maximum reduction of 50%) and recovery is just fast enough for the lower 10th quantile (within 150 days). Note that for the figure the runs were shifted by the difference in *ttox* to the default value of 400 to make it possible to compare all to the same threshold for recovery time. Additionally, we can have a closer look at the distribution of maximum effect and recovery time; this also shows, that most effects are smaller than 50% and most recovery times are shorter than 150 days after treatment:

```
# minimum population size
me <- sort(1- apply(N_para, 2, min))
f <- ecdf(me)
plot(f, ylab = "cummulative probability", xlab = "maximum effect", main = "
")
abline(v = 0.5, lty = 2, col = "darkorange")
abline(h = f(me)[min(which(me >= 0.5))], col = "royalblue1", lty = 2)
legend(x = 0.1, y = 0.9, col = c("darkorange", "white","royalblue1"),
       lty = 2, bty = "n", legend = c("threshold for\nmaximum effect","",
                  "fraction of sample\nruns below threshold"))
```

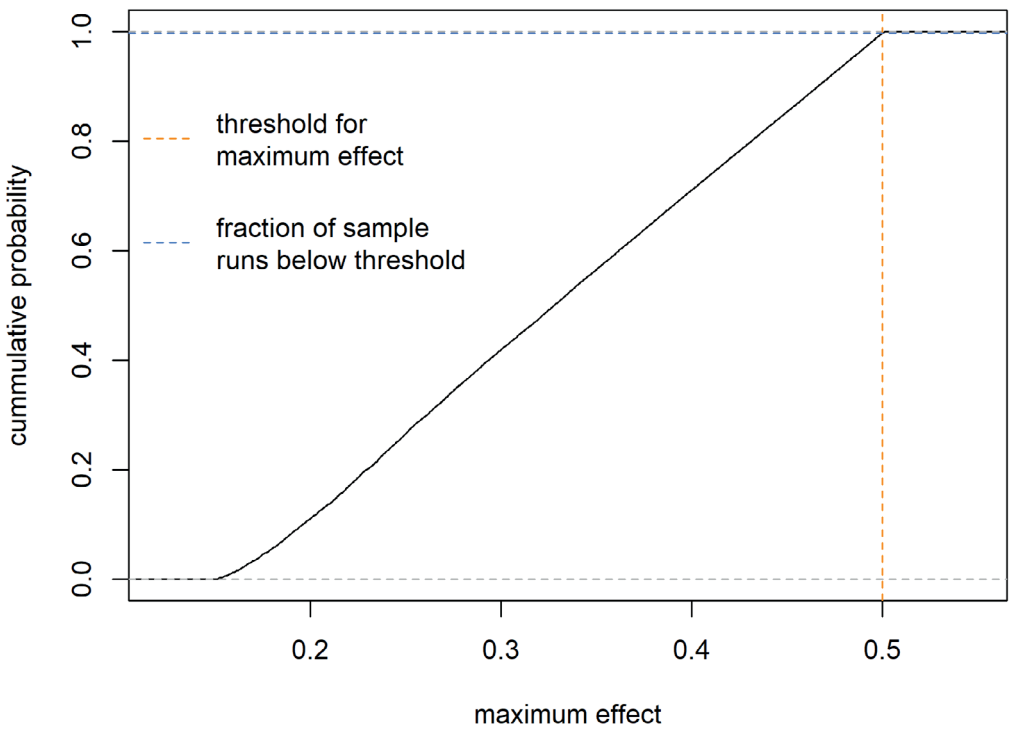


Figure 8.15: Empirical cumulative distribution function of the maximum effect for 5000 sample runs. Note that the dashed blue horizontal line is not exactly at y = 1 but at approx. 0.997; hence, only 0.3% of all runs have maximum effects above the chosen threshold of 0.5.

```
# recovery time
rt <- vector(length = nsamp)
for(i in 1:nsamp){
  rt[i] <- which(N_para[para[i,5]:length(x),i] > 0.9)[1] - (para[i,5] - t_t
ox_val)
}
f <- ecdf(rt)
plot(f, ylab = "cummulative probability", xlab = "recovery time", main = ""
)
abline(v = 0150, lty = 2, col = "darkorange")
abline(h = f(rt)[min(which(rt >= 150))], col = "royalblue1", lty = 2)
legend(x = 0.1, y = 0.9, col = c("darkorange", "white", "royalblue1"),
       lty = 2, bty = "n", legend = c("threshold for\nmaximum recovery
       time", "", "fraction of sample\nruns below threshold"))
```

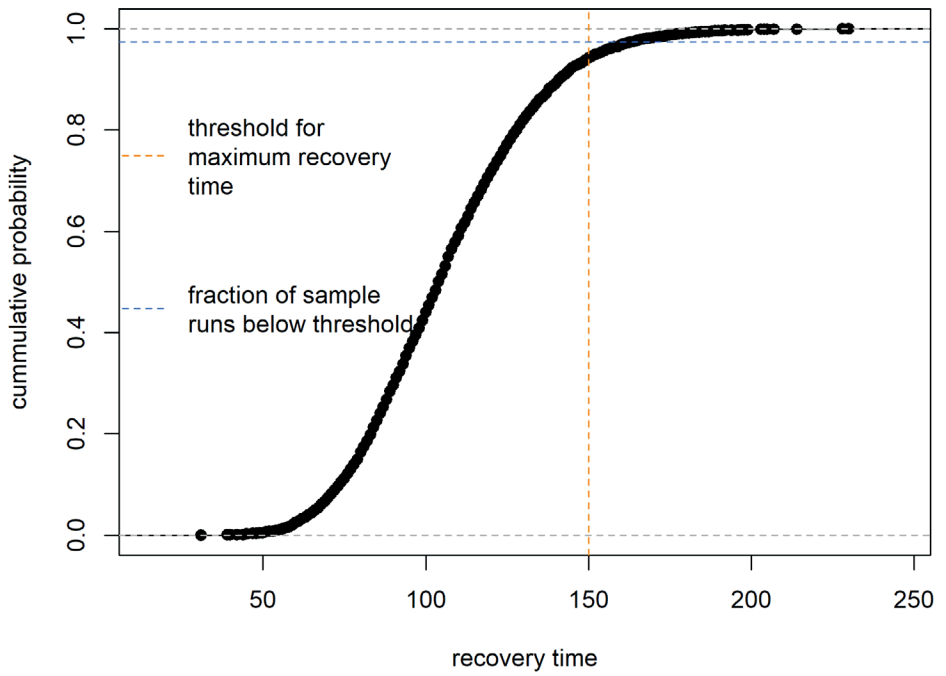


Figure 8.16: Empirical cumulative distribution function of the recovery time for 5000 sample runs.

Because the quantiles (blue dashed lines) are calculated for each time point it might be that a different 10% of the runs fulfil the recovery requirement than those fulfilling the maximum effect requirement. Therefore, we have a closer look on how many of the sample runs fulfill both criteria:

```
# minimum population size 50% of control population
me <- apply(N_para, 2, min) >= 0.5
# when is population size at least 90% of control population
rt <- vector(length = nsamp)
for(i in 1:nsamp){
  rt[i] <- which(N_para[para[i,5]:length(x),i] > 0.9)[1]
}
# should be equal or smaller than 150
rt <- rt <= 150
# how often are both criteria met
check_all <- sum(data.frame(me + rt) == 2, na.rm = TRUE)/nsamp
print(check_all)
## [1] 0.945
```

Hence, approx. 94% of all runs fulfill both requirements. Therefore, for the given sources of uncertainty here, it might be decided that the uncertainty is acceptable, and the modeling results could be taken as evidence in the risk assessment.

Stronger uncertainty in the input parameters

As an example, when input uncertainty might be too high, we show here the results for very high uncertainty in the parameters describing the toxicant effects.

```
param_dist <- data.frame(
  parameter = c("N0", "K", "b", "d",
                "t_tox", "e_tox_mort", "e_tox_rep", "ts_tox_rep"),
  dist = c("uniform", "lognormal","lognormal","lognormal",
           "uniform", "uniform", "uniform", "lognormal"),
  par1 = c(1, 100, 0.02, 0.005, 370, 0.1, 0.2, 50),
  par2 = c(100, 20, 0.004, 0.001, 430, 0.6, 0.7, 0.5)
)

# create the samples
para <- randomLHS(nsamp, nrow(param_dist))
for(i in 1:nrow(param_dist)){
  para[,i] <- q_dist(para[,i], dist = param_dist$dist[i],
                     par1 = param_dist[i, "par1"],
                     par2 = param_dist[i, "par2"])
}
# round to integer for t_tox
para[,5] <- round(para[,5],0)

# run the model with all parameter combinations
N_para_c <- matrix(ncol = nsamp, nrow = length(x))
N_para_t <- matrix(ncol = nsamp, nrow = length(x))
for(i in 1:nsamp){
  Ni <-  pop_model_tox_control(para[i,1],para[i,2],para[i,3],para[i,4],len
gth(x),
                       para[i,5], para[i,6], para[i,7], para[i,8])
  N_para_c[,i] <-Ni[,1]
  N_para_t[,i] <-Ni[,2]
}

N_para <- N_para_t/N_para_c
x_para <- matrix(ncol = nsamp, nrow = length(x), byrow=F, x)
t_diff <- para[,5] - t_tox_val
x_para <- sweep(x_para, 2, t_diff)
matplot(x = x_para, y = N_para, type = "l", lty = 1, ylim = c(0.2,1),
        col = rgb(0,0,0,alpha = 0.03), las = 1, xlab = "time [days]",
        ylab = "treatment population size / control population size")
lines(x,N_tox/N, col = "red", lwd = 2)
quant <- t(apply(N_para, 1, quantile, probs = c(0.1,0.9)))
lines(x, quant[,1] , col = "royalblue1", lty = 2, lwd = 2)
lines(x, quant[,2] , col = "royalblue1", lty = 2, lwd = 2)
abline(h = 0.5, col = "magenta", lty = 2)
abline(h = 0.9, col = "darkorange", lty = 2)
abline(v = t_tox_val + 150, col = "darkorange", lty = 2)
```

```
legend("bottomright", col = c("red", "grey30", "royalblue1"),
       lty = c(1,1,2), lwd = c(2, 1, 2), bty = "n",
       legend = c("default parameter values",
                  paste(nsamp, "sample runs"), "10th and 90th quantile"))
legend("bottomleft", col = c("darkorange", "magenta"), lty = 2, lwd = 1,
       bty = "n", legend = c("treatment population size = 90%\nof control
      population size,\n150 days after treatment", "maximum effect = 50%")
)
```

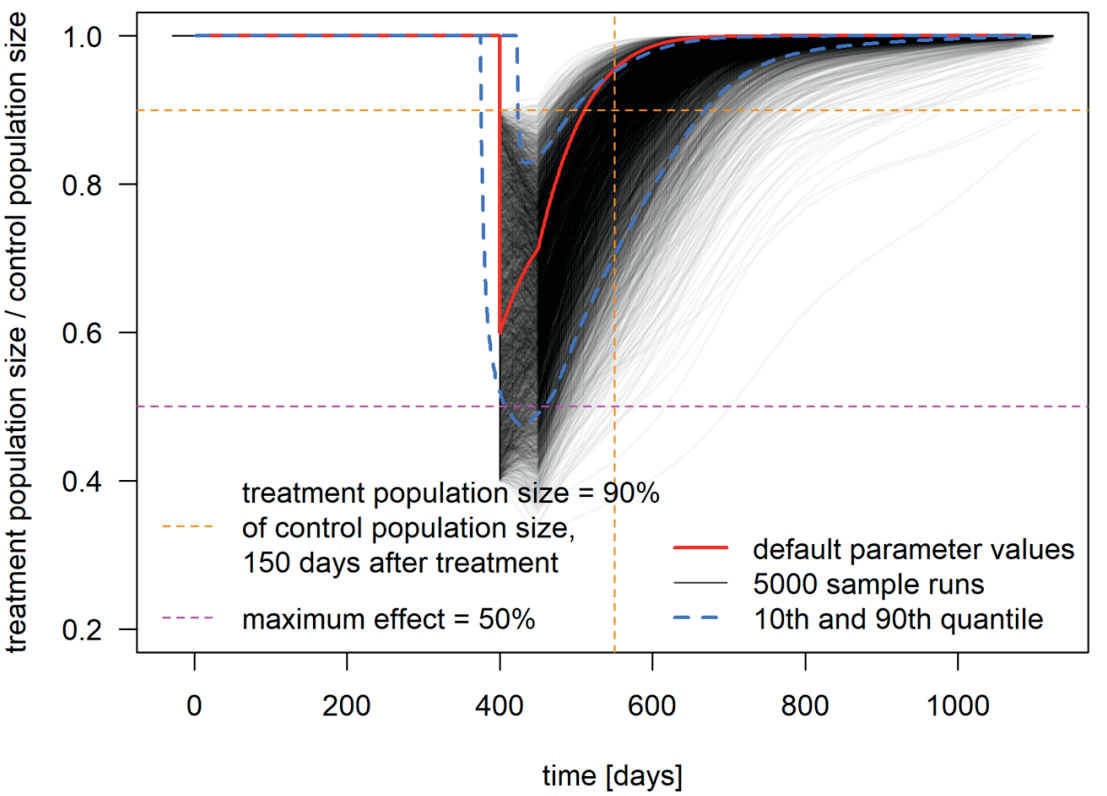


Figure 8.17: Relative population size for 5000 sample parameter sets including uncertainty in the parameters defining the toxicant effect with wider distributions for the parameters (i.e., larger uncertainty in the parameters values).

```
# minimum population size
me <- sort(1- apply(N_para, 2, min))
f <- ecdf(me)
plot(f, ylab = "cummulative probability", xlab = "maximum effect", main =
"")
abline(v = 0.5, lty = 2, col = "darkorange")
abline(h = f(me)[min(which(me >= 0.5))], col = "royalblue1", lty = 2)
legend(x = 0.1, y = 0.9, col = c("darkorange", "white", "royalblue1"), lty
= 2, bty = "n",
       legend = c("threshold for\nmaximum effect",  "",
                  "fraction of sample\nruns below threshold"))
```

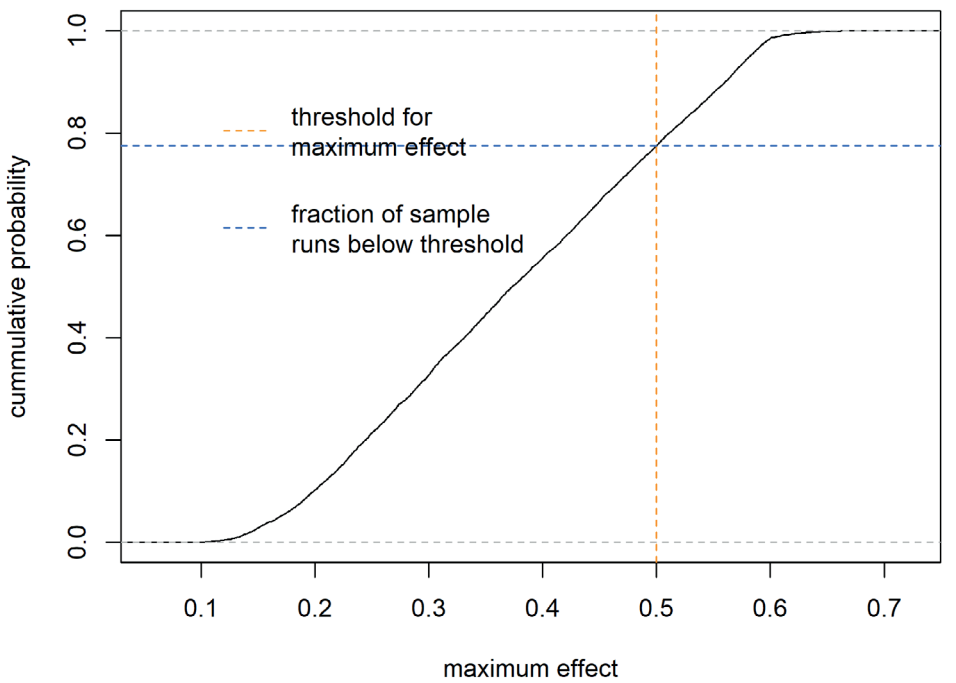


Figure 8.18: Empirical cumulative distribution function of the maximum effect for 5000 sample runs for more uncertain parameter ranges.

```
# recovery time
rt <- vector(length = nsamp)
for(i in 1:nsamp){
  rt[i] <- which(N_para[para[i,5]:length(x),i] > 0.9)[1]
}

f <- ecdf(rt)
plot(f, ylab = "cummulative probability", xlab = "recovery time", main = ""
)
abline(v = 0150, lty = 2, col = "darkorange")
abline(h = f(rt)[min(which(rt >= 150))], col = "royalblue1", lty = 2)
legend(x = 0.1, y = 0.9, col = c("darkorange", "white", "royalblue1"),
       lty = 2, bty = "n", legend = c("threshold for\nmaximum recovery
       time", "", "fraction of sample\nruns below threshold"))
```

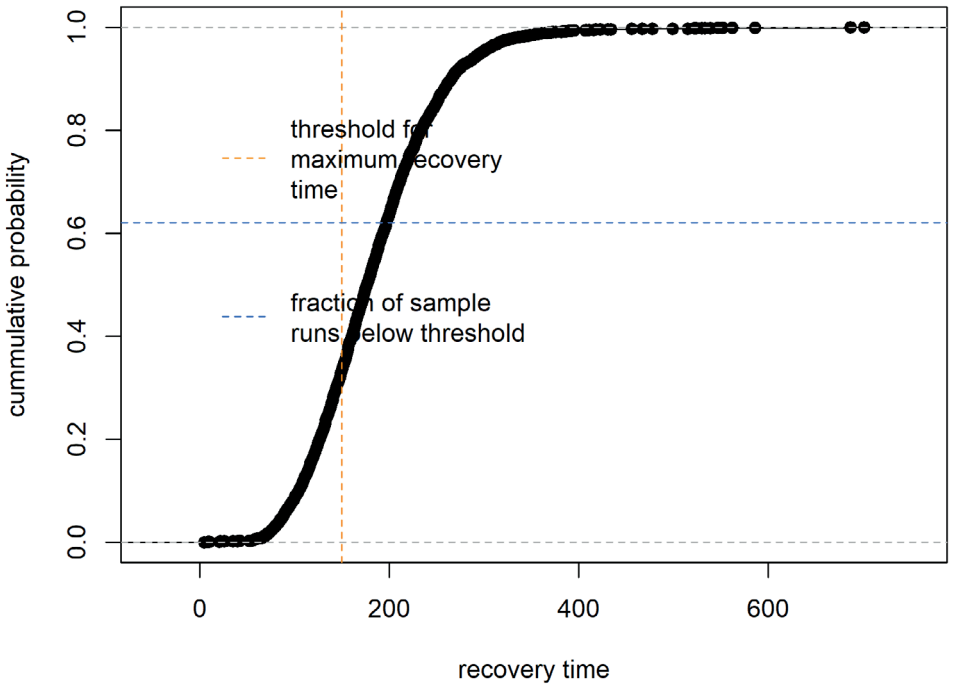


Figure 8.19: Empirical cumulative distribution function of recovery time for 5000 sample runs for more uncertain parameter ranges.

```
# minimum population size 50% of control population
me <- apply(N_para, 2, min) >= 0.5
# when is population size at least 90% of control population
rt <- vector(length = nsamp)
for(i in 1:nsamp){
  rt[i] <- which(N_para[para[i,5]:length(x),i] > 0.9)[1] - (para[i,5] - t_
tox_val)
}
# should be equal or smaller than 150
rt <- rt <= 150
# how often are both criteria met
check_all <- sum(data.frame(me + rt) == 2, na.rm = TRUE)/nsamp
print(check_all)
## [1] 0.3326
```

Hence, in this case only 33% of all model runs fulfill the requirements for acceptable population reduction and recovery time. Also, in the figure it is visible that many runs show larger effects than allowed and take much longer for recovery. In this case, accepting the modeling as evidence in the risk assessment would not appear reasonable due to the high uncertainty.

8.5.2.4. Sensitivity analysis

Rearrange the model

First, the function defining the model needs to be slightly rearranged to be able to use the functions provided in the R package 'sensitivity.'

```
pop_model_sensi <- function(X){
  n <- nrow(X)
  max.effect <- vector(length = n)
  rec.time <- vector(length = n)
  for(j in 1:n){
    N0 <- X[j,"N0"]
    K <- X[j,"K"]
    b <- X[j,"b"]
    d <- X[j,"d"]
    t_tox <- X[j,"t_tox"]
    e_tox_mort <- X[j,"e_tox_mort"]
    e_tox_rep <- X[j,"e_tox_rep"]
    ts_tox_rep <- X[j,"ts_tox_rep"]
    ts <- X[j,"ts"]
    min_rec_p <- X[j,"min_rec_p"]

    N_c <- N_t <- vector(length = ts)
    N_c[1] <- N_t[1] <- N0
    etox <- 1
    for(i in 2:ts){
      if(i == t_tox+1){
        N_t[i-1] <- N_t[i-1] * (1 - e_tox_mort)
      }
      if(i %in% seq(t_tox, t_tox + ts_tox_rep, 1)){
        etox <- 1 - e_tox_rep
      } else {
        etox <- 1
      }
      N_t[i] <- N_t[i-1] * (1 + b * etox  * (1 - N_t[i-1]/K)) *
        (1 - d * (sin(i * 2*pi/365) + 1))
      N_c[i] <- N_c[i-1] * (1 + b  * (1 - N_c[i-1]/K)) *
        (1 - d * (sin(i * 2*pi/365) + 1))
    }
    N_rel <- N_t/N_c

    # max effect
    max.effect[j] <- 1- min(N_rel)
    # recovery time
    rec.time[j] <- min(which(N_rel[t_tox:ts] > min_rec_p))
  }
  return(cbind(max.effect, rec.time))
}
```

Global sensitivity analysis using GSA

In this example, we use the function soboljansen, which calculates Sobol's sensitivity indices for both first-order and total indices (see Table 7.1). The function draws parameter samples from two provided matrices (due to the mode of operation of the algorithm). Hence, we first create two uniform matrices of parameter values. Because the range for each parameter we use the 0.01 and 0.99 quantiles of the parameter distributions (if not specified as uniform) as defined for the uncertainty analysis (the narrower distributions, i.e., lower uncertainty):

```
# parameter distributions
param_dist <- data.frame(
  parameter = c("N0", "K", "b", "d",
                "t_tox", "e_tox_mort", "e_tox_rep", "ts_tox_rep"),
  dist = c("uniform", "lognormal","lognormal","lognormal",
           "uniform", "uniform", "uniform", "lognormal"),
  par1 = c(1, 100, 0.03, 0.005, 385, 0.15, 0.2, 50),
  par2 = c(100, 20, 0.004, 0.001,415, 0.5, 0.5, 0.2)
)
nparam <- nrow(param_dist)
# get min / max for uniform distribution for sensitivity sample
# as 0.01 and 0.99 quantiles of chosen distributions
param_range <- matrix(nrow = nrow(param_dist), ncol = 2)
for(i in 1:nparam){
  if(param_dist[i, "dist"] == "uniform"){
    param_range[i,] <- unlist(param_dist[i, c("par1", "par2")])
  } else{
    param_range[i,1] <- q_dist(0.01, dist = param_dist[i, "dist"],
                               par1 = param_dist[i, "par1"],
                               par2 = param_dist[i, "par2"])
    param_range[i,2] <- q_dist(0.99, dist = param_dist[i, "dist"],
                               par1 = param_dist[i, "par1"],
                               par2 = param_dist[i, "par2"])
  }
}
# draw two  samples
nsamp <- 5000
X1 <- X2 <- matrix(ncol = nparam, nrow = nsamp)
for(i in 1:nparam) {
  X1[,i] <- runif(nsamp, min = param_range[i,1], max = param_range[i,2])
  X2[,i] <- runif(nsamp, min = param_range[i,1], max = param_range[i,2])
}
#round values for t_tox to integer
X1[,5] <- round(X1[,5],0)
X2[,5] <- round(X2[,5],0)

colnames(X1) <-  colnames(X2) <- param_dist[,"parameter"]
```

Next, we setup the algorithm. Because the model is a bit more complex, we cannot plug it into the function directly, but must use 'tell' (see help soboljansen and below).

```
library(sensitivity)
# setup an object, that will later calculate the sensitivities
nboot <- 1000
sensi_sj <- soboljansen(model = NULL, X1, X2, nboot = nboot)
# get the sample of parameter values created by the algorithm
xx <- sensi_sj$X
# add fixed values to the sample (number of simulation time steps,
# minimum ratio of treatment to control population size for recovery)
xx <- cbind(xx, 3 * 365, 0.9)
colnames(xx)[(nparam+1):ncol(xx)] <- c("ts", "min_rec_p")
```

Now, we can run the model and calculate the sensitivity indices (note that this can take a while if the parameter sample is large):

```
# run the model with the parameter sample
yy <- pop_model_sensi(xx)
# calculate sensitivity indices, i.e., tell the soboljansen-object to do so
tell(sensi_sj , yy)
```

Now we can plot the first order and total indices:

```
library(plotrix)
cap <- c("maximum effect", "recovery time")
for(i in 1:2){
  output <- dimnames(sensi_sj$S)[[3]][i]
  plotCI(x = 1:nparam, y = sensi_sj$S[,"original",output],
         col = "gray40",
         ylim = c(-0.5,1.5), xlim = c(1, nparam + 0.5), pch = 16,
         li=as.vector(sensi_sj$S[,"min. c.i.",output]),
         ui=as.vector(sensi_sj$S[,"max. c.i.",output]),
         sfrac = 0, las = 1, cex = 1.2, cex.axis = 1, xaxt = "n",
         main = cap[i],
         ylab = "first-order and total indices",
         xlab = "")
  plotCI(x = (1:nparam) + 0.4,  y = sensi_sj$T[,"original",output],
         col = "black",
         li=as.vector(sensi_sj$T[,"min. c.i.",output]),
         ui=as.vector(sensi_sj$T[,"max. c.i.",output]),
         pch = 17, sfrac = 0, cex = 1.2,
         add = T)
  axis(1, at = (1:nparam)+0.2, cex.axis = 1, labels = FALSE)
  mtext("parameters", side = 1, line = 3.5)
  text(x = (1:nparam)+0.2,
       y = par("usr")[3] - 0.15,
       labels = param_dist[,1],
       xpd = NA,
       ## Rotate the labels by 35 degrees.
       srt = 30,
       cex = 1,
       adj = 1)
  legend("topleft", c("first-order","total", "0.95 confidence intervals"),
bty = "n",
       pch = c(16,17, 124), col = c("gray40", "black", "gray20"),cex = 1)
}
```

Maximum effect is very sensitive toward the parameter e_{tox_mort}, the effect on survival. On the other hand, this model output is nearly totally insensitive toward the other model parameters. Recovery time is most sensitive toward birth rate (b), followed by the effect on survival (e_{tox_mort}) and mortality rate (d). Hence, those are also the parameters causing the largest part of the output uncertainty. If the output uncertainty is too large, one should try to quantify them with larger certainty. For both endpoints the total indices (total contribution to model output variance; direct effect + interactions) are all only very slightly larger than the first order indices (direct effect only, no interactions) meaning that there is hardly any interaction in the parameters of the model (i.e., the model is additive).

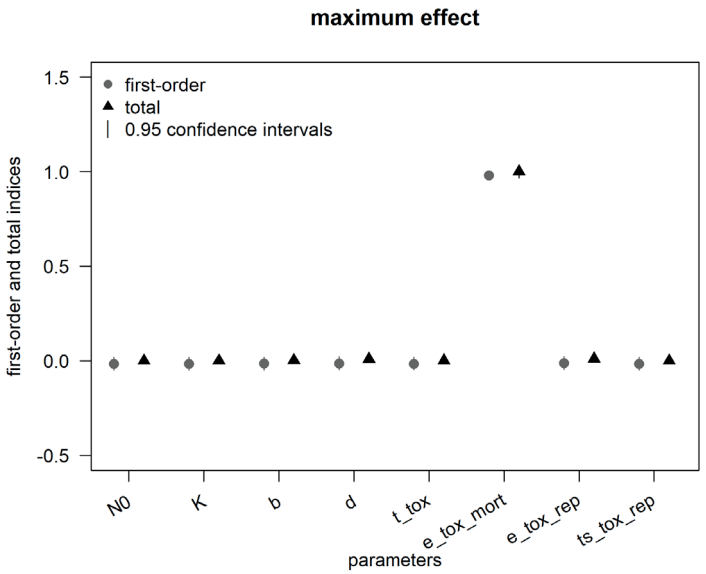


Figure 8.20: First order and total sensitivity indices for maximum effect.

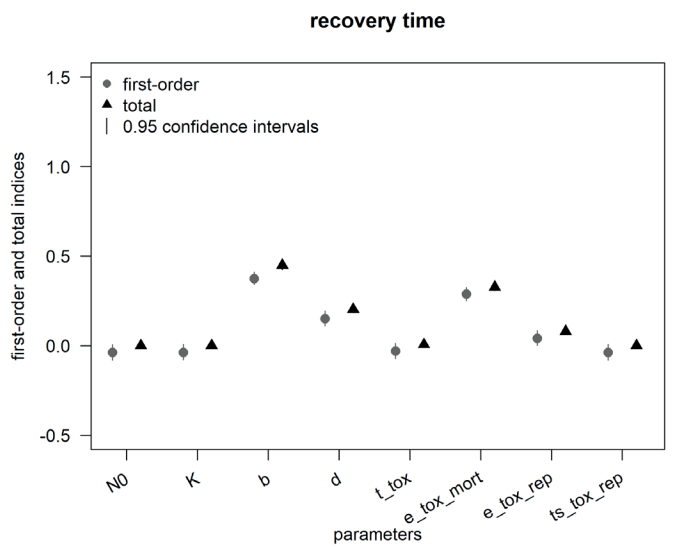


Figure 8.21: First order and total sensitivity indices for recovery time.

9. Glossary

Accuracy

The closeness of the measurements or predictions to the correct value. A high accuracy therefore indicates a low bias of predictions or measurements.

Calibration

The process of adjusting model parameters within physically defensible ranges until the resulting predictions give the best possible fit to the observed data. Often, calibration is realized using optimization algorithms. In some disciplines, calibration is also referred to as parameter estimation (EFSA PPR 2014).

Conceptual model

A hypothetical structure regarding important factors and internal connections that govern the behavior of an object or process of interest. This can be an interpretation or working description of the characteristics and dynamics of a physical system (EFSA PPR, 2014).

Deterministic model

A model that does not includes variability in model parameters or stochastic influences. The solution obtained by the model or output is therefore completely defined by a given set of parameter values and initial conditions (EFSA PPR, 2014).

Domain of applicability

Domain in which the model can be considered as "valid" (good enough) and is able to address the hypotheses that are to be tested or to make predictions that can be used for risk assessment.

Ecology module

Part of an effect model describing the behavior of the modeled system without influence of the chemical stressor.

Ecosystem model

A model describing the interactions between a community of living organisms in a particular area and its non-living environment (e.g., air, water, and soil).

Effect model

Predicts ecotoxicological effects of pesticide exposure in a modeled system.

Empirical model

Statistical model where, in contrast to mechanistic model, the model parameters have no meaning and cannot be directly measured. For example, classical concentration response models with slope and EC50.

Emergence or emergent property

Behavior in model output that is neither hard-coded, nor can be derived from model formulation without explicit simulation.

Fate model

Model that can simulate the behavior of a chemical in the environment, and possibly the spatial or temporal distribution (EFSA PPR, 2014).

General effect model

A substance-independent effect model, for example GUTS, or a model describing the dynamics of a population not parameterized for a specific toxicant. Except for substance related parameters, no fundamental recalibration of the model for a specific application is needed (see Figure 2.3).

Generic model

A model with transferable model structures (offers the possibility to re-use a model structure that has been validated before).

Inference, Bayesian

This is a way to estimate parameter values from data. This could be in calibration or for estimating particular parameter values from dedicated experiments. The estimation is based on prior knowledge on the parameter (e.g., taken from the literature, expert judgement, or previous experiments). In a nutshell, this approach uses the prior knowledge and the comparison between measured data and model results to get a new probability distribution of the parameter value called posterior distribution. Often, the parameter value with the highest posterior probability is used for further modeling. The posterior distribution also gives a measure of the uncertainty of the parameter value; the wider it is, the less certain is the value (EFSA PPR, 2014).

Inference, Frequentist

This is a way to estimate parameter values from data. This could be in calibration or for estimating particular parameter values from dedicated experiments. The goal is to find one best parameter value, defined as the one that shows the closest fit between measurements and the outcome of the simulation or statistical model. The uncertainty of this value can be described by the confidence interval (or confidence region). Most often, the error distribution (i.e., the distribution describing the random variation in the measurements) is assumed to be normal and independent for the different measurements. Prior knowledge of the parameter value (e.g., the biologically reasonable range or values from previous experiments) is not taken into account during the estimation (EFSA PPR, 2014).

Influential parameter

Parameters for which small variations significantly affect the results (Ciric et al., 2012). Non-influential parameters can be fixed at a nominal value without significantly reducing the variance of the output, thereby making the calibration step less complex (Saltelli et al., 2006).

Model

A simplification of reality that is constructed to gain insights into select attributes of a physical, biological, economic, or social system. A formal representation of the behavior of system processes, often in mathematical or statistical terms. The basis can also be physical or conceptual (NRC, 2007, cited in EFSA PPR, 2014).

Mechanistic effect model

A model describing how a biological entity (individual or population or community) develops over time and how it responds to a toxicant (EFSA PPR 2014) based on plausible ecological principles and physical laws.

Modeling framework for risk assessment

Combination of effect and fate model as well as the environmental scenario.

Modules

Parts of models that can be developed and tested independently and then combined in different models. Models can be used as modules in larger models, for example, GUTS within an IBM model.

Parameter

Constants within a simulation. Examples are uptake rate, growth rate, assimilation rate, temperature optimum, consumption rate, or prey preference. However, depending on the complexity of a model, for example, a growth rate can be a variable depending on environmental factors and/or the state of the organisms.

Parameterization

Assignment of specific values (or distributions) to model parameters. This can be done by use of literature data, measurements, plausible estimations, or calibration.

Pattern-oriented modeling

Multi-criteria design, selection, and calibration of models of complex systems (Grimm and Railsback, 2012).

Pesticide

Substance used to kill or control pests, including disease-carrying organisms and undesirable insects, animals, and plants. (https://www.efsa.europa.eu/en/search?s=pesticide). Pesticides include Plant Protection Products (PPPs) and biocides. For this book, we do not differentiate further between plant protection products and the active substance(s) they include.

Precision

Scatter (e.g., coefficient of variation) of measurements or predictions.

Problem formulation

Problem formulation is a systematic approach that identifies all factors critical to a specific risk assessment and considers the purpose of the assessment, scope, and depth of the necessary analysis, analytical approach, available resources and outcomes, and overall risk management goal.

Recovery (ecological)

The return of the perturbed ecological endpoint (e.g., species composition, population density) to its normal operating range (EFSA SC, 2016).

Regulatory model

A package consisting of the following components (i) the mechanistic exposure-effects model, (ii) programs for pre – and post-processing, often made available in the form of graphical user interfaces, (iii) model parameters, and (iv) environmental scenarios. By combining the computer model with scenarios, the model will address a certain goal and can therefore be used regulatory purposes (EFSA PPR, 2014).

Specific (regulatory) model

A model parameterized for a risk assessment for a given toxicant. Thus, including the substance-specific parameterization; for example, a GUTS model sufficiently well calibrated and validated and therefore considered suitable to predict effects of different exposure profiles.

Stochastic model

A model that relies on sampling random numbers from a probability distribution to produce model output. Stochasticity can induce individual variability or be used to simulate inherent randomness in a process (e.g., stochastic death mechanism, directed random walk).

Scenario, ecological

The combination of biotic and abiotic factors and conditions that affect the development of the modeled organisms or population, for example, weather conditions, habitat structure, food availability, or predation pressure.

Scenario, exposure

The sum of external factors affecting the exposure; for example, use pattern of a pesticide, weather conditions, velocity of a stream, or organic matter content of soil or sediment.

Scenario, environmental

The representation of the environmental context in which a mechanistic effect model is run, i.e., the exposure and ecological scenario together with the agronomic, abiotic, and biotic factors including the initial conditions.

Sensitivity (of a model)

The degree to which the model outputs are affected by changes in selected input parameters (EFSA PPR, 2014).

Sensitivity analysis

The quantification of the effect of changes in input values or assumptions (including boundaries and model functional form) on the outputs. By investigating the relative sensitivity of model parameters, a user can become knowledgeable about the relative importance of parameters in the model. (EFSA PPR, 2014).

Specific protection goals

An explicit expression of the environmental components that need protection, including a quantification of maximum impacts that can be tolerated, where and over what time period (EFSA SC, 2016).

Toxicology module

Part of an effect model that describes how the chemical stressor affects an organism.

Tiered approach

A way of organizing toxicology assessments to maximize efficiency where the lower tiers are simpler and more conservative while higher tiers are more complex and more realistic. Higher tiers can be conducted if lower tiers indicate no acceptable risk. For example, step 1 to 4 of the FOCUS exposure models refine exposure calculations if lower tiers show no acceptable risk. Another example is the assessment of aquatic effects, which ranges from standard tests with a few standard species, to species sensitivity distributions with at least eight species, to refined exposure tests and finally to mesocosm studies.

Uncertainty (of a model)

Lack of knowledge about models, parameters, constants, data, and beliefs. There are many sources of uncertainty, including the science underlying a model, uncertainty in model parameters and input data, observation error, and code uncertainty. Additional studies and collecting more information can reduce (or eliminate) uncertainty. In contrast, variability (see definition) is irreducible but can be better characterized or represented with further study (EFSA PPR, 2014).

Uncertainty analysis

Description of the variability of model outputs resulting from the uncertainty on model structure, parameter uncertainty, as well as variability. Characterizing "the uncertainty in model prediction, without identifying which assumptions [i.e., input factors, sources of uncertainty] are primarily responsible" (Saltelli et al. 2019).

Validation

The process of establishing that a model is a sufficiently accurate representation of the real world to be used for regulatory decisions. Validation assesses how well the model fits relevant data patterns, and if the model provides predicted endpoint and output values with an acceptable error for risk assessment. This last step is performed through the comparison of model or sub-model outputs with empirical data that were preferably not used for parameter estimation. (EFSA PPR, 2014).

Variable

A measured or estimated quantity that describes an object that can be observed in a system and that is subject to change (EFSA PPR, 2014). Thus, in contrast to the model parameters, variables can change over time. Two kinds of variables can be distinguished. State variables (e.g., population size or individual body mass) are dependent variables that are calculated in a model. In contrast, forcing variables (or forcing functions) correspond to time variable input data to the model (e.g., temperature, environmental concentrations of the toxicant, simple function representing predation pressure) that are not affected by the model. These input data are part of the environmental scenario.

Variability

Observed differences attributable to true heterogeneity or diversity. Variability is the result of natural random processes and is usually not reducible by further measurement or study (although it can be better characterized); (EFSA PPR, 2014).

Verification

In this context, establishing the correctness of the model implementation (i.e., that the code does what [the equations in] the theoretical model defines).

Vulnerability (ecological)

A vulnerable species is a species with a relatively high sensitivity to a specific stressor, a high chance of exposure, and/or high risks of indirect effects, plus a poor potential for population recovery (EFSA SC, 2016).