Environmental Toxicology and Chemistry—Volume 43, Number 1—pp. 197–210, 2024
Received: 24 March 2023 | Revised: 26 May 2023 | Accepted: 5 October 2023

Hazard/Risk Assessment

A Toxicokinetic–Toxicodynamic Modeling Workflow Assessing the Quality of Input Mortality Data

Barbara Bauer,[a] Alexander Singer,[a,*] Zhenglei Gao,[b] Oliver Jakoby,[a] Johannes Witt,[b] Thomas Preuss,[b] and André Gergs[b]

[a]RIFCON, Hirschberg, Germany
[b]Crop Science Division, Bayer, Monheim, Germany

Abstract: Toxicokinetic–toxicodynamic (TKTD) models simulate organismal uptake and elimination of a substance (TK) and its effects on the organism (TD). The Reduced General Unified Threshold model of Survival (GUTS-RED) is a TKTD modeling framework that is well established for aquatic risk assessment to simulate effects on survival. The TKTD models are applied in three steps: parameterization based on experimental data (calibration), comparing predictions with independent data (validation), and prediction of endpoints under environmental scenarios. Despite a clear understanding of the sensitivity of GUTS-RED predictions to the model parameters, the influence of the input data on the quality of GUTS-RED calibration and validation has not been systematically explored. We analyzed the performance of GUTS-RED calibration and validation based on a unique, comprehensive data set, covering different types of substances, exposure patterns, and aquatic animal species taxa that are regularly used for risk assessment of plant protection products. We developed a software code to automatically calibrate and validate GUTS-RED against survival measurements from 59 toxicity tests and to calculate selected model evaluation metrics. To assess whether specific survival data sets were better suited for calibration or validation, we applied a design in which all possible combinations of studies for the same species–substance combination are used for calibration and validation. We found that uncertainty of calibrated parameters was lower when the full range of effects (i.e., from high survival to high mortality) was covered by input data. Increasing the number of toxicity studies used for calibration further decreased parameter uncertainty. Including data from both acute and chronic studies as well as studies under pulsed and constant exposure in model calibrations improved model predictions on different types of validation data. Using our results, we derived a workflow, including recommendations for the sequence of modeling steps from the selection of input data to a final judgment on the suitability of GUTS-RED for the data set. *Environ Toxicol Chem* 2024;43:197–210. © 2023 Bayer AG and The Authors. *Environmental Toxicology and Chemistry* published by Wiley Periodicals LLC on behalf of SETAC.

Keywords: Toxicodynamics; Toxicokinetics; Ecological risk assessment; Calibration; Validation; General Unified Threshold model of Survival

INTRODUCTION

There is an increasing need to better capture environmental risks of chemical products in spatiotemporally heterogeneous ecosystems, particularly in freshwater habitats, where their residues are ubiquitous (Schneeweiss et al., 2022). Therefore, international institutions and the scientific community are promoting the use of mechanistic modeling approaches for environmental risk assessment (ERA), such as toxicokinetic–toxicodynamic (TKTD) effect models that show toxic effects at the individual level (European Food Safety Authority [EFSA] et al., 2018; Jager & Ashauer, 2018). Toxicokinetics (TK) encompass processes influencing the time course of toxicant concentration at the site of toxic action, such as uptake and elimination. Toxicodynamics (TD) describe how the toxicant affects the organism. The TKTD modeling framework combines a mechanistic representation of TKTD processes (the TKTD model) with advanced statistical tools for parameter estimation and uncertainty analysis.

Environmental risk assessment in the European Union uses a tiered approach whereby Tier I risk is assessed based on laboratory experiments and conservative assumptions, while at Tier II and higher, refinements can be applied. For such higher tier refinements TKTD effect models can potentially be used. They are promising tools to support an understanding of species' sensitivities (Nickisch Born Gericke et al., 2022;

This article includes online-only Supporting Information.

* Address correspondence to alexander.singer@rifcon.de
Published online 11 October 2023 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/etc.5761

Sardi et al., 2019; Singer et al., 2023), to extrapolate toxic effects between different exposure regimes (Ashauer & Escher, 2010), and to predict mixture toxicity (Bart et al., 2022; Hansul et al., 2021; Vlaeminck et al., 2021).

One of the most well-established theoretical frameworks for the TKTD modelling of lethal effects, especially for aquatic organisms, is the General Unified Threshold model for Survival (GUTS; Jager & Ashauer, 2018; Jager et al., 2011). It unifies TK and TD processes in a few equations. Toxicokinetics include the uptake and elimination of the substance, and TD covers damage accrual and repair as well as subsequent organismal effects. The representation of TK processes is often simplified because internal substance concentrations are not measured in standard toxicity tests, and the lack of information impedes the separation of uptake and elimination. Instead, in the so-called reduced GUTS (GUTS-RED), the accrual of "scaled damage" is modeled, which is dominated by the fastest process that affects scaled damage dynamics. There are two basic alternative assumptions on death mechanisms within the GUTS framework. One assumes that death is a stochastic process and that all individuals have the same chance of dying at a given damage level (stochastic death [SD]). The alternative assumption assumes that there is intraspecific variability in the tolerance to a toxicant and that individuals with low tolerance will die at low exposure (individual tolerance [IT]).

The GUTS-RED model for survival can be parametrized (calibrated) based on exposure concentrations and the number of surviving organisms over time. Such data are routinely available from regulatory ecotoxicological experiments, even though sometimes at low temporal resolution. For the application of such a model for risk assessment, a reliable calibration is needed, obtained in a well-documented and reproducible manner, including documentation of parameter uncertainty. Reliability of a model can be tested by comparing its predictions with data sets that were not used for model calibration, a process usually referred to as validation. Validation is important to increase our confidence that the model with its calibrated parameters adequately represents the actual processes in the system and to test the model's transferability to new conditions (Schuwirth et al., 2019). Several software implementations exist for automated calibration and validation of GUTS-RED (e.g., openGUTS, 2021; and "morse" [Baudrot & Charles, 2021]).

One of the main purposes of calibrating TKTD models is to make predictions for effects under a number of different exposure scenarios (Ashauer et al., 2016). In a case study, Nyman et al. (2012) found that the model calibrated on data from a constant experiment predicted the effects observed in a pulsed experiment better than vice versa. Others suggested that models calibrated on time-variable data are best suited to make predictions on time-variable data (Jager, 2014; Larras et al., 2022). These studies suggest that the type of calibration data used itself, besides the quality of the calibration, influences the predictive ability (validation performance) of GUTS-RED.

In the present study we have calibrated and validated a large number of GUTS-RED models on a comprehensive database of ecotoxicological experiments covering several substances, taxa, and experimental designs, conducted by different laboratories. This unique real-world data set allows for a more systematic investigation of the effects of experimental design factors proposed to be influential on GUTS-RED model quality in previous studies, such as temporal resolution of measurements, range of observed toxic effects across treatments, number of concentrations, and duration of the whole study as well as pulses tested (Ashauer et al., 2016; EFSA et al., 2018; Focks et al., 2018). In addition, we addressed novel questions related to how different aspects of model performance (calibration fit, parameter uncertainty, and validation performance) relate to input data and to each other. In total, we investigated five questions, also listed in Table 1: Q1: Which experimental design factors of the calibration data impact the precision of parameter estimates the most? Q2: Do experiments with short pulses provide sufficient information for parameter estimation? Q3: What is the influence of the number of experiments included in the calibration data on the uncertainty of parameter estimates? Q4: What is the relative influence of calibration quality and the similarity of the calibration and validation data in duration and exposure on validation per-

TABLE 1: Questions investigated in the study, including our initial expectations and related figures

No.	Question	Expectation	Figure
Q1	Which experimental design factors of the calibration data impact the precision of parameter estimates the most?	None (no previous comparisons available that included all factors tested here).	1
Q2	Do experiments with short pulses provide sufficient information for parameter estimation?	We expect that when toxicokinetics are slow, pulse duration must be sufficiently long to generate informative effect data.	—
Q3	What is the influence of the number of experiments included in the calibration data on parameter uncertainty?	Increasing the number of experiments may increase the precision of parameter estimation due to high-information content or may increase uncertainty of parameter estimation if the information content of the studies is not consistent.	2
Q4	What is the relative influence of calibration quality and the similarity of the calibration and validation data in duration and exposure on validation performance, measured by different GoF metrics?	A successful validation (high GoF) can be expected 1) if calibration quality is good, regardless of the type of calibration and validation data; or 2) if validation data is of a similar type as was used for calibration, regardless of calibration quality; or 3) if both of the previous conditions are satisfied.	3, 4
Q5	What is the effect of combining studies from different labs/years on GoF for 1) calibration, or 2) validation?	We expect 1) lower calibration GoF when the input data come from different labs/years; or 2) lower validation GoF when the validation and the calibration data stem from different labs/years, due to a higher chance of inconsistencies.	5

GoF = goodness-of-fit.

formance, measured by different goodness-of-fit (GoF) metrics? Q5: What is the effect of combining studies from different laboratories/years on GoF for 1) calibration, and 2) validation?

METHODS

We collected a large data set, which contained ecotoxicological survival experiments. We characterized the experiments according to their experimental design, considering factors such as experiment duration or number of tested concentration levels. The total data set contained several subdata sets, each of which was a collection of at least three ecotoxicological experiments differing in exposure patterns for a species–substance combination, as detailed in the next paragraph.

For each subdata set with a species–substance pair, we built several GUTS-RED (both GUTS-RED-SD and GUTS-RED-IT). To build a GUTS-RED for a subdata set, we calibrated the model on one part of its survival experiments (calibration data) and validated it against another independent part (validation data). This calibration/validation procedure was repeated for all possible combinations of experiments in the subdata set (Table 2), resulting in many GUTS-RED simulations. The GUTS-RED simulations were analyzed with respect to their quality. Quality metrics described the precision of estimated GUTS-RED parameters in terms of a parameter uncertainty index (PUI) as well as the GoF to calibration or validation data.

To learn how the quality of GUTS-RED is impacted by the design of ecotoxicological experiments, we related GUTS-RED quality metrics to experimental design factors. Relations were assessed visually and if possible underpinned by statistical tests.

TABLE 2: An example of the calibration–validation setup for a subdata set containing three experiments, coded as A, B, and C

Calibration	Validation
A	B
A	C
A	BC
B	A
B	C
B	AC
C	A
C	B
C	AB
AB	C
AC	B
BC	A
ABC	—

Two- or three-letter combinations mean that data were combined from several experiments to calibrate or validate a model. There are two GUTS models corresponding to each row of the table, an IT and an SD variant. When all data sets were used for calibration (here: ABC), the models were not validated, but the parameter estimates and their uncertainty were analyzed together with those of the other calibrated models.
GoF = goodness-of-fit; GUTS = General Unified Threshold model of Survival; IT = individual tolerance; SD = stochastic death.

Data

Experimental data were selected from aquatic study reports owned by Bayer and based on the criteria that for a given substance–species combination at least three different studies must be available, with at least one of them being a pulse exposure study. For the data set we aimed at a comprehensive representation of study types. The final data set used for the present study included 14 subdata sets, each comprising three to seven experiments for a specific substance–species pair, in total 59 experiments, conducted between 1991 and 2020. The experimental data set covered nine substances (two herbicides, four insecticides, and three fungicides) and nine taxonomic species (four crustaceans, three insects, and two fish).

In terms of experimental design, a roughly equal number of experiments were acute (conducted over a few days) or chronic (spanning several weeks), under constant or pulsed exposure, and conducted under static, semistatic, or flow-through conditions. The number of concentration levels tested besides control ranged from 1 to 10, and the number of individuals per concentration level ranged from 10 to 80. For the pulsed studies, one to three pulses were applied. Pulse duration was variable, ranging from short (4–8 h), to intermediate (20–24 h), to long pulses (3–7 days). No-exposure periods between pulses varied from 3 h to 13 days. Temporal resolution of the survival data was variable, with 2–88 measurement time points available for each concentration level assessed. The experiments were performed by 10 different laboratories. For more details on experimental characteristics, see Appendix 1.1 and data in the Supporting Information.

Effects observed under different exposure levels were substantially different across subdata sets, indicating differences between substance–species combinations in their TK and TD properties (Supporting Information, Figure A.I.4 and Appendix 1.2).

GUTS calibrations and validations

GUTS describe TK and lethal effects due to damage caused by a substance (Jager et al., 2011). We used the GUTS-RED variant, which links the exposure concentration directly to the damage in an individual, without explicitly modeling internal concentrations. Damage leads to lethal effects according to the assumption of IT or, alternatively, SD.

The GUTS models were calibrated using data from every single experiment included in a subdata set and in all possible combinations with other experiments in the subdata set (Table 2). This was done for both GUTS variants (GUTS-RED-IT and GUTS-RED-SD), resulting in 772 calibrated models. Validations of a GUTS-RED were conducted with studies in the subdata set that were not used for its calibration. Validations were conducted with each of these remaining studies individually or in all possible combinations, resulting in 6852 validations.

For GUTS modeling, we used an existing software implementation, the R package morse Ver. 3.3.1 (Baudrot & Charles, 2021). This package uses a Bayesian approach for parameter fitting (Billoir et al., 2008). For the IT variant, the

parameters dominant rate constant ($k_{d\text{-}IT}$), the median of the distribution of threshold values (α), and the shape parameter (β) were calibrated; for the SD variant, the dominant rate constant ($k_{d\text{-}SD}$), the threshold (z), and the killing rate constant (k_k) were calibrated. The dominant rate constant governs the speed of toxicokinetic/-dynamic processes, the parameters α and z govern the threshold concentration at which survival is affected, and β and k_k govern the strength of the effects. Background mortality rate (h_b) was fitted together with other parameters for calibrations and refitted from the controls for validations. The morse package automatically determines priors for the parameter estimate, based on the experimental design of toxicity tests under constant exposure. For other than constant exposure profiles (as contained in our data set), priors might not be optimally chosen. Because this could have introduced bias, we have refrained from a thorough analysis of prior influence on model predictions. In preliminary tests, we did not find a clear influence of priors on the prediction performance of the models. Details of the model are given in the Supporting Information, Appendix 2.

Simulations were conducted using the function "predict_Nsurv_ode" of the morse package, which uses the predicted survival probability in a stochastic binomial process to simulate the number of surviving individuals. Simulations were repeated for each of 1000 randomly selected parameter sets from the calibration posterior distribution to generate 95% credible intervals.

The exposure profiles in the input data were derived from measured concentrations in all cases, except for a study from 1991 in which concentration measurements were not available for all treatments, so nominal concentrations were used. For many arthropod species, immobility is commonly used as a proxy for mortality, but in some experiments immobile individuals were found to be mobile at a later stage. Such recovery was disregarded, because in the GUTS approach the number of dead individuals can only increase or stay constant over time. In line with standard GUTS-RED modeling practice, we used the data as provided and did not apply any rescaling of the data or parameters to account for potential differences in experimental conditions among toxicity tests.

GoF metrics

We quantitatively captured a wide range of fit quality aspects with a comprehensive set of GoF metrics (Supporting Information, Table AI.2 and Appendix 5). These were: mean squared error (MSE), normalized root-mean-square-error (NRMSE), survival probability prediction error (minimum and maximum across treatments, reflecting under- and over-prediction error, respectively [$SPPE_{min}$ and $SPPE_{max}$]), posterior predictive check (PPC), and four pseudo-R^2 metrics. We integrated information across GoF metrics into one measure of fit quality, the Average GoF. This metric ranges between 0 and 1, and larger values express a better model fit. To calculate the Average GoF, we used all metrics, except those that were redundant to others (Supporting Information, Appendix 5).

Statistical analysis of parameter uncertainty

We introduce PUI, a metric that describes the average uncertainty of parameter estimation for a given calibrated model across all three TKTD parameters (GUTS-RED-IT: $k_{d\text{-}IT}$, α, β; GUTS-RED-SD: $k_{d\text{-}SD}$, z, k_k).

$$PUI = \frac{1}{3}\sum_{j=1}^{3} \log_{10} \frac{Q97.5_j}{Q2.5_j}$$

The $\log_{10}$ of the ratio of the 97.5th to the 2.5th quantile (Q) of the marginal posterior distribution of a given parameter j expresses how many orders of magnitude are spanned by the 95% of the posterior distribution of that parameter. The larger the PUI is, the more uncertain are the parameter estimates. The PUI implicitly accounts for correlations between calibrated parameters. If parameters are correlated, their joint distribution is determined, but their marginal (i.e., separate) estimates are uncertain and increase the PUI. The background mortality h_b is not considered in PUI calculation, because its uncertainty is significantly lower than that of the other parameters. Including h_b in the calculation of PUI did not change the major results presented but decreased PUI values overall (results not shown).

The relation between PUI and experimental characteristics (Q1) was analyzed visually and with a generalized linear model (GLM) approach (gamma distributed error and log-link to account for the positive-valued response). As explanatory factors, we included independent experimental characteristics that directly relate to the information content of the effect data: study duration, number of survival measurements, number of treatments (i.e., levels of tested concentrations), fraction of treatments with intermediate mortality, and a factor specifying whether the full effect range was covered (i.e., there was at least one treatment with ≤25% and one with >75% mortality, including the control). Number of pulses were not included in this analysis due to lack of variability, because most of the treatments had two pulses (Supporting Information, Figure A.I.3). For conducting the GLM analysis, continuous explanatory variables (e.g., study duration) were scaled to reduce exceeded discrepancies among their ranges and leverages. We applied R-package glmulti (Calcagno, 2020) to conduct a model selection among all possible GLMs with first- and second-order (pairwise parameter interactions) effects, based on the Akaike information criterion (AIC). We excluded models from the automatic selection if the fitting algorithm failed. The importance of experimental characteristics and their interactions was analyzed according to their contributions to the set of the 100 best fitting models (weighted by AIC). In addition, we analyzed the influence of the ratio of exposure pulse duration and study duration (as a proxy for the ratio of exposure to nonexposure time) for studies conducted under pulsed exposure using linear regression (using the function stat_poly_eq from the R-package ggpmisc; Aphalo, 2022).

To address the impact of calibration quality and input data on validation performance (Q4), we classified calibrations into "well-calibrated" models (calibration average GoF of ≥0.6 and PUI of ≤2) and "poorly calibrated" models (calibration average GoF <0.6 or PUI >2). Validations were considered

"good" if the average GoF exceeded 0.6. These threshold values were chosen because they provided good discrimination to identify the influence of types of input data on calibration and validation. In a sensitivity analysis, slight variation of the threshold values did not affect the qualitative patterns of the analysis, indicating that the thresholds provide robust results.

The fraction of successful validations was first calculated for each calibration, then averaged for each combination of calibration and validation data type (acute/chronic/acute + chronic and pulsed/constant/pulsed + constant) for each subdata set separately, and then averaged across all data sets. This way each subdata set was weighted equally, regardless of how many experiments it contained. However, the averaged fractions are uncertain, if sample sizes are small, which was the case for some calibration/validation input data combinations for which only a few examples were available.

To compare the ability of single GoF metrics to select calibrated models with high predictive abilities, we repeated the analysis using only single GoF metrics to distinguish between "well-calibrated" and "poorly calibrated" models, calculated the average success of models falling in the two categories by data set, and compared the fraction of successful (GoF >0.6) validations between the two categories. To have comparable thresholds, we selected the threshold for "well-calibrated" models such that a fixed percentage of calibrations were accepted for each criterion, the same percentage as when the criteria calibration average GoF of ≥0.6 and PUI ≤2 were used.

Finally, we investigated the impact of using data from different laboratories or different years on calibration and validation quality (Q5). Because all experiments conducted by different laboratories were also conducted in different years, we cannot separate the effect of these two factors. For the test on calibration quality, we selected models that were calibrated on at least two experiments of the same type, conducted either in the same or different laboratories and years (only available in the subdata sets imidacloprid × *Cloeon* and trifloxystrobin × *Daphnia*). We compared average calibration GoF and PUI between calibrations based on data from different or the same laboratories. For testing the effect on validation quality, we selected validations in which the calibration data and validation data came from different or the same laboratories and years but were otherwise the same type (available in the subdata sets imidacloprid × *Cloeon*, trifloxystrobin × *Daphnia*, and fluoxastrobin × *Americamysis*), and we compared average validation GoFs between the two cases. Comparisons were done using a two-sided Wilcoxon test (stat_compare_means function from the ggpubr package; Kassambara, 2022).

On all figures including boxplots, dots are considered as outliers, lines around boxplots range from minimum to maximum values, boxplot edges are 25th and 75th percentiles, and the middle line is the median. Analysis and visualization of results was done using the package tidyverse (Wickham et al., 2019) from the software R (R Core Team, 2022), Ver. 4.2.0 in the IDE RStudio (RStudio Team, 2022).

RESULTS

First we present results related to parameter uncertainty for calibrations based on one experiment only (Q1 and Q2). Then we discuss how parameter uncertainty and calibration GoF change when data sets are added to calibration input data (Q3). Furthermore, we analyze the relationship between calibrations and validations (Q4). Finally, we investigate the effect of using experimental data from different laboratories on calibration and validation quality (Q5).

Most model calibrations could be considered successful, because 96% of calibrated models reached a GoF of ≥0.6 and 91% reached a PUI of ≤2. All models, including those with low GoF and high parameter uncertainty, were used in the following analysis because our aim was to understand which factors led to well-calibrated and which to poorly calibrated models.

Q1: Effects of experimental design on parameter uncertainty

The most important and significant experimental characteristic relating to parameter uncertainty in our data set was the range of effects observed across treatments (Figure 1A and Supporting Information, Figure AI.5 and Table AI.1). The effect strength of the factor "full effect range" in the best fitting GLM was 0.81 (Supporting Information, Table AI.1). This means that calibrating models on experiments with a full effect range compared with an incomplete effect range reduced PUI by 81%, that is, 0.81 orders of magnitude narrower posteriors.

The number of treatments was the second most important factor, although it mostly contributed via its interactions with number of measurements per treatment, study duration, and effect range (Supporting Information, Figure AI.5). For experiments with an incomplete effect range, a decrease in PUI could be observed when the number of treatments was increased (Figure 1B). Models calibrated on experiments with only two treatments and an intermediate number of measurements, which only showed an incomplete effect range, had a high PUI (Figure 1C). The PUI became lower when increasing the number of measurements above 20 (possible in chronic experiments), without increasing the number of treatments, and when increasing the number of treatments above two without increasing the number of measurements in the calibration input data. Above these thresholds PUI had no clear relationship with either number of treatments or measurements. We found a weak relationship between PUI and relative pulse duration when pulse experiments only were considered in a linear model, but only when effect range was incomplete (Figure 1D). The uncertainty ranges around the relationships shown in Figure 1B and D are wide due to the relatively small number of cases in our data set with an incomplete effect range. The explanatory variable "number of treatments with intermediate mortality" was not relevant in the model (Supporting Information, Figure AI.6).

When the GLM analysis was repeated considering only IT or SD models, there were no major differences between the results; only the importance order of the secondarily important

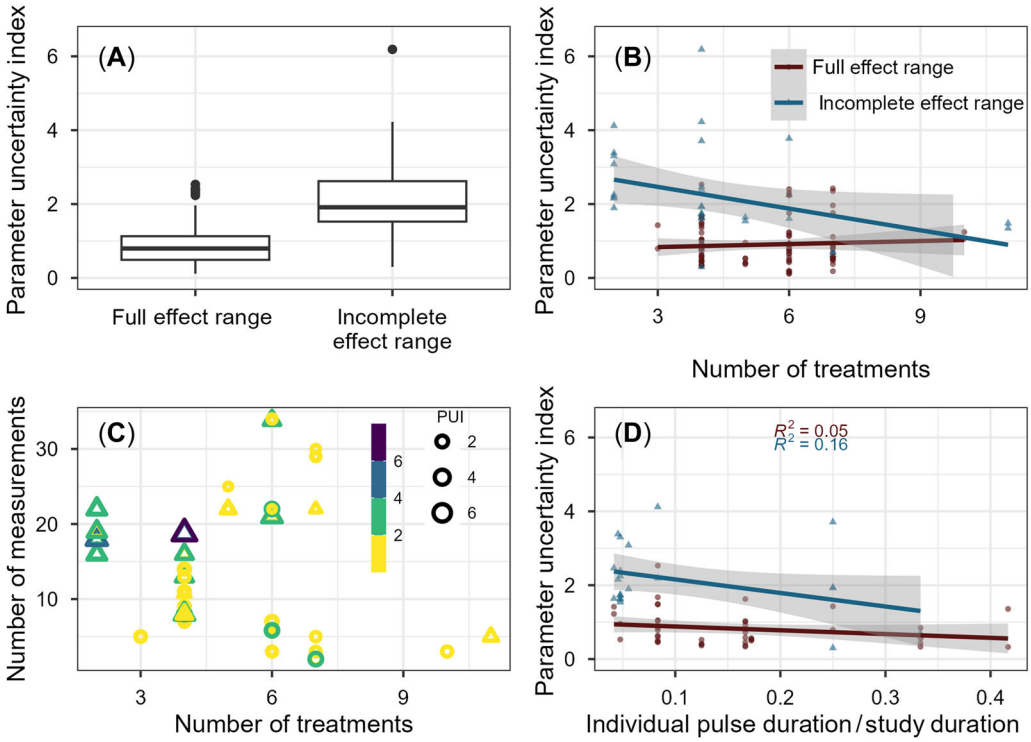

FIGURE 1: Parameter uncertainty index (PUI) as a function of (**A**) effect range in the experimental data: full effect range (at least one treatment with ≤25% and one with >75% mortality, including the control), and incomplete effect range: the lack of a treatment with high (28 cases) or low (4 cases) mortality; (**B**) number of concentration levels tested, and effect range (full: brown circles; incomplete: blue triangles); (**C**) number of concentration levels tested (*x*-axis) and temporal resolution of the data, that is, the number of survival measurements/treatment (*y*-axis). Area and color of the shapes (full effect range: circles; incomplete: triangles) are relative to PUI; (**D**) only for pulsed studies: the duration of individual pulses divided by the total study duration (shapes and colors as in [**B**]). Each dot represents one calibrated model, based on data from one experiment. Lines are linear regression lines fitted by the geom_smooth function in the ggplot2 package (Wickham, 2016); the gray areas show standard error. Two models (stochastic death [SD] and individual tolerance [IT] variants) based on the same experiment with full effect range and 75 measurements, with corresponding PUI values between 0 and 1, were removed from (**C**) for a better visibility of the lower range of values.

characteristics (number of treatments, measurements, and study duration) changed, whereas effect range remained the most important experimental characteristic to determine PUI. Therefore the two model variants were pooled in the analysis.

Q2: The impact of pulse duration on parameter uncertainty under slow kinetics

In agreement with our previous expectation, the data suggested that pulses must be long enough for a precise parameter estimation when TKTD processes are slow. We assumed slow TKTD when the median k_d was too low for damage to reach 5% of steady state by the end of experiment. Calibrated models with such low k_d estimates based on input data with a pulse duration of ≤24 h had a high parameter uncertainty (PUI >3 for SD and PUI ≥1.5 for IT), as opposed to PUI <1 for pulse durations between 72 and 168 h.

Q3: Effects of combining studies for calibration on parameter estimates and their uncertainty

Average GoF generally decreased with increasing number of experiments included in the calibration data (Figure 2A). This means that as the number of input data sets increases, it becomes harder to find a parameter set that results in a perfect fit. Nevertheless, GoF for higher numbers of experiments was still in the range of GoF for lower numbers of experiments. Furthermore, the median GoF for models based on seven experiments, the maximum in our data set, was still approximately 0.8, which was in the higher range of average GoF values overall (Supporting Information, Appendix 5). The probability for an extremely low GoF decreased when more studies were included in the input data.

Parameter uncertainty also decreased with increasing number of experiments included in the calibration data (Figure 2B and Supporting Information, Figure A.I.5). Although a very high parameter uncertainty was possible when the calibration was based on a single or just a few studies, this became more unlikely when more studies were combined. There were also a few exceptions, mostly seen when a study with already uncertain parameters was combined with other studies (Supporting Information, Appendix 3.1) or when experiments suggested significantly different parameter values (Supporting Information, Appendix 4).

Q4: Effect of calibration and validation data on validation GoF

The criteria average GoF ≥0.6 and PUI ≤2 to classify a model as "well calibrated" gave 676 (88%) "well-calibrated"

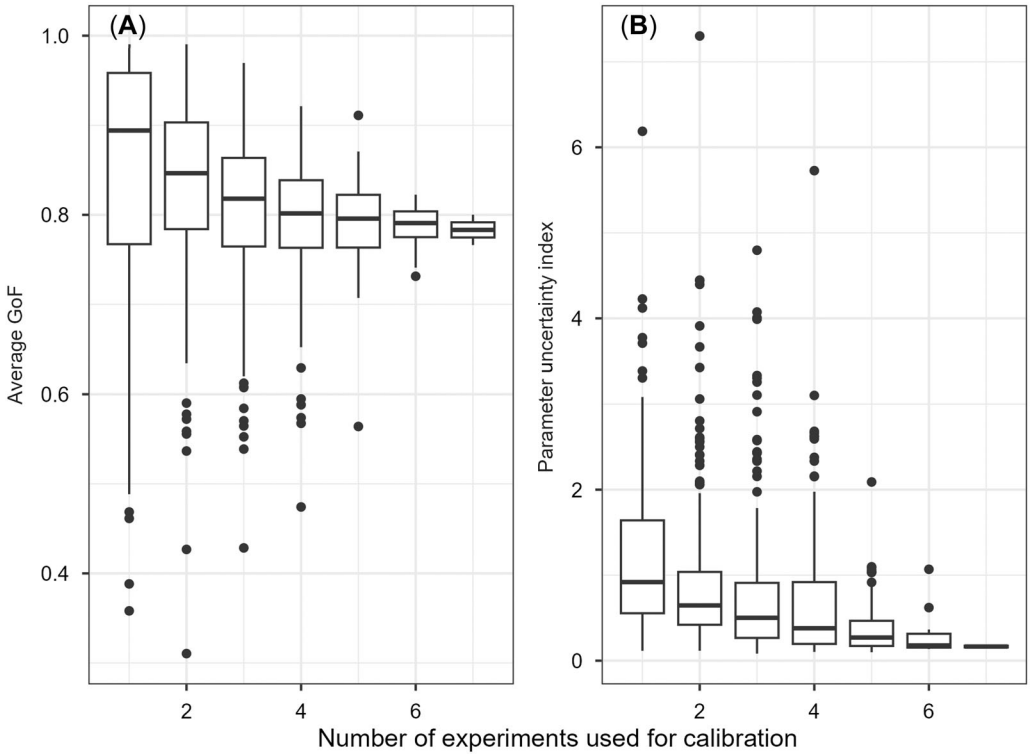

FIGURE 2: (**A**) Average goodness-of-fit (GoF) and (**B**) parameter uncertainty index (PUI) as a function of increasing the number of experiments used for calibration of a single model. Individual tolerance (IT) and stochastic death (SD) models are pooled in the figures.

and 96 (12%) "poorly calibrated" models. Thus fewer models were available to test predictive abilities of "poorly calibrated" models. "Well-calibrated" models could be generated with all combinations of input data (acute or chronic; constant or pulse exposure), in contrast to "poorly calibrated" models. When input data consisted of the combination of acute constant and pulsed studies, no resulting models were classified as "poorly calibrated" (Figure 3).

Validation success depends on both the quality of the calibrated model and the match of calibration and validation data type, that is, exposure and timescale (Figure 3). "Well-calibrated" models tended to make successful predictions (average validation GoF ≥0.6) on validation data overall, but especially when the same type of data was used for validation and for calibration (Figure 3A). Looking at single-study calibrations and validations (Figure 3A lower left corner), models calibrated on acute constant or chronic constant experiments had the same fraction of success in predicting acute pulse or chronic pulse data, respectively, as the corresponding pulsed studies, but not vice versa. Calibrations with high parameter uncertainty or poor fit to calibration data tended to lead to unsuccessful validations (Figure 3B).

Under Q4 we further expected that combining different types of data for calibration would increase the chance for a good validation, because the overlap between calibration and validation data types would be increasing. Indeed, the predictive ability of models based on only one type of data (left side of Figure 3A) was generally lower than those of combinations (right side of Figure 3A). However, the predictive ability of models based on chronic constant and pulsed data was not higher than those based on chronic constant data alone (cf. Ch_c and Ch_c+p in Figure 3A).

When only single GoF metrics were used to distinguish between "well-calibrated" and "poorly calibrated" models, the thresholds for calibrated model acceptance that resulted in 88% models classified as "well calibrated" were: NRMSE: 0.78; Nagelkerke-pseudo-R^2: 0.99; PPC: 0.65; $SPPE_{max}$: 0.62; $|SPPE_{min}|$: 0.4. In all cases, "well-calibrated" models had a higher validation success. However, the extent of difference in the predictive ability between "well-calibrated" and "badly calibrated" models varied among individual metrics used to classify the models, being the highest for NRMSE and the lowest for $SPPE_{max}$ (Figure 4).

Q5: Effects of using data from different laboratories and years on calibration and validation quality

Median calibration GoF was lower and median PUI was higher when data from different laboratories were combined for calibration compared with those based on several experiments from the same laboratory. However, the difference was not significant, at the $p < 0.05$ level (Figure 5A and B). Nevertheless, the relatively small sample size ($n = 30$) precludes drawing robust conclusions. There was also no significant difference between validation GoFs when calibration and validation data were produced by the same or different

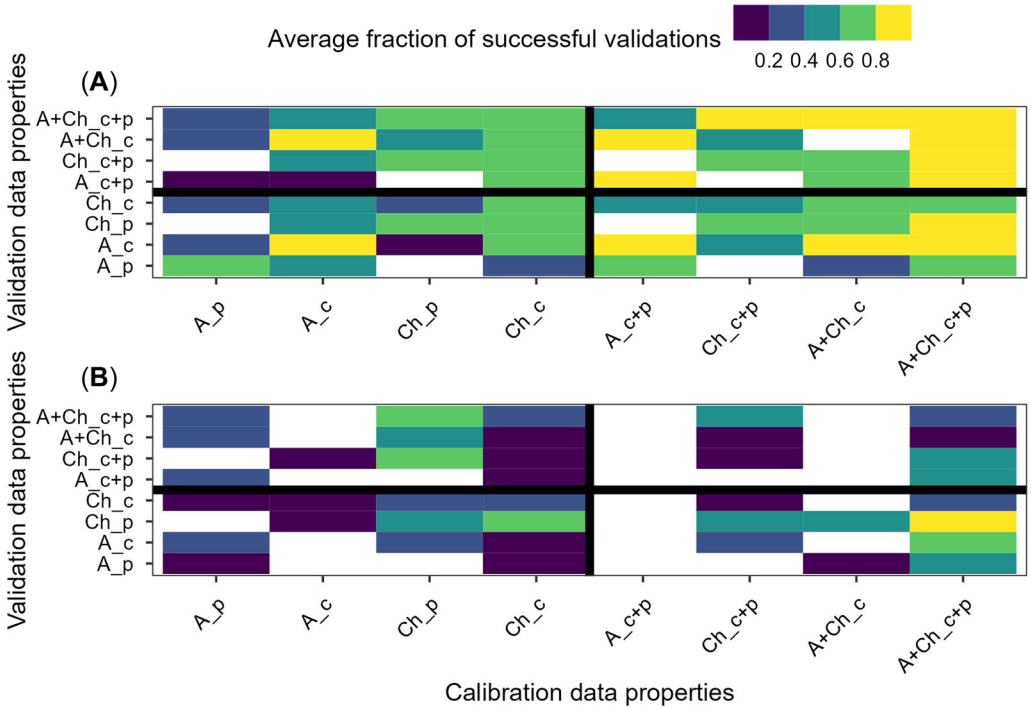

FIGURE 3: Fraction of successful validations by (**A**) "well calibrated" models (calibration average goodness-of-fit [GoF] of ≥0.6 and parameter uncertainty index [PUI] of ≤2) and (**B**) "poorly calibrated" models (calibration average GoF <0.6 or PUI >2). The horizontal and vertical thick black lines visually separate cases when the input data for calibration (*x*-axis) or validation (*y*-axis) comprised only one type of data (A: acute, Ch: chronic, c: constant exposure, p: pulsed exposure) or combinations (indicated by the "+" symbol). For example, A + Ch_c + p means that both acute and chronic, and constant and pulsed data were used in any combination.

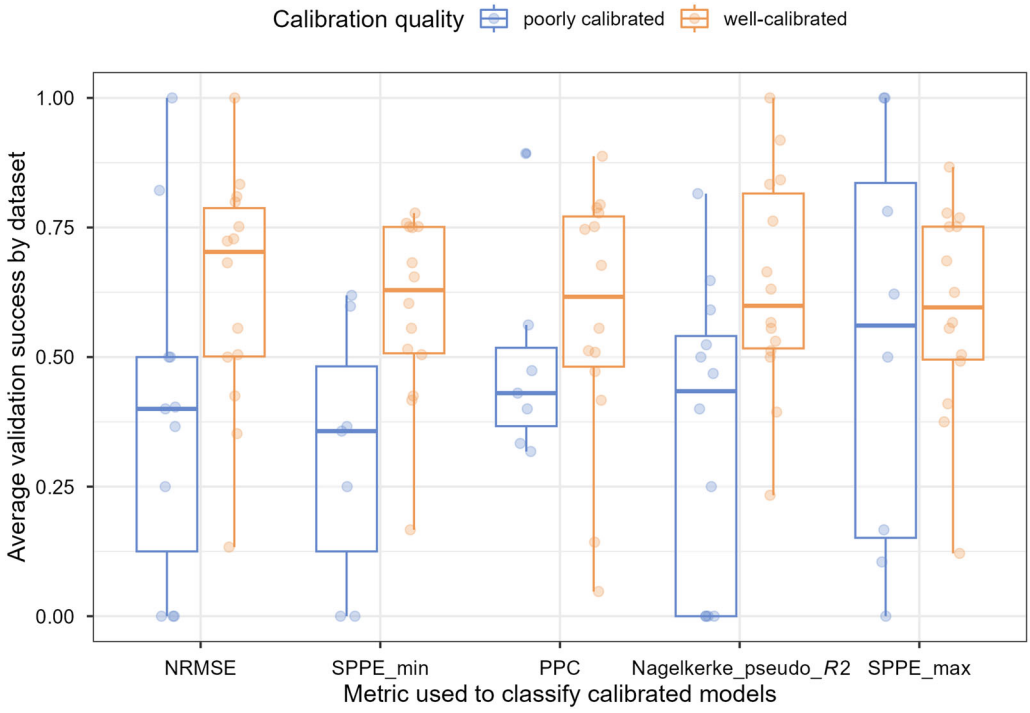

FIGURE 4: Distribution of the fraction of successful validations (fraction of validations with validation goodness-of-fit [GoF] of ≥0.6 by each calibrated model, averaged across models for each data set) by calibrated models classified as "well calibrated" (orange) and "poorly calibrated" (blue) depending on the metric used to classify models. NRMSE = normalized root-mean-square-error; PPC = posterior predictive check; SPPE = survival probability prediction error.

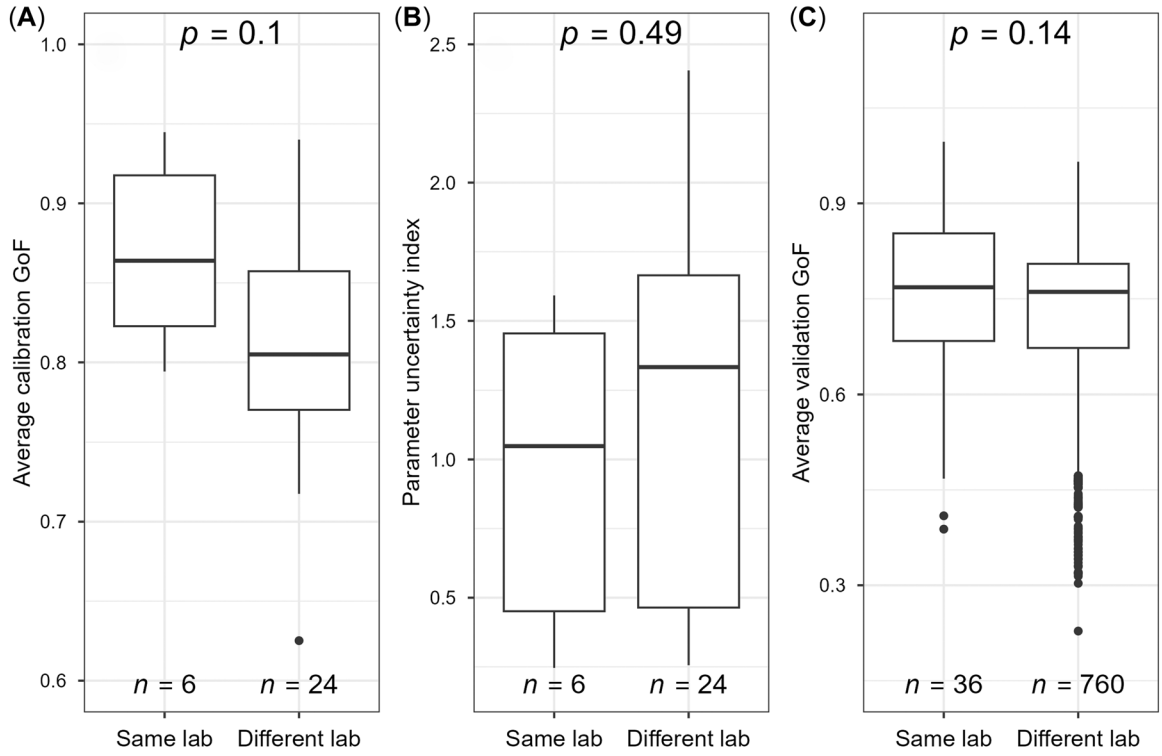

FIGURE 5: (**A**) Average calibration goodness-of-fit (GoF) and (**B**) parameter uncertainty index (PUI) for calibrated models based on input data from same or different laboratories but otherwise of the same type (i.e., combination of acute/chronic and pulsed/constant). (**C**) Average validation GoF when calibration and validation data were generated in the same or different laboratories but were otherwise the same type.

laboratories, but validation GoF seemed slightly lower in the latter case (Figure 5C).

DISCUSSION

We automatically calibrated and validated a large number of GUTS-RED on a large number of experiments and developed indices to capture parameter uncertainty and GoF. This allowed us to answer our initial five questions (Table 1) about the influence of calibration and validation data characteristics on model performance, as follows:

Q1: The most important determinant of the precision of parameter estimates was the range of effects in the input data used for calibration.

Q2: Longer pulses were associated with lower parameter uncertainty when TK was slow.

Q3: Average GoF somewhat decreased with increasing number of experiments included in the calibration input data, but generally remained high, and calibrations with extremely bad GoF became less likely; parameter uncertainty also decreased when the number of experiments was increased.

Q4: Well-calibrated models tended to make more successful predictions, especially if validation data came from a similar experimental setup as the calibration data, or at least a part of the calibration data; nevertheless, pulsed data were predicted with the same chance of success by models calibrated on constant or pulsed experiments.

Q5: No major effect of combining data from different laboratories for calibration was detectable or when calibration and validation data were generated in different laboratories compared with using data from the same laboratories.

What kind of data are needed for model calibration?

In standard experiments at Tier 1 risk assessment, a static exposure regime is used. However, at Tier 2 peak exposure can be considered, which involves a time-variable exposure whereby one or more substance pulses of constant concentrations are applied alternating with no-exposure periods. In semistatic designs, concentrations are allowed to decline or might be experimentally diluted after application.

An ideal experimental design for modeling depends on model purpose. It is common practice in ERA to apply statistical models to test for differences between control and treatment at the end of toxicity tests. This purpose is supported by the design of standard toxicity tests. However, this approach ignores temporal variation in exposure, even though time-variable exposure is closer to realistic exposure (Brock, 2009). Toxicokinetic–toxicodynamic models allow for the consideration of time-variable exposure in model predictions for ERA. Nevertheless, it has been questioned whether experiments that were designed for static exposure (like standard toxicity tests) provide suitable data to construct models that consider temporally variable exposure.

The purpose of calibration input data is to estimate parameter values that are transferable to exposure situations beyond the laboratory studies used for calibration. The more informative the data is, the less uncertain are the resulting parameter estimates, decreasing the uncertainty of model predictions as well. The EFSA recommends that for calibration of GUTS-RED using input data from aquatic toxicity studies, the number of treatment levels and experimental duration are set so that ideally zero to full effects are observed across treatments (EFSA et al., 2018). In accordance with this recommendation, our results stress the importance of a large effect range, which may not be realized in experiments conducted to define a no-effect concentration based on hypothesis-testing statistical approaches. Nevertheless, we found that parameter uncertainty was already reduced when at least one treatment showed <25% and another >75% mortality. Thus, effects from 0% to 100% are not necessary. In fact, the apparent decrease in parameter uncertainty while increasing the number of treatments may be a consequence of widening the effect range, even when the range does not reach the 25% and 75% thresholds.

Furthermore, our results suggest that if a full effect range is realized, it is not crucial for parameter estimation that other criteria such as number of treatments fully satisfy the recommendations. From a practical perspective, it is advantageous that it does not seem to be necessary to conduct experiments at many treatment levels to have a sufficient number of treatments resulting in intermediate mortality, as is necessary for dose–response modeling (Jager, 2014). The necessary condition for a precise estimation of parameters is that the data contain enough information for the model to infer the internal accumulation of damage over time. This condition can already be satisfied with few treatment levels when there are some measurement points before a large mortality is reached in at least one of the treatments, and when the concentrations applied across treatments are not too widespread. Thus, it is the interplay of various factors, most importantly effect size, as well as the resolution of concentrations and temporal measurements, that will decide the suitability of the experimental data for GUTS calibration.

The EFSA et al. (2018) suggest at least five observation points over time, based on scientific studies emphasizing the role of temporal resolution in the data (Ashauer et al., 2016; Jager, 2014). We did not find this condition to be either necessary or sufficient for precise parameter estimation. No clear relationship appeared between number of measurements over time and precision of parameter estimation. It can be argued that such a relationship would be expected, because a higher time resolution provides a more exact dose (internal concentration)–response within a treatment. On the other hand, as mentioned previously, timing of the effects can already be captured with a few measurements, and further measurements do not add precision.

When the dominant rate constant (k_d) is very small compared with the exposure time, a compound shows slow kinetics in the test species (Jager & Ashauer, 2018; Kooijman & Bedaux, 1996). This case is often difficult to handle with the GUTS-RED approach, because the data hardly contain enough information to constrain parameter k_d. Our results suggest that if data from pulsed experiments are used for model calibration, it is advisable to use long pulses to inform k_d.

How to select calibration and validation studies for a meaningful model validation

Good predictive performance on validation data is an indicator of transferability of the model, which is crucial if the model is to be used to extrapolate effects to untested conditions (Schuwirth et al., 2019). In our study, calibrated GUTS-RED models with high parameter uncertainty or low GoF generally had poor predictive ability. However, predictive ability of well-calibrated models still varied depending on the validation data.

One of the perceived advantages of TKTD models is the ability to extrapolate from effects observed under constant experimental conditions, as applied in standard toxicity tests, to potential effects under more realistic, fluctuating conditions (Ashauer & Escher, 2010). Indeed, several studies have demonstrated that models calibrated on data from experiments under constant exposure made good predictions on time-variable data (Focks et al., 2018; Nyman et al., 2012). However, some authors have suggested that data collected under time-variable exposure should be used for both calibration and validation (Jager, 2014; Larras et al., 2022). In our study, models based on constant experiments were just as successful in predicting pulsed experiments as those based on pulsed studies. This finding supports the potential usefulness of standard experiments for model calibration. Nevertheless, models calibrated on a combination of both acute and chronic data as well as constant and pulsed data had the highest rate of success in predicting effect data from all types of experimental setups. This indicates that parameter estimates of these models best approached the real underlying values. This is in line with the results of Albert et al. (2012) for the Threshold Damage Model, a precursor of GUTS-RED-SD; these authors suggested that because different experimental designs are useful to gain information about different parameters, it is best to use a combination of different experiments to inform all parameters. An additional advantage of using a combination of data for calibration is that, besides being able to predict a wider range of situations, uncertainty decreases with an increasing number of data sets used for calibration.

Despite the advantages in combining multiple experiments for calibration, typically the number of available studies is limited, and conducting new studies for the purpose of model calibration may have ethical and financial obstacles. Arguably, in data-limited cases it is important to get the best possible parameter estimates, which is achievable by using most experiments for calibration and fewer for validation. In the most extreme case, one could argue that validation is not necessary, and it is better to use all available information for calibration. Our results do not support this suggestion, because even well-calibrated models failed to successfully predict effects in independent data in some cases. It can be very informative to

investigate such cases to be able to gain insights into model transferability and its limits. It needs to be pointed out that demonstrating the transferability of endpoints derived from statistical hypothesis-testing approaches would be similarly desirable; however, it is not currently a requirement in ERA, as opposed to TKTD models.

In the context of the European ERA of plant protection products, it is especially important to stress a careful consideration of the model purpose (risk prediction for time-variable environmental exposure) when selecting the data for both calibration and validation. The EFSA et al. (2018) have suggested that validation data need to include several time-variable exposure regimes, whereas they did not specify exposure regimes for calibration data. However, our results show that using all experiments with one type of exposure (constant or pulse) for validation and all the others for calibration may result in a calibration data set that does not provide enough information for the calibration routine to find the parameter estimates that correctly describe the dynamics of the system. As just discussed, data from standard experiments may be sufficient for model calibration especially if few data are available that were collected under other exposure regimes, which then need to be used for validation. However, when possible, calibration data should include experiments that are representative of the purpose of the model, to increase the chances that the calibrated model is truly suitable for use.

Different GoF metrics are expected to capture different aspects of fit (EFSA et al., 2018). Nevertheless, metrics were similar in their ability to identify calibrated models that were more successful in validations. Only the metric measuring underestimation error, $SPPE_{max}$, performed less well. Individual metrics reflect different aspects of the model fit, which depends on accuracy and parameter uncertainty, which in turn are impacted by different aspects of the input data, as discussed previously. Therefore, it is possible that individual metrics have specific relationships to the input data. Because metrics are used in ERA to assess model suitability, it would be important to investigate whether certain experimental designs favor or handicap certain metrics. This would be possible using a data set like ours combined with a detailed analysis of the metrics' mathematical properties, which is beyond the scope of our study.

Guidance for GUTS-RED calibration and validation

Based on the insights discussed, we have developed a decision tree to aid transparent decision-making during the GUTS-RED modeling cycle. Our results provide guidance at several steps of a typical GUTS-RED modeling cycle for ERA (Figure 6). As a first step, the above considerations can help in choosing suitable calibration and validation data for the model (Figure 6, B1), in terms of number (Q3) and type (duration and exposure pattern; Q2 and Q4) of studies included. When calibration is high quality based on selected criteria reflecting GoF and parameter uncertainty, and validation results in a high GoF, the model can be considered acceptable for risk assessment (Figure 6, B2 and B3).

A calibration that results in an insufficient fit or high uncertainty in the parameter estimates could be the result of technical issues as well as problems with the data themselves

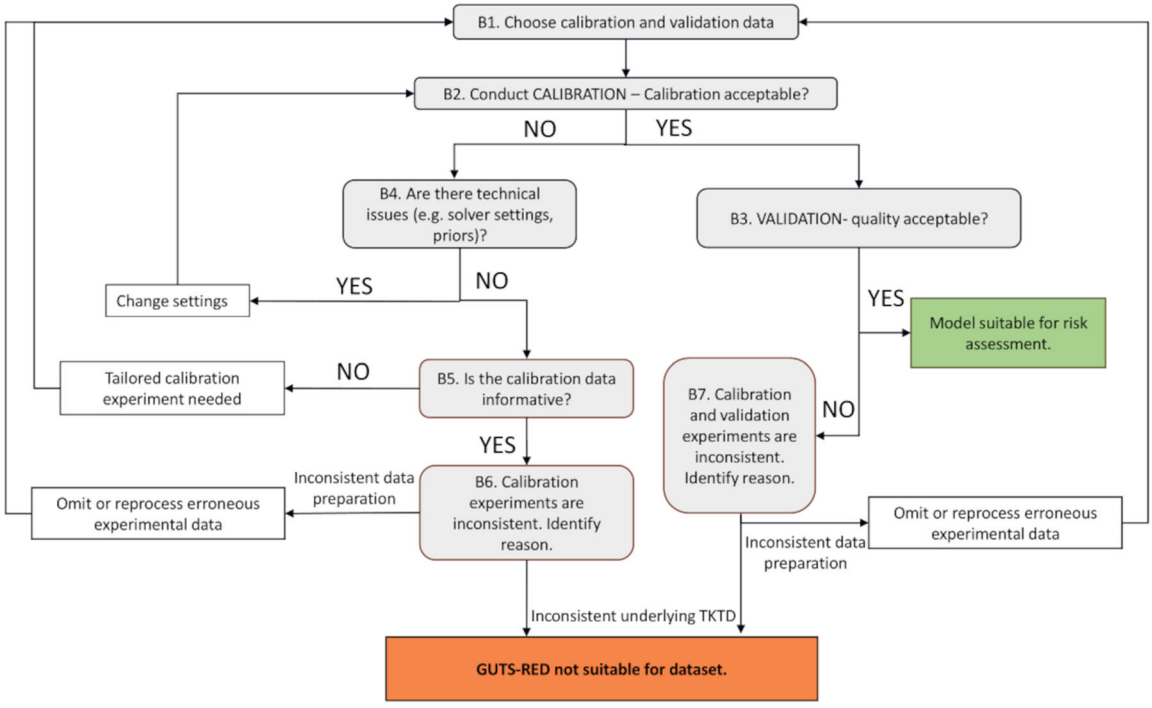

FIGURE 6: Schematic representation of the decision process for toxicokinetic–toxicodynamic (TKTD) modeling. Arrows indicate the sequence of decision steps on modeling. Rounded boxes describe questions, and rectangular boxes describe insights about the model. See explanation in the text for a more detailed explanation of the steps. GUTS-RED = Reduced General Unified Threshold model of Survival.

(Figure 6, B4 and B5). When technical issues can be excluded, an in-depth evaluation of the calibration data, the model fit, and the parameter posteriors can reveal whether the whole data set is not informative enough (Figure 6, B5). This is likely when effect range is incomplete (Q1), pulses are too short compared with kinetics (Q2), parameter estimates are uncertain, and predictions have wide uncertainty bands. In this case, calibration data need to be extended by an additional data set. If necessary and possible, new experiments can be conducted to inform the parameters that were previously uncertain. Please note that although precise parameter estimates are generally desirable, biological processes have an inherent variability, and too precise parameter estimates may give predictions with a false precision.

When the data set appears informative, but the calibration routine is still not able to find a good fit and precise parameter estimates, it is likely that individual experiments are inconsistent (Figure 6, B6). This could be the result of variability between laboratories (Q5). For example, for some taxa in aquatic risk assessment, immobility is considered as mortality. Immobility signs may be interpreted differently between laboratories, leading to relatively more or less apparent mortality at similar exposures compared with one another. Such issues can only be identified by going back to the original experimental protocols after inspecting the model fits. In these cases, the inconsistent experiments can be removed from the calibration data, or, if possible, the raw data can be reprocessed in a manner consistent with other studies. However, if variability in data processing and other human errors can be excluded and the standard GUTS-RED still fails to calibrate well on the input data, then this approach is not suitable to model the data set. A simple reason can be variability in conditions, such as temperature (Huang et al., 2023) or size of individuals used, which would need adjustments to the model structure (Gergs et al., 2015; Mangold-Döring et al., 2022) or additional assumptions on the mathematical relationship among temperature, size, and TKTD parameters to rescale the data (Gergs et al., 2019; Rakel et al., 2022). Other reasons are more complex, but modeling can help to pinpoint them. For example, the substance may have different modes of action for the species when exposure is short versus long (see Gergs et al., 2021). This would be indicated if the input data contain experiments with different exposure durations, and if the parameter estimation gives very different results when experiments are used as calibration input in isolation and fails to converge or gives a bad fit to some of the studies when they are combined (see the Supporting Information, Appendix 4, for an example). Some cases of TKTD are hard to capture using GUTS-RED, such as a fast uptake coupled with slow damage repair dynamics. In these cases, more complex TKTD models are needed for modeling the environmental risk of the substance. Similar considerations apply to identifying potential reasons for inconsistency between calibration and validation data (Figure 6, B7).

CONCLUSIONS

We tested GUTS-RED calibrations and validations based on realistic variability of study designs, substances, species, laboratories, and experimental data (dose–effect relationships). To our knowledge this is the first time that such extensive research has been performed on this topic. The systematic analysis of GUTS-RED performance allows data-driven conclusions on the impact of calibration and validation data on model performance, complementing previous insights based on expert knowledge. However, because design factors usually covaried, it was difficult to completely disentangle their effects, for example, pulse duration and experimental duration. This precluded strong conclusions about the effect of one specific aspect of experimental design on parameter estimation. Creating calibrations and validations based on artificial data would be a promising approach to investigate the effects of one variable systematically, while keeping variation in other factors minimal (see Albert et al., 2012; Ashauer et al., 2016).

During the last decade, the use of models in pesticide risk assessment has been increasing (Forbes et al., 2011; Schmolke et al., 2010), but is still hindered by mistrust (Hunka et al., 2013). Following a decision tree such as the one we present makes data selection and modeling decisions more transparent, especially in combination with documentation of the steps (Ayllón et al., 2021), and thus will lead to more reliable and interpretable models.

Supporting Information—The Supporting Information is available on the Wiley Online Library at https://doi.org/10.1002/etc.5761.

Acknowledgments—The work of B. Bauer, A. Singer, and O. Jakoby was funded by Bayer. We thank D. Nickisch for help with data curation, T. Martin for his comments on the manuscript, and anonymous reviewers for their helpful suggestions.

Conflict of Interest—B. Bauer, A. Singer, and O. Jakoby are employees of RIFCON. Z. Gao, J. Witt, T. Preuss, and A. Gergs are employees of Bayer. Bayer is a manufacturer of the active substances investigated in the present study.

Author Contributions Statement—**Barbara Bauer:** Methodology; Software; Formal analysis; Data curation; Writing—original draft; Visualization; Project administration. **Alexander Singer:** Methodology; Software; Formal analysis; Data curation; Visualization; Project administration. **Zhenglei Gao:** Methodology. **Oliver Jakoby:** Supervision; Project administration; Funding acquisition. **Johannes Witt:** Methodology; Visualization. **Thomas Preuss:** Supervision; Visualization; Funding acquisition. **André Gergs:** Data collection; Methodology; Resources; Supervision; Visualization; Project administration; Funding acquisition. All the authors contributed to Conceptualization and to Writing—review & editing.

Data Availability Statement—All input data as well as the R packages calvalr and calvalrmorse are provided in the Supporting Information. Other scripts necessary to reproduce the figures and analyses presented in the manuscript are available from the corresponding author on request (alexander.singer@rifcon.de). The experimental study reports

and corresponding M-numbers are listed in the Supporting Information, Appendix and are available on request by sending an email specifying the requested M-numbers to transparency@bayer.com.

REFERENCES

Albert, C., Ashauer, R., Künsch, H. R., & Reichert, P. (2012). Bayesian experimental design for a toxicokinetic–toxicodynamic model. *Journal of Statistical Planning and Inference*, *142*(1), 263–275. https://doi.org/10.1016/j.jspi.2011.07.014

Aphalo, P. J. (2022). *ggpmisc: Miscellaneous extensions to ggplot2*. https://github.com/aphalo/ggpmisc

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., Schäfer, 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*(1), 29178. https://doi.org/10.1038/srep29178

Ashauer, R., & Escher, B. I. (2010). Advantages of toxicokinetic and toxicodynamic modelling in aquatic ecotoxicology and risk assessment. *Journal of Environmental Monitoring*, *12*(11), 2056–2061.

Ayllón, D., Railsback, S. F., Gallagher, C., Augusiak, J., Baveco, H., Berger, U., Charles, S., Martin, R., Focks, A., Galic, N., Liu, C., van Loon, E. E., Nabe-Nielsen, J., Piou, C., Polhill, J. G., Preuss, T. G., Radchuk, V., Schmolke, A., Stadnicka-Michalak, J., … Grimm, V. (2021). Keeping modelling notebooks with TRACE: Good for you and good for environmental research and management support. *Environmental Modelling & Software*, *136*, 104932. https://doi.org/10.1016/j.envsoft.2020.104932

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*, 157048. https://doi.org/10.1016/j.scitotenv.2022.157048

Baudrot, V., & Charles, S. (2021). "morse": An R-package to analyse toxicity test data. *Journal of Open Source Software*, *6*(68), 3200. https://doi.org/10.21105/joss.03200

Billoir, E., Delignette-Muller, M. L., Péry, A. R. R., & Charles, S. (2008). A Bayesian approach to analyzing ecotoxicological data. *Environmental Science & Technology*, *42*(23), 8978–8984. https://doi.org/10.1021/es801418x

Brock, T. C. (2009). *Linking aquatic exposure and effects: Risk assessment of pesticides* (1st ed.). CRC Press. https://doi.org/10.1201/9781439813492

Calcagno, V. (2020). *glmulti: Model selection and multimodel inference made easy*. https://CRAN.R-project.org/package=glmulti

European Food Safety Authority (EFSA), Plant Protection Products and their Residues (PPR), Ockleford, C., Adriaanse, P., Berny, P., Brock, T., Duquesne, S., Grilli, S., Hernandez-Jerez, A. F., Bennekou, S. H., Klein, M., Kuhl, T., Laskowski, R., Machera, K., Pelkonen, O., Pieper, S., Smith, R. H., Stemmer, M., Sundh, I., Tiktak, A., … Teodorovic, I. (2018). 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. https://doi.org/10.2903/j.efsa.2018.5377

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. https://doi.org/10.1007/s10646-018-1940-6

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: An International Journal*, *17*(2), 287–299. https://doi.org/10.1080/10807039.2011.552391

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), Article 16. https://doi.org/10.1021/acs.est.9b01690

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. *Environmental Toxicology and Chemistry*, *40*(11), 3034–3048. https://doi.org/10.1002/etc.5176

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*. *Science of the Total Environment*, *856*, 158886. https://doi.org/10.1016/j.scitotenv.2022.158886

Hunka, A. D., Meli, M., Thit, A., Palmqvist, A., Thorbek, P., & Forbes, V. E. (2013). Stakeholders' perspective on ecological modeling in environmental risk assessment of pesticides: Challenges and opportunities. *Risk Analysis*, *33*(1), 68–79. https://doi.org/10.1111/j.1539-6924.2012.01835.x

Jager, T. (2014). Reconsidering sufficient and optimal test design in acute toxicity testing. *Ecotoxicology*, *23*(1), 38–44. https://doi.org/10.1007/s10646-013-1149-7

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., & Ashauer, R. (2018). *Modelling survival under chemical stress—A comprehensive guide to the GUTS framework* (Ver. 1.0). leanpub https://leanpub.com/guts_book

Kassambara, A. (2022). *ggpubr: "ggplot2" based publication ready plots*. https://CRAN.R-project.org/package=ggpubr

Kooijman, S. A. L. M., & Bedaux, J. J. M. (1996). *The analysis of aquatic toxicity data*. Vrije University Press.

Larras, F., Charles, S., Chaumot, A., Pelosi, C., Le Gall, M., Mamy, L., & Beaudouin, R. (2022). A critical review of effect modeling for ecological risk assessment of plant protection products. *Environmental Science and Pollution Research*, *29*(29), 43448–43500. https://doi.org/10.1007/s11356-022-19111-3

Mangold-Döring, A., Huang, A., van Nes, E. H., Focks, A., & van den Brink, P. J. (2022). Explicit consideration of temperature improves predictions of toxicokinetic–toxicodynamic models for flupyradifurone and imidacloprid in *Gammarus pulex*. *Environmental Science & Technology*, *56*(22), 15920–15929. https://doi.org/10.1021/acs.est.2c04085

Nickisch Born Gericke, D., Rall, B. C., Singer, A., & Ashauer, R. (2022). Fish species sensitivity ranking depends on pesticide exposure profiles. *Environmental Toxicology and Chemistry*, *41*, 1732–1741. https://doi.org/10.1002/etc.5348

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

openGUTS. (2021). openGUTS. http://www.openguts.info/

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Comput ing. https://www.R-project.org/

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. *Archives of Environmental Contamination and Toxicology*, *83*(4), 349–360. https://doi.org/10.1007/s00244-022-00963-2

RStudio Team. (2022). *RStudio: Integrated development environment for R*. RStudio, PBC. http://www.rstudio.com/

Sardi, A. E., Augustine, S., Olsen, G. H., & Camus, L. (2019). Exploring interspecies sensitivity to a model hydrocarbon, 2-methylnaphtalene, using a process-based model. *Environmental Science and Pollution Research*, *26*(11), 11355–11370. https://doi.org/10.1007/s11356-019-04423-8

Schmolke, A., Thorbek, P., Chapman, P., & Grimm, V. (2010). Ecological models and pesticide risk assessment: Current modeling practice. *Environmental Toxicology and Chemistry*, *29*(4), 1006–1012. https://doi.org/10.1002/etc.120

Schneeweiss, A., Juvigny-Khenafou, N. P. D., Osakpolor, S., Scharmüller, A., Scheu, S., Schreiner, V. C., Ashauer, R., Escher, B. I., Leese, F., & Schäfer, R. B. (2022). Three perspectives on the prediction of chemical effects in ecosystems. *Global Change Biology*, *29*(1), 21–40. https://doi.org/10.1111/gcb.16438

Schuwirth, N., Borgwardt, F., Domisch, S., Friedrichs, M., Kattwinkel, M., Kneis, D., Kuemmerlen, M., Langhans, S. D., Martínez-López, J., & Vermeiren, P. (2019). How to make ecological models useful for environmental management. *Ecological Modelling*, *411*, 108784. https://doi.org/10.1016/j.ecolmodel.2019.108784

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

Vlaeminck, K., Viaene, K. P. J., Van Sprang, P., & De Schamphelaere, K. A. C. (2021). Development and validation of a mixture toxicity implementation in the dynamic energy budget–individual-based model: Effects of copper and zinc on *Daphnia magna* populations. *Environmental Toxicology and Chemistry*, *40*(2), 513–527. https://doi.org/10.1002/etc.4946

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

