# Robust likelihood-based approach for automated optimisation and uncertainty analysis of toxicokinetic-toxicodynamic models

Tjalling Jager*†

† DEBtox Research, The Netherlands. tjalling@debtox.nl
* To whom correspondence may be addressed.

This is the peer reviewed version of the following article:

Jager, T (2021). Robust likelihood-based approach for automated optimization and uncertainty analysis of toxicokinetic-toxicodynamic models. Integrated Environmental Assessment and Management 17(2):388-397.

which has been published in final form at https://doi.org/10.1002/ieam.4333. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions (see https://authorservices.wiley.com/author-resources/Journal-Authors/licensing/self-archiving.html).

## ABSTRACT

Toxicokinetic-toxicodynamic (TKTD) models offer a mechanistic understanding of individual-level toxicity over time, and allow for meaningful extrapolations from laboratory tests to exposure conditions in the field. Thereby, they hold great potential for ecotoxicological studies, both in a regulatory context as well as for basic research. In contrast to mechanistic effect models at higher levels of biological organisation, TKTD models can be, and generally are, parameterised by fitting them to data (results from toxicity tests). Fitting models comes with a range of statistical and numerical challenges, which may hamper the application of TKTD models in a practical setting. Especially in the context of environmental risk assessment (ERA), there is a need for robust and user-friendly software tools to automatically extract the best-fitting model parameters, and quantify their uncertainty, from any data set. This paper presents a general outline for TKTD model analysis, rooted in likelihood-based ('frequentist') inference. The general outline is followed by a presentation of the specific algorithm that has been implemented into software for the robust and automated analysis of toxicity data for survival. However, the presented approach is more broadly applicable to low-dimensional problems.

**Keywords:** TKTD modelling, risk assessment, uncertainty analysis, error propagation, statistical inference

## INTRODUCTION

Mechanistic effect models have been identified as essential tools to move towards more meaningful environmental risk assessment (ERA) of chemicals (Grimm and Martin 2013; Hommen et al. 2016). At the level of the individual organism, mechanistic effect models are classified as toxicokinetic-toxicodynamic (TKTD) models (Jager et al. 2006; Ashauer and Escher 2010), which aim to replace current descriptive methods for dose-response analysis, such as hypothesis testing and curve fitting. Rather than describing the toxic effects, at a certain time point, TKTD models attempt to explain the effects as a function of both exposure (which could be time varying) and time. These models have gained concrete interest for ERA. In particular, the General Unified Threshold model for Survival (GUTS, Jager et al. 2011; Jager and Ashauer 2018b) was judged to be "ready for use" in ERA of pesticides in Europe (EFSA 2018).

In contrast to many other models with (potential) applications in ERA (e.g., fate and population models), TKTD models are always parameterised by fitting them to data, data that are specific for the compound of interest (i.e., results from a toxicity test following effects over time). Model application thus requires consideration of optimisation algorithms, but also a different approach towards uncertainty analysis than what is usually proposed for "good modelling practice" (EFSA 2014), as the uncertainty in the model parameters (and thereby in model predictions) follows from the fit to a data set (see Jager and Ashauer 2018a). A proper model optimisation and uncertainty analysis poses specific challenges for the users of (results from) TKTD models. Firstly, we need to assert that the optimisation routine has indeed found the best possible fit for a data set: the global optimum. Optimisation requires starting values for the model parameters, and, with an unfortunate choice of starting values, the optimisation method may end up in a local optimum. Secondly, we need to derive meaningful confidence intervals (CIs) on model parameters, and propagate the parameter uncertainty to CIs on model predictions. A model will never fit perfectly to the data for several reasons: measurement errors, variability between individuals, but most importantly because every model is a simplification of reality and thus 'wrong'. This uncertainty needs to be quantified as much as possible, and propagated to model predictions.

To facilitate application of TKTD models in ERA, the challenges outlined above would need to be supported by robust and user-friendly software. Some might argue that these complex model analyses are best left to experts. However, automated software tools are already widely used in ERA for other models (e.g., for fitting species-sensitivity distributions, dose-response curves, and degradation kinetics), and for good reason. Experts are scarce, and there is concern from regulatory authorities about their independence. Authorities therefore request the ability to check calculations with effect models, which is reflected in EFSA (2014): "there is a need for … robust, user-friendly and freely available software, which can routinely be used by industry and regulators in all Member States." This aim requires automation: reducing the human factor in the model analysis, such that the data analysis can be performed by users with limited expertise in modelling, numerical methods, and statistics. Automation will increase reproducibility and consistency of model analyses, which, in turn can build trust in TKTD modelling amongst the stakeholders in ERA. Automation also brings TKTD models within reach of a wider audience to apply them, and more importantly, critically test their performance.

In this contribution, I present a general outline for TKTD model analysis, rooted in likelihood-based ('frequentist') inference, which lends itself to robust automation, followed by a specific implementation for GUTS analyses and a case study. In this paper, the focus will be on the concepts of the methods, and not on the theoretical and technical details (which are included in the supporting information). Users of (results from) TKTD models should

understand the concepts behind the statistical inference, but not necessarily the details. Furthermore, the focus in this paper lies on TKTD models although the presented approach is more broadly applicable to low-dimensional problems.

## THEORETICAL BACKGROUND

### *Different views on statistical inference*

Several schools of thinking can be distinguished in statistical inference, often simplified as 'Bayesian versus frequentist'. The various schools have different strengths and weaknesses, which has proven to be fertile ground for discussion (for an example, see Efron 1986). This paper is not the right place for an in-depth discussion of the pros and cons of different statistical frameworks. However, some points on Bayesian inference need to be made because, recently, an on-line software tool was presented, containing an automated approach for fitting GUTS models in a Bayesian context (Baudrot et al. 2018). This begs the question of why an additional 'frequentist' tool would be needed. The short answer is that, in my opinion, frequentist methods are easier to robustly automate (see Efron 1986) and are better suited to deal with identifiability problems (see Raue et al. 2013). Identifiability problems arise when the information content in the data set is insufficient to identify all model parameters. Some parameters run away to zero or infinity, and, as a result, their CIs are not finite (an example is provided in the case study). Lack of information is a common problem for TKTD analysis since standard protocols for toxicity testing are not geared towards the needs of TKTD modelling. To allow for Bayesian inference in an automated manner, in light of potential identifiability issues, the on-line GUTS tool (Baudrot et al. 2018) relies on a set of rules for constructing 'weakly-informative' priors from test design (Delignette-Muller et al. 2017). If these priors always have negligible impact on the conclusions of the analysis, the discussion would be rather academic. However, the case study in this paper represents a situation where these priors will severely bias the risk estimates. The corresponding Bayesian analysis, and more discussion on this topic, is presented in the supporting information. Despite these problems with one specific tool, it is beneficial to have multiple optimisation frameworks, independently developed by different groups, and based on different statistical principles. Furthermore, it is good to stress that Bayesian and frequentist approaches can strengthen each other: likelihood-based approaches may serve as initial exploration of parameter space to see whether the more elegant and coherent Bayesian analysis is feasible, or whether more information (either in terms of additional data or informative priors) is necessary (Vanlier et al. 2012; Raue et al. 2013).

Frequentist methods are well suited for automation, and to deal with identifiability issues, but the price to pay is a less intuitive, and in fact rather tortuous, series of calculations to derive meaningful CIs. Profiling of the likelihood function offers a powerful approach for constructing CIs and detecting identifiability problems (Meeker and Escobar 1995; Kreutz et al. 2013; Raue et al. 2013). This is, however, a rather cumbersome procedure, requiring large numbers of sequential optimisations, whereby each optimisation runs the risk of getting stuck in a local optimum. The solution put forward in this paper is to make use of a sample from parameter space: evaluate a large number of parameters sets, and map the likelihood function. If this mapping is detailed enough, it will allow location of the global optimum, derivation of CIs on model parameters, and a sample to be used for error propagation to model predictions, all in one procedure.

The next section, presents a general outline for TKTD model analysis, based on the idea of mapping the likelihood function with a sample from parameter space. This is followed by a more specific presentation of the algorithm recently implemented in the openGUTS software (see http://openguts.info).

***Using a sample in likelihood-based inference***

In this section, a conceptual overview of the proposed statistical framework is given; a more extensive treatise is available in the supporting information. The theoretical background is best illustrated with a two-parameter model; parameter space is then two-dimensional (a surface area). Mapping parameter space involves calculating a goodness-of-fit measure (here, the minus log-likelihood, MLL) for a large number of parameter sets on this surface. Calculating the MLL adds a dimension to parameter space, so our two-dimensional surface becomes a three-dimensional landscape. Figure 1 (left panel) shows the three possible 2-D projections of this 3-D landscape. The best-fitting parameter set (resulting in the lowest MLL) is shown as a yellow point. The MLL for each set has been recalculated relative to the best value (minus log-likelihood ratio, or MLLR), such that the best fitting parameter set sits at MLLR=0. The central projection (top view) shows the surface of parameter space, with a line of equal MLLR (outline of the green ellipse). There are two side views, where one parameter is on the x-axis and the MLLR (i.e., the height in the landscape) is on the y-axis. The lower edge of such a side-view plot is the profile likelihood for the parameter: it is the lowest possible value of the MLLR, given that the parameter of interest is fixed to the value on the x-axis. Where the profile likelihood crosses a critical value, based on the $\chi^2$-distribution (1 degree of freedom, $\alpha$=0.05), marks the edges of the 95% CI for the model parameter (Meeker and Escobar 1995). The CI is thus defined as all values of a parameter that are not rejected in a likelihood-ratio test.

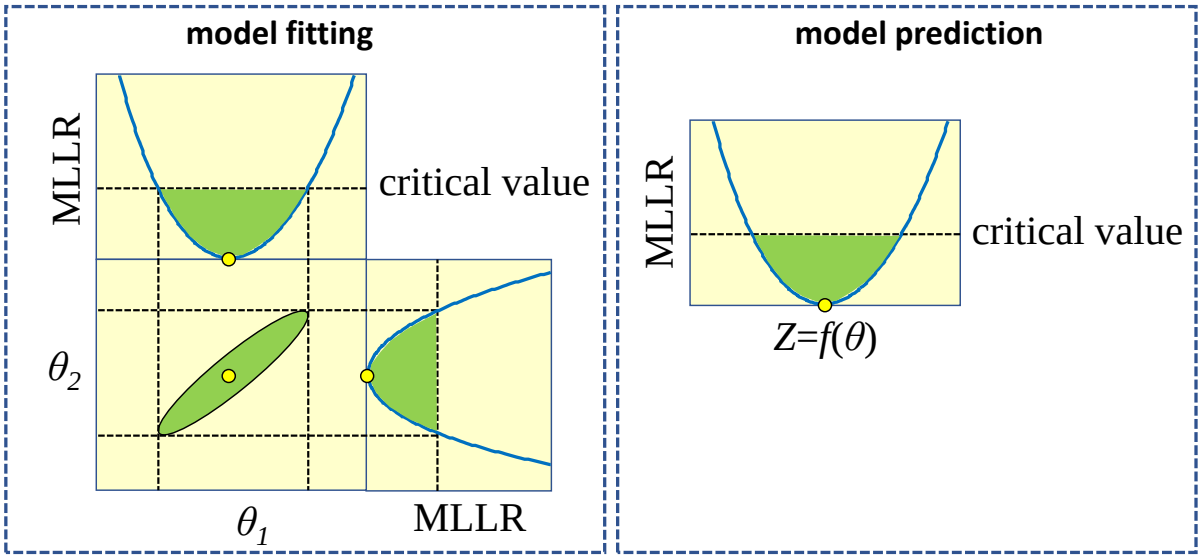


**Figure 1.** Schematic representation of the relationship between parameter space and profile likelihoods for a case with two free model parameters ($\theta_1$ and $\theta_2$). MLLR stands for minus log-likelihood ratio. The yellow dot indicates the best-fitting parameter set (lowest MLL; MLLR=0). The green area contains the parameter sets that yield an MLLR that is not rejected in a likelihood-ratio test (*df*=1). The model prediction (*Z*, right panel) is a function of the two model parameters. The same MLLR is plotted here, so the green area contains the same parameter sets in all plots.

Usually, profile likelihoods are generated by sequential optimisation: fixing a parameter and fitting the remaining parameters using an optimisation routine (Meeker and Escobar 1995). However, from Figure 1, we can see that, with a detailed map of parameter space, the profiles come for free. For large data sets (asymptotically), the line of equal MLLR in the parameter space plot will be an exact ellipse, the profiles will be parabolas, and the CIs will be

symmetrical. However, we are usually far removed from the asymptotic situation, so more oddly-shaped parameter landscapes will occur. In some cases, a profile will not increase above the critical value on one side of the best value, or even on both sides. In that case, the data set does not contain enough information to constrain the model parameters, and we have a situation of non-identifiability (Raue et al. 2009). The fact that oddly-shaped landscapes and identifiability problems occur regularly in practical cases with TKTD models is the main reason why a quadratic approximation of the log-likelihood function should be avoided (see Meeker and Escobar 1995; Raue et al. 2009). Profiling the likelihood is an excellent way to identify these problems, and to construct robust CIs. However, it should be noted that, strictly speaking, the $\chi^2$-distribution only applies asymptotically, so for very small data sets, the coverage of the CI (as defined here) may be somewhat more or less than 95%.

The same map of parameter space can also be used for constructing CIs on model predictions. A model prediction is some function of the model parameters. What we like to construct is the so-called prediction profile likelihood (see Kreutz et al. 2012). This profile can be calculated by constrained optimisation of the model parameters on the data set: fitting all model parameters under the constraint that the model prediction from those parameters yields a certain value. However, we can also use the sample from parameter space to derive it, very similar to the approach used for the profile likelihood. The MLLR of each parameter set is fixed; it is only determined by the goodness-of-fit of the parameter set to the data (relative to the best fit). The profile likelihood is the lower edge of the parameter cloud, when viewed from the side, with the parameter of interest on the x-axis. The lower edge can be seen as an optimisation under a constraint, as one parameter is fixed to a certain value. Similarly, we could view the sample, from the side, with a model prediction on the x-axis (e.g., the 4-day LC50). This is illustrated in the right panel of Figure 1; this is a reshuffled version of the sample, such that the prediction is on the x-axis. Constructing such a plot thus requires the associated model prediction to be calculated for each element in the sample. The lower edge of this cloud represents a constrained optimisation: the best model fit (lowest MLLR) that can be achieved under the constraint that the prediction has the value on the x-axis. The edges of the CI again can be taken as the points where the profile likelihood crosses the critical value, again using the $\chi^2$-distribution with 1 degree of freedom, $\alpha$=0.05.

The sequential (constrained) optimisations needed for profiling the likelihood can thus be replaced by a map of parameter space, whereby MLLR and model predictions are calculated for each point in that space. In practice, we will need to approximate such a continuous map by using a discrete sample. However, obtaining a sufficiently detailed sample from parameter space, and calculating MLLRs and predictions, can be very calculation intensive, especially for higher-dimensional problems. We can limit this task in practice, as we are only interested in the best-fitting parameter set and the sets close to the critical value (those sitting on the ellipse in Figure 1), since they determine all CIs. In practice, efficiently mapping parameter space with a sample is probably restricted to lower-dimensional problems. Since reduced GUTS models have only 3-4 model parameters, such mapping is feasible in principle.

It is good to stress that a sample from parameter space in the likelihood-based context has a different interpretation than the Markov-Chain Monte Carlo (MCMC) sample used by Bayesians. MCMC attempts to obtain a random sample from a joint probability distribution (the posterior). Therefore, quantiles from that sample are meaningful as CIs. For likelihood-based inference, the parameter landscape does not represent a probability distribution, so taking quantiles from a sample is meaningless. Instead, we are looking for lines-of-equal-likelihood in that landscape to use for our CIs; CIs are based on likelihood-ratio testing of parameter sets against the best-fitting set. This also implies that it does not matter *how* this sample is generated. There is no need for it to be random in any sense, we just need a good coverage of the interesting volume of parameter space.

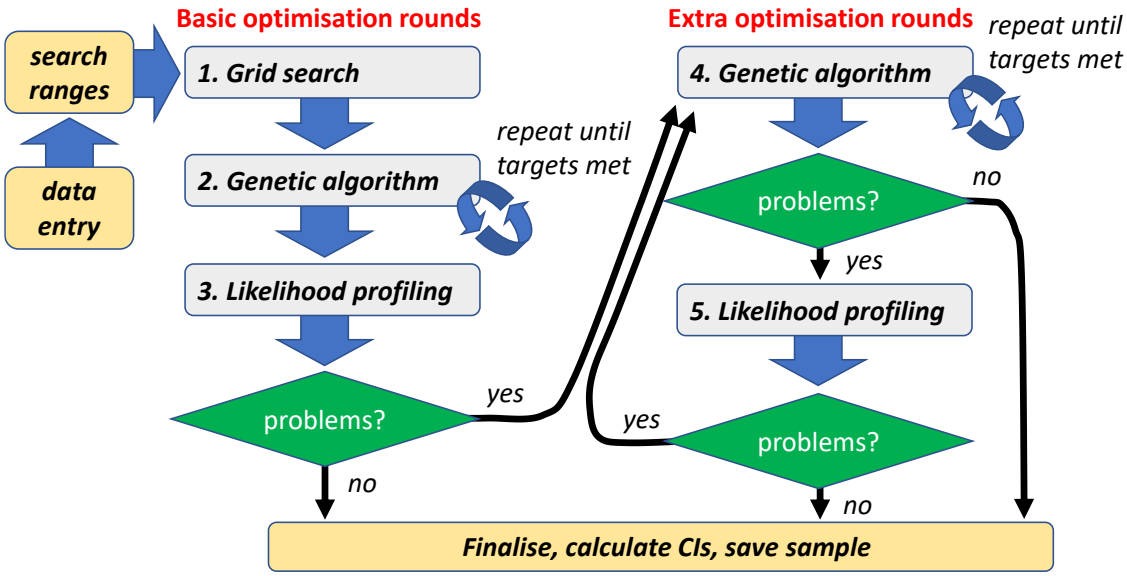


**Figure 2.** Schematic workflow of the openGUTS algorithm for automated optimisation and uncertainty analysis. The targets for the genetic algorithm are a minimum number of samples found within a certain distance from the best fit, or a maximum number of rounds at this stage.

## IMPLEMENTATION IN OPENGUTS

The general framework laid out in the previous section was used in the openGUTS project, which produced software for automated GUTS analysis. This free and open-source software is available online (http://openguts.info) as a standalone executable and as Matlab code. Since this software is specifically intended to support ERA, the focus lies on automation and robustness, and to a lesser extent on calculation speed (mapping parameter space requires many model evaluations). Nevertheless, calculation time is still very reasonable since the reduced GUTS models can largely be computed analytically, increasing both speed and accuracy (the reduced GUTS models are explained in the supporting information). More recently, the algorithm has also been implemented into the generic BYOM framework for Matlab (Bring Your Own Model, see http://debtox.info/byom.html) so it can be used for any model. Figure 2 presents a flow chart for the algorithm of openGUTS. The user enters a data set, but otherwise does not interact with the calculation process. End result is the best-fitting parameter set (the global optimum), CIs on those parameters, and a sample to be used for error propagation to model predictions. Even in 3 or 4 dimensions, mapping parameter space is very much helped by reducing the (hyper)volume of space that needs to be searched. Large parts of parameter space will produce extremely bad fits, and it is efficient to *a priori* avoid sampling in the worst sections by setting minimum-maximum boundaries. Some bounds are rather trivial; for example, background mortality and the threshold for effects can be almost zero, but they cannot be negative. Other bounds require more careful consideration. In setting the search ranges, we need to consider to what extent parameters can be identified from the (short-term) toxicity test. For example, a 4-day acute test will not allow parameter identification in case of very slow dynamics of the toxic response (as demonstrated in the case study of the next section). However, we also need to consider the intended model extrapolations. It may be impossible to accurately identify a very low rate constant with a short toxicity test, but that does not mean that low values are irrelevant for extrapolating to a 485-day exposure profile from an environmental fate model, as used for pesticide ERA. In openGUTS, the bounds of the search ranges are set automatically, such that we always obtain a representative sample for the relevant

foreseen model predictions. In other words: using even wider bounds will not lead to a different fit, nor to different (CIs on) model predictions in this context.

The algorithm proceeds through several modules (more detail in the supporting information):

1. Grid search. Create a regular grid in parameter space, covering the ranges of the (log-transformed) parameter values. Evaluate all parameter sets from the grid; each set gets an MLL. Find the best fit (lowest MLL) so far, and select the parameter sets whose MLLs are within a certain distance from the best fit (MLLR). These are the candidate sets to continue to the next module.
2. Genetic algorithm. Over several steps, the candidate sets are mutated randomly, MLLs calculated, and it is checked whether the relevant part of parameter space contains sufficient sets. If not, candidate sets are selected for another step of mutation. With each round, the settings of the algorithm (such as maximum mutation step size) are changed such that the sample can contract. There is no need to discard sets, as their MLL will not change. However, the total set is pruned to the most relevant part of parameter space: only keep parameter sets whose MLL is not too far from the best MLL found so far. At the end of every step in this module, a quick simplex optimisation is performed to improve on the best value found.
3. Likelihood profiling. The edges of the parameter cloud are refined by classic profiling: sequential simplex optimisations, using the sample to provide starting values. In principle, there is no need for this step if the sample is large enough. However, parameter space for GUTS analyses is sometimes oddly-shaped, which makes it difficult to sample properly. It turns out that explicit profiling is an excellent way to provide robustness, forcing the optimisation routine to explore other parts of parameter space. If the refined profile is not close enough to the sample, this area is marked for additional sampling. If no problems are identified, jump to Module 6.
4. Extra rounds of sampling. If profiling identified problems, or if the initial sampling rounds failed to find sufficient sets, additional rounds of mutation are performed, targeted at the problem area(s). Since the final sample does not need to be random in any sense, targeted extra sampling is possible. In every step, the sample is tested against the profile to select new candidates for mutation. If no problems are flagged, jump to module 6.
5. Extra rounds of profiling. When there are sample points below the refined profile line, there is a need to create a new refined profile by sequential optimisations. This may flag problems that need additional sampling, and hence a jump back to module 4.
6. Finishing up and reporting. Calculate CIs from the profile by interpolating in the refined profile curves, and check whether the CIs run into our bounds of parameter space. Save the sample to file for use in model predictions.

Module 1 and 2 form the basis of the algorithm. Starting with a wide grid, and refining with a genetic algorithm, constitutes a global optimisation approach, which is more robust than local optimisation (e.g., simplex search). Module 3-5 are additional steps to increase robustness for more extreme data sets. In this likelihood-based framework, identifiability problems become immediately apparent from the likelihood profiles: one (or more) profile curves run into a bound of parameter space. A rather common problem is that of ‘slow kinetics’, in which the duration of the experimental test is too short to identify the dynamics of the damage that is driving the toxic effect. This is illustrated with the case study in the next section.

## CASE STUDY WITH OPENGUTS

To illustrate the working of the algorithm described in the previous section, this section provides a case study with the reduced GUTS-SD model, where the data present identifiability problems; a more well-behaved example is treated in step-by-step detail in the supporting information. The case study is for fathead minnows exposed to fluorophenyl (see Russom et al. 1997). The background hazard rate was fixed to the lowest allowed value in openGUTS ($10^{-6}$ $d^{-1}$) since there was no mortality observed in the control. This leaves three parameters to be fitted to the toxicity data: the dominant rate constant ($k_d$), the threshold for effects ($m_w$), and the killing rate ($b_w$). It is of course also possible to fit all four parameters simultaneously. Calculations were performed with the BYOM implementation of GUTS in Matlab.

### *Analysis of the data set*

Figure 3 shows the parameter-space plot, with the various 2-D projections of the 4-D landscape (three fitted parameters plus the MLLR for each parameter set), comparable to the left panel of Figure 1. The profile likelihoods are shown on the diagonal. There is a clear best-fit set (yellow point), but the profiles are not the parabolas as expected for the asymptotic case. In fact, the profile for the dominant rate constant ($k_d$, top left) hits its lower boundary (vertical blue line). This is the identifiability problem referred to as 'slow kinetics': the dynamics of the damage in this species, for this chemical, is simply too slow to be identified from a 4-day toxicity test. We cannot measure damage, so its dynamics (determined by $k_d$) must be inferred from the effect patterns over time. However, if these patterns indicate that dynamics is very slow (relative to the test duration), several model parameters will become extremely correlated and non-identifiable. This is inherent to TKTD modelling unless we can measure the damage driving the effect directly: the structure of the model prevents us from identifying all three model parameters in such a data set (see Jager and Ashauer 2018b, Appendix C). For GUTS-SD, all three parameters become strongly correlated, as can be seen from the three binary panels in Figure 3. Without bounds to parameter space, the parameters will simply continue to run away to minus infinity ($k_d$ and $m_w$, on log-scale) or plus infinity ($b_w$). This can be seen from the profile likelihood of $k_d$, which becomes flat as $k_d$ goes to its lower bound: the goodness-of-fit remains the same when the parameter becomes smaller.

In Figure 3, $k_d$ is running into its lower bound. Therefore, its CI should be reported as a half-open interval ($k_d < 1.0$ $d^{-1}$). As $m_w$ and $b_w$ are strongly correlated to $k_d$, there are, in this case, consequences for their identifiability as well. The only reason that these parameters do not hit their minimum-maximum bounds is because $k_d$ runs into its lower bound first. The lower CI of $m_w$ and upper CI of $b_w$ in Figure 3 are thus artefacts caused by the lower bound of $k_d$. Even though these parameters do not run into their bounds, their CIs should be reported as half-open intervals as well ($m_w < 2.7$ µM and $b_w > 0.066$ $\mu M^{-1}$ $d^{-1}$). This is clearly one of those issues where some expertise from the user is needed to recognise what is going on. Nevertheless, failing to spot this issue only affects the CIs on the model parameters (which will not be used in ERA), not those on the model predictions (see next section). It is good to realise that the non-open edge of each CI is well defined, and does not depend on how low we allow $k_d$ to go. As explained in the section "*Using a sample in likelihood-based inference*", inference is only based on the MLLR of each parameter set, and that value is unaffected by the choice of minimum-maximum bounds (under the condition that the best fit, the overall lowest MLL, has been established). Identifiability issues thus have limited consequences for the statistical inference. These complexities all result from a rather innocuous looking data set, and a very reasonable model fit (Figure 4).

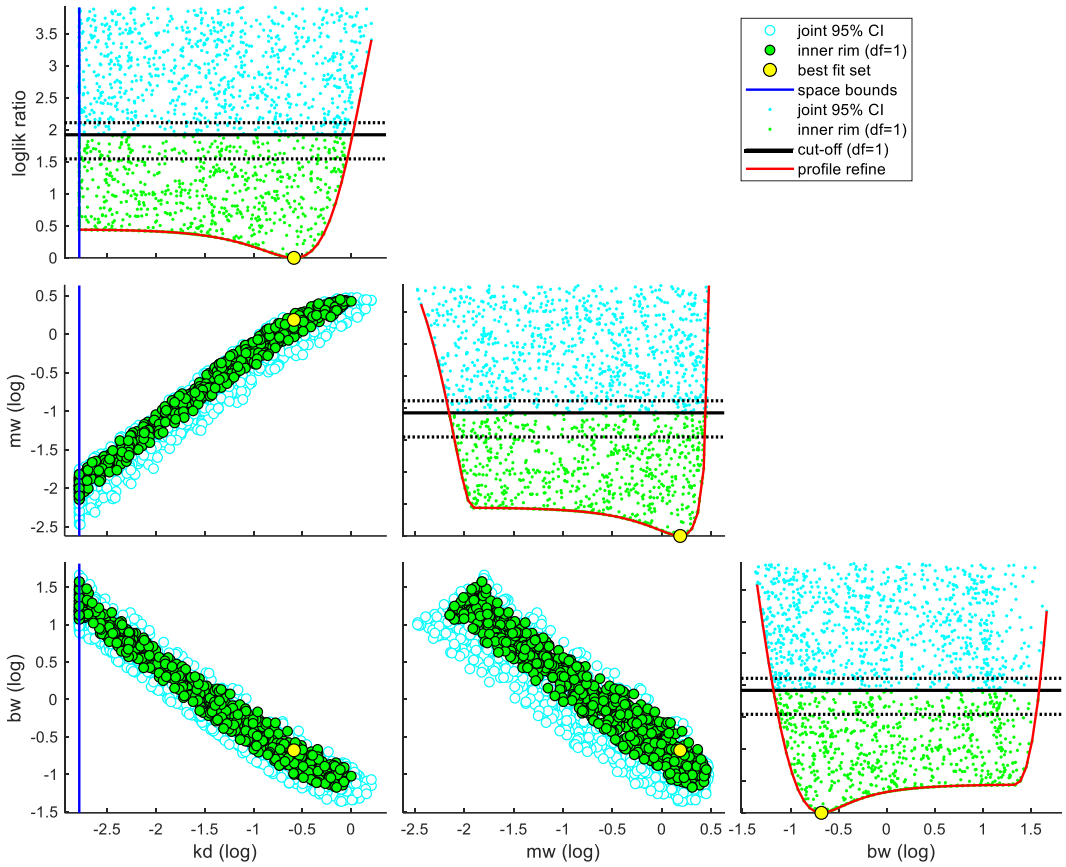


**Figure 3.** Parameter-space plot for the fit of GUTS-SD to survival data for fathead minnows exposed to fluorophenyl. Parameters are the dominant rate constant ($k_d$), the threshold for effects ($m_w$), and the killing rate ($b_w$). Yellow points mark the best fit. Green points are within the critical value (horizontal black line). The parameter sets between the dotted lines, above and below the critical value, are used for error propagation.

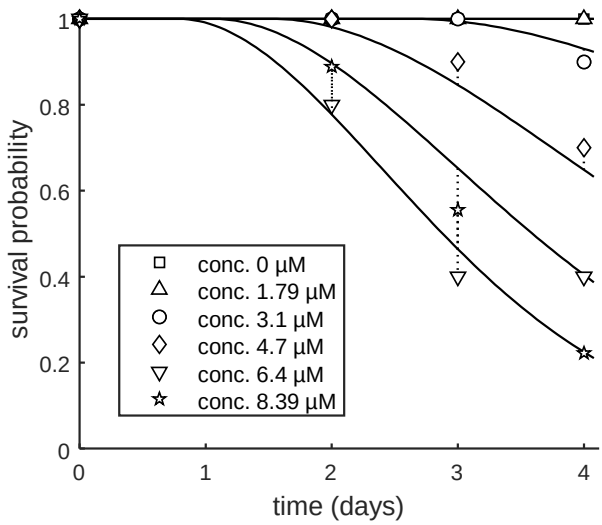


**Figure 4.** fit of GUTS-RED-SD to survival data for fathead minnows exposed to fluorophenyl under constant exposure. Model curves result from the best-fit parameters (yellow point in Figure 3).

***Consequences for model predictions***

All three CIs are non-finite, which implies that none of our model parameters can be identified from the data. Nevertheless, the data set still contains considerable information, owing to the strong correlations between the model parameters. At low $k_d$, the model behaviour is completely determined by the product $b_w$ x $k_d$ and the fraction $m_w$ / $k_d$ (see Jager and Ashauer 2018b, Appendix C). And, that information *is* present in the sample of Figure 3. We can therefore use the sample to make model predictions, which ensures that the information and its uncertainty are propagated to the model predictions. Based on the model predictions, it can then be judged whether further experiments are needed, and how they should be designed. An example of a model prediction is shown in Figure 5. The LC50 is a function of time, a model output calculated from the model parameters, which can thus be extrapolated beyond the duration of the toxicity test (4 days). The CIs are calculated from the sample in Figure 3, according to the procedure explained in the section "*Using a sample in likelihood-based inference*". Even though the parameters themselves cannot be identified, the CI on the LC50 remains well defined due to the strong correlations between the parameters. The LC50 decreases over time, which is a consequence of the slow kinetics. In fact, since there is no lower edge of the CI for $k_d$, the lower edge for the CI of the LC50 will go to zero for long exposure.

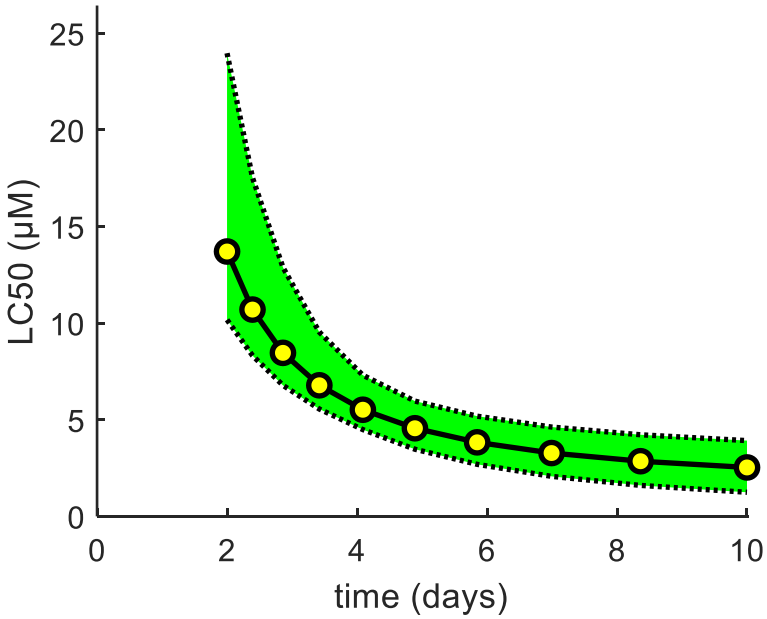


**Figure 5.** Model predictions from the best fit and parameter-space sample of Figure 3. The prediction is for the LC50 as function of time, with 95% CI. The CI on the LC50 is calculated from a limited part of the sample: the sets between the horizontal dotted lines in the plots on the diagonal of Figure 3.

***Relevance of slow kinetics for ERA***

The case study illustrated the situation of slow kinetics: $k_d$ running away to zero. Even though other identifiability problems may occur in GUTS analyses, this specific situation warrants special attention in ERA. A very low $k_d$ signals that the compound has an irreversible, or only slowly reversible, toxic action in this species. Since $k_d$ lumps toxicokinetics and damage dynamics, this could result from a lack of elimination of the chemical from the body, or a lack of damage repair. As a consequence, very low levels of exposure (much lower than those tested in the toxicity test) may lead to considerable effects after prolonged exposure (much longer

than the test duration). Furthermore, under time-varying exposure, effects may carry over from one exposure event to the next (Ashauer et al. 2010). If organisms in the field are exposed to a large number of exposure events over a long period of time, the effects will build up and may be much more severe than expected on the basis of the maximum peak concentration and the 4-day LC50. The case study shows that, based on this data set, an irreversible action cannot be excluded for fluorophenyl in fathead minnow, as the lower edge of the CI for $k_d$ is effectively zero. The sample does not contain zero, but does include very low values. Figure 6 shows an extrapolation to a long exposure profile, as used for the GUTS ring test (Jager and Ashauer 2018b, Appendix A). The exposure scenario is multiplied by a factor (the LP10) such that there is 10% effect on survival probability due to the chemical at the end of the profile (Ashauer et al. 2013; EFSA 2018). Clearly, the uncertainty in that extrapolation explodes for extrapolations well beyond the duration of the toxicity test. The best estimate for the survival probability results from the best estimate of the parameters (the yellow point in Figure 3). The lower edge of the CI on survival probability (right panel) relates to the lower edge of the CI on scaled damage (left panel). It is perhaps counter intuitive that low values of damage lead to strong effects, but this relates to the strong correlations between the parameters: low values for $k_d$ imply a low threshold $m_w$ and high values of the killing rate $b_w$. The lower edge of the CIs on survival probability thus relates to slow kinetics: the possibility that this chemical has an irreversible action, implying that effects will build up over the entire exposure profile. It is therefore advisable to consider the CIs on model predictions as well, and not just the best estimate. It remains to be seen whether complete irreversibility is actually realistic, but that question can only be resolved by dedicated experimentation.

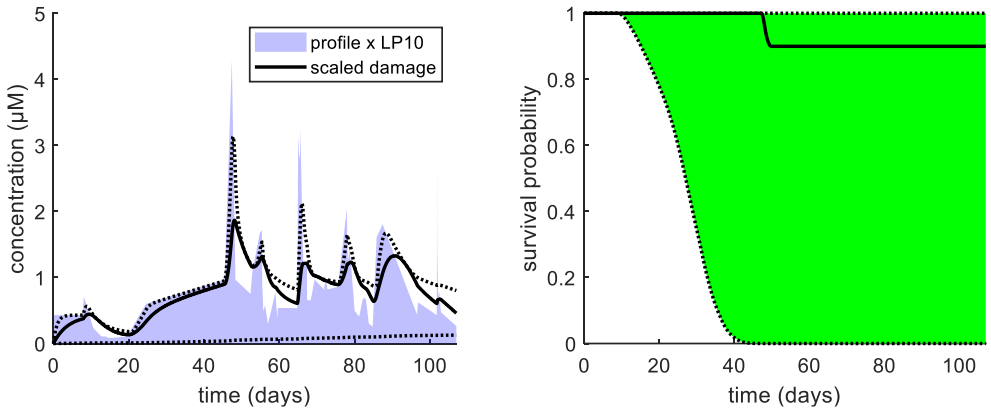


**Figure 6.** Model predictions for a long exposure profile. The profile is multiplied by a factor (LP10) such that the best estimate for the survival probability at the end of the profile is 90%. Left panel shows the exposure profile (area) and the predicted damage dynamics (line, with 95% CI as dotted lines). Right panel shows the corresponding predicted survival probability due to the chemical, with 95% CI (green area).

## DISCUSSION

### *A general framework for frequentist inference*

Application of TKTD models in ERA, and in ecotoxicology in general, is served by automated and robust methods to perform the model optimisation and the quantification of uncertainties. In this paper, I present a general framework that addresses this issue from the likelihood-based perspective, and a specific algorithm that is implemented in the openGUTS software and the BYOM platform. The most important feature of this algorithm is that it combines tasks that are usually treated separately in frequentist inference: finding the global optimum, constructing CIs on model parameters, and obtaining the information needed to

calculate CIs on model predictions. By mapping parameter space using sampling and sample-based profiling, these tasks can be automated in a robust manner. The specific algorithm implemented in openGUTS and BYOM combines grid search, a genetic algorithm, and likelihood profiling. The likelihood-based framework is well equipped to deal with issues of parameter identifiability (Raue et al. 2013), as illustrated in the case study. Such identifiability issues are quite common for the rather limited data sets that are currently available for TKTD modelling in ERA.

Here, the algorithm is illustrated with a GUTS example. However, in conjunction with BYOM, it can now be used for any model, or at least, any model that can be implemented in BYOM, with an appropriate likelihood function. A limitation of the algorithm is that it is, in this form, restricted to low-dimensional problems; preliminary testing suggests that five parameters may be the practical maximum. However, the same general framework could be used in conjunction with smarter sampling schemes, and even with MCMC sampling. A further limitation is that the algorithm requires many evaluations of the model against the data set. For reduced GUTS models, this is not a huge problem since the model evaluations can be made cheaply, in terms of calculation time. However, GUTS models only deal with effects on survival; for TKTD analysis of sub-lethal effects, DEBtox models should be used (Jager et al. 2006). It is interesting to explore whether the same algorithm can be used for DEBtox models. Even in their simplest form (Jager 2020), these models have many more parameters than GUTS models. These parameters fall into two classes: the basic parameters that govern the life history of the species, and the parameters that govern the response to the toxicant. The first class of parameters can be fitted to the control data (Jager 2020) or taken from a dedicated parameter library (Marques et al. 2018). This leaves just 3-5 toxicity parameters to be fitted, which makes this problem suitable for the algorithm described in this paper. A remaining issue is that DEBtox models are much more computationally expensive to evaluate than GUTS due to the need for numerical approximations of the model's differential equations. However, if the entire procedure can be automated, this only means considerable computer time, which is not such a big issue nowadays.

### *Limitations of automated procedures*

Application of models in ERA first and foremost requires an evaluation that the model is 'fit for purpose', i.e., that the model structure and its underlying assumptions are a good match to the question at hand. If that hurdle is taken, routine application is served by automated and robust tools that take away the technical complexities from the user as much as possible. For experts in TKTD modelling, automation will allow them to use their time more efficiently and reduce human errors. However, to what extent can such automated tools be used by persons without modelling expertise? This obviously depends on the complexity of the model, but also on the context in which the model is applied.

In case of the openGUTS software, the model is very simple (3-4 parameters). When the model is applied in the intended context (as defined by EFSA 2018), and when the data are entered correctly, the algorithm described in this paper is robust enough to provide automated data analysis, in almost all cases. Some cases require knowledge on the model structure to fully interpret; an example was provided in the case study. In rare cases, the algorithm has trouble sampling the relevant part of parameter space. Examples of these cases, how to recognise and how to deal with them, are provided in a dedicated interpretation document from the openGUTS web site. Some additional expertise is needed from the user. Firstly, ecotoxicological expertise is needed to scrutinise the input data (toxicity test results and exposure definitions), and to judge the toxicological/biological relevance of the observed/predicted effects. However, this type of expertise is more readily available in the

ERA community. Secondly, expertise with the GUTS model is needed to judge the goodness-of-fit on a data set. It is the task of algorithm to extract the best-fitting parameters and a sample for error propagation, given the model, for any data set. However, judging whether the correspondence between model and data is 'good enough' is left to the user.

For simple models, used within a well-delimited context, automated tools can be developed that can be effectively used by non-experts. However, when such a tool is used in a different context than the one for which it was designed, more expertise will be needed to ensure proper functioning. For example, openGUTS can be used with data for terrestrial invertebrates and mammals, but more expertise will be needed to judge whether the underlying assumptions of the GUTS model apply (e.g., with regard to exposure pathways), and whether the algorithm needs modification. The algorithm described in this paper is not restricted to GUTS analyses but could be adapted to other models. However, it is not at all evident that this will be possible in an automated fashion, such that it can be used by non-experts. The already-mentioned DEBtox models are more complex than GUTS, and their application involves greater numerical and statistical challenges. The application of this algorithm to DEBtox is therefore, at this moment, best left to experts.

## CONCLUSION

In this paper, I present a general likelihood-based framework that lends itself to robust automation, and a specific algorithm based on this framework. This algorithm was intended for use with simple GUTS models in the context of pesticide ERA, but the framework is generic enough to allow adaptation to other models and modelling questions. There is certainly room for improvement; hopefully, this work will inspire development of more efficient successors.

***Acknowledgment*** – The author thanks Cefic-LRI for funding the development of openGUTS, including the algorithm described here (project ECO39.2). Furthermore, thanks are due to Syngenta for funding the translation of the algorithm to BYOM to make it more generally usable for other models, and for supporting the writing of the present study. Finally, Roman Ashauer and Benoit Goussen are thanked for commenting on drafts of the manuscript.

***Disclaimer*** – The author declares no conflicts of interest.

***Data Availability Statement –*** Calculation tools are available online from http://openguts.info and http://debtox.info/byom.html. The data set used for the case study is included as an example file in openGUTS.

## SUPPLEMENTAL DATA

Consists of a PDF with detailed explanation of the statistical background, the algorithm in openGUTS, and greater analysis of the case study.

## REFERENCES

Ashauer R, Escher BI. 2010. Advantages of toxicokinetic and toxicodynamic modelling in aquatic ecotoxicology and risk assessment. *J Environ Monit* 12:2056-2061.

Ashauer R, Hintermeister A, Caravatti I, Kretschmann A, Escher BI. 2010. Toxicokinetic and toxicodynamic modeling explains carry-over toxicity from exposure to diazinon by slow organism recovery. *Environ Sci Technol* 44:3963-3971.

Ashauer R, Thorbek P, Warinton JS, Wheeler JR, Maund S. 2013. A method to predict and understand fish survival under dynamic chemical stress using standard ecotoxicity data. *Environ Toxicol Chem* 32:954-965.

Baudrot V, Veber P, Gence G, Charles S. 2018. Fit reduced GUTS models online: from theory to practice. *Integr Environ Assess Manag* 14:625-630.

Delignette-Muller ML, Ruiz P, Veber P. 2017. Robust fit of toxicokinetic−toxicodynamic models using prior knowledge contained in the design of survival toxicity tests. *Environ Sci Technol* 51:4038-4045.

Efron B. 1986. Why isn't everyone a Bayesian? *Am Stat* 40:1-5.

[EFSA] European Food Safety Authority. 2014. Scientific Opinion on good modelling practice in the context of mechanistic effect models for risk assessment of plant protection products. *EFSA journal* 12:3589.

[EFSA] European Food Safety Authority. 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* 16:5377.

Grimm V, Martin BT. 2013. Mechanistic effect modeling for ecological risk assessment: where to go from here? *Integr Environ Assess Manag* 9:E58-E63.

Hommen U, Forbes V, Grimm V, Preuss TG, 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. *Integr Environ Assess Manag* 12:21-31.

Jager T. 2020. Revisiting simplified DEBtox models for analysing ecotoxicity data. *Ecol Modell* 416:108904.

Jager T, Albert C, Preuss TG, Ashauer R. 2011. General Unified Threshold model of Survival - a toxicokinetic-toxicodynamic framework for ecotoxicology. *Environ Sci Technol* 45:2529-2540.

Jager T, Ashauer R. 2018a. How to evaluate the quality of toxicokinetic-toxicodynamic models in the context of environmental risk assessment. *Integr Environ Assess Manag* 14:604-614.

Jager T, Ashauer R. 2018b. Modelling survival under chemical stress. A comprehensive guide to the GUTS framework. Toxicodynamics Ltd., York, UK. Available from Leanpub, https://leanpub.com/guts_book, Version 2.0, 8 December 2018

Jager T, Heugens EHW, Kooijman SALM. 2006. Making sense of ecotoxicological test results: towards application of process-based models. *Ecotoxicology* 15:305-314.

Kreutz C, Raue A, Kaschek D, Timmer J. 2013. Profile likelihood in systems biology. *FEBS J* 280:2564-2571.

Kreutz C, Raue A, Timmer J. 2012. Likelihood based observability analysis and confidence intervals for predictions of dynamic models. *BMC Syst Biol* 6:120.

Marques GM, Augustine S, Lika K, Pecquerie L, Domingos T, Kooijman SALM. 2018. The AmP project: comparing species on the basis of dynamic energy budget parameters. *PLOS Comput Biol* 14:e1006100.

Meeker WQ, Escobar LA. 1995. Teaching about approximate confidence regions based on maximum likelihood estimation. *Am Stat* 49:48-53.

Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmüller U, Timmer J. 2009. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics* 25:1923-1929.

Raue A, Kreutz C, Theis FJ, Timmer J. 2013. Joining forces of Bayesian and frequentist methodology: a study for inference in the presence of non-identifiability. *Philos Trans R Soc A* 371:20110544.

Russom CL, Bradbury SP, Broderius SJ, Hammermeister DE, Drummond RA. 1997. Predicting modes of toxic action from chemical structure: acute toxicity in the fathead minnow (*Pimephales promelas*). *Environ Toxicol Chem* 16:948-967.

Vanlier J, Tiemann CA, Hilbers PAJ, Van Riel NAW. 2012. An integrated strategy for prediction uncertainty analysis. *Bioinformatics* 28:1130-1135.