Outcome of the consultation on the Draft Opinion with the Pesticide Steering Network

	Organization	Chapter	Comment	Reply
1.	NL	Chapter 1	No particular comment on this chapter	Noted
2.	NL	Chapter 2	126: standard DEB parameter should read standard DEB model parameter	That is true that the AmP database refers not only to parameters of the standard DEB model, but only of its different variants. The text was modified accordingly.
			730: Baas et al., (2018). Articles of Tjalling Jager and Bas Kooijman also explain the potentials of DEB models for ERA, they should not be ignored.	The following reference was added: Jager T, Barsi A, Hamda NT, Martin BT, Zimmer EI, Ducrot V. 2014. Dynamic energy budgets in population ecotoxicology: Applications and outlook. Ecol. Modell. 280:140–147
			799: Chapter 2.4.2: It is expected to get more info on this model than it is shown currently. Most of the explanations seem to relate to toxicity test design than to the conceptual model.	Yes, the text has been clarified

	672: please note that parameters and parametr values in AmP database might not be applicable to every type of DEB models. It should be clarified to which kind of DEB models is refereed here. Is it a standard DEB model, or derivations of the standard model, e.g. DEBtox, DEBkiss or any other.	
	908: -Strengths: Both GUTS and DEBtox make use of all available standard and non-standard toxicity test data. The same probably applies to primary producer models. - Please provide a better description of the meaning "strengths"and "weaknesses". For example "Assumes homogeneous mixing of toxic chemical within an organism." This could be considered as an advantage, so the strength, not the weakness. In many cases it may be the conservative assumption, and therefore not necessarily a “weakness” per se. - DEBtox: "Several model formulations..." This should refer only to the toxicity module, not to the physiological model itself. On the other hand, (physiological) model formulations for the same species could still differ depending on the choice of the species and a type of DEB model. - "calibration requires combination of time series..." These are usually recorded in toxicity tests on a regular basis, so this might not represent a weakness. - "advanced knowledge in statistics..." This is not a matter of TKTD models, but to a particular expertise at hand. The same applies to GUTS and Primary producer models.	Table 1 was completely revised following the comments. Nevertheless, assuming homogeneous mixing of toxicants within an organisms is a simplifying hypothesis of biological reality that is chosen for ensure the mathematical handling of the models; so the sentence in the weaknesses column has been retained. Growth is not always recorded over time, so that was kept in the weaknesses column.

3.	DE	Chapter 2	Figure 3: Numbers 1 to 5 in figure 3 do not correspond to assumptions 1 to 5 as in lines 666, but to 1 to 5 appearing further in the text (lines 704-708). This is a bit confusing and could be improved.	The numbering 1 to 5 was removed when not referring to the DEB modes of actions within DEBtox models.
			Section 2.2.: For a better understanding, it would be desirable to have more information or a short description to illustrate how the dominant rate constant KD is derived since it is unclear how this is achieved without measurements of internal concentrations.	The § explaining parameter k_D has been rephrased and a reference to chapter 4, where details are given on its estimation, has been added
			It would be also relevant to add information explaining why the individuals do not need to have reached a steady state to parametrize the model even when internal concentrations are not available, as we understand that the model can be calibrated on the basis of data from tests performed under variable or non- constant exposure	Steady state is no precondition for the use of GUTS modelling. The whole GUTS concept is based on dynamic modelling.
4.	UK	Chapter 3	Section 3.1, line 993 - It would be useful to add a footnote here referencing the EFSA aquatic guidance document section 9.1.3 for further info on how to use TKTD models to assess the toxicological independence of peaks. The suggested method in the aquatic guidance document requires knowledge of internal concentrations in individual organisms.	A reference to the EFSA PPR Panel 2013 has been added
			Section 3.3.3, line 1155 - Please change 'demonstrated' to 'demonstrate'.	Modified
			Section 3.4.3, line 1221 - It would be helpful to clarify why a factor of 3 difference specifically has been selected to assess whether to conduct modelling based on biomass or shoot length/frond number.	The factor 3 was deleted and replaced by considering the confidence intervals around the ErC50 values. If they overlap, then biomass-related endpoints should be selected. If not, the most sensitive relevant endpoint should be selected.

5.	AU	Chapter 3	Chapter 3.4.1 Line 1218-1223: „If in experimental studies the effects of pesticide exposure on growth inhibition of biomass and shoot length/frond number endpoints result in more or less similar ErC50 values (i.e. deviating not more than factor of 3), then it is proposed to select biomass-related endpoints in TKTD modelling. If not, the TKTD modelling approach should be able to predict the response for the most sensitive relevant endpoint.”	See reply to comment 4
			- TKTD models in the context of time variable exposure of aquatic macrophytes still require thorough testing before implementation. A model is either able to predict the relevant endpoint or it is not. The closing sentence, “TKTD modelling approach should be able to predict”, highlights the uncertainty the authors still attribute to the model.	The sentence was modified for clarity; it is now “TKTD modelling approach should be used to predict the response for the most sensitive relevant endpoint”
			- For the purpose of risk assessment the models need more testing to reduce the uncertainty of the model output to a minimum.	The text has been modified for clarification.
			- We don’t quite understand where the factor 3 comes from	The factor 3 was deleted and replaced by considering the confidence intervals around the ErC50 values. If they overlap, then biomass-related endpoints should be selected. If not, the most sensitive relevant endpoint should be selected.

6.	NL	Chapter 3	973-974: Could also any appropriate data set be used fro a TKTD model calibration, not specifically Tier1 and Tier 2A,B data sets? Many data sets could be available and it might be useful to include any available information to calibrate a model.	Text was rephrased for clarification: “For this, Tier-1, Tier-2A and/or Tier-2B toxicity data sets can be used but also can use dedicated refined exposure tests with the selected species of concern.”
			980: Why 2 conc. profiles are recommended, why not 1 or 7 or any other possible number? 980-987: Why at least 2 pulses are required? What if the exposure profile contains only 1 distinctive peak concentration? Why do we need to know wheter carry-over toxicity occurs or not in situations where 1 concentration peak occurs?	Text was rephrased for clarification: “The validation experiment, however, should include at least 2 different profiles with at least 2 pulses each (each tested at least at 3 concentrations leading to low, medium, and strong effects) (see section 7**Error! Reference source not found.** for more details). This is to address phenomena related to toxicological dependence/independence, the modelled internal concentration or damages states (e.g. dynamics between internal and external exposure concentrations) and possible repair of effects”

			1242-1244: But we sometimes do not know if exposure during a more sensitive stage would propagate effects further to a less sensitive stage. Therefore, toxicity test data for the full life cycle or even 2 generation studies might be needed to test the assumption, right?	Text rephrased to avoid confusion: Experimental data sets for calibration and validation of TKTD models may concern a specific life stage (size class) of individuals of a specific species, particularly in acute laboratory-toxicity tests. It is assumed that if the most sensitive life-stage is tested, the calibrated/validated TKTD model most likely will result in a more conservative prediction than when the experimental data set concerns a less sensitive life-stage.
			1311-1312: Maybe this should be left open since for some substances it might still be possible to extrapolate model parameters.	The text was slightly modified to address this issue. It reads now: "In a regulatory context, TKTD models should not be used for extrapolations from one substance to another with the same MoA".
7.	DE	Chapter 3	Line 955: "TKTD models used as tools in Tier-2 assessments need to be calibrated. For this, Tier-1, Tier-2A and/or Tier-2B toxicity data sets can be used but also may require dedicated refined exposure tests with the selected species of concern." Please clarify the second part of the sentence, since in our understanding dedicated refined exposure tests refer to the validation step and not to the calibration step.	The sentence was slightly modified to clarify this issue.

	Lines 961-967. From information in section 4.1.4.5, and for clarification, we suggest to modify/complement the following sentence, as follows: “The validation experiment, however, should include 2 different profiles with at least 2 pulses each (each tested at least at 3 concentrations) to address phenomena related to the modelled internal concentration or damages states (e.g. dynamics between internal and external exposure concentrations) and repair of possible effects. The individual depuration and repair time (DRT95) should be calculated and considered for the timing of the pulses; one of the profiles should show a no-exposure interval shorter than the DRT95 (where toxicological dependence is suspected), the other profile clearly larger than the DRT95 (where toxicological independence is suspected) (see section 4.1.4.5 for more details)”.	The sentence was modified to clarify this issue
	Line 1005-1007: Suggest to shift the sentence of line 1032 (iteration with Tier-1 data for organisms groups) at the end of the sentence from lines 1005-1007 (iteration with Tier-1 data for species within a group).	These sentences were kept as they were since the “iteration with all Tier-1 data” refers to various species within one group of organisms in the first place and refers to various taxonomic groups in the second place

	Line 1025: We are concerned about the following : "… the Tier-2C ERAs need to be calibrated with the RACs and associated exposure profiles derived from the (surrogate) reference tier (see EFSA PPR Panel, 2010…". Indeed in the tiered approach, the RA needs to be calibrated in order to increase certainty on which level of protection is achieved. It is questionable at this stage in which extent a RA based on a Tier 2C approach is sufficiently conservative/ enables a sufficient protection level. Indeed when considering the calibration of risk assessment for aquatic invertebrates exposed to insecticides based on tier 2 A or B- RACs (against RAC of micro-/mesocosm data), it appears that the margin of security is overall not so high. Van Wijngaarden et al., 2015 show that a number of cases are borderline or insufficient (especially when using an SSD HC5 with Af3) although these calibrations are performed with datasets from standard (i.e. constant/ worst case) exposure conditions. Under refined exposure conditions, it is clear that this trend is then shifted towards even less margin of security. In addition, for the effect assessments of vertebrates that should benefit of a higher level of protection, there are high uncertainties since such a calibration exercise is not possible (no surrogate reference tier available for vertebrates).	This issue is mentioned now at the end of section 3.2
	This issue is not specific to TKTD model but applies to the overall Tier 2C approach ; i.e. exposure refinements either addressed with experimental or modelling approaches. However with the introduction of TKTD models , this concern of a sufficient level of protection is amplified as with the TKTD tool, the use of the Tier 2C approach in RA will be emphasized (in particular for vertebrates as it enables a better consideration of animal welfare issues). We would appreciate this issue to be addressed in a revision of the GD and to be mentioned in this ScOp.	Noted

			Line 1275-1278: "Specific regional ecological scenarios, including regional focal species, …and Tier-4 assessments for risks at larger spatial and temporal scales. Extrapolation in space is implicitly done when exposure time series of different locations are evaluated by compound and species-specific TKTD models". Please clarify as when including the component spatial and temporal scales, regional ecological scenarios will not only include focal species but also mixtures of compounds applied simultaneously or successively. Thus the link to the last sentence "evaluated by compound" is unclear.	This issue was addressed by adding this sentence: "It should be noted that if landscape scale is addressed, relevant exposure scenarios including multiple stressors may become relevant."
			Line 1286-1287: "TKTD models can be applied independently from the MoA, but for their application it has to be checked whether unexpected mortality is observed under long-term (chronic) exposure". This means that any species modelled and tested acutely should be also tested chronically (not only for validation , i.e. under refined exposure conditions) but under constant/ standard conditions in order to check for unexpected mortality. It would be interesting to mention that some substances have a quasi-irreversible/ irreversible MoA and the implications for the dynamic and repair of damage processes in TKTD models.	This issue was addressed by modifying the next sentence :" In such cases, toxicity cannot be predicted based on acute testing only and calibration and validation experiments should be available on longer time scales in order to detect such potential delayed lethal effects."
8.	UK	Chapter 4	Section 4.1.2.3.0, line 1541 - Please change 'what' to 'which'.	Thank you. The text was modified accordingly.

9.	AU	Chapter 4	Chapter 4.1.2.4 Figure 11: Clearly all parameters are influential. As a consequence small deviations from the reference parameter values are hugely influential and therefore great care has to be taken choosing parameter values to avoid biasing the results. It is therefore very important to comprehensibly calibrate the parameters to allow the assessor to gain a clear and informed understanding of the model at work.	Small deviations and huge influence are relative. Of course it is important to perform a sound parameter estimation (see paqe 46: 'all parameters can be considered to be important and have to be calibrated carefully').The sense of a sensitivity analysis is, however to discriminate whether some parameters are more influential, than others, and this is not the case in the case of the GUTS models. Hence, no changes in the document.

	Chapter 4.2.1.1 Line 2208-2215 For a calibration dataset this is on the small side. Given the importance of calibration due to the influence of the parameters we would like to see a larger dataset implemented to assure good, reliable calibration. We don't think 2 replicates per concentration can catch the variation in the data appropriately, also the requirements state that there should be at least 5 time points (line 1934-1939) measured whilst here we are dealing with 4 measured time points only (not considering the 0-start measure). For the purpose of testing the model it would also be desirable to calibrate with more than 1 dataset to get a comparison of the variation in parameter output.	Again, 'small' is relative. A sentence was inserted in the text: 'The relatively small number of only 2 replicates is counterbalanced by testing 7 treatment levels. The data set was used in a ring test for GUTS models and is considered realistic and sufficient for the parameterisation of GUTS modelling by more than 10 scientists from different affiliations. The WG intended to enable the use of observations over time from standard toxicity testing, hence what is meant in lines 1943ff is to use initial abundance plus four observations over time. This has been clarified in the text there.
	Chapter 4.2.2.1 Figure 16: Replicate beakers have been lumped into one; therefore the variation between replicates has been lost. We would like to see a "black dot" for each replicate in Figure 16 to get an understanding for the variance in the raw data (this is why more replicates would be useful to give 95%CI (confidence intervals) in the first place)	Done

	Chapter 4.2.2.2 The parameter estimates from the calibration-dataset are based on an entirely different setup than the validation-dataset (single regime per beaker vs. time variable exposure, respectively). We would like to see the performance of the model after training on a comparable training set (one could probably use randomly chosen replicates of the time variable dataset (e.g. 2 out of 7) for calibration and test the performance on the remaining (in this case 5) replicates; not optimal, but better than the situation at the moment) as well as dataset used at the moment for parametrisation to be able to compare the relevance of an appropriate trainingset for parametrisation.	The intention of GUTS modelling is to calibrate on one experimental test profile and to extrapolate to another. The wish of the MS appears interesting, but the WG does not see why this is urgently required. The prediction of the experts in the WG is that one can play around with bootstrapping etc a lot, but this will not at all change the model predictions, and also not the confidence in the model results. It would be desirable to invest research time and money to clarify this question.

	The time variable dataset does not comply with the minimum requirements for a validation dataset as set out in 4.1.4.5 (specifically line 2111-2115 has been violated: "The individual depuration and repair time (DRT95; see below) has been calculated, and the duration of the no-exposure intervals was defined accordingly; one of the profiles shows a no-exposure interval shorter than the DRT95, the other profile clearly larger than the DRT95; In case DRT95 values are larger than it can be realised in validation experiments, or even exceed the lifetime of the considered species, the second tested exposure profile may be defined independent from the DRT95"). The DRT95 has been determined to be 10h, whilst the time intervals between the pulses are 2 and 6 days respectively. For a better reference a shorter time interval than DRT95 should be chosen to understand additive/synergistic effects.	The WG has also acknowledged the mismatch between the minimum requirements and the data set used. The time variable dataset used for validation was not produced for this SO, but in the scope of earlier PhD theses. The failure of full compliance with the minimum requirements in the SO does in the eyes of the WG not prohibit the use in the SO, since these data sets are the best data that are available., Moreover, just because these data could be used, the WG was able to formulate the minimum requirements.
	Figure 21: This figure highlights the shortcoming of the observed data. There is no difference between A+B and C+D. The second pulse shows no effect in the observed data which argues for a resistence/tolerance mechanism or some other design flaw. Accordingly the model predicts the data poorly with the surprising exception of B-IT, where the model predicts a negligible effect of the second pulse. 95%CI (confidence intervals) on the observed data would help in the interpretation of the model performance tremendously. Figure 22: More than a single observed multiplication factor would be desirable!	In the eyes of the WG, there is a difference between C+D (constant low level exposure) and D (control).The WG wonders on what basis the MS identifies a 'poor prediction'? 95% CI are included now. The data are discussed in more depth in the revised text The WG agrees; this is why it is included in the minimum requirements.

			Line 2439-2441: “Overall, the quality of the model predictions appears as acceptable, when accepting a maximum level of 50% deviation between predicted and observed numbers, especially when assuming that toxicological effects change on a logarithmic scale rather than on a nominal one.” 50% deviation seems a lot in the context of risk assessment especially when the model might underestimate mortality (e.g GUTS-RED-IT).	The choice of the 50% is a proposal by the WG, which could be revised based on further experience and information. The choice of 50% appears protective because the GUTS model output is used in combination with Tier 1 assessment factor. The recommendation in the cited section is clearly to use the model that does not underestimate mortality.
			Chapter 4.2.3.1 + 4.2.3.2 I don’t think one can come to any conclusion other than the model catches the TER under constant conditions and fails under time variable conditions (see my comments before).	As expected, the GUTS model reflects TER under constant conditions. This is a convincing confirmation of the conservatism and appropriateness of GUTS. In case GUTS predictions would result in same risk estimates as Tier 1, the whole modelling would be redundant. Instead, it is the declared aim of GUTS modelling to obtain refined risk estimates in case of considerably variable exposure patterns in time.
10.	NL	Chapter 4	1696: subheading: "... goodness-of-fit" but not much is said in this section about how to judge the goodness-of-fit	Has been changed into ‘Parameter optimisation and likelihood’

	1934: How to evaluate the model if less than 5 time points are available, why is that important, how it affects our trust in the model, how to deal with such data sets? This is of a great importance, especially for 48h toxicity tests. Therefore, we think that this issue should highlighted and maybe a bit more explained.	In lines 1937ff it reads 'If a standard 48 hour-study is only available, a calibration might still be attempted but the quality of the fit (convergence, uncertainty limits and visual fit) should be carefully checked' This is the maximum advice that the WG can give in this SO, which is no guidance. In limit cases, expert knowledge and probably consultation of experts is necessary.
	1940-1941: Another useful information would be whether the steady state has been reached and what is that important (if it is)? What if the steady state has not been reached, what would be the concern of risk assessors?	The application of GUTS is not depending on having reached steady state.
	1975: "...and validated GUTS model" It is a bit confusing because the validated model would be the one being used for predictions. It would not be apriori validated.	A clarifying sentence was inserted.
	1978: "target time-point" This might better be defined to conform guidelines.	The text is clear enough here in the view of the WG no further definition is needed.

	2010: Could you please clarify what is meant with "the former being due to insufficient information within the model"? For example GUTS works only if it is fully parameterized (not partly), so only when all information is provided. Does this relate to a conceptual model, methods for parameters estimation and statistical treatment of uncertainties, or something else? Or it relates to ecotox test design? It is because uncertainties in model predictions are supposed to propagate from model parameters and uncertainties in model parameters are supposed to come mainly from biological observations and test design.	One example for insufficient information is the case where calibration data do not fully capture effects, e.g. because the dose-response curve was not fully covered. Of course all parameters are needed, but the estimates vary in their quality caused by the quality in the underlying data.
	2011: Could you please clarify what is meant by structural uncertainty of a GUTS model?	The question whether the processes included into GUTS is sufficient to capture the observed effects over time.

	2103: Does it really need to be only time-variable exposure, or constant exposure would also suffice?	The application of GUTS modelling is intended to assess the case of time-variable concentrations in the environment; hence the WG is of the opinion that it is of critical importance that the validation is performed under time-variable exposure. The described case of 1 application and FOCUS exposure profile with 1 peak only does not prevent that under environmental conditions exposure to more than 1 peak appears realistic. There is no reason to allow for special cases for validation, depending on the exposure. This would mean, that in other cases also 3,4, 5,... pulses would need to be tested..
	2108: Why is it insisted on (multi)pulsed exposure? Would constant exposure of different time duration and/or different test concentration suffice?Or test with only one pulse?	See reply to comment 6

	If a GAP assumes 1 application and FOCUS profile gives only one distinctive peak, does the model need to be validated with predictions on pulsed exposure? What is the importance for risk assessors of having "pulsed" test in this particular case? If it does not bring much	See reply to comment 6
	into the RA, a test design with pulsed exposure might bring unnecessary testing or costs for applicants.	
	Another point is what duration of pulses are recommended and what in case of the substances with fast degradation rates, within minutes and hours? Should pulse durations as simulated by fate models be used? 2130: Why would it be important to demonstrate toxicological dependence/independence model validation? Especially in case of one application? 2132: Even for calibration test it would be important to have a test of a sufficient duration. It is because the model cannot predict latency of effects if it was calibrated with the data that do not indicate the latency. 2139: Model performance criteria. Do quantitative model criteria apply regardless of the approach of frequentist or Batesian approach? Another question: in Jager and Ashauer 2018 (book), it was mentioned that Akaike information criterion (AIC) can be used. What is the opinion about it?	Pulses should be long enough to see effects during the exposure, but mimicking simulated exposure (Focus) is not required True, but this is the reason it is recommended to use observed effects under both acute and chronic exposure. Yes, a corresponding sentence was inserted there. AIC values are informative, but values cannot be compared with absolute cut-off values. AIC informs mostly about performance of different models for one data set, not of a specific model for different data sets.

11.	FR	Chapter 4	In Figure 7, parameters that must be determined from experimental data are marked in red. While the description of each parameter is explained throughout the chapter it is not clear how these parameters can be derived from experimental data. More details on the procedure to derive such parameters from the experimental data should be provided.	A sentence referring to the respective part of the document is inserted now: Calibration of model parameters is explained in section 4.1.3.
12.	DE	Chapter 4	Line 2162: the datasets are taken from an example from the literature. They do not correspond to the recommendations made in this Sc Op (e.g. for validation, the various profiles are not tested at various concentrations (it should be a min. of 3); they tests are only of 10 days duration; there is no chronic test under constant/ standard conditions in order to check for unexpected mortality). These “deviations” should be clearly stated.	A respective sentence was inserted: “The dataset was not produced specifically for this SO, but in the scope of an earlier PhD thesis. The failure of full compliance with the minimum requirements as formulated in this SO (section 4.1.4.5) does not prohibit the use as demonstration data set, but the use of the data helped to derive the minimum requirements instead.”
			Line 2446: “LP50 are compared with a Toxicity Exposure Ratio (TER) as equivalent to the lower tier RAC calculation”: a TER is different than a RAC, please reword; in this case the RACacute is 0.192 µg/L	The text has been reformulated.

13.	AU	Chapter 5	Chapter 5: DEBtox models show a lot of promise, but have not been validated yet. It would be desirable for the assessor in the future to have a set of OECD guidelines on how to perform standard tests to gain the necessary data for model parametrisation in an user-friendly DEBtox testing environment. If everybody would be allowed to use “homemade” scripts on “homemade” experimental setups it would become impossible for the assessor to evaluate the quality and reliability of the data presented and hence could not be used in risk assessments. We need standardized setups and tools for a fair and reliable assessment.	Noted. Thank you. The comment is agreed. This is also one of our conclusions.
14.	NL	Chapter 5	2722: Chapter 5.3. It should be clarified what is meant under a DEBtox model. Would that be any DEB model, e.g. standard DEB, DEBtox or DEBkiss, or any model that follows the DEB framework?	The beginning of section 5.3 was revised to clarify that the example uses a DEBtox model (with reserve in steady state ad compound parameters, based on revised equations from Billoir et al. (2008).
			2728: DEBtox models. It would be more appropriate to refer to 5 DEBMoA than to 5 DEBtox models. In this case, the physiological model is the same, but the TKTD module consist of 5 possible toxicity MoAs.	The sentence was changed as follows: “[…] involved one of the five DEB modes of action (DEBMoA) that can be tested […]”.
15.	NL	Chapter 6	No particular comment on this chapter	Noted

16.	UK	Chapter 7	Section 7.7.2, line 3593 - It is indicated that the second tested exposure profile may be defined independent from the DRT95 if the DRT95 exceeds the lifetime of the considered species. It would be helpful to clarify this statement regarding algae and macrophytes. Given the discussion in section 2.4.1 of the difficulties assessing such organisms in TKTD modelling at an individual level, what is the relevant lifetime that should be considered?	A comment that the lifetime relates to animals has been added. A sentence has been added for algae and macrophytes but specific guidance on the interval between peaks can be provided.
			Section 7.7.2, line 3665 - 'Qualitative' should be replaced by 'quantitative'.	Agreed and corrected.
			Section 7.8, lines 3694 & 3697 - It is stated that in the absence of standard software, it would still be of benefit to both the applicant and the assessor if an implementation of the model can be provided. From the perspective of a regulator, having access to an implementation of the model is considered essential rather than beneficial. Having access to the source code is unlikely to be sufficient.	The comment is in principle agreed. This is one of our conclusions however this is not considered practical at this stage.
			Section 7.8, line 3701 - It is stated that for GUTS background mortality as observed in the calibration or validation data sets should be noted, but for simulations, background mortality is assumed to be 0. It may	

			be helpful to add that the background mortality rate constant may have been calibrated to the observed mortality in the controls and be fixed in the calibration of the other GUTS parameters to data from the treatments, as discussed in line 3494	This section is about simulations from TKTD models, not about calibration. When simulating the survival rate over time (SOT) under a given exposure profile, the output of interest for RA is the *relative* decrease in SOT in comparison with the control. As the background mortality rate applies in both control and contaminated cases, its value can be fixed, for example at 0, without change in the prediction of the relative decrease in SOT.
17.	AU	Chapter 7	Chapter 7.7.1: Line 3537-3538: "In general, the effort for an analysis of model uncertainty should be in balance with its importance and use in the regulatory risk assessment." The assessor can only make an informed assessment on a model if he knows about the uncertainty of the model and hence the analysis of model uncertainty should be an integral part of the regulatory risk assessment	This sentence has been deleted
18.	NL	Chapter 7	3287: Chapter 7.2. We suggest to also consult the article regarding test design for TKTD models: Jager 2014. Reconsidering sufficient and optimal test design in acute toxicity testing.	The text has been modified after the consultation of the suggested paper

	3295: "...the most closely related guideline should be followed as far as possible." It should be considered that the most appropriate way to deal with the data quality would be to leave the test design flexible and fit for purpose. This way a test design would be tailored to the modeling needs and thus the obtained data would be of a great value for model calibration/validation. Standard toxicity tests are generally not fit for TKTD modeling purposes, therefore, the data might not be the best match regarding e.g. estimation of model parameters. We advise this to be reconsidered in the final version of the SO. When considering toxicity tests for model calibration, the test design should also conform the conceptual model. For example, if data are to be used for a DEB model, animals should be fed in the toxicity test.	Wording has been added to clarify this. The WG thinks this is covered since animals are fed in the chronic tests.
	It has to be stressed that all available data can be used for a model calibration (GUTS and DEBtox at least), even those form other toxicity tests (after scrutinizing the data).	Agree, all acceptable data should be used, unless it is used for the validation.
	3313: "...not recommended." For this matter we advise this to be stated more strongly. If screed data has been used the model itself would not be considered viable (unless a good reason for exclusion of some studies was provided).	Agreed – this has been changed to not acceptable.

	3319-3324: Model parameters for the animal in control conditions might differ from test to test because of differences in test conditions and animal strains. Therefore, even data for physiological part of the model might need checking every time and new model parameter values estimated for control conditions (and possibly compared to historical information or data from other labs). Only when this has been done, toxicity module parameters should be evaluated. Furthermore, we do not think that MORE EXTENSIVE evaluation for physiological parameters would be needed - we think that an evaluation of the parameters is always needed, but would not phrase it as an extensive evaluation.	The wording has been adapted to clarify that it is a separate evaluation rather than a more extensive one.
	3335-3340: Chapter 7.3, general question: Why is it not considered useful to apply GUTS for refinements of birds and mammals risk assessment?	The remit for this SO was only aquatic and so does not cover terrestrial organisms, but it is agreed that TKTD is also useful for terrestrial organisms.
	3342: Maybe a clarification is needed to which model concept needs checking. It is because the concept of the DEB theory has been successfully tested for more than 30 years of research. Therefore, the DEB theory concept has been well tested, but might not be so for a particular DEB model in question.	This has been clarified in the text.
	This particularly holds since the concepts for the models for primary producers are considered sufficiently addressed in this SO, but they have been significantly less explored and tested than the DEB concept.	The WG agrees that more data are needed. However the WG is confident that the conceptual model is good from what available.

	3352: We are of the opinion that there is no sufficient evidence presented in this SO for the acceptability of the concept for Lemna and algae models. It seems unusual to consider the concepts acceptable based on relatively little information provided. Is there more information available in public literature or elsewhere regarding the (proof of the) concepts of the Lemna and lagae models?	The DEB modelling is applied for many species while the model for *Lemna* sp. and algae is just applied to few species.
	3391: The link leads to an error page	Thank you, this has been corrected
	3419-3422: Could you please clarify what is meant by these two sentences? Environmental scenarios can also be simulated in the lab. Furthermore, it is not clear why temperature is considered important in DEBtox, but unimportant in GUTS, please clarify?	The conditions should match between the experiments conducted and the environmental scenario modelled. This isn’t relevant for the GUTS models because there are not inputs relating to environmental conditions.
	3422-3425: If this is demand for TKTD models, why not put the same requirements for the endpoints derived from standard toxicity tests (and also when using them for SSDs and Geomean approach). Why should much more stringent criteria be put on more robust methods of analysis such as TKTD (DEB) modeling?	The Opinion is not suggesting more stringent criteria in TKTD models, just that the model and data used should match and allowing for additional factors to be included if there is the data.

	3426: Figure 38. The scheme seems confusing. Could you please explain what is meant exactly with inclusion of abiotic parameters via the physiological part of DEB/plants model? Abiotic factors should be part of the environmental scenario, and TKTD models have no place in that box.	Abiotic parameters include factors such as nutrients for the plant models. The physiological parts of some models might require information that is not covered in the environmental exposure models, so there needs to be an option to include these factors. Figure 38 has been revised.
	3457: It has been stated in one of the publications of Tjalling Jager that even a toxicity test with a small effects can be used to calibrate a TKTD model if enough data points are available. Could you maybe comment on it in relation to the requirement of minimum 50% effect as proposed in this SO?	This is may be possible however to have a robust response the full dose response should be covered.
	3464: "... since all available reliable data sets (apart from those used for validation) should be used." We would add "... since all available reliable data sets (apart from those used for validation) should be used (even non-standard test data when scrutinized)."	This has been added.
	3561:"... there is no "gold standard." It might be more accurate to state that test design should be tailored to the question at hand.	This has been done.
	3589-3590: It is not clear is it always necessary to have at least two pulses and why it is needed to know in every case whether or not carry-over toxicity occurs?	See replies to comment 6.

19.	NL	Chapter 7	3736: Table 5. It should be highlighted that the observation of the endpoints in the toxicity tests should be as frequent as possible to reduce uncertainty in model outputs. It also should be highlighted that validity criteria as defined in the OECD do not necessarily need to be the same criteria for TKTD models. TKTD models are mechanistic models and as such they place different demand on test design as compared to standard dose-response fitting methods.	The comment is agreed – the table headings have been changed to make it clear that the second column just describes standard test guidelines and the third column discussed the relevance for studies supporting TKTD modelling.
			- "The mortality in the control(s) should not exceed 10 percent (or one fish if less than ten are used) at the end of the test." This is also relevant for calibration of background mortality	This has been emphasised.
			- "There must be evidence that the concentration of the substance being tested has been satisfactorily maintained, and preferably it should be at least 80 per cent of the nominal concentration throughout the test. If the deviation from the nominal concentration is greater than 20 per cent, results should be based on the measured concentration." Keeping exposure concentration constant should not be a prerequisite for dynamic models as long as the actual concentrations are reported. This should be clearly stated because it might happen that risk assessors reject the study only because the concentration was not maintained within a desirable range. Is there a recommendation how to deal with e.g. fast dissipating substances, should the exposure be assumed as initial measured conc. or as mean measured conc. over the experimental duration? Would that have a consequence on TKTD model parameters (values and uncertainties)?	The actual measured concentrations over time should be used (so this could be a single short peak or several short peaks in semi static systems). A sentence reflecting this has been added to the SO.

			-"The water temperature should not differ by more than + 1.5°C between test chambers or between successive days at any time during the test, and should be within the temperature ranges specified for the test species (Annex 2 of the guideline)." We think that as long as the conditions are reported and the model is capable to account for the variations, the temperature might not need to remain constant. Actually a great advantage of mechanistic models that the test design can be flexible.	Since the GUTS models do not include temperature, it is suggested that constant conditions should be maintained so the effects of the toxicant are not confused with effects from the temperature variation. For DEBtox or plant models if the model includes temperature, this could be varied in the study design.
			-"The coefficient of variation of average specific growth rates during the whole test period in replicate control cultures must not exceed 7% in tests with Pseudokirchneriella subcapitata and Desmodesmus subspicatus. For other less frequently tested species, the value should not exceed 10%." and "The mean coefficient of variation for yield based on measurements of shoot fresh weight (i.e. from test initiation to test termination) and the additional measurement variables (see paragraph 37 of this guideline) in the control cultures do not exceed 35% between replicates." This seems relevant for standard dose-response statistics. Would TKTD modeling require different criteria in this respect?	It is considered that this should still apply because it relates to the ability of the test system to detect effects. The WG and the Panel is not aware of any reason why more variable controls would be a benefit in studies used for TKTD modelling.
20.	FR	Chapter 7	L.3228. More details and/or a reference to section of the document (chapter 7.5?)might be needed to define "sufficiently well documented".	Reference to chapters 4, 5, and 6 added.

	L.3252. The reference "(see recommendation section)", should be clarified.	The cross-reference was added.
	7.5 The GUTS and primary producers models seem to be the most advanced modelling tools in comparison of DEB. Still, for harmonization purpose of the EU risk assessment, based on the restricted diversity of designs of both models, "easy to use" (meaning routinely used) software or toolboxes should be foreseen for models considered suitable for the risk assessment.	The comment is agreed – see recommendations chapter 9.2.
	Recommendation regarding the use of modelling for regulatory purposes should be mentioned. For instance, for regulatory purpose and practical case, repeating systematic checking of the computer model and computer code when a new model implementation would be submitted does not seems feasible and pragmatic (time, skills, resources constrains).	The comment is agreed – see recommendations chapter 9.2.
	For efficiency purpose of the regulatory risk assessment and for facilitating TKTD approach as higher tier, an "user friendly" agreed tool box should be elaborated. In addition, it would facilitate the peer-review process of the risk assessment at EU and zonal levels.	The comment is agreed – see recommendations chapter 9.2.

			7.7.2 The regulation about animal welfare of vertebrates included in 283/2011 and 284/2011 states that « In order to minimise fish testing, a threshold approach to acute toxicity testing on fish shall be considered. ». Regarding the vertebrates, some case in context of regulated substances could be addressed with relevant Tier 1 studies and endpoints. Higher Tier would ideally require generating new studies, especially for validation of calibrated models. Therefore for further applications of TK/TD modelling, consideration regarding acceptability of additional vertebrate studies should be elaborated. Should additional vertebrates studies be required in case of: - No safe use being identified at lower tiers (without modelling). - TKTD Modeling needed to demonstrate a safe use for some intended uses while other intended uses are acceptable with lower tier (without modelling). - Safe uses identified at lower tier (without modelling), but modelling used for reducing mitigation measures. - They are optional to ensure a better modelling design?	The text has been expanded to better cover the issue of the reduction of vertebrate testing.
21.	UK	Chapter 8	Sections 8.3 & 9.2 - The use of a detailed example is helpful, especially as it includes the lower tier/2A/B approaches which allows the reader to consider the use of TKTD modelling in the context of the current aquatic risk assessment approach. It is suggested that the addition of similar examples using DEBtox and primary producer models are added with the future development of these approaches. This suggestion could perhaps be added to the Recommendations and future perspectives in chapter 9.2?	The suggestion given in the comment is already included in section 9.2
22.	NL	Chapter 8	3766-3769: This is difficult to follow, please clarify it. Would it be considered useful to apply TKTD modelling even when PECsw and PECsw-twa are above teh Tier1 RACsw,ch?	The text has been clarified.

	3775: Could you please explain why this criterion is selected (one order of magnitude)? It might be useful to just select the most sensitive species for which tha data are available.	The most sensitive species in a Tier 1 experiment might not be the same species under time variable exposure due to e.g. different repair mechanism. Therefore, an order of magnitude difference is proposed to cover possible shift in the sensitivity ranking.
	3789 and 3792: Since this is a refined risk assessment, why does the same AF as in Tier 1 assessment remain?	The use of the same AF as in Tier 1 is in line with the Aquatic Guidance document (EFSA PPR Panel, 2013)
	3801: 8 aquatic arthropods/or primary producers. It will probably be rare that TKTD models are available or well calibrated/validated for all 8 species. The same applies for fish.	Noted
	3809: SSD approach. Would it be considered appropriate to combine endpoints (LC/EC50, not EP/LP50) calculated by using standard dose-response models and TKTD models, both derived from standard toxicity tests? Theoretically, there should be no objections on it, and even TKTD derived endpoints could be considered more robust.	The WG and the Panel consider that this is possible.
	3839: Can these endpoints (LP50/EP50) really be used for calculations of HP5 and consider it comparable to HC5 and SSD? It is because the assumptions and statistics behind the SSD are based species sensitivities (LCx/ECx) and not on exposure multiplication factors LPx/EPx? Could you please elaborate on this?	It is considered that the same assumptions still apply to SSD built using LPx/ECx.

23.	DE	Chapter 8	Line 3784: please see comment from line 1025 regarding the protection level actually achieved when using tests performed under standard conditions for the calibration of lower towards higher Tier	The protection level provided by Tier-2C assessments should be calibrated with results of the (surrogate) reference tier. In the example data set, presented in the Opinion, it is shown that the Tier-2C approach resulted in realistic worst-case predictions when compared with results of a mesocosm experiment. This calibration should of course be conducted for more compounds deviating in exposure characteristics and toxic mode-of-action. This is recommended in the Scientific Opinion.
			Line 3793- 3800: the recommendations of AF ≥ 10 and ≥ 4 for acute and chronic RA under step 5 would need to be better justified. In our view, these are not sufficient: for acute risk assessment, assuming that the AFoverall of 100, the AF species and AF others have an equal weight (i.e. 10 and 10) (as stated in EFSA 2006), the AF overall recommended here should be ≥ 20 to account for a partly (but not fully) reduced AFspecies, remaining (>1 and <10). Also referring to line 4259	The recommendations already formulated in the sediment scientific opinion (EFSA PPR, 2015) were followed. Furthermore, it is clearly recommended in this TKTD Scientific Opinion that a transparent WoE procedure should be developed (coordinated by EFSA) with criteria for the reduction of the AF if toxicity data for additional species are available (dependent on the quality and number of these data)

24.	NL	Chapter 9	4371: It should be clearly stated which DEB model is preferable, standard DEB, DEBtox, DEBkiss or any model that is built within DEB framework.	At this stage this question is considered difficult to be answered. The DEB part of the model needs to be approved for regulatory purposes and the TKTD part for a specific active substance should be sufficiently calibrated and validated given the criteria in Chapter 7
25.	Sweden	General comment	Our experience in using TKTD modelling in environmental risk assessments is very limited and therefore we have not prepared any detailed comments related to the scientific or technical aspects of the draft opinion. From a Member State perspective we have a general concern with the increasing complexity in guidance for ERA. Substance evaluations are becoming resource consuming and it is a challenge to cope with the timelines in the EU regulation. Further, the development of complex environmental risk assessments also results in a knowledge gap between risk assessors and risk managers/decision makers, which in turn has a negative effect on the confidence in the authorisation process.	Noted.

			In this particular case, we do not find it acceptable to disregard identified risks at Tier 2A/2B just because a risk at a specific exposure profile has been deemed acceptable at Tier 2C by using a validated TKTD model. The risk in other potential exposure profiles occurring under realistic and representative natural conditions may still be unacceptable. We cannot see that introducing TKTD models would reduce the uncertainties of current standard ERA.	The conclusion on low risk for a single exposure profile cannot be extrapolated to other exposure profiles. In addition, any fate exposure models or monitoring data could be used as input exposure profiles. However, at EU level the FOCUS exposure model is the basis for the PPP environmental risk assessment. Although the use of TKTD does reduce uncertainties of risk assessment in terms of more realistic exposure, the same AF as Tier 1 or Tier 2A or 2B are still used.
26.	AU	General comment	General comments: In our opinion the models, as they are right now, are not fit for implementation in risk assessment procedures. We do see the great benefit and think the models are on the right track, but right now clarity and appropriate validation are lacking for all of the model implementations.	The WG has concluded that the 3 types of models covered in this opinion are in a different stage of development. See abstract and section 9.1.

			We would welcome clear guidelines for standardized experimental procedures to get reliable data for parametrisation. Also the experimental procedures for the validation dataset set would need to be standardized in order to assess the goodness of fit of the model reproducibly (here clear guidelines on appropriate cut-off values would need to be provided), only if all of these preconditions would pass the mark we would deem a model good enough for extrapolating reliably to experimentally untested scenarios.	Recommendations have been given on how to evaluate data for calibration and validation in section 7. It is essential to gain experience with the use of TKTD models before clearer guidance can be given and in line with other areas of the risk assessment improvements can be made over time.
27.	FR	General comment	Anses wishes to thank EFSA PPR for this opportunity to comment the Scientific opinion of the state of the art of Toxicokinetic/Toxicodynamic effect models. This document is well explained and provides a valuable overview of the state of the art of TK/TD modelling possibilities and limitations.	Noted, Thank you.
28.	UK	General comment for chapter 2 and 4-7	General comment for chapters 2 and 4-7 - Suggestion on the layout of chapters: Chapter 2 contains some useful foundation knowledge on topics covered in chapters 4-7. Chapters 4-7 read as though the author should already have a good understanding of these foundation topics. For example, chapter 2.2 explains concepts like Stochastic Death (SD) and Individual Tolerance (IT) in basic terms, which are not reiterated in chapter 4 (which describes the GUTS model in great detail). It is suggested that some of the information presented in chapter 2 may be more usefully presented as introductions to the relevant topics in chapters 4-7. Each chapter would therefore be more self-contained and easier to reference by individuals looking up a certain topic. The topics may also be easier to grasp for someone who is new to TKTD modelling as the whole document would be more modular and basic principles would be fresher in the mind than if they were introduced in a separate chapter. Section 2.2, line 617 - Please delete the word ‘be’.	The Opinion has been revised accordingly by putting a guide for readers in the summary. In addition, the concepts of GUTS-RED-SD and GUTS-RED-IT has been reiterated in chapter 4.

29.	FR	General comments	Anses wonders how TK/TD modelling should be considered if the risk for aquatic organisms is acceptable when mitigation measures are applied. Anses acknowledges that TK/TD modelling is a potential refinement option, but should these be used in order to reduce and/or remove mitigation measures when the risk assessment could already be acceptable with actual risk assessment guidelines?	Thank you. TKTD models can be used as refinement options. The use of TKTD models for removing and/or reducing mitigation measures is discouraged in case of vertebrates if this would require additional vertebrate testing. See chapter 7.
30.	DE	General comments	Thank you very much for this comprehensive and very well written Opinion, which we think will be very valuable in a regulatory context	Noted, Thank you
31.	UK	Summary	Line 196 - Please remove “Lemna” at start of sentence.	Thank you for spotting. The text has been amended.
32.	NL	Summary	126: we suggest to re-word "user-friendly" to "more user-friendly" or "easier-to use" interfaces are available, or similar	The Opinion was not amended in line with the comment because it is considered that user-friendly tools are not yet available (e.g. expert knowledge is needed with current tools).