User guide MOSAIC$_{\text{bioacc}}$ (Gamma version)

A. Ratier, C. Lopes, G. Multari, S. Charles

April 22, 2021

Contents

Introduction

Objectives

The purpose of this document is to introduce you how to use the MOSAIC$_{\text{bioacc}}$ application. This application is based on the R software[1] and especially the `rjags` library (version 4.10)[2], to estimate parameters of Toxico-Kinetic (TK) models under a Bayesian framework. MOSAIC$_{\text{bioacc}}$ is developed as an R-Shiny interface (version 1.6.0)[3].
If you want to be kept informed, please email us: sandrine.charles@univ-lyon1.fr.

Context

The MOSAIC$_{\text{bioacc}}$ application is a turn-key web tool providing bioaccumulation metrics (BCF/BMF/BSAF) from a TK model fitted to accumulation and depuration data. It is designed to fulfill the requirements of regulators when examining applications for market authorization of active substances.

Toxico-Kinetic/Toxico-Dynamic (TKTD) models are used to describe and predict the toxicity and the effects of chemical substances on individual traits based on experimental data. The TK part describes the relationship between the exposure medium and the organism, considering various processes such as ADME (accumulation, depuration, metabolization and excretion)[4]. Regulation No 283/2013 (EU)[5] defines the data requirements for active substances of plant protection products in marketing authorization applications. In particular, a bioaccumulation study on fish is required following OECD guideline 305[6]. Achieved in agreement with EFSA's scientific opinion on good modeling practices[7,8], this ready-to-use on-line service allows to easily estimate BCF/BMF/BSAF as required in a regulatory framework, accounting for bioaccumulation of parent compounds and their metabolites through biotransformation. MOSAIC$_{\text{bioacc}}$ does not expect any input besides the accumulation-depuration data sets according to exposure concentrations. The service automatically fits the TK model, initially defined from the appropriately user data and optimizes the estimation of its parameters. Then, the service provides the corresponding bioaccumulation metrics, as well as all goodness-of-fit criteria required to carefully check the reliability of the results[9]. All calculations are based on the JAGS software and its companion R packages `rjags`[2,10] and `jagsUI`[11].

Installation

If you use the web interface (https://mosaic.univ-lyon1.fr/bioacc), you don't need to install anything.

However, if you want to run the R script (downloadable from the application) by yourself, you need to install:

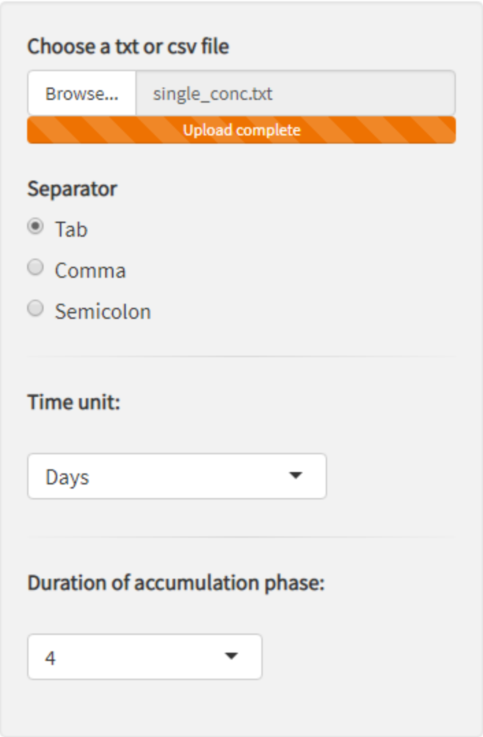
- the JAGS software[2]. Refer to http://sourceforge.net/projects/mcmc-jags/ to proceed.
- the R software[1]. Refer to https://cran.r-project.org/ to proceed.
- the `rjags` package[2]. You can install it directly from the R software > Tools > Install Packages > rjags or from the CRAN website http://cran.r-project.org/web/packages/rjags/index.html.
- the `jagsUI` package[11]. You can install it directly from the R software > Tools > Install Packages > jagsUI or from the CRAN website http://cran.r-project.org/web/packages/jagsUI/index.html.
- Others R packages necessary to run the application: `tidyverse`, `gridExtra`, `ggmcmc`, `GGally`, `ggmcmc`, `stringr` and `DT`.

Here is an example of the R code to install the required packages:

```r
if(is.element('rjags', installed.packages()[,1]) == FALSE)
{install.packages('rjags')}

if(is.element('jagsUI', installed.packages()[,1]) == FALSE)
  {install.packages('rjagsUI')}

if(is.element('tidyverse', installed.packages()[,1]) == FALSE)
  {install.packages('tidyverse')}

if(is.element('gridExtra', installed.packages()[,1]) == FALSE)
  {install.packages('gridExtra')}

if(is.element('ggmcmc', installed.packages()[,1]) == FALSE)
  {install.packages('ggmcmc')}

if(is.element('GGally', installed.packages()[,1]) == FALSE)
{install.packages('GGally')}

if(is.element('stringr', installed.packages()[,1]) == FALSE)
{install.packages('stringr')}

if(is.element('DT', installed.packages()[,1]) == FALSE)
{install.packages('DT')}
```

1 Step 1: Data uploading

When using MOSAIC$_{\text{bioacc}}$, the first step is to upload input data (**Fig.** 1):



Choose a txt or csv file

Browse... single_conc.txt

Upload complete

Separator

● Tab

○ Comma

○ Semicolon

Time unit:

Days

Duration of accumulation phase:

4

Figure 1. *Data uploading and user information to enter.*

1.1 Format

You can upload your own data (click on "Browse") by taking care about the format specification of your file. MOSAIC$_{\text{bioacc}}$ expects to receive data as a .txt file or a .csv file (comma, semicolon or tabular separator). Each line of the table corresponds to a time point for a given replicate and a given exposure concentration of the contaminant. The table must contain the four following columns, with exact header names (**Table** @ref(tab:table2)):

- "time": the time point of the measurement at the exposure concentration;
- "expw," "exps," "expf," "exppw": the concentration of the contaminant in the exposure medium (expw: water, exps: sediment, expf: food, exppw: pore water);
- "replicate": a number or a string that is unique for each replicate;
- "conc": the concentration measurements of the contaminant within the organism.

According to your data, further columns can be added in the file:

- "concmℓ": the concentration measurements of metabolite ℓ from the parent compound within the organism ($e.g.$ concm1, concm2, …). Please note that only metabolites of phase I (deriving directly from the parent compound) in the metabolization process are considered;
- "growth": the growth measurements ($e.g.$ weight, size) of the organisms.

Then be careful to the units:

- The time points must be in hours, minutes, days or weeks;
- The exposure concentration in the medium must be in $\mu g.mL^{-1}$ or in $\mu g.g^{-1}$;

- The numbering of replicates is dimensionless;
- The concentration measurements (parent compound and metabolite(s)) within the organisms must be in $\mu g.g^{-1}$;
- The growth measurements must be in g, mg, cm, mm or other.

Table 1. *Example of a data set ready to be uploaded.*

time	conc	expw	replicate
0	0.000	0.0044	2
3	0.225	0.0044	2
7	0.355	0.0044	2
14	0.553	0.0044	2
21	0.658	0.0044	2
28	0.785	0.0044	2

Once the upload is complete, you have to manually select the corresponding separator, the appropriate time unit and the duration of the accumulation phase.

1.2 Example data

Instead of using your own data, you can try MOSAIC$_{\text{bioacc}}$ from several example files that are provided. Three data sets use a simple TK model (one exposure route and one elimination process), whereas two data sets consider a more complex TK model, accounting for metabolization and growth (**Fig.** 2).

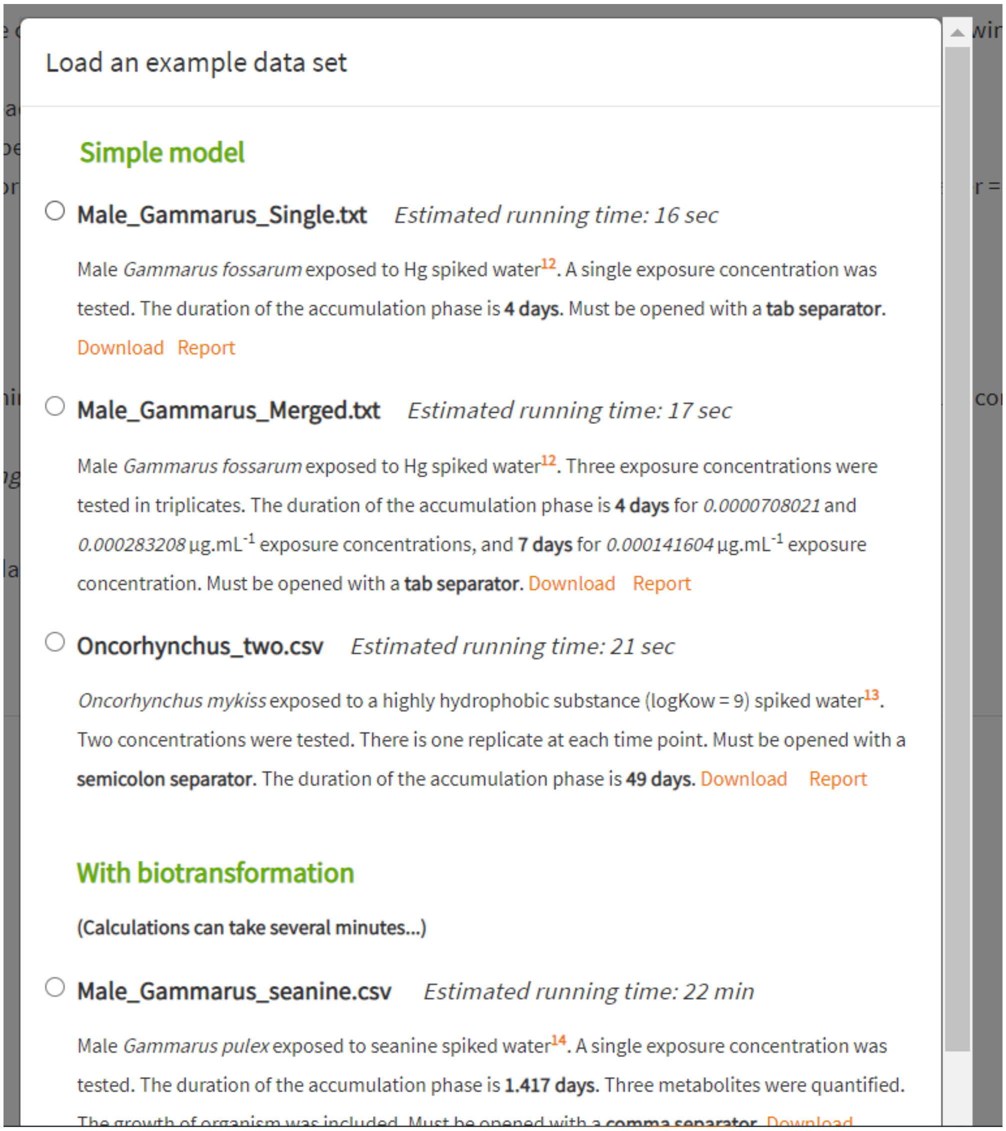
Please note that more data are, the more the TK model is complex, and the more the calculations take time. Thus, we indicated the approximative mean time calculation for each example data sets (**Fig.** 2).

1.3 Visualization of the data

In case you upload a data set with several exposure concentrations, select the one for which you want to see the results. We propose two types of visualization to check if the file has correctly been uploaded: as a table or a plot (**Fig.** 3 and 4).

According to your data, you can visualize the plot for the parent compound, the metabolite(s) and/or for growth (**Fig.** 5).

Don't forget to select the appropriate duration of the accumulation phase and the growth unit (if required) before to continue (**Fig.** 3 and 4, left side).



Load an example data set

Simple model

○ **Male_Gammarus_Single.txt** *Estimated running time: 16 sec*

Male *Gammarus fossarum* exposed to Hg spiked water[12]. A single exposure concentration was tested. The duration of the accumulation phase is **4 days**. Must be opened with a **tab separator**. Download Report

○ **Male_Gammarus_Merged.txt** *Estimated running time: 17 sec*

Male *Gammarus fossarum* exposed to Hg spiked water[12]. Three exposure concentrations were tested in triplicates. The duration of the accumulation phase is **4 days** for *0.0000708021* and *0.000283208* µg.mL^{-1} exposure concentrations, and **7 days** for *0.000141604* µg.mL^{-1} exposure concentration. Must be opened with a **tab separator**. Download Report

○ **Oncorhynchus_two.csv** *Estimated running time: 21 sec*

Oncorhynchus mykiss exposed to a highly hydrophobic substance (logKow = 9) spiked water[13]. Two concentrations were tested. There is one replicate at each time point. Must be opened with a **semicolon separator**. The duration of the accumulation phase is **49 days**. Download Report

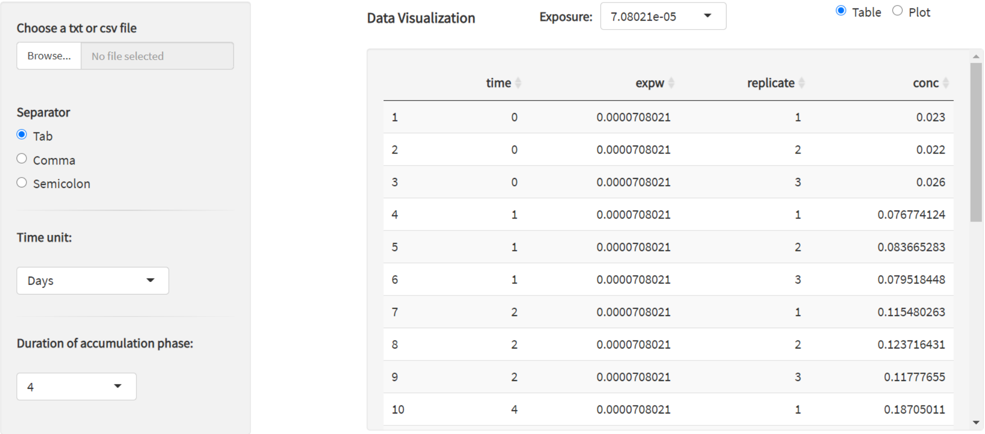
With biotransformation

(Calculations can take several minutes...)

○ **Male_Gammarus_seanine.csv** *Estimated running time: 22 min*

Male *Gammarus pulex* exposed to seanine spiked water[14]. A single exposure concentration was tested. The duration of the accumulation phase is **1.417 days**. Three metabolites were quantified. The growth of organism was included. Must be opened with a **comma separator**. Download

Figure 2. *Example files available in MOSAIC$_{bioacc}$.*



Choose a txt or csv file

Browse... No file selected

Separator

● Tab
○ Comma
○ Semicolon

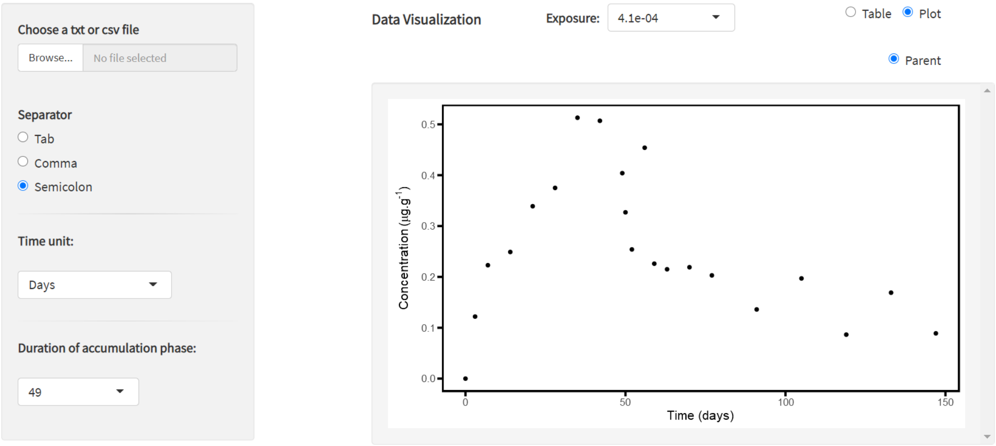
Time unit: Days

Duration of accumulation phase: 4

Data Visualization Exposure: 7.08021e-05 ● Table ○ Plot

	time	expw	replicate	conc
1	0	0.0000708021	1	0.023
2	0	0.0000708021	2	0.022
3	0	0.0000708021	3	0.026
4	1	0.0000708021	1	0.076774124
5	1	0.0000708021	2	0.083665283
6	1	0.0000708021	3	0.079518448
7	2	0.0000708021	1	0.115480263
8	2	0.0000708021	2	0.123716431
9	2	0.0000708021	3	0.11777655
10	4	0.0000708021	1	0.18705011

Figure 3. *Table of the uploaded data at the selected exposure concentration.*



Choose a txt or csv file

Browse... No file selected

Separator

○ Tab

○ Comma

● Semicolon

Time unit:

Days

Duration of accumulation phase:

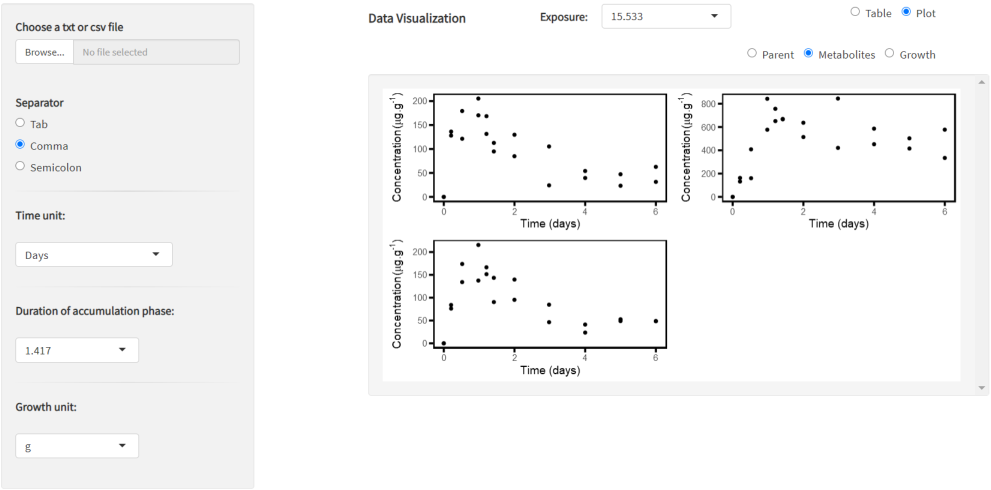
49

Data Visualization Exposure: 4.1e-04

○ Table ● Plot

● Parent

Figure 4. *Plot of the uploaded data at the selected exposure concentration.*



Choose a txt or csv file

Browse... No file selected

Separator

○ Tab

● Comma

○ Semicolon

Time unit:

Days

Duration of accumulation phase:

1.417

Growth unit:

g

Data Visualization Exposure: 15.533

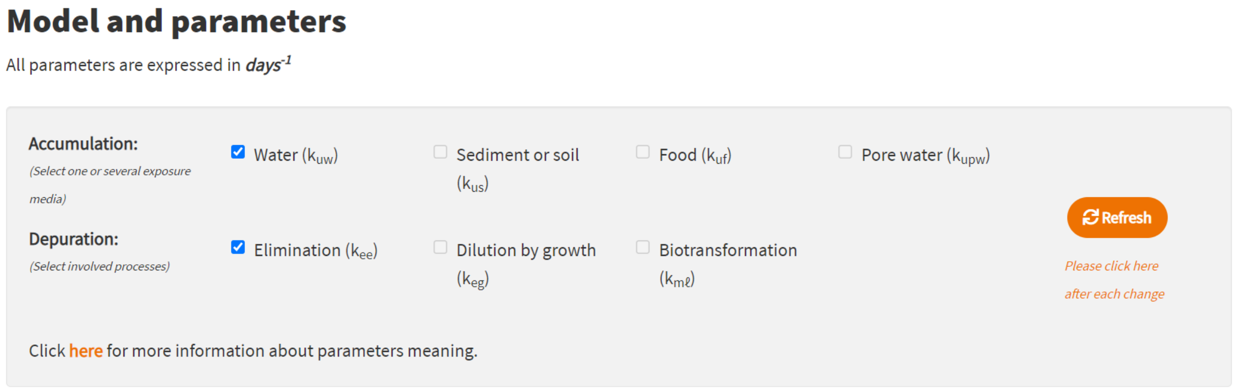
○ Table ● Plot

○ Parent ● Metabolites ○ Growth

Figure 5. *Plot of the uploaded data (metabolites) at the selected exposure concentration.*

2 Step 2: Model and parameters

The most complete model and its corresponding parameters are automatically selected according to the experimental design as given within the uploaded data set (**Fig.** 6). Users can also deselect some of the parameters (based on biological hypotheses related to the most probable exposure route or by neglecting one elimination process, for example). These choices lead to the automatic building of a nested TK sub-model to fit again on the data in clicking on the 'Refresh' button. The equations of the most complete model are provided to the users on-line (**Fig.** 7).



Model and parameters

All parameters are expressed in ***days***$^{-1}$

Accumulation: *(Select one or several exposure media)*

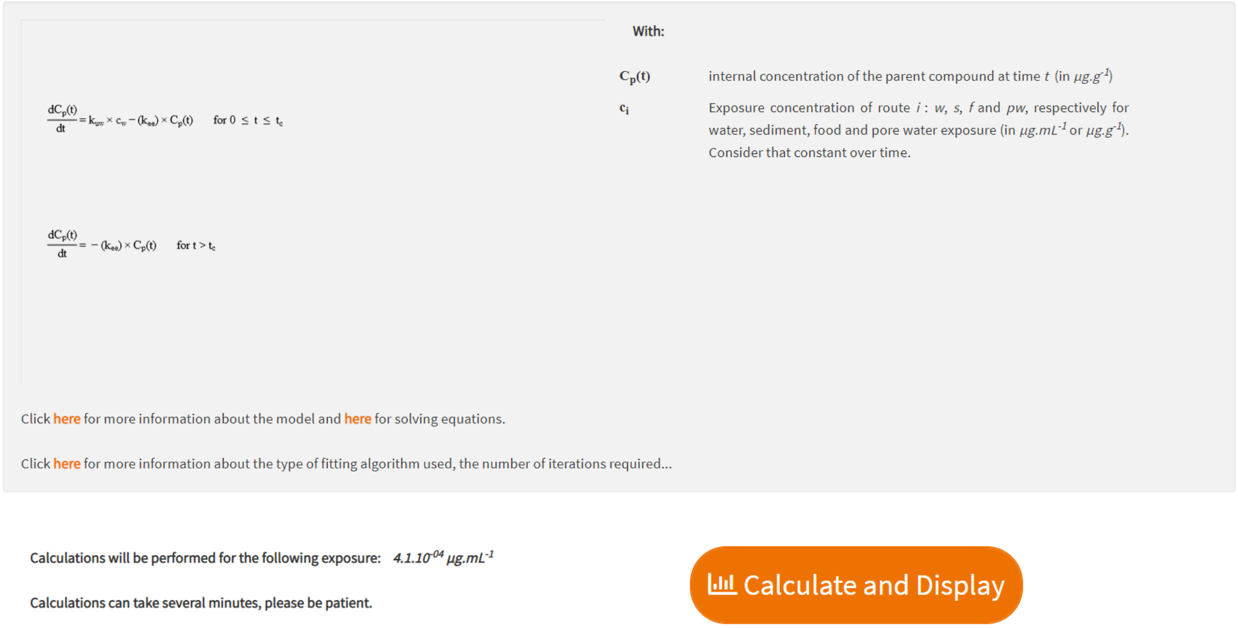
☑ Water (k_{uw}) ☐ Sediment or soil (k_{us}) ☐ Food (k_{uf}) ☐ Pore water (k_{upw})

Depuration: *(Select involved processes)*

☑ Elimination (k_{ee}) ☐ Dilution by growth (k_{eg}) ☐ Biotransformation ($k_{m\ell}$)

Click here for more information about parameters meaning.

Figure 6. *Example of a model choice.*



$$\frac{dC_p(t)}{dt} = k_{uw} \times c_w - (k_{ee}) \times C_p(t) \quad \text{for } 0 \leq t \leq t_c$$

$$\frac{dC_p(t)}{dt} = -(k_{ee}) \times C_p(t) \quad \text{for } t > t_c$$

With:

$C_p(t)$ internal concentration of the parent compound at time *t* (in *µg.g*$^{-1}$)

c_i Exposure concentration of route *i* : *w, s, f* and *pw*, respectively for water, sediment, food and pore water exposure (in *µg.mL*$^{-1}$ or *µg.g*$^{-1}$). Consider that constant over time.

Click here for more information about the model and here for solving equations.

Click here for more information about the type of fitting algorithm used, the number of iterations required...

Calculations will be performed for the following exposure: *4.1.10*$^{-04}$ *µg.mL*$^{-1}$

Calculations can take several minutes, please be patient.

Figure 7. *Example of equations of the model corresponding to Fig. 6.*

For the accumulation phase:

- k_{u_w}, water exposure;
- k_{u_s}, sediment exposure;

- k_{u_f}, food exposure;

- $k_{u_{pw}}$, pore water exposure.

For the depuration phase:

- k_{e_e}, excretion process;

- k_{e_g}, if the weight of the organism varying;

- k_{m_ℓ}, if there are metabolites, where ℓ is the metabolite number.

Also, σ are the expected standard deviations of the measured contaminant concentration or growth of the organisms:

- σ_p the standard deviation of the measured parent compound concentration in the organisms;
- $\sigma_{met\ell}$ the standard deviation of the measured metabolite ℓ concentration in the organisms;
- σ_G the standard deviation of the measured growth of the organisms.

2.1 Theoretical model

A mathematical model is composed of two parts: the deterministic and the stochastic parts. In the case of TK models fitted to concentration measurements, it can be written as follows (Eq. (1)):

$$y = f(x, \theta) + \varepsilon \quad (1)$$

with f the function describing the mean relationship between x and y, y the observed variable, x the controlled variable, θ the parameter vector to estimate and ε the random variable describing the variability of the data around the mean tendency.

The deterministic part

The organisms are here considered as single compartments for which a generic first-order kinetic bioaccumulation model can be expressed as follows[12] (Eqs. (2) and (3) for the accumulation phase and Eqs. (4) and (5) for the elimination phase):

$$\left\{\begin{array}{l} \frac{dC_p(t)}{dt} = U - (E+M)C_p(t) \quad (2) \\ \frac{dC_{m_\ell}(t)}{dt} = k_{m_\ell}C_p(t) - k_{e_\ell}C_{m_\ell}(t) \quad \forall \ell = 1 \dots M \quad (3) \end{array}\right. \qquad \text{for } 0 \leqslant t \leqslant t_c$$

$$\left\{\begin{array}{l} \frac{dC_p(t)}{dt} = -(E+M)C_p(t) \quad (4) \\ \frac{dC_{m_\ell}(t)}{dt} = k_{m_\ell}C_p(t) - k_{e_\ell}C_{m_\ell}(t) \quad \forall \ell = 1 \dots M \quad (5) \end{array}\right. \qquad \text{for } t > t_c$$

with:

Symbol	Meaning
I	total number of exposure sources
J	total number of elimination processes
L	total number of metabolites
i	index of exposure sources, $i = 1 \dots I$
j	index of elimination processes, $j = 1 \dots J$

Symbol	Meaning
ℓ	index of metabolites, $\ell = 1 \dots L$
t	time (expressed in time units)
c_i	exposure concentration of route i (in $\mu g.mL^{-1}$)
$C_p(t)$	internal concentration of the parent compound at time t (in $\mu g.g^{-1}$)
$C_{m_\ell}(t)$	internal concentration of metabolite ℓ (in $\mu g.g^{-1}$)
k_{u_i}	uptake rate of exposure source i (expressed per time units)
k_{e_j}	elimination rates of elimination process j (expressed per time units)
$k_{e_{m\ell}}$	elimination rates of metabolite ℓ (expressed per time units)
k_{m_ℓ}	metabolization rate of metabolite ℓ (expressed per time units)
t_c	duration of the accumulation phase
$U = \sum_{i=1}^{I} k_{u_i} c_i$	sum of all uptake terms
$E = \sum_{j=1}^{J} k_{e_j}$	sum of all elimination terms for the parent compound
$M = \sum_{\ell=1}^{L} k_{m_\ell}$	sum of all elimination terms for metabolite ℓ

The simplest model in MOSAIC$_{\text{bioacc}}$ application considers only one exposure route (for example by water, parameter k_{u_w}) with the corresponding elimination rate (excretion, parameter k_{e_e}), as given by Eq. (6):

$$\begin{cases} \frac{dC_p(t)}{dt} = k_{u_w} c_w - k_{e_e} C_p(t) & \text{for } 0 \leqslant t \leqslant t_c \\ \frac{dC_p(t)}{dt} = -k_{e_e} C_p(t) & \text{for } t > t_c \end{cases} \qquad (6)$$

where k_{u_w} is the uptake rate from water ($time^{-1}$), c_w is the mean contaminant concentration in water ($\mu g.mL^{-1}$) and k_{e_e} is the rate related to the excretion process ($time^{-1}$).

The more complex model in MOSAIC$_{\text{bioacc}}$ application considers a maximum of four exposure routes (water: k_{u_w}, pore water: $k_{u_{pw}}$, sediment: k_{u_s} and food: k_{u_f}) and a maximum of three elimination processes (excretion, k_{e_e}, biotransformation: $k_{m\ell}$, growth dilution: k_{e_g}), as given by Eq. (7) for parent compound, Eq. (8) for metabolite ℓ and Eq. (9) for growth (Von Bertalanffy equations, classically used for fishes):

$$\begin{cases} \frac{dC_p(t)}{dt} = k_{u_w} c_w + k_{u_{pw}} c_{pw} + k_{u_s} c_s + k_{u_f} c_f - (k_{e_e} + k_{e_g} + k_{m_\ell}) C_p(t) & \text{for } 0 \leqslant t \leqslant t_c \\ \frac{dC_p(t)}{dt} = -(k_{e_e} + k_{e_g} + k_{m_\ell}) C_p(t) & \text{for } t > t_c \end{cases} \qquad (7)$$

$$\frac{dC_{m_\ell}(t)}{dt} = k_{m_\ell} C_p(t) - k_{e_{m\ell}} C_{m_\ell}(t) \qquad (8)$$

$$\frac{dG(t)}{dt} = k_{e_g}(g_{max} - G(t)) \qquad (9)$$

with $G(t)$ the measured growth of the organism at time t (in growth unit) and g_{max} the asymptotic growth ℓ (in $\mu g.g^{-1}$).
More details and especially the exact solutions of these equations are given **here**.

The stochastic part
We assumed a Gaussian (normal) probability distribution of the concentration measurements within the organism as follows (Eqs. (10) to (12)):

$$C_{obs_p}(t) \sim \mathcal{N}(C_p(t), \sigma_p^2) \quad (10)$$

$$C_{obs_{m_\ell}}(t) \sim \mathcal{N}(C_{m_\ell}(t), \sigma_{m_\ell}^2) \quad (11)$$

$$G_{obs}(t) \sim \mathcal{N}(G(t), \sigma_g^2) \quad (12)$$

where $\mathcal{N}$ stands for the normal law, $C_{obs_p}(t)$ and $C_{obs_{m\ell}}(t)$ correspond to the measured parent and metabolite ℓ concentrations within the organism measured at time t, $C_p(t)$ is the parent compound concentration at time t predicted by the model based on to Eqs. (2) and (4), $C_{m_\ell}(t)$ is the concentration of metabolite ℓ at time t predicted by the model based on Eqs. (3) and (5), $G_{obs}(t)$ is the measured growth of the organism at time t, $G(t)$ is the growth at time t predicted by the model based on Eq. (9), σ are the expected standard deviations of the measured contaminant concentration or growth of the organisms ($\mu g.g^{-1}$ or growth unit).

2.2 Directed Acyclic Graph

A Directed Acyclic Graph (DAG) is given on **Fig.** 8, which symbolize the deterministic links between parameters and variables for the generic TK model (Eqs.(2) to (5)) and the stochastic links between the observed and predicted data (Eqs. (10) to (12)).

2.3 Choice of prior distributions

In MOSAIC$_{\text{bioacc}}$, prior choice is hidden to the user. However, we give here some information to help the user to understand the model behind. Before conducting an experimental study, prior distributions are defined for each parameter according to information available from the literature, expert knowledge and/or previous experiments. Depending on the sources where the information come from, informative, semi-informative or non-informative prior distributions can be used. If a parameter was already estimated in previous studies or if previous data are available, a prior distribution can easily be characterized with an appropriate probability distribution. However, if no information is available but an order of magnitude is (positive only, for example), it is possible to use a weakly informative prior. If any information is available on the order of magnitude of a parameter, its prior is preferably defined on a decimal logarithm scale in order to consider with equal probability both low and high expected values.

As MOSAIC$_{\text{bioacc}}$ application has to be the most generic as possible, priors were assumed to be non-informative log10-uniform distributions within [-5, 5] for all uptake and elimination rate constants. For growth, a uniform prior distribution $[10 \times mean(G)/4,\ 10 \times mean(G)]$ was assumed for parameter g_{max} and a uniform prior distribution $[0,\ (10 \times mean(G))/(8 - 10^{-10})]$ for parameter g_0. We assumed a non-informative $(0,A)$ uniform prior for all σ with a large A, here defined as five times the maximum internal measured concentration, which is then removed from the data set, as usually proceeded[13].

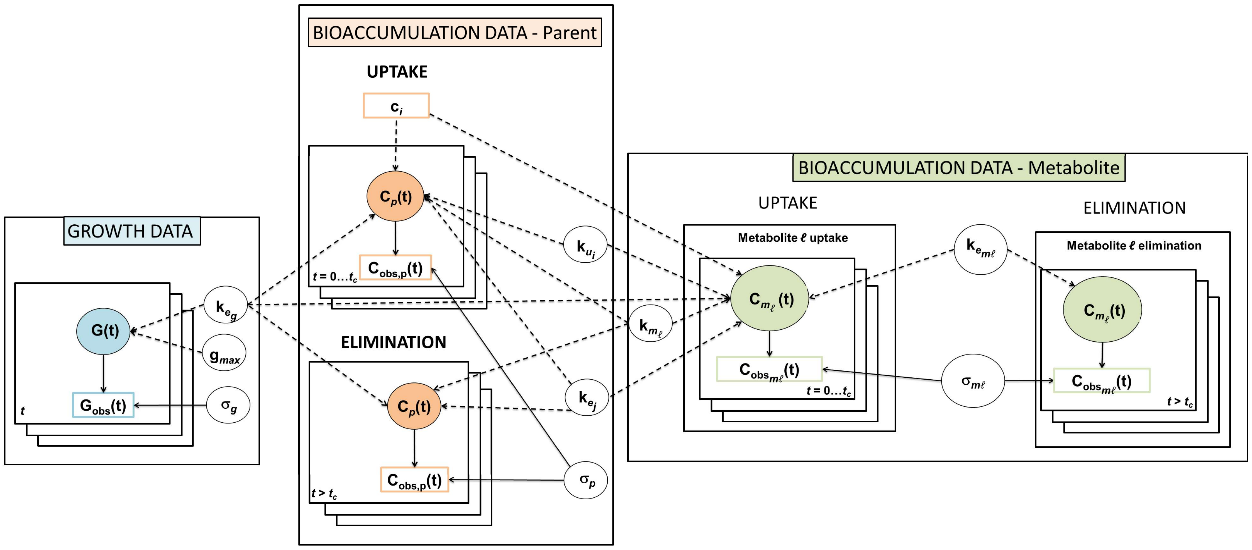


Figure 8. *Directed acyclic graph of the generic TK model. Observed variables, such as the contaminant concentration in organisms and medium i and growth data, are represented by rectangle nodes. Model parameters and variables are represented by circular nodes. Dotted arrows represent deterministic links (Eqs. (2) to (5)), while solid arrows represent stochastic links between predicted and observed data (Eqs. (10) to (12)).*

2.4 Bayesian inference

The Bayesian approach considers that data are fixed but that the parameters are unknown random variables following a probabilistic distribution. This leads to the following practical implications: *(i)* the Bayesian process optimises the probability of parameter vector θ given the data set Y used for calibration (the so-called posterior distribution) not only the likelihood (see below); *(ii)* there is a need to provide reasonable prior information, then updating this information by accounting for the data. Below is a short introduction to Bayesian principles[14].

In short, the Bayesian approach requires the following steps:
- Choose the prior distributions on parameters based on previous results, literature or expert knowledge (without looking at the data to fit): $\mathrm{P}(\theta)$;
- Define the probabilistic model from the data, that is the random variable whose data would be one realisation assuming known values of parameters, namely the likelihood: $\mathrm{P}(\mathrm{Y} \mid \theta)$;
- Calculate the joint posterior distribution of the parameters given the data via the Bayes formula: $\mathrm{P}(\theta \mid \mathrm{Y})$;
- Provide statistical summaries of parameter estimates (namely, appropriate quantiles);
- Get any function of the parameter estimates as posterior probability distribution, like for example BCF calculations or predictions of new observations.

Basic principles
The keystone of the Bayesian approach is the Bayes formula (Eq. (13):

$$\mathrm{P}(\theta \mid \mathrm{Y}) = \frac{\mathrm{P}(\theta)\mathrm{P}(\mathrm{Y} \mid \theta)}{\mathrm{P}(\mathrm{Y})} \quad (13)$$

where Y are the observed data; $\mathrm{P}(\theta \mid \mathrm{Y})$ is the joint posterior distribution of parameter vector ; $\mathrm{P}(\mathrm{Y} \mid \theta)$ is the likelihood of the data given the parameters; $\mathrm{P}(\theta)$ is the joint prior distribution of parameter vector

θ. Given that $\mathrm{P(Y)}$ is known and fixed, it is often not considered as it does not depend on and will not influence the posterior distribution. Hence (Eq. (14)):

$$\mathrm{P}(\theta \mid \mathrm{Y}) \simeq \mathrm{P}(\theta)\mathrm{P}(\mathrm{Y} \mid \theta) \quad (14)$$

with $\mathrm{P}(\theta)\mathrm{P}(Y|\theta)$ the unormalised posterior density leading to (Eq. (15)):

$$\mathrm{P(Y)} = \int \mathrm{P}(\theta)\mathrm{P}(\mathrm{Y} \mid \theta)d\theta \quad (15)$$

The prior distribution $\mathrm{P}(\theta)$ expresses the available parameter information without knowing the observed data, while the posterior distribution $\mathrm{P}(\theta \mid \mathrm{Y})$ combines this prior information (which may be more or less informative depending on what is known about the value of the parameters beforehand) with evidence from the data (expressed through the likelihood) into a joint posterior density probability distribution for the parameters. The overall expectation is to get a narrower posterior distribution compared to the prior one after the computing of the Bayesian algorithm: the difference between the two distributions reflects the information provided by the data. When the non-informative prior is vague (translated for example into a flat uniform distribution), and the data sufficiently informative, the results are similar than those obtained under a frequentist approach.

Joint posterior distribution

The joint posterior distribution has the dimension of the number of parameters times the number of iterations within the Monte Carlo Markov Chain (MCMC) chains. It can be plotted in planes of parameter pairs to visualise correlations between parameters. In an example case with two binormally distributed parameters, the joint posterior distribution can be plotted in the 2D-parameter space as illustrated by ellipses on **Fig.** 9; in this example, parameters θ_1 and θ_2 appear slightly positively correlated. From the joint posterior distribution, the marginal posterior distributions for each parameter (as illustrated by grey normal distributions on bottom and left sides of **Fig.** 9) can be extracted. Then, from the marginal posterior distributions, some statistical summaries on parameter estimates can be extracted, usually the median (illustrated by vertical and horizontal plain grey lines on **Fig.** 9) as well as 2.5% and 97.5% quantiles to serve as 95% credible intervals (illustrated by vertical and horizontal dotted grey lines on **Fig.** 9). Another advantage of having the joint posterior distribution is that the posterior distribution of any function of the parameters can also be obtained.

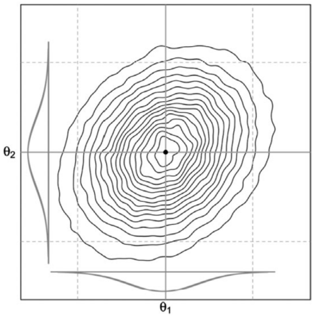


Figure 9. *Theoretical binormal joint posterior distribution of parameter vector* (θ_1, θ_2)*. Ellipses correspond to isoclines of the joint posterior distribution; grey distributions are marginal posterior distributions of both parameters; solid horizontal and vertical lines correspond to the medians of these marginal distributions; dashed horizontal and vertical lines correspond to the 2.5% and 97.5% quantiles of the marginal distributions.*

Parameter uncertainties

One implication of adopting a Bayesian approach is that the uncertainty on a parameter can easily be expressed as a probability distribution from which a credible interval (also called a Bayesian confidence interval) can be extracted. For example, the 95% credible interval delimits a range of values where the parameters lies with a 95% probability.

Numerical computation

Many numerical methods have been developed to approximately compute the joint posterior distribution, mainly based on simulations by Monte Carlo Markov Chain (MCMC) sampling methods used to generate random numbers from complex joint distributions. MCMC algorithms are a general method based on drawing values of parameter vector θ from approximate distributions and then correcting those draws to better approximate the target posterior distribution $\mathrm{P}(\theta \mid \mathrm{Y})$. The sampling is done sequentially, with the distribution of the sampled draws depending only on the last value drawn; hence, the draws form a Markov Chain. The key to the method success, however, is not the Markov property but rather the fact that the approximate distributions are improved at each step of the simulation, in the sense that it finally converges to the target posterior distribution after an enough number of iterations. Indeed, with MCMC algorithms, the simulation process must run long enough so that the distribution of the current draws is close enough to the desired target posterior distribution.

MCMC algorithms use random walk algorithms, among which the Metropolis algorithm (and its generalisation, the Metropolis–Hasting algorithm) as an adaptation of a random walk with an acceptance/rejection rule to converge to the specified target distribution[15,16]. The Gibbs sampler is a special case of the Metropolis–Hastings algorithm applicable when the joint distribution is not known explicitly, or when it is difficult to directly sample from, while the conditional distribution of each parameter is known and it is easy (or at least, easier) to sample from[17].

Several tools are available to automatically perform these computations. In MOSAIC$_{\text{bioacc}}$, JAGS[10] (version 4.3.0. (2017-08-10)) and R software[1] (version 4.0.2 (2020-06-22)) are used. The models are fitted to bioaccumulation data using Bayesian inference via Monte Carlo Markov Chain (MCMC) sampling based on a Gibbs-type algorithm. For each model, we start by running a short sampling phase with three chains (5,000 iterations after a burn-in phase of 10,000 iterations) then using the Raftery and Lewis[18] method to set the necessary thinning and number of iterations to reach an accurate estimation of the joint posterior distribution.

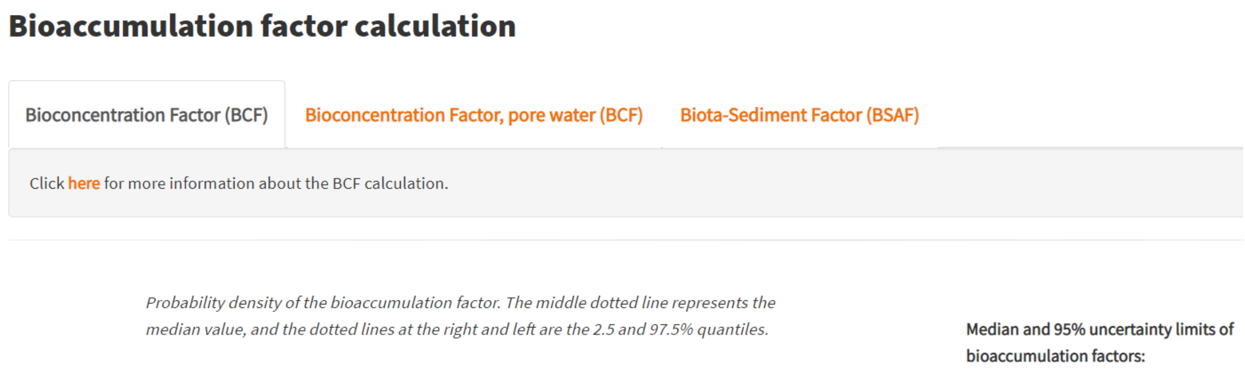
3 Step 3: Results

Once data uploaded and the model stated, you can click on button "**Calculate and Display**." We recommend you to pay attention to the selected exposure concentration before to proceed to calculations. Besides, calculations can take several minutes for the most complicated models.

You will find at the end of this user guide an appendix (**section** 5) which gathers together other types of results which can be obtained with other data sets and how to interpret them.
To illustrate the result section, here we used the example file 'Pimephales_two.csv,' where *Pimephales promelas* are exposed to a highly hydrophobic substance (logKow = 9) spiked water at $0.0044\ \mu g.mL^{-1}$ and where the duration of the accumulation phase is equal to 49 days.

3.1 Bioaccumulation metrics

As a first result, we provide the bioaccumulation metrics (BCF, BCF$_{\text{pw}}$, BSAF and BMF). Please, note that this section is reactive according to your data. If required, change tab to get the other bioaccumulation metrics (**Fig.** 10).



Bioaccumulation factor calculation

Bioconcentration Factor (BCF)
Bioconcentration Factor, pore water (BCF)
Biota-Sediment Factor (BSAF)

Click here for more information about the BCF calculation.

Probability density of the bioaccumulation factor. The middle dotted line represents the median value, and the dotted lines at the right and left are the 2.5 and 97.5% quantiles.

Median and 95% uncertainty limits of bioaccumulation factors:

Figure 10. *Example of reactive tabs for bioaccumulation metric section.*

3.1.1 BCF

As a first result, we provide the kinetic bioconcentration factor (BCF_k). The BCF at steady-state (BCF_{ss}) can be asked if you consider a steady-state is almost reached at the end of the accumulation phase. These factors are mathematically given by Equations (16) and (17) respectively:

$$BCF_k = \frac{k_{u_w}}{\sum_{j=1}^{J} k_{e_j} + \sum_{\ell=1}^{L} k_{m_\ell}} \quad (16)$$

$$BCF_{ss} = \frac{C_p(t_c)}{c_w} \quad (17)$$

where $C_p(t_c)$ is the contaminant concentration within the organism at steady-state that is at the end of the accumulation phase ($\mu g.g^{-1}$) and c_w is the contaminant concentration in water ($\mu g.mL^{-1}$).

BCF are given as probability distributions (**Fig.** 11) and summarized with their median and their 95% uncertainty limits (95% credible intervals, **Table** 3).

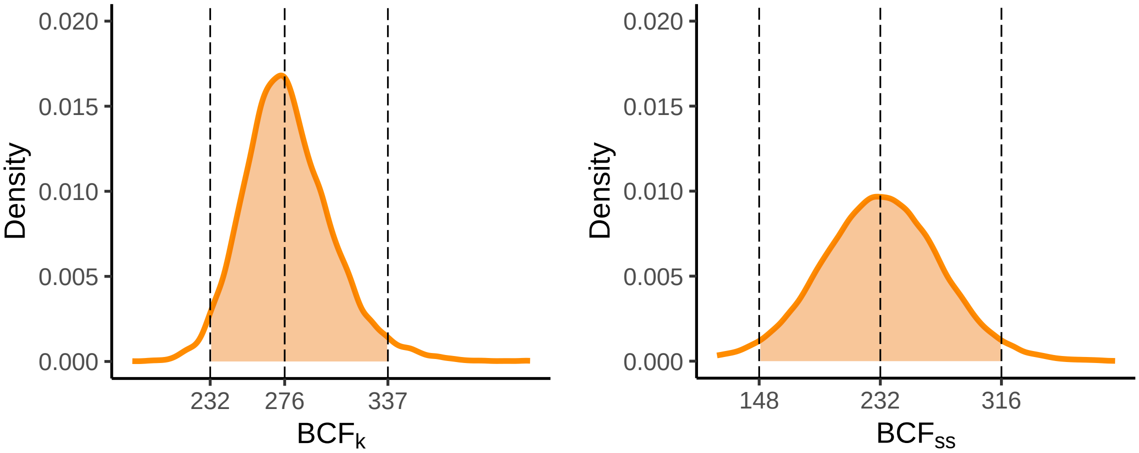


Figure 11. *Probability density of BCF_k and BCF_{ss}. The middle dotted line represents the median value, and the dotted lines on left and right sides are the 2.5 and 97.5% quantiles. These results were obtained with the example file Pimephales_two.csv.*

Table 3. *Median and 95% uncertainty limits of BCF_k and BCF_{ss}. These results were obtained with the example file Pimephales_two.csv.*

	2.5%	50%	97.5%
BCF_k	232	276	337
BCF_{ss}	148	232	316

In the above example, the steady-state is almost reached at time t_c (**Fig.** 12), thus it is reasonable to ask for the BCF_{ss}.
Note that the OECD guideline 305[6] reports that "greater emphasis must be on kinetic BCF estimate (when possible) next to estimating the BCF at steady-state." Thus we recommend to preferentially consider the BCF_k rather than the BCF_{ss}.

3.1.2 BCF$_{\text{pw}}$

When appropriate, we provide the kinetic pore water bioconcentration factor (BCF_{pw_k}).Again, the corresponding BCF at steady-state ($BCF_{pw_{ss}}$) can be asked. These factors are mathematically given by Equations (18) and (19) respectively:

$$BCF_{pw_k} = \frac{k_{u_{pw}}}{\sum_{j=1}^{J} k_{e_j} + \sum_{\ell=1}^{L} k_{m_\ell}} \quad (18)$$

$$BCF_{pw_{ss}} = \frac{C_p(t_c)}{c_{pw}} \quad (19)$$

where $C_p(t_c)$ is the contaminant concentration within the organism at time t_c ($\mu g.g^{-1}$) and c_{pw} is the contaminant concentration in pore water ($\mu g.mL^{-1}$). BCF$_{\text{pw}}$ are given as probability distributions and summarized with their median and their 95% uncertainty limits (95% credible intervals).

3.1.3 BSAF

When appropriate, we provide the kinetic biota-sediment factor ($BSAF_k$) and the BSAF at steady-state ($BSAF_{ss}$) can be asked. These factors are mathematically given by Equations (20) and (21) respectively:

$$BSAF_k = \frac{k_{u_s}}{\sum_{j=1}^{J} k_{e_j} + \sum_{\ell=1}^{L} k_{m_\ell}} \quad (20)$$

$$BSAF_{ss} = \frac{C_p(t_c)}{c_s} \quad (21)$$

where $C_p(t_c)$ is the contaminant concentration within the organism at time t_c ($\mu g.g^{-1}$) and c_s is the contaminant concentration in sediment ($\mu g.g^{-1}$). BSAF are given as probability distributions and summarized with their median and their 95% uncertainty limits (95% credible intervals).

3.1.4 BMF

When appropriate, we provide the kinetic biomagnification factor (BMF_k) and the BMF at steady-state ($BSAF_{ss}$) can be asked. These factors are mathematically given by Equations (22) and (23) respectively:

$$BMF_k = \frac{k_{u_f}}{\sum_{j=1}^{J} k_{e_j} + \sum_{\ell=1}^{L} k_{m_\ell}} \quad (22)$$

$$BMF_{ss} = \frac{C_p(t_c)}{c_f} \quad (23)$$

where $C_p(t_c)$ is the contaminant concentration within the organism at time t_c ($\mu g.g^{-1}$) and c_f is the contaminant concentration in food ($\mu g.g^{-1}$). BMF are given as probability distributions and summarized with their median and their 95% uncertainty limits (95% credible intervals).

3.1.5 Coefficient of variation

A coefficient of variation (CV) for each bioaccumulation metric can be calculated as follows (Eq. (24)):

$$CV = \frac{Q_k 97.5 - Q_k 2.5}{4 \times Q_k 50} \quad (24)$$

with $Q_k 2.5$, $Q_k 50$ and $Q_k 97.5$ the 2.5%, 50% and 97.5% quantiles of the kinetic bioaccumulation metric.
Based on our experiment, the coefficient of variation is expected to not exceed 0.5.
If the bioaccumulation metric at steady-state is asked, the corresponding CV is given (Eq. (25)):

$$CV = \frac{Q_{ss} 97.5 - Q_{ss} 2.5}{4 \times Q_{ss} 50} \quad (25)$$

with $Q_{ss} 2.5$, $Q_{ss} 50$ and $Q_{ss} 97.5$ the 2.5%, 50% and 97.5% quantiles of the steady-state bioaccumulation metric.

3.2 Predictions

We first provide the fitted curve superimposed to the observations (**Fig.** 12, black dots): the orange plain line is the median curve, the gray zone is the uncertainty band delimited by 2.5% and 97.5% quantiles of predictions in orange dotted lines. This section is reactive according to your data: if there is biotransformation or growth, the fitted curve superimposed to the observations for parent compound, metabolite(s) and/or growth data are given (for example from the example file "Male_Gammarus_seanine.csv," **Fig.** 13).

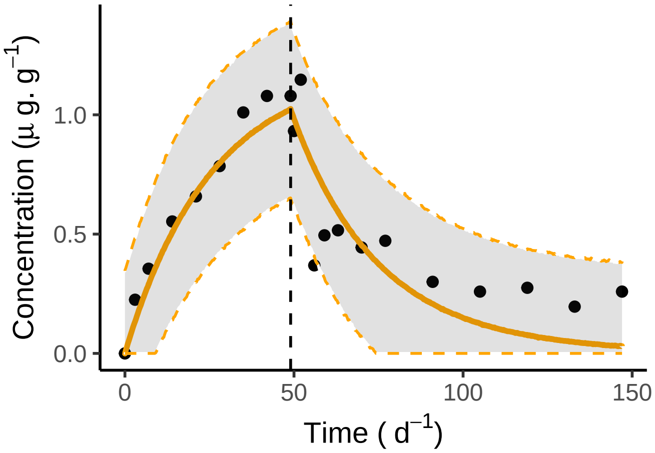


Figure 12. *Measured (black dots) and predicted contaminant concentrations in the organism (µg.g^{-1}). Median predictions are symbolized by the orange plain line and the uncertainty bands by the gray zone which is delimited by the 2.5% and 97.5% quantiles in orange dotted lines. The black dotted vertical line indicates the separation between the accumulation phase and the depuration phase. These results were obtained with the example file Pimephales_two.csv.*

From the joint posterior distribution, we can obtain the marginal posterior distributions for each parameter, which can be summarized by their medians and their 95% credible intervals (**Table** 4).

Table 4. *Example of parameter medians (50% quantile) and 95% credible intervals (2.5% - 97.5% quantiles). These results were obtained with the example file Pimephales_two.csv.*

	2.5%	50%	97.5%	
k_{u_w}	7.437	10.61	15.55	d^{-1}
k_{e_e}	0.0233	0.03873	0.06168	d^{-1}
σ_p	0.1246	0.1679	0.244	$\mu g.g^{-1}$

3.3 Goodness-of-fit criteria

Goodness-of-fit criteria are given below in our prioritised order; the Posterior Predictive Check (PPC) and the prior-posterior comparison are the most important to check; if they do not correspond to the expectation, you must consider your results with an even more particular attention. As an indication, if at least two criteria are fulfilled, the results obtained can be considered as good enough. We suggest that you refer to the appendix at the end of the document for more information about cases not in accordance with the classical expectations (**section** 5).

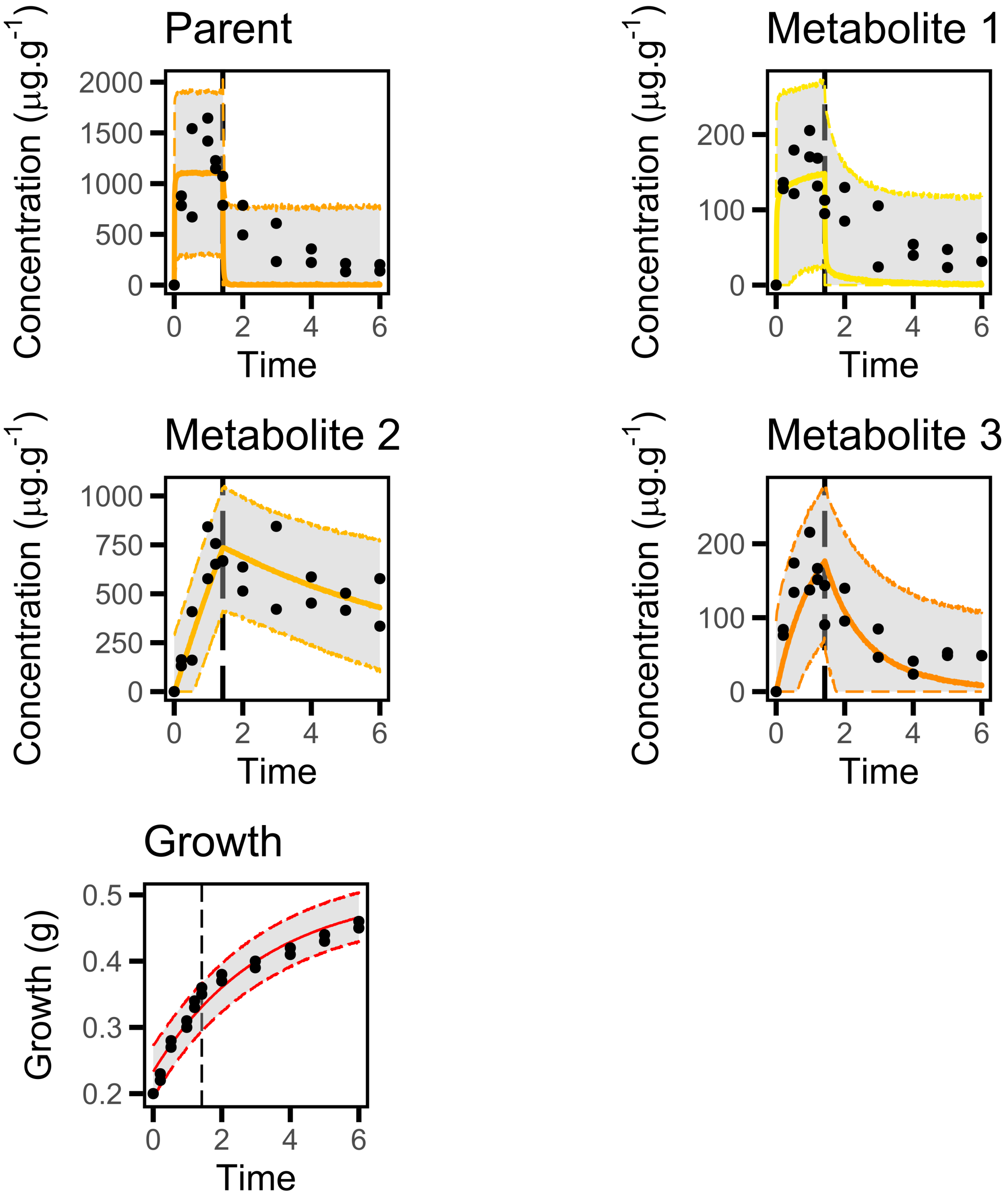


Figure 13. *Measured (black dots) and predicted contaminant concentrations in the organism (µg.g^{-1}) or growth (in g). Median predictions are symbolized by the orange plain line and the uncertainty bands by the gray zone which is delimited by the 2.5% and 97.5% quantiles in orange dotted lines. The black dotted vertical line indicates the separation between the accumulation phase and the depuration phase. These results were obtained with the example file Male_Gammarus_seanine.csv.*

3.3.1 Posterior Predictive Check (PPC)

The PPC shows the observed values against their corresponding estimated predictions (black dots), along with their 95% credible interval (vertical segments). If the fit is correct, we expect to see 95% of the data within the intervals. Ideally, observations and predictions should coincide, so we would expect to see black dots along the first bisector $y = x$ (plain black line). The 95% credible intervals are colored in green if they overlap this line, in red otherwise. In the following example (**Fig.** 14), 95.24% of the measured concentrations ($n = 20/21$) are in the 95% credible intervals of their predictions.

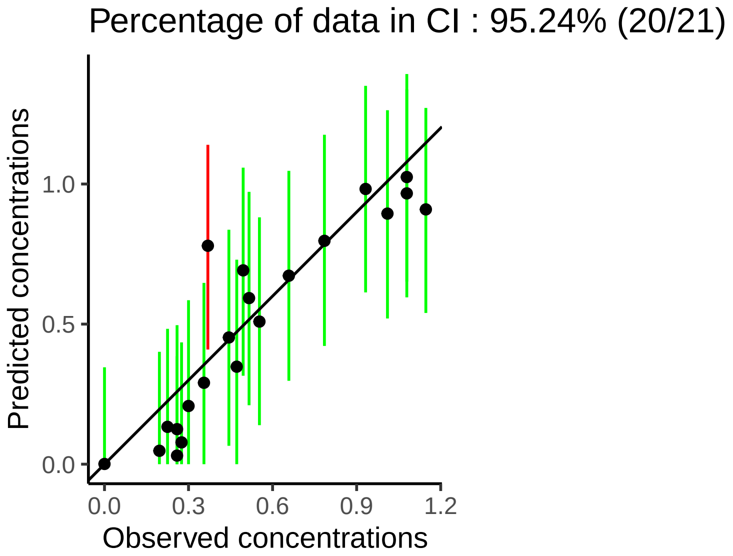


Figure 14. *Example of a PPC plot: predicted against observed concentrations (black dots) and predicted 95% credible intervals (vertical green and red segments). These results were obtained with the example file Pimephales_two.csv.*

3.3.2 Prior and posterior distributions

An example of prior and posterior distributions is illustrated in **Fig.** 15. The prior distribution is represented by the gray area and the posterior distribution by the orange area. The precision of the model parameter estimation can be visualized by comparing prior and posterior distributions: the overall expectation is to get a narrower posterior distribution compared to the prior one, what reflects that data contributed enough to precisely estimate parameters. In the example of **Fig.** 15, marginal posterior distributions for k_{u_w} and k_{e_e} are narrower (orange area) than their respective prior distributions (grey area). Within the application, you can chose to visualize either the deterministic or the stochastic parameters by selecting the corresponding tab.

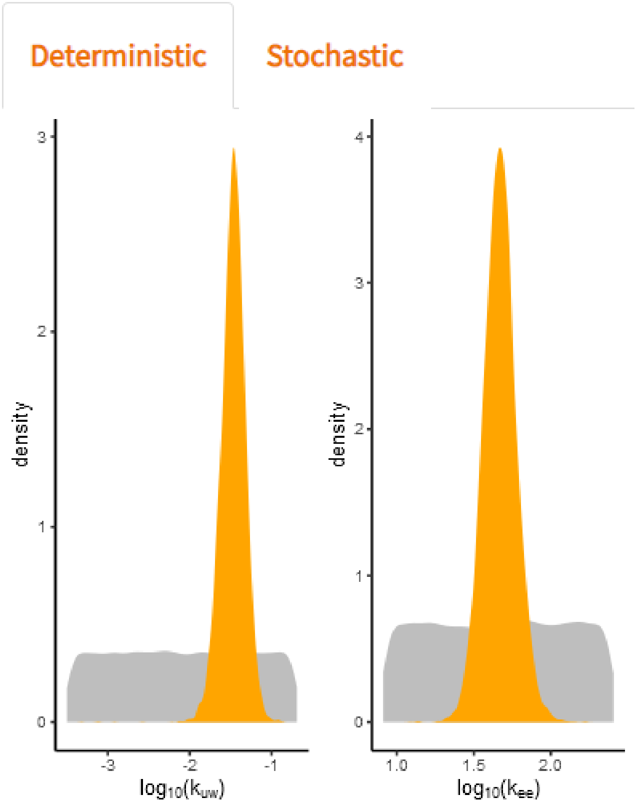


Figure 15. *Example of prior (gray) and posterior (orange) probability distributions for a set of parameter. These results were obtained with the example file Pimephales_two.csv.*

3.3.3 Correlations between parameters

MOSAIC$_{\text{bioacc}}$ gives a colored matrix in order to highlight correlations between parameters (**Fig.** 16, 17). This output allows you to see at a glance the most correlated or anti-correlated parameters, in order to diagnose potential problems of precision due to highly correlated parameters.

Correlations between parameters can also be visualized by projecting the joint posterior distribution in a plot matrix with planes of parameter pairs (**Fig.** 18, lower triangular elements), marginal posterior distribution of each model parameter (**Fig.** 18, diagonal) and Pearson correlation coefficients (**Fig.** 18, upper triangular elements). Correlations are expected to be low (reflected by "potatoid" shapes of density lines in orange, *e.g.*, k_{e_e} and σ_p in **Fig.** 18); a leaning elliptical shape translates high correlations (positive if leaning to the right, *e.g.*, k_{u_w} and k_{e_e} in **Fig.** 18, negative if leaning to the left).

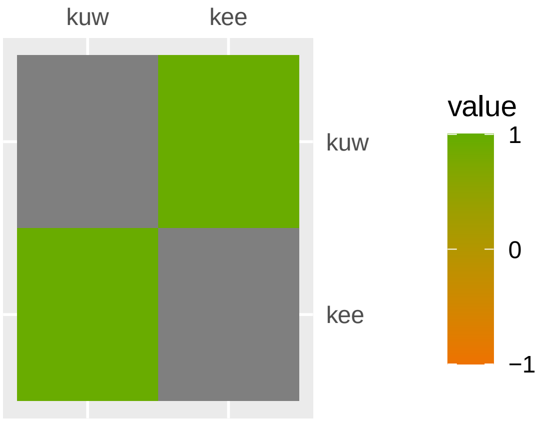


Figure 16. *Example of a cross correlation colored matrix between two parameters. These results were obtained with the example file Pimephales_two.csv.*

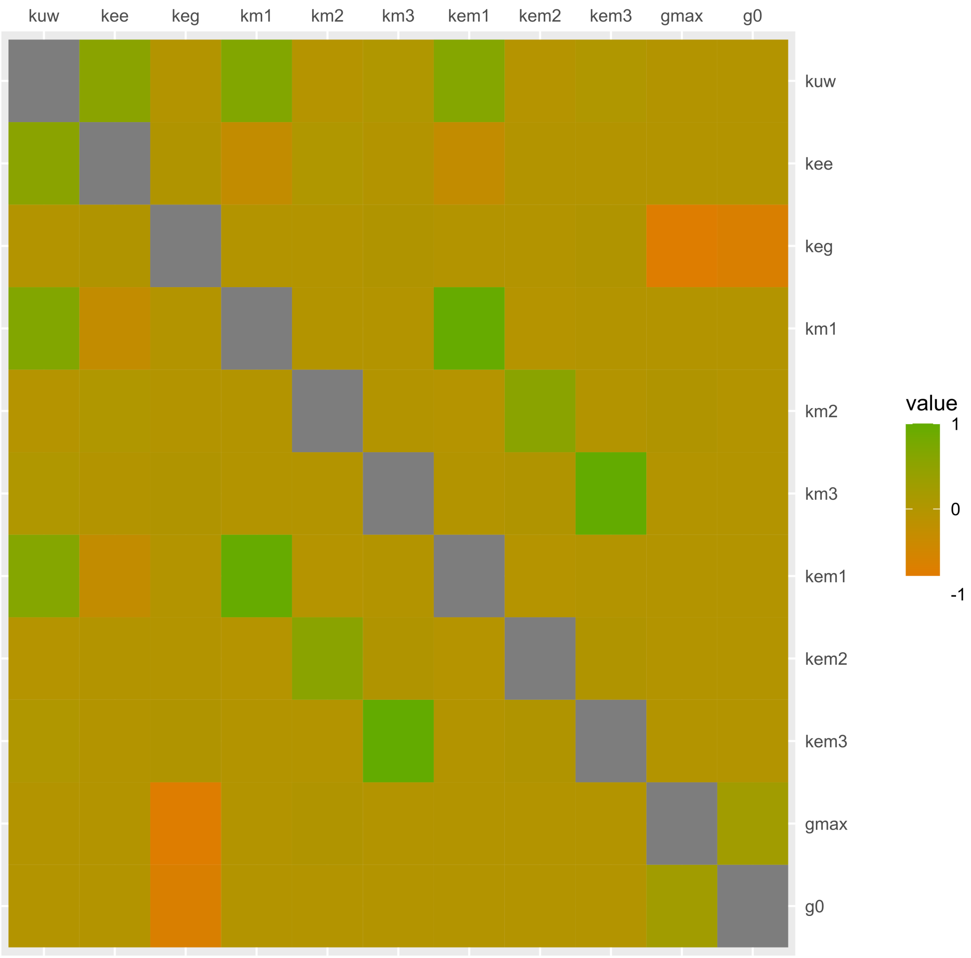


Figure 17. *Example of a cross correlation matrix, from the example file Male_Gammarus_seanine.csv.*

In the example of **Fig.** 18, k_{u_w} and k_{e_e} are highly positively correlated ($r = 0.941$), meaning that the estimate obtained for one of these two parameters will strongly influence the estimate of the other parameter. This high correlation may be due to the model structure (*e.g.*, by definition k_{u_w} and k_{e_e} are correlated), or comes from data not fully appropriate to precisely estimate the parameters.

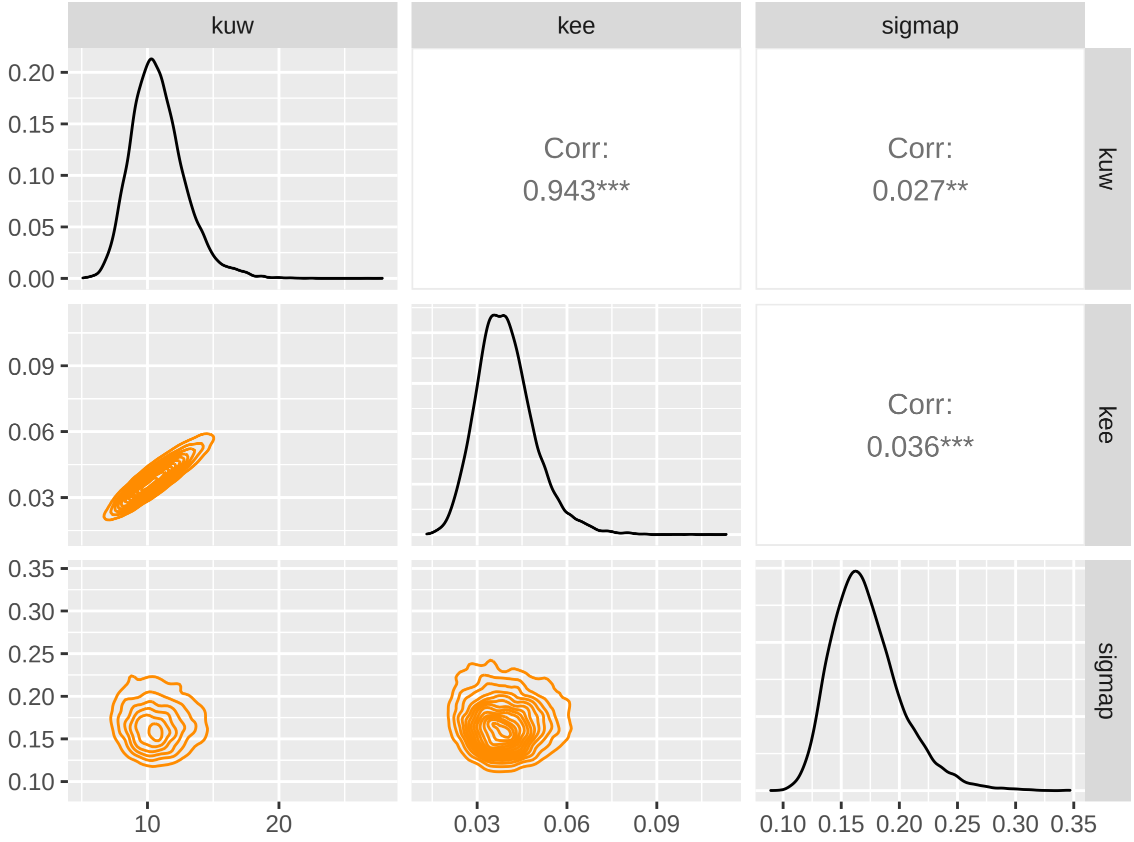


Figure 18. *Example of parameter correlations. These results were obtained with the example file Pimephales_two.csv.*

3.3.4 Potential Scale Reduction Factors (PSRF)

Convergence of the MCMC can be checked with the Gelman-Rubin diagnostic expressed with the potential scale reduction factor (PSRF). Approximate convergence is diagnosed when the PSRF is close to 1.00 (**Fig.** 19)^{19}. In the example of **Fig.** 19, the PSRF is equal to 1.0 for each model parameter, thus the convergence of the MCMC was correctly achieved.

3.3.5 Watanabe-Akaike Information Criterion (WAIC)

Information criteria offer a computationally appealing way of estimating the generalization performance of the model. A fully Bayesian criterion is the widely applicable information criterion (WAIC) by Watanabe a penalized deviance statistics accounting for the uncertainty in the parameters and can be used also for singular models. WAIC is widely used in model comparison for a same dataset (e.g., with or without k_{e_e}). Sub-models with lower WAIC values will be preferred.

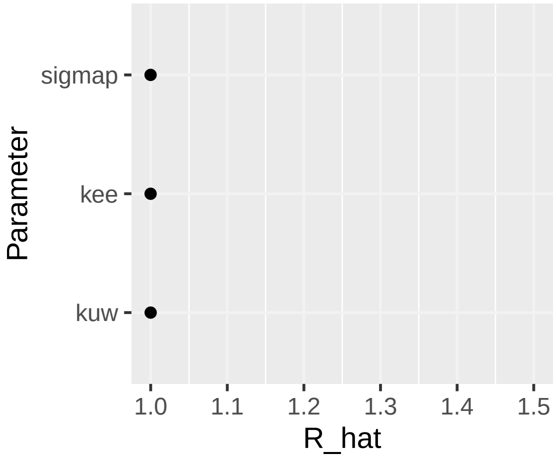


Figure 19. *Example of a PSRF. These results were obtained with the example file Pimephales_two.csv.*

For example, for highly hydrophobic substances, two models can be compared, one considering water and food exposure (which we call here model (a)) and one model considering food-only exposure due to the physico-chemical properties of the substance (model (b)). If the WAIC of model (a) is smaller than the WAIC of model (b), then model (a) will be preferred.

In this version of MOSAIC$_{\text{bioacc}}$, the users can deselect some of the parameters (based on biological hypotheses related to the most probable exposure route or by neglecting one elimination process, for example). In these specific cases, tested TK sub-models ($e.g.$, with all selected parameters corresponding to the complete TK model and sub-model with some parameter(s) deselected) can be compared with the WAIC criterion for each tested sub-model.

3.3.6 Deviance Information Criterion (DIC)

This criterion, denoted DIC, is a penalized deviance statistics accounting for the number of parameters for use in **model comparison** for a same data set ($e.g.$, with or without k_{e_e}). Sub-models with lower DIC values will be preferred[18]. DIC value can be negative. However, DIC value itself is not important, what matters is the difference between two DICs, what can help in deciding which model is the most appropriate.

For example, for highly hydrophobic substances, two models can be compared, one considering water and food exposure (which we call here model (a)) and one model considering food-only exposure due to the physico-chemical properties of the substance (model (b)). If the DIC of model (a) is smaller than the DIC of model (b), then model (a) will be preferred.

In this version of MOSAIC$_{\text{bioacc}}$, the users can deselect some of the parameters (based on biological hypotheses related to the most probable exposure route or by neglecting one elimination process, for example). In these specific cases, tested TK sub-models ($e.g.$, with all selected parameters corresponding to the complete TK model and sub-model with some parameter(s) deselected) can be compared with the DIC criterion.
You can get more information of the use of the DIC in Ratier $et\ al.$ (2019)[12].

3.3.7 Traces of MCMC iterations

A traceplot is also an essential plot for assessing convergence and diagnosing of MCMC. It shows the time series of the sampling process leading to the joint posterior distribution. Different colors are used for each of the chains (here three) to assess the within-chain convergence. The user must check whether all MCMC converge towards the same distribution limit (overlapping of the chains). This can be verified visually by observing the simulated values for each node of interest as a function of the number of iterations (**Fig.** 20). In the following example, the three MCMC overlap and converge towards the same distribution limit for each model parameter. Thus, the algorithm has suitably converged.

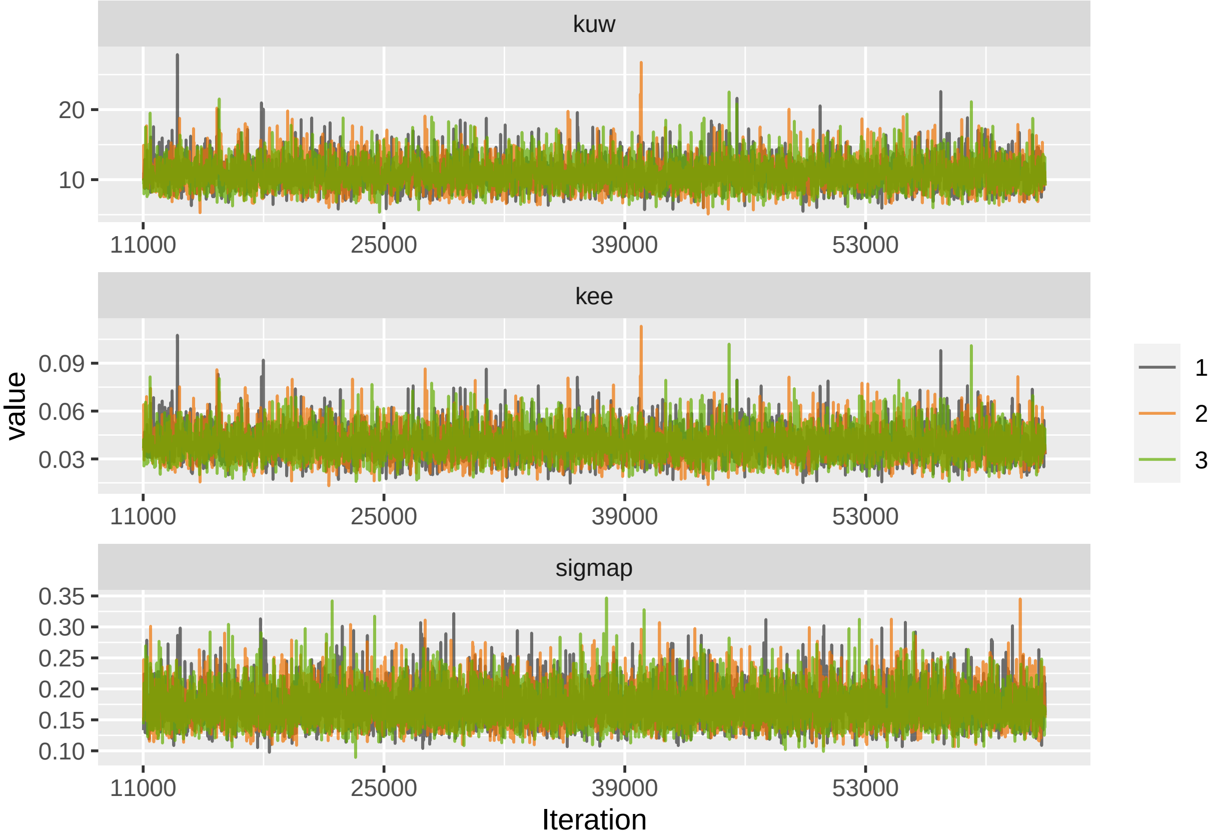


Figure 20. *Example of the overlapping of the MCMC. These results were obtained with the example file Pimephales_two.csv.*

4 Step 4: Downloads

4.1 Plots

You can download all plots as displayed by the application in several formats (.png, .jpg, .pdf, .svg, .tiff and .eps): equations, bioaccumulation metrics, fitting results and goodness-of-fit criteria. They can be downloadable separately or simultaneously in a .zip file.

4.2 Tables

You can also download table results in .txt or .csv, as for example the BCF numerical values and the joint posterior distribution for all parameters (columns) and all iterations of the MCMC algorithm (lines).

4.3 Report

You can easily download a full report of your calculations, which summarize all the results in a .pdf, .html or .docx file. We warn you that the creation of the report may take some time depending on your data set.

4.4 R Script

The R script can be downloaded as a gateway to identically reproduce all calculations provided within the application (this guarantees the full reproducibility of the results) and to perform further calculations directly within the R software with your own computer.

5 Appendix: how to interpret not "ideal" results

In practice, you may encounter situations where the results are not "ideal," conversely to those presented in sections 1 to 5 of this document. This appendix will allow you to better interpret the results in such cases. For the goodness-of-fit criteria, we suggest to consider as enough the fact that two criteria as given by $\text{MOSAIC}_{\text{bioacc}}$ are receivable.

5.1 Bioaccumulation metrics

If you asked for the BCF_{ss}, $BCF_{pw_{ss}}$, $BSAF_{ss}$ or BMF_{ss} but the median value and the 95% credible interval are not of the same order of magnitude than for the BCF_k, BCF_{pw_k}, $BSAF_k$ or BMF_k, then ensure the accumulation phase really reached the steady-state (*i.e.*, at least three successive measured concentrations with no statistical differences).

On **Fig.** 21, you can see a counter-example for asking for the BCF_{ss}:

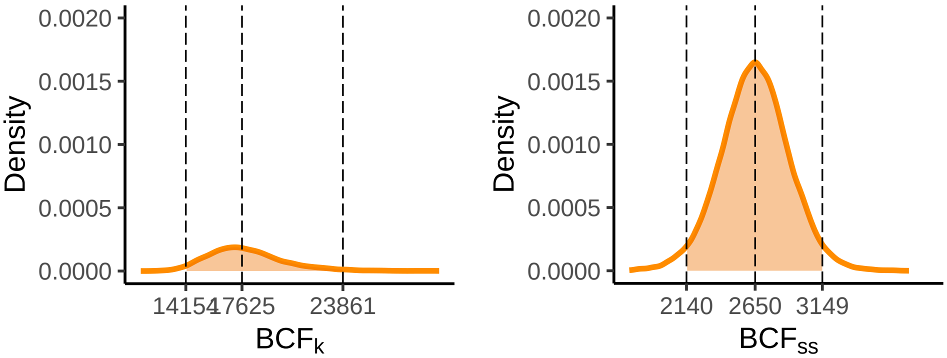


Figure 21. *Counter-example for asking for the* BCF_{ss}.

Indeed, the steady-state was not reached at the end of the accumulation phase (day 4, **Fig.** 22). Thus, the median value for BCF_{ss} is totally different from the one of BCF_k. The same consideration can be applied for the other bioaccumulation metrics.

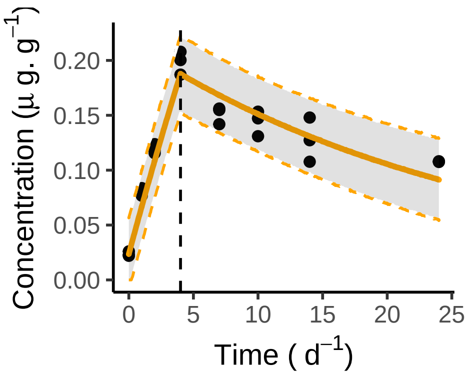


Figure 22. *Example of fitting results when the steady-state was not reached at the end of the accumulation phase.*

5.2 Parameter estimates

Large 95% credible intervals can sometimes be obtained for some parameters, especially for parameter k_{e_e} (**Table** 5). Such a situation leads to non precise estimate of these parameters, what should question their use for predictions. This can be due to the model structure or the experimental data themselves. Thus, it is an information to consider when you interpret the fitting results.

Table 5. *Example of parameter medians (50% quantile) with their 95% credible intervals (2.5% - 97.5% quantiles) with unprecise estimates for* ke_e.

	2.5%	50%	97.5%	
k_{u_w}	641.4	726.5	836.2	d^{-1}
k_{e_e}	1.605e-05	0.005828	0.02001	d^{-1}
σ_p	0.04666	0.06305	0.09025	$\mu g.g^{-1}$

5.3 PPC

If the fit is correct, it is expected to get 95% of the data within the 95% credible intervals of their predictions. So, if the range of the percentage of data within the credible intervals is between 92 and 96%, calculations and predictions can be considered as good enough. If the percentage is under 92%, predictions are considered as underestimated. If the percentage is upper 96%, predictions are considered as overestimated.

On **Fig.** 23, you can see a counter-example with large uncertainties of the model predictions leading to 100% of the data within their credible intervals.

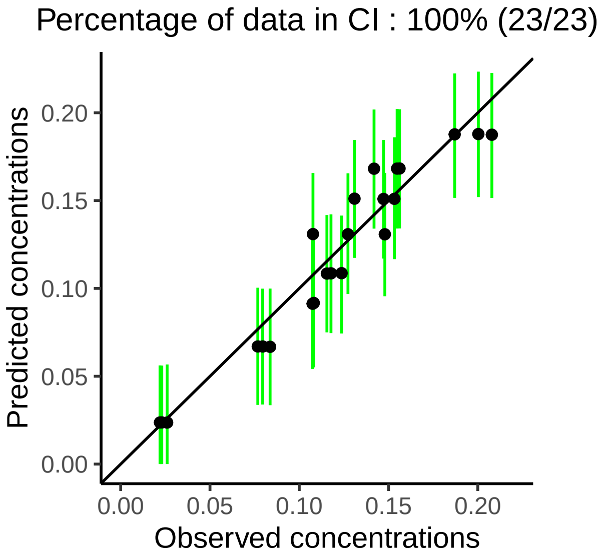


Figure 23. *Example of a PPC where there are large uncertainties of the model predictions leading to 100% of the data within their credible intervals.*

5.4 Prior and posterior distributions

We remind you that prior distributions are defined by default to be the most generic as possible. However, it can happen that your data would require other prior distributions (*e.g.*, inspired by literature, expert knowledge or by a previous study leading to parameter estimations outside of the default values as used in MOSAIC$_{\text{bioacc}}$).

The precision of the model parameter estimation can be visualized by comparing prior and posterior distributions: the overall expectation is to get a narrower posterior distribution compared to the prior one, what reflects that data contributed enough to precisely estimate parameters.
If one of the posterior distribution for a model parameter has bounds close to the lower or the upper bound of the prior distributions (*e.g.* $log10(\theta) \simeq -5$ or $log10(\theta) \simeq 5$ with θ being one of the model parameter), then the prior distribution may be not well defined. For example, if a distribution tail is observed at these bounds as illustrated on the left of **Fig.** 24 (here the distribution tail of $log10(k_{e_e})$ is too close to -5, the lower bound), then the inference process needs to be questioned. Conversely, the fit can be considered as correct if you obtain prior and posterior distributions as illustrated on the right of **Fig.** 24.

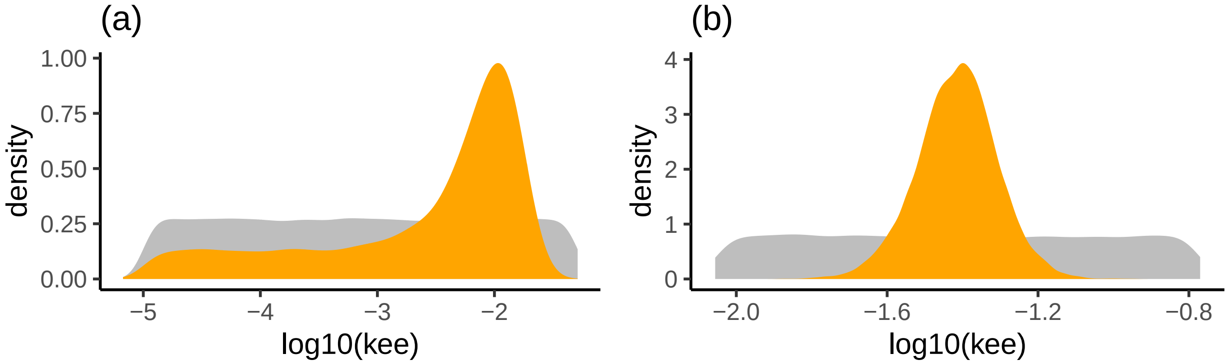


Figure 24. *Questionable posterior distribution (a distribution tail on $log10(k_{e_e})$ which tends to -5) on the left (a), and a posterior distribution as expected on $log10(k_{e_e})$ on the right (b).*

In MOSAIC$_{\text{bioacc}}$, it is not possible to change the prior distributions of parameters directly within the application. To do this, we suggest you to download the R code and to change the prior distributions directly in the R software. We remind you that to define the prior distributions you should not have a look at your data, but only on previous experiments, literature data or expert knowledge.

You can also check this goodness-of-fit criterion through the estimated values of the model parameters. If one of the model parameter is closed to the lower or the upper bound of its 95% credible interval ($\theta \simeq 10^{-5}$ or $\theta \simeq 10^{5}$ with θ one of the model parameter), the prior distribution may be not appropriate. As for example illustrated in **Table** 5, the distribution tail of $log10(k_{e_e})$ is too closed to the lower bound -5.

5.5 Correlations between parameters

If a high correlation is obtained between two parameters (*e.g.*, more than 0.7 or less than -0.7 for the Pearson correlation coefficient), it is an information to consider, not necessarily a bad result. It means that the estimate obtained for one of these two parameters will strongly influence the estimate of the other one. Such a high correlation may be due to the model structure itself (for example, by definition k_{u_w} and k_{e_e} are correlated, **Fig.** 25, or to the data so that it cannot be avoided).

Sometimes, you may get a bimodal posterior distribution for one or several parameters what translates into a double maximum on density plots (*e.g.*, parameters k_{u_w} and k_{e_e} in **Fig.** 25). To make the application as generic as possible, we defined priors for each parameter the most global as possible. However, depending on the experimental conditions, the parameters may not be really included within the prior chosen range of values. In such a case, we recommend you to contact sandrine.charles@univ-lyon1.fr if you are not experimented with Bayesian inference and R software, or to change the downloadable R script by yourself.

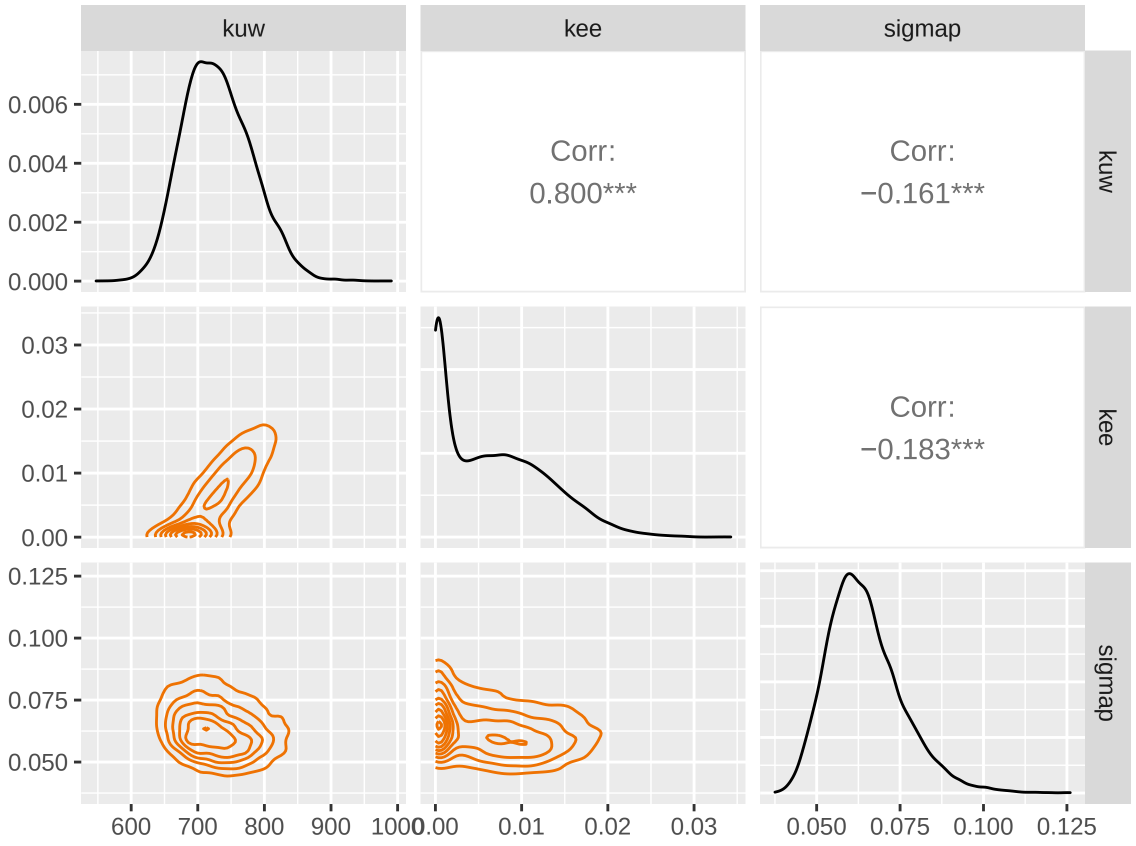


Figure 25. *Example of a parameter correlation result which raises questions.*

5.6 PSRF

This criterion must be as close as possible to 1 for each model parameter to ensure that the between-chain variability is small compared to the within-chain variability. Based on our experience, from a value of 1.03, the results should be questioned. Most often, such a case appears when priors are not fully appropriate or when the data do not contain enough information. One of the solution may be to increase the number of iterations in the MCMC by using the R script directly.

5.7 WAIC and DIC

The WAIC or the DIC are not criteria to consider to check the goodness-of-fit. However, they are crucial criteria to consider when two models or more are compared after fitting on a same data set.

In practice, users just need to choose the parameters they want to appear in sub-models. According to the experimental conditions, several sub-models can indeed be considered and compared depending on the hypotheses to test either on the exposure routes or on the elimination processes. Organisms may have been exposed via several media (water and sediment in the following case study. By default, MOSAIC$_{\text{bioacc}}$ fits the full TK model. Then users can test different TK sub-models, for example sub-models with only one exposure route (*e.g.*, water or sediment, and compare them to the full model based on both the Deviance Information Criteria (DIC) and the Watanabe-Akaike information criterion (WAIC) delivered by MOSAIC$_{\text{bioacc}}$. In order to illustrate this framework, the data set of *Physa acuta* exposed to AgNO3 spiked water and clean sediment for 7 days[20] was uploaded in MOSAIC$_{\text{bioacc}}$.

First hypothesis: complete TK model

A first run of analyses was performed where parameters were automatically selected according to the corresponding experimental design (Eqs. A1 and A2), as given within the uploaded data set (*i.e.*, water and sediment for the exposure routes).

$$\begin{aligned}
\tfrac{dC_p(t)}{dt} &= k_{u_w} \times c_w + k_{u_s} \times c_s - \left(k_{e_e}\right) \times C_p(t) && \text{for } 0 \leq t \leq t_c \quad (A1)\\
\tfrac{dC_p(t)}{dt} &= -\left(k_{e_e}\right) \times C_p(t) && \text{for } t > t_c \quad (A2)
\end{aligned}$$

Second hypothesis: only account for water exposure

A second run of analyses was performed on the same data set that for the first run, but without considering the uptake rate from sediment (Eqs. A3 and A4), because the concentration of the chemical in the sediment is at environmental level.

$$\begin{aligned}
\tfrac{dC_p(t)}{dt} &= k_{u_w} \times c_w - \left(k_{e_e}\right) \times C_p(t) && \text{for } 0 \leq t \leq t_c \quad (A3)\\
\tfrac{dC_p(t)}{dt} &= -\left(k_{e_e}\right) \times C_p(t) && \text{for } t > t_c \quad (A4)
\end{aligned}$$

Comparison between the nested TK sub-models

For each analyses, the parameter estimates are summarized in Table 6. The WAIC and the DIC are penalized deviance statistics accounting for the number of parameters for use in model comparison for a same data set[19]. In this example (Table 6), the WAIC and DIC are similar between the two runs, as well as the estimations of (k_{e_e}), considering or not the sediment exposure route. Thus, for this data set, and applying the parsimony principle, it can be deduced that *P. acuta* accumulate AgNO3 principally by water. Silva et al. (2020)[20] also concluded that when accounting for double exposure via both water and clean sediment, water was likely to be the main route.

5.8 Traces of MCMC

You must check whether the MCMC converge towards the same distribution limit (overlapping of the chains). As shown in **Fig.** 26, the three MCMC do not overlap and do not converge towards the same distribution limit for two model parameters (k_{u_w} and k_{e_e}). If at least two of the other criteria are good enough, this can

Table 6. *Summary of parameters estimated for the Physa_AgNO3_Silva2020.csv data set for run 1 (complete TK model) and for run 2 (without considering sediment exposure route, $k_{u_s} = 0$). For parameters, the median value in given with its 95% credible interval into bracket.*

	run 1	run 2
k_{u_w} (days^{-1})	1.492 [$1.874 * 10^{-5}$;60480]	679[568.9;972.6]
k_{u_s} (days^{-1})	624.6[30.6;903.6]	0
k_{e_e} (days^{-1})	$5.605 * 10^{-3}$ [$1.421 * 10^{-5}$;0.07212]	$3.952 * 10^{-3}$ [$1.382 * 10^{-5}$;0.07106]
WAIC	177.9	178
DIC	178.3	178.6

be disregarded. If not, your experimental data could be not sufficient to performed bioaccumulation metric calculations and to estimate parameters of a TK model.

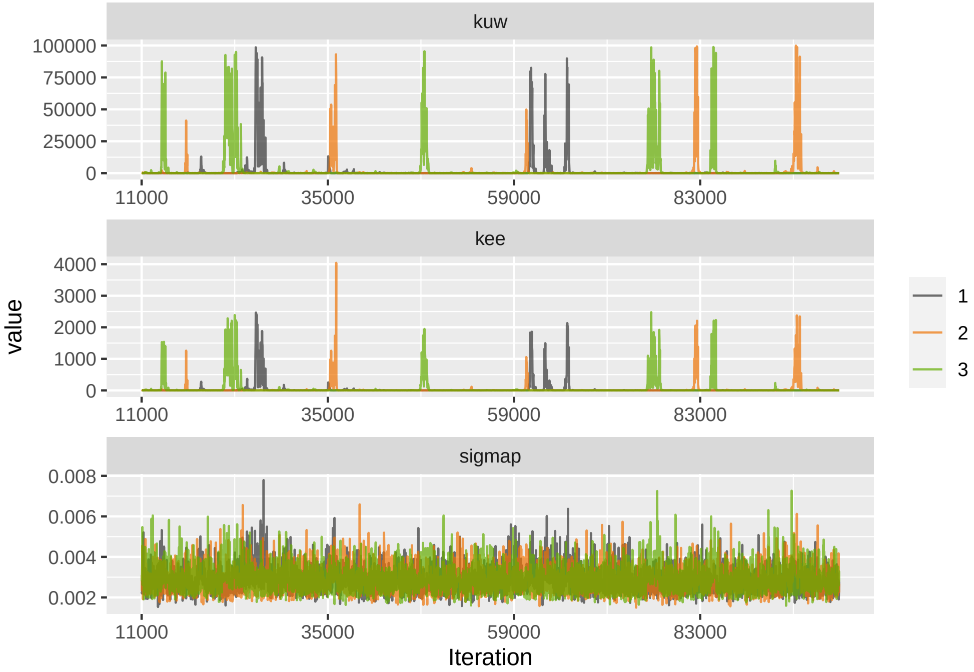


Figure 26. *Example of problematic MCMC traces because the three chains have not converged.*

6 Glossary

Bioconcentration factor (BCF): BCF is a parameter describing bioaccumulation of water-associated organic compounds or metals into tissues of ecological receptors.

Biomagnification factor (BMF): BMF is a parameter describing bioaccumulation of food (or in the predator's prey)-associated organic compounds or metals into tissues of ecological receptors.

Biota-Sediment Accumulation Factor (BSAF): BSAF is a parameter describing bioaccumulation of sediment-associated organic compounds or metals into tissues of ecological receptors.

Credible Interval (CI): A credible interval is the interval in which an parameter has a given probability. It is the Bayesian equivalent of the confidence interval.

Directed Acyclic Graph (DAG): It symbolize the deterministic links between parameters and variables for a model and the stochastic links between the observed and predicted data.

Monte Carlo Markov Chain (MCMC): A method which comprise a class of algorithms for sampling from a probability distribution. By constructing a Markov chain that has the desired distribution as its equilibrium distribution, one can obtain a sample of the desired distribution by recording states from the chain.

Potential Scale Reduction Factor (PSRF): Gelman-Rubin diagnostic to check the convergence of the MCMC.

Toxico-Kinetic model (TK model): Toxico-kinetics is the mathematical description of the uptake and disposition of a chemical in the organism. TK modeling is usually implemented by describing the time course of the amount or concentration of the parent substance and its metabolites in all the organism.

References

(1) R Core Team. (2020) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

(2) Plummer, M. (2019) rjags: Bayesian Graphical Models using MCMC. R package version 4-10.

(3) Chang, W., Cheng, J., Allaire, J. J., Xie, Y., and McPherson, J. (2020) Shiny: Web Application Framework for R.

(4) MacKay, D., and Fraser, A. (2000) Bioaccumulation of persistent organic chemicals: Mechanisms and models. Environmental Pollution 110, 375–391.

(5) European Commission. (2013) COMMISSION REGULATION (EU) No 283/2013 of 1 March 2013 setting out the data requirements for active substances, in accordance with Regulation (EC) No 1107/2009 of the European Parliament and of the Council concerning the placing of plant protection produc.

(6) OECD. (2012) Test No. 305: Bioaccumulation in Fish: Aqueous and Dietary Exposure. Paris.

(7) Charles, S., Veber, P., and Delignette-Muller, M. L. (2018) MOSAIC: a web-interface for statistical analyses in ecotoxicology. Environmental Science and Pollution Research 25, 11295–11302.

(8) EFSA. (2014) Scientific Opinion on good modelling practice in the context of mechanistic effect models for risk assessment of plant protection products. EFSA Journal 12.

(9) Fernández-i-Marín, X. (2016) ggmcmc: Analysis of MCMC samples and Bayesian inference. Journal of Statistical Software 70(9), 1–20.

(10) Plummer, M. (2003) JAGS: A program for analysis of Bayesian models using Gibbs sampling. Proceedings of the 3rd International Workshop on Distributed Statistical Computing; Vienna, Austria.

(11) Kellner, K. (2019) jagsUI: A Wrapper Around 'rjags' to Streamline 'JAGS' Analyses. R package version 1.5.1.

(12) Ratier, A., Lopes, C., Labadie, P., Budzinski, H., Delorme, N., Quéau, H., Peluhet, L., Geffard, O., and Babut, M. (2019) A Bayesian framework for estimating parameters of a generic toxicokinetic model for the bioaccumulation of organic chemicals by benthic invertebrates: Proof of concept with PCB153 and two freshwater species. Ecotoxicology and Environmental Safety 180.

(13) Gelman, A. (2006) Prior Distribution for Variance Parameters in Hierarchical Models. Bayesian Analysis 3, 5901–5906.

(14) Ockleford, C., Adriaanse, P., Berny, P., Brock, T., Duquesne, S., Grilli, S., Hernandez-Jerez, A. F., Bennekou, S. H., Klein, M., Kuhl, T., Laskowski, R., Machera, K., Pelkonen, O., Pieper, S., Smith, R. H., Stemmer, M., Sundh, I., Tiktak, A., Topping, C. J., Wolterink, G., Cedergreen, N., Charles, S., Focks, A., Reed, M., Arena, M., Ippolito, A., Byers, H., and Teodorovic, I. (2018) Scientific Opinion on the state of the art of Toxicokinetic/Toxicodynamic (TKTD) effect models for regulatory risk assessment of pesticides for aquatic organisms.

(15) Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953) Equation of state calculations by fast computing machines. The Journal of Chemical Physics 21, 1087–1092.

(16) Hastings, W. K. (1970) Monte carlo sampling methods using Markov chains and their applications. Biometrika 57, 97–109.

(17) Geman, S., and Geman, D. (1984) Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6, 721–741.

(18) Raftery, A. E., and Lewis, S. M. (1992) [Practical Markov chain Monte Carlo]: Comment: one long run with diagnostics: implementation strategies for Markov chain Monte Carlo. Statistical Science 7, 493–497.

(19) Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002) Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society. Series B: Statistical Methodology 64, 583–639.

(20) Silva, P. V., van Gestel, C. A. M., Verweij, R. A., Papadiamantis, A. G., Gonçalves, S. F., Lynch, I., and Loureiro, S. (2020) Toxicokinetics of pristine and aged silver nanoparticles in Physa acuta. Environmental Science: Nano 7, 3849–3868.