Ecotoxicological HC20-values and their statistical distribution. A nonlinear weighted regression applied to thousands of chemicals

Oskar Nyberg, Reinout Heijungs, Patrik Henriksson

Department of Ecology, Environment and Plant Sciences, Stockholm University, Stockholm, Sweden  
 Department of Operations Analytics, Vrije Universiteit, Amsterdam, The Netherlands  
 Institute of Environmental Sciences, Leiden University, Leiden, The Netherlands  
 Stockholm Resilience Centre, Stockholm, Sweden  
 WorldFish, Jalan Batu Maung, Penang, Malaysia  
 Beijer Institute of Ecological Economics, The Royal Swedish Academy of Science, Stockholm, Sweden

Corresponding author: [oskar.nyberg@su.se](mailto:oskar.nyberg@su.se)

**Keywords** Ecotoxicological impacts, Probabilistic uncertainty distributions, Species sensitivity distribution, Life cycle impact assessment

## Abstract

*Introduction:* The agriculture sector is a major emitter of toxic chemicals into the environment. However, the ecotoxicological effect data which form the basis of evaluations of ecological impacts from chemical emissions are incomplete, and completely absent for some chemicals, which result in risks being overlooked. Where data are available, they tend to be heterogeneous and accompanied with large uncertainties. In the present research we present a methodology for quantifying the variability and uncertainty of agri-food chemicals and evaluate its implications for environmental assessment frameworks, such as life cycle assessments.

*Methods:* Starting from a list of 16,797 chemicals from the online agri-food environmental framework HESTIA, we assigned ecotoxicological effect data at NOEC, EC10, and EC50 endpoints from the following repositories: ECOTOX, Aquatic OASIS, Aquatic ECETOC, Aquatic Japan MoE, and Food TOX Hazard EFSA. These data were converted into EC10 equivalent endpoints using regression coefficients from the literature, which allowed us to calculate the concentration response slope factors corresponding to the slope on the SSD curve at the 20% response level of organisms exposed to a chemical (). In turn, we fit effect data to a cumulative normal distribution using a nonlinear least square model, to estimate and of the . The modelled and are used to fit a normal distribution for 100,000 Monte Carlo simulations of the , from which the 95% percentile distribution for the can be extracted and inform on the variability in this point.

*Results:* The outcomes from our analysis resulted in a database detailing 118,443 curated records that span 1,792 species and 3,692 chemicals suitable for calculations. From these data we are able to calculate values and 95% percentile distributions of the for 2,350 and 684 chemicals respectively. The chemical class pesticides is the most data rich category of chemicals, but also the category with largest uncertainty attached to the .

*Discussion:* We show that the variance among toxicity estimates for the same species and chemical can be used in weighted nonlinear model fitting to generate an uncertainty range attached to a value. This allows for uncertainties related to ecotoxicological impact characterization in environmental frameworks to be estimated. Data scarcity is an omnipresent issue when it comes to characterizing toxicity of chemicals, where only 63.7 % of all chemicals with effect data records have enough data to calculate a value, and 18.5 % have enough data to fit a weighted nonlinear least squares model. Our recommendation is to incorporate toxicological variance in estimations of ecotoxicity impacts and LCIA impact categories, to reduce ambiguity and allow for verification when comparing ecotoxicological impacts.

## Introduction

Increasing chemical use is a major concern for operating within the safe space of planetary boundaries, and agriculture is a major driving force behind this increase, through pesticide application to commercial crops, veterinarian drugs applied to livestock and farmed aquatic organisms, and disinfectants applied in post-harvest food processing stages (Gordon et al., 2017; Persson et al., 2022; Simpson and Mitch, 2022). Chemical use in agriculture is also expected to continue to increase in the coming decade, as a result of larger production volumes and more intensive production systems (Schreinemachers and Tipraqsa, 2012). The agrifood sector is a major emitter of potentially toxic chemicals (Persson et al., 2022), mainly through the use of pesticides, with an estimated annual use of close to 3 million metric tonnes (FAO, 2023). However, data on which potentially toxic chemicals are used for different farming systems, toxicity evaluations for different compounds, and variability in toxicity potentials are incomplete and fragmented, hampering accurate global comparisons of ecotoxicity impacts of different food products (Der Werf et al., 2020). While the majority of pesticides emitted are biodegraded, a fraction of pesticides does reach rivers and water catchments where they degrade at slower rates, with potential to bioaccumulate in aquatic organisms (Maggi et al., 2023). Accurate toxicological characterization of chemicals is essential for ensuring the safety of human health and the environment, where proper toxicological characterization involves the identification and evaluation of the potential adverse effects of chemicals, determining the level of exposure that is safe for humans and the environment, and assessing the risk posed by exposure to chemicals (Krewski et al., 2010).

Life cycle assessment (LCA) is increasingly being used as a framework for benchmarking toxicological impacts of agrifood products (Der Werf et al., 2020). LCA is an ISO-standardized environmental framework that seeks to aggregate the emissions and resource use throughout a value chain of a product, and characterize these towards one or more environmental impact categories (Standardization, 2006). The results are subsequently scaled to a predefined unit of reference (functional unit), to allow for comparisons between products (e.g. per of kg food), functions (e.g. per kcal), or services (e.g. per washing of a plate). It is most commonly encountered as the methodology behind carbon footprints, but is also used to quantify freshwater consumption, land occupation, eutrophication, biodiversity loss, or toxicity impacts (Hauschild and Huijbregts, 2015). Toxicity impacts, in turn, are commonly evaluated in terms of human toxicity, either including or excluding cancer cases, and ecotoxicity impacts on freshwater, marine, or terrestrial ecosystems (Hauschild and Huijbregts, 2015).

LCA captures toxicological impacts throughout whole value chains, i.e., mercury emissions from coal-fired power plants to therapeutants used in aquaculture ponds (Bohnes et al., 2019; Malode et al., 2023). This allow LCAs to provide useful insights into where in value chains the largest toxicity reductions can be achieved, such as that most freshwater ecotoxicological impacts related with prawn farming in Bangladesh are related to the production of certain agricultural materials used for feeds, and not the prawn grow-out (Henriksson et al., 2015). To assess ecotoxicological impacts throughout the value chain of a product, ecotoxicological characterization factors (CF) for chemicals are used to evaluate potential impact of different chemicals in the LCIA phase of an LCA. Toxicological effect data for chemicals provide, together with data on fate and exposure, the basis for CF calculations within LCA. Regrettably, none of the available life cycle impact assessment (LCIA) methodologies provide a complete set of CFs for pesticides and therapeutants, which often leads to neglecting or inserting of a proxy value (Nemecek et al., 2022). Additionally, ecotoxicological effect data are notoriously heterogeneous since they are reported across hundreds of tested species at variable concentrations, measured effects, and at various empirical or modeled endpoints, resulting in large uncertainty and variability (which we will from here on simply refer to as “uncertainty”) in CF calculations.

The evaluation of toxicological impacts in LCA is limited by the number of chemical compounds characterized by impact assessment methodologies. The CF is the value used to translate the amounts of chemicals used to their potential toxicological impacts. There are several different impact assessment methodologies to derive these characterization factors, including USES-LCA v1&2 (Huijbregts et al., 2000; Van Zelm et al., 2009a), IMPACT 2002 (Pennington et al., 2005), and UNEP-SETAC’s USEtox (Fantke et al., 2017), but ultimately, they all rely on similar fundamental data on toxicity and physicochemical properties. Among these impact assessment methodologies, the USEtox model is most widely used at present, and also the one promoted by the European Platform on Life Cycle Assessment (ILCD, 2010).  
The USEtox framework is designed to generate CFs for human toxicological and aquatic ecotoxicity impacts by linking models which estimate environmental distribution and fate of an emitted chemical, ecosystem population exposure to this chemical, and potential toxicological effects to populations being exposed to the chemical (Fantke et al., 2017). The present study focuses on the ecotoxicological impacts and the toxicological effect data that are used to generate effect factors (EF) in the latter model. The EF is defined by the the concentration-response slope factor (), which corresponds to the slope on a species sensitivity distribution curve (SSD) at the point of the percentile response level (HC20) derived from toxicological EC10-equivalent () effect data. Each toxicological data point used to construct an SSD curve is an estimated toxicological response of an organism from exposure to a chemical. However, there is variability in the toxicity estimates for every chemical, for the same species. Additionally, depending on the amount of data, which type of chemical, a specific toxic mode of action, and which organisms that are tested will affect the uncertainty of the HC20 value of the SSD curve. When applying the current methodology to derive the HC20 value from an SSD curve, variability of these toxicity data are not accounted for. To investigate the uncertainties attached to values, we propose a method for applying weighted means to species-specific averages when constructing SSD curves by applying a non-linear least squares fit model to a cumulative normal distribution of data and solve for the mean () and standard deviation (). The estimated and can subsequently be used in Monte Carlo simulations of a normal distribution of data in the from which a 95% percentile distribution will inform on the probabilistic uncertainty at this data point.

In order to generate the above mentioned variables, we gathered a large set of ecotoxicological effect data using the OECD QSAR Toolbox (Dimitrov et al., 2016) across 16,797 chemicals included in the HESTIA inventory (<https://www.hestia.earth>). HESTIA a free open-access platform that provides a data repository for life cycle inventory (LCI) data using a harmonized schema and glossary of terms, and provide calculation models for various emissions and impact assessments.

The objectives of this article are to 1) gather freely available ecotoxicological records for chemicals present in the HESTIA inventory and curate these records into a database adapted for EF calculations, 2) calculate values according to recommendations for all chemicals possible within the database, 3) present a methodological approach that can estimate the probabilistic distribution of the value as 95% percentile distribution, and lastly 4) report statistical uncertainty at the value as a geometric standard deviation for as many chemicals possible in the database.

## Methods

### Toxicological data collection

Several databases and software were used to collect data and construct an ecotoxicological database for CRF and uncertainty estimations with detailed descriptions for each step of data collection and curation procedures provided in supplementary information (SI) along with a schematic diagram of the methodology (SI, Figure S1). The starting point for ecotoxicological effect data collection is the substance inventory of HESTIA ([http://HESTIA.earth](http://hestia.earth)); a list of 16,797 CAS registry numbers (CASRN) with matching chemical names of potentially harmful substances (SI 1.1). To ensure the correct identity of chemicals, CASRN and chemical names were matched to Simplified Molecular Input Line Entry System (SMILES) notations by querying the NCBI PubChem database using the R package Webchem (Szöcs et al., 2020) (SI 1.2). Then, two data queries were made based on the CASRN-SMILES associations to OECD QSAR Toolbox v4.5 (Dimitrov et al., 2016) for: a) physicochemical properties required by the USEtox model (SI 2.1), and b) all openly available aquatic ecotoxicological records accessible through the software (SI 2.2).  
Next, chemical use-classification information was gathered from USEPA CompTox v2.2 (<https://comptox.epa.gov/dashboard/>, accessed 2023-01-20), USEPA ECOTOX database (downloaded in its entirety 2022-03-10), The British Compendium of Pesticide Common Names (<http://www.bcpcpesticidecompendium.org/>, accessed 2022-10-13), The Chemical Entities of Biological Interest (ChEBI) database (<https://www.ebi.ac.uk/chebi/init.do>; accessed 2023-01-18), The Anatomical Therapeutic Chemical (ATC) Classification System (<https://www.whocc.no/>), followed by substance type classification into the following chemical groups: Organic, Chemical mixture, Inorganic, and Unknown (SI 3.1).  
Taxonomic information attached to effect data were updated to the most recent taxonomic classification, common names were converted into Latin names, and spelling errors were corrected (SI 3.2).  
Data curation was performed according to the following steps (SI 3.3):  
1) Ecotoxicological effect data are generally reported as a mean effect value at a certain endpoint. Additionally, a substantial amount of data (74,216 records) are reported with a qualifier; “<”, “>”, =“, or”ca.”, and in some of these cases records have effect data only reported as a range with a minimum and maximum value (33,560 records). We adjust test concentrations reported as a range for records where a lower range value is annotated with “ca.”, “=” or no annotation to the lowest reported value;  
2) Records where control type is described as “Insufficient” or “Unsatisfactory” are removed. Additional data based on QSAR estimations or *in vivo* effect based tests are removed in this step by excluding records containing the strings “QSAR”, “bioassay”, “quantitative” in the title, and the strings “QSAR” and “bioassay” in the experimental design description;  
3) selecting freshwater and culture media records only;  
4) selecting records with effect criterions relevant for ecological impacts;  
5) remove records missing taxonomic information at the species level;  
6) harmonizing effect concentrations and effect concentration units into ;  
7) remove records where test duration is <24 h or missing;  
8) acute or chronic definitions can be assigned to records based on test duration and taxonomic group; and  
9) effect data from multiple endpoints are combined and converted to endpoint according to recommendations (Owsianiak et al., 2019). EC10, EL10, IC10, LC10, LOEC are combined into “EC10”, EC50, EL50, IC50, LC50 are combined into “EC50”, and EC0, LC0, NOAEC, NOEC, NOER, NOEL are combined into “NOEC”. Endpoint conversion using regression coefficients are applied for converting acute EC50, acute NOEC, chronic EC50, and chronic NOEC into chronic into chronic respectively using the suggested regression coefficients from (Aurisano et al., 2019). Admittedly, this will group endpoints that are measured differently, for instance EC0 corresponds to the measured concentration of a chemical at which there is no detectable effect, while NOEC corresponds to the highest measured concentration at which no statistically significant effects are observed compared to a control population.

### Exploring weighted means of averages and uncertainty through nonlinear least square fit modelling

Based on the assumption of lognormal distribution of ecotoxicological effect data, we investigate the uncertainties of the value by fitting all records per substance to a nonlinear least squares model to explore and at the 20% response level on the SSD curve. First, we define the raw data. For a specific substance , we have data for species , done in a number of experiments, labeled as . The data are thus indicated as , Throughout the analysis, we will work with the 10-logarithm of the data, and for conciseness the data are indicated as , thus:

The per-species average over all samples are found as:

and the corresponding standard deviation as:

The average values have a standard error that is given by:

Next, we make the assumption that species sensitivity follows a lognormal distribution in agreement with Posthuma et al. (2019). From this distribution of the -values, which is characterized by two parameters, the mean () and the standard deviation (), that are traditionally estimated as follows:

Also, the log is found as the 20-percentile value of the distribution:

where is the inverse of the standard normal distribution, which approximates -0.84.  
The value is then derived from the log by finding the average slope of the SSD curve at the percentile response level in Eq.8:

We follow Owsianiak et al. (2023) and Sala et al. (2022) in defining the CRF-values according to Eq. 8, while emphasizing that this represents an average factor, not a marginal factor as described by Owsianiak et al. (2023). We refer to Posthuma et al. (2001) for the introduction of marginal factors, to Huijbregts et al. (2011) for an argument to switch to marginal factors, and to Heijungs (2021) for pointing out some problems with average factors.  
The uncertainty of this estimate basically depends on two elements: the uncertainty in and the uncertainty in . Both of these depend on the degree of fit with the normal distribution, and with the intra-species variation. Our approach will be to fit a normal distribution to the vector of mean values , weighted with the reciprocal of the variance, . Suppose we fit a function to the cumulative distribution of -values, each of which is associated with a standard error . We can find the optimal values of and through a least-squares minimization of the residual:

Here, denotes the order of appearance in the cumulative form. More precisely:

For F, we take the cumulative normal distribution, given by:

To model estimates for and , we fit a weighted nonlinear least-squares (nls) model to the toxicological dose-response data using R-programming language. By assuming a cumulative normal distribution of the data (Eq. 11) and defining a self-starting function for the cumulative normal distribution to start solving for the least squares residual of and , programming in R allow the use of the nls() function for this purpose. In cases where only one toxicological record per species is available, is not obtainable, and an arbitrary value of is used. Starting points for the model are defined as and respectively. If there are sufficient records for each chemical, the function models the (Eq. 5) and (Eq. 6) for each chemical, based of the and per species. We then calculate the log using weighted means for each chemical (Eq. 7) as well as extracting the value in the 0.2 quantile of the model results.

To calculate the variability in the , we extract and from the model results and fit data to a normal distribution using Monte Carlo simulations with 100,000 iterations to populate vectors with simulated and at the 20% percentile level respectively. These vectors are then used to calculate 100,000 log values, from which we subsequently derive the (Eq. 8). By selecting the 2.5% and 97.5% quantile, the 95% probability distribution of the value is computed to quantify the uncertainty (Brömssen and Röös, 2020). Since the is no longer expressed as logarithmic data, we can compare the magnitude of variability in the across multiple chemicals expressed as the geometric standard deviation, calculated using the “GeoSD” function in the R package EnvStats (Millard, 2013).

In cases where species-specific effect is based on only one record, no variance is obtainable in this point. Since variance is required for weighting, data lacking variance are excluded from calculations, along with data where the variance is 0, which occurs if a species-specific effect is based exclusively on two or more identical values. The calculations are then run for all chemicals with effect data (and variance) available for >1 species. Data are flagged with “Data insufficient” when SSDs are based on <5 species across <3 taxonomic groups because of three reasons: 1) the USEtox v2.0 manual requires data from three taxonomic groups, van Zelm et al., (2007) points out that uncertainties increase dramatically when the ecotoxicological is calculated using <4 species, and Owsianiak et al., (2019) points out that at least five species are required for deriving SSDs, to ensure that the HC20 point is not extrapolated below any species’ .

### Data availability in the HESTIA Ecotox Explorer:

To make the results from this study easily accessible, we have created the an online Ecotox Explorer (*URL being generated*), a repository where users can generate a summary of the model results for each eligible chemical, create dose-response curves for chemicals, view a histogram over the Monte Carlo result’s data distribution, and download the underlying ecotoxicological data.

### Software used

OECD QSAR Toolbox v.4.5 (Dimitrov et al., 2016) was used for collecting ecotoxicological data and physicochemical data, R version 4.1.2 (2021-11-01) (R Core Team, 2023) was used for data curation and visualization, text and data was published using Pandoc v. 2.19.2, and Shiny R package version 1.7.4.9002 (Chang et al., 2023) was used to construct the interactive web-based Ecotox Explorer .

## Results

### Ecotoxicological database construction

The present study has gathered physicochemical properties required for freshwater aquatic toxicity potential characterization in the USEtox model for 13,680 chemicals. Ecotoxicological effect data are gathered for 3,692 chemicals using OECD QSAR Toolbox software. An overview of the dataset curation process is presented in Figure 1.  
Substance use categories are defined as Antibiotic, Other inorganic chemicals, Other organic chemicals, PPCP, and Pesticide, with 943, 3134, 17288, 5712, and 91366 records, respectively across 48, 108, 1888, 212, and 1436 substances, respectively (Table 1). Pesticides is the most data rich substance use category, with 77.1 % of the total number of records belong to this category, but is represented by only 38.9 % of the total number of chemicals in the dataset.

Table 1: Number of toxicological records and chemicals respecively categorized into defined chemical use categories. PPCPs = Pharmaceuticals and Personal Care Products.

| Chemical use category | Number of records | Number of chemicals |
| --- | --- | --- |
| Antibiotic | 943 | 48 |
| Other inorganic chemicals | 3,134 | 108 |
| Other organic chemicals | 17,288 | 1,888 |
| PPCP | 5,712 | 212 |
| Pesticide | 91,366 | 1,436 |
| Total | 118,443 | 3,692 |

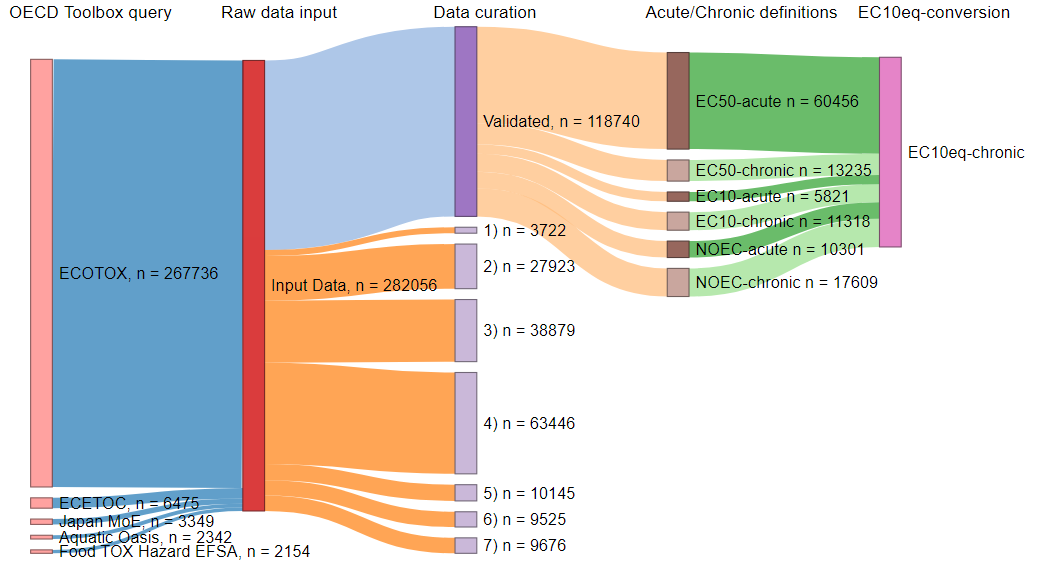


Figure 1: Overview of the data curation process from respective number of records (n) per database, stepwise removal of records, and the counts of acute or chronic records per endpoint in the ‘validated data’-category. Data removal occured in the following steps: 1) Data reported as a range are conditionally excluded, 2) Controls marked as insufficient or unsatisfactory, or effect data are based on QSAR estimations or bioassays, 3) Non-fresh water data, 4) Effect criterions irrelevant, 5) Poor taxonomic descriptions, 6) Effect unit or effect value missing or reported as 0, 7) Test duration missing or < 24h

The curated dataset contains toxicological data with 118,443 records across 1,792 species divided into 10 taxonomic groups (Table 2). Data availability varies greatly across species, with a few species being over-represented in the dataset: *Daphnia magna, Oncorhynchus mykiss, and Pimephales promelas* corresponds to 28 % of the dataset.

Table 2: Number of records found in the HESTIA Environmental Toxicology dataset by taxonomic group and endpoint.

| Taxonomy group | EC10 acute | EC10 chronic | EC50 acute | EC50 chronic | NOEC acute | NOEC chronic | Total per taxa |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Algae | 82 | 2,278 | 454 | 6,080 | 141 | 3,044 | 12,079 |
| Amphibian | 553 | 539 | 2,183 | 272 | 631 | 1,326 | 5,504 |
| Annellidae | 76 | 39 | 676 | 32 | 33 | 80 | 936 |
| Crustacean | 915 | 2,897 | 14,466 | 2,114 | 2,269 | 4,515 | 27,176 |
| Fish | 2,936 | 3,422 | 31,598 | 1,432 | 5,903 | 5,222 | 50,513 |
| Insect | 249 | 279 | 6,316 | 347 | 358 | 709 | 8,258 |
| Mollusca | 208 | 319 | 2,596 | 185 | 295 | 650 | 4,253 |
| Others | 88 | 547 | 787 | 1,528 | 130 | 766 | 3,846 |
| Plant | 708 | 728 | 945 | 889 | 526 | 912 | 4,708 |
| Rotifera | 5 | 270 | 288 | 235 | 7 | 365 | 1,170 |
| Total | 5,820 | 11,318 | 60,309 | 13,114 | 10,293 | 17,589 | 118,443 |

values are calculated for 2,350 chemicals using chronic EC10 equivalents as the underlying effect data according to Eq. 1-8, i.e the methodology of (Owsianiak et al., 2023), while probabilistic uncertainty in the value is assessed for only 684 chemicals due to the large number of data required for the nls model to start and converge.

### Uncertainty estimations

By fitting a weighted nls model to a cumulative normal distribution of ecotoxicological effect data, 1,134 chemicals have enough data to calculate , albeit 450 of these chemicals are flagged with “Data insufficient”, implying that the SSD model is calculated using either too few species or taxonomic groups and generally yield high uncertainties (Figure 3). 1,563 chemicals in the dataset contain effect data for only one species or one taxonomic group and are not run in the model. Since the weighting of the model depends on the standard error, species-specific data with only one represented effect value (e.g no mean or standard error present) are also discarded and flagged as “One sigma or fewer”. The results from applying the nls model across the entire dataset is summarized in figure 2.

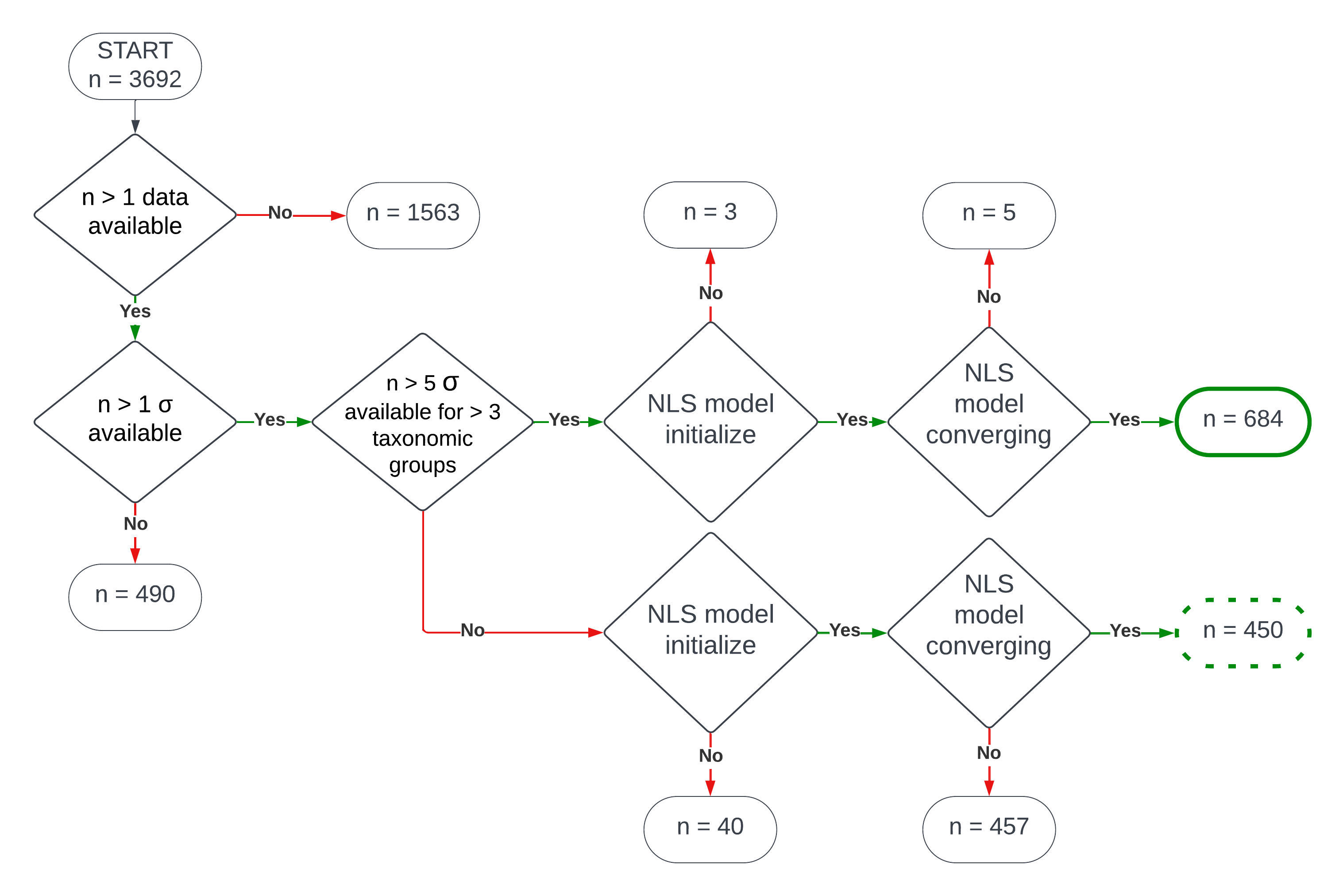


Figure 2: Decision tree for data selection and status of model results per chemical. NLS = nonlinear least squares model. n = number of chemicals. Dashed lines imply that results are based on cases with <5 species across <3 taxonomic groups.

A difficulty of the nls model fitting is to define starting parameters for the model that are suitable across the entire dataset. Each chemical data point should ideally be inspected and have starting parameters for the nls model adapted for each case individually, but this is a tedious optimization process and lies outside the scope of this study. 43 chemicals failed to initialize because of sub-optimal starting values, and the model failed to converge for 462 chemicals because the step factor was reduced below the model’s automatic cut-off value, implying that, in total, 13.7 % of chemicals within the dataset failed to produce uncertainty estimates.

We identified 8 chemicals that fulfilled the data requirements of >5 species across >3 taxonomic groups, but failed to either start or converge. For all of these chemicals, effect data are spanning several orders of magnitude across both negative an positive values, due to the log-transformation of effect data (data not shown). We tried resolving this issue by adjusting the unit measurement of by a factor of and change the measured concentrations to , resulting in “problematic” concentrations being . However,this did not resolve starting or convergence issues. Some chemicals are seemingly not open for analysis using the current weighted nls method, or choice of starting parameters, for reasons unknown.

Limited data availability is an issue here, and only 18.5% of all chemicals have enough data to fit the nonlinear least squares model (Table 3). The nls modelling approach allows estimations of uncertainties in the value, but requires a substantial amount of effect data input for the model to run optimally.

Table 3: Summary overview of the nonlinear least square model fit for the HESTIA ecotoxicological database.

| Output status | Number of records | Minimum number of species per chemical | Maximum number of species per chemical | Minimum number of taxonomic groups per chemical | Maximum number of taxonomic groups per chemical |
| --- | --- | --- | --- | --- | --- |
| Convergence | 684 | 5 | 402 | 3 | 10 |
| Convergence; Data insufficient | 450 | 3 | 26 | 2 | 6 |
| Fail.Conv; data insufficient | 457 | 2 | 13 | 2 | 6 |
| Fail.Conv; data present | 5 | 6 | 10 | 3 | 5 |
| Fail.Init; data insufficient | 40 | 2 | 12 | 2 | 5 |
| Fail.Init; data present | 3 | 7 | 11 | 3 | 5 |
| Not enough data | 1,563 | 1 | 9 | 1 | 1 |
| One sigma or fewer | 490 | 2 | 14 | 2 | 5 |

In order to compare the probability distribution in the value across all chemicals, we calculate the geometric standard deviation for an artificial vector containing 100,000 values generated using Monte Carlo simulations. The geometric standard deviation is decreasing as the number of effect data increases per chemical, and when considering chemicals that fulfill the data requirements of 5 species across 3 taxonomic groups, the highest geometric standard deviation belong to chemicals classified as pesticides, with a range up to 6.5 (Figure 3). 48 of the top 50 chemicals with highest geometric standard deviation of the probability distribution in the value belongs to pesticides. Chemicals classified into other use categories have far lower geometric standard deviations.

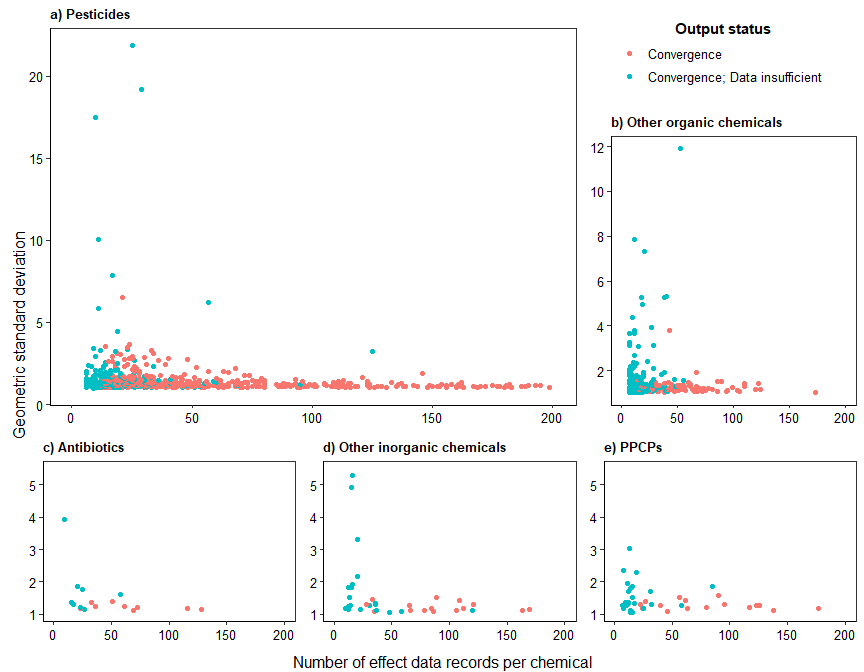


Figure 3: Geometric standard deviation of probability distribution ranges which are calculated from 100,000 Monte Carlo simulations for chemicals defined as a) Pesticides, b) Other organic chemicals, c) Antibiotics, d) Other inorganic chemicals, and e) Pharmaceuticals and personal care products (PPCPs). Chemicals with more than 200 effect data records not shown, since geometric standard deviation is <1.4 with 200+ records present. Convergence = The model successfully identified a global minimum sum of squares. Data insufficient = chemicals do not fulfill the data requirements of = 5 species across = 3 taxonomic groups.

### Discussion

It is enabling uncertainty calculations of toxicological data

This study has investigated a set of chemicals limited to the HESTIA inventory, listing 16,797 chemicals, which is far from a fully comprehensive database for all potentially harmful agrochemical substances. Moreover, ecotoxicological effect data are available for 5,027 chemicals (30%) of the 16,797 queried chemicals. After curating the data, 3,692 chemicals have reliable effect data, of which 638 chemicals only have one record with reliable effect data. While the focus of this study is not to gather data on as many chemicals as possible, the numbers do illustrate the omnipresent problem of ecotoxicological characterization: limited data availability. Since uncertainty in the decrease as underlying effect data availability increases, we did investigate other potential sources for additional toxicological effect data and found the EnviroTox database (Connors et al., 2019) and the EU environmental footprint v3.1 (EF3.1) database. In October 2022, Sala et al. (2022) published EF3.1 (<https://eplca.jrc.ec.europa.eu/ecotox.html>) containing values for 6,038 chemicals calculated from toxicological effect data gathered from the ECHA REACH database (European Chemicals Agency, <http://echa.europa.eu/>). The EF3.1 contains 140 % more characterized chemicals than the, now legacy, USEtox v2.1 organic substance database (Sala et al., 2022; Saouter et al., 2019). However, a substantial part of the toxicological effect data in EF3.1 are marked as proprietary substance registrations within REACH, and cannot be shared (S. Sala, personal communication, 14 Nov 2022). This is unfortunate, since one of the 19 key recommendations of the Ecotoxicity Task Force and the Pellston workshop held in Valencia, Spain in 2018 is to “use data that has a traceable origin” (Owsianiak et al., 2019). Toxicological data contained within the EnviroTox database overlap to a large extent with the data sources of the present database and, after inspection, only >1 % of records (data not shown) in the EnviroTox database could be added without the risk of introducing duplicate data, which prompted us to refrain from including data from this source.

Finding a good balance in maximizing data inclusion and sorting out improper toxicological effect records is difficult. During data curation we removed 58 % of the original records. The step of removing improper effect criterion descriptions of effect data excluded 27.4 % of the starting dataset (step 4 in Figure 1). There are surprisingly little guidance available in the literature on this data selection; relevant effect criterions are mentioned only briefly in Saouter et al. (2019) and Sala et al. (2022). We note that for several toxicological effect records, taxonomic details are given at levels above species. When calculating , species-specific effect values (Eq. 5), records defined at genus level without a species name will influence the calculation by representing an artificial species. For example *Daphnia sp.* will count as one species alongside *D. magna*. Taxonomic identifiers above species-level have been removed from this dataset resulting in the exclusion of 803 taxonomic definitions (not counting common names or erroneous names). However, records defined at genus level are present in the taxonomic records of the EF3.1 with no comment on how these data are treated (Sala et al., 2022; Saouter et al., 2018). While data scarcity is an issue, we argue that imposing an artificial data point is an incorrect methodological choice.

The present database contains data at both acute and chronic effect data from EC50, EC10 and NOEC endpoints (and their respective analogues, like LC50, EC0, LOEC etc.), which have been extrapolated using regression coefficients from Aurisano et al. (2019). A substantial part of the effect data (51 %) are extrapolated from Acute EC50 effect data to Chronic EC10eq where large confidence interval ranges are present (e.g., the regression coefficient for fish is estimated to 7.44, with a 95% confidence interval 2.92-18.95; see SI Table 6), and these confidence intervals have not been considered when constructing SSD models. It can be argued that with sufficient numbers of records, a large variation in one of the data points included in the species-specific average is negligible. How much this uncertainty influences the final stands to be investigated in future studies. A problem for probabilistic uncertainty assessment of pesticides is the large range of effects between target and non-target organisms, due to the specific mode of action (e.g. insecticides are predominantly toxic for insects) (Warren-Hicks and Hart, 2010). With a distribution of effect data at two extremes, we can assume that pesticides will have large uncertainties in the value, which, indeed, is the case for the insecticide “Novaluron” and herbicide “Pyroxsulam”, representing the highest geometric standard deviation in the dataset (5.1 and 4.8 respectively), despite fulfilling data requisites of 5 species across 3 taxonomic groups (figure 3**a)**).

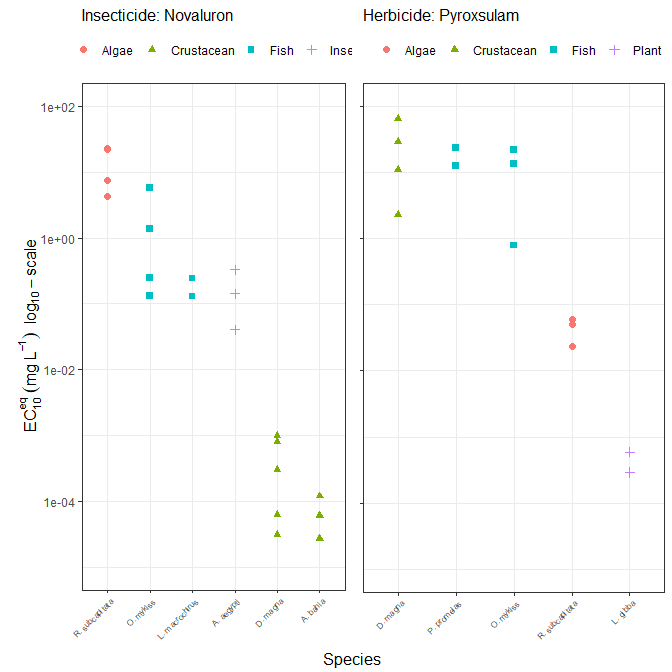


Figure 4: Examples of how specific toxic mode of action creates large uncertainty in the CRF. Target organism groups are displaying a response at low concentrations to the insecticide Novaluron and the herbicide Pyroxulam respectively, while non-target organism groups show effect at several orders of magnitude higher concentrations. Dots represent individual toxicological records at a converted EC10eq concentration.

The role of uncertainty in ecotoxicological characterizations of chemicals varies across studies, for broad LCAs covering hundreds or thousands of chemicals, for example a complex agri-food system with many processes aggregated into one functional unit, large uncertainty in one of the characterizations will be masked by the sheer amount of data. On the other hand, for niche LCAs that focus on a single process or non-LCAs (ERAs) that focus on a single chemical, uncertainty in the toxicological characterization will be very relevant and the effort of estimating uncertainty can be highly motivated. Applying the proposed method by using nonlinear least squares and estimating the probabilistic ranges in the is now easily achieved.

Uncertainties in the value are not present only due to variability within each point of the SSD curve, in particular, many pesticides have an inherently variable effect to different organisms due to their specific toxic mode of action (Van Zelm et al., 2009b). Uncertainty is attached to the output from all steps in the LCIA, including the models used to estimate chemical fate and chemical exposure. Fate models have recently been investigated for uncertainties by Nemecek et al. (2022), and uncertainty distributions for human toxicity impacts have been investigated by Aurisano et al. (2023), while other parameters connected to exposure of chemicals, cocktail effects, chemical application methodology still needs to be investigated. Unfortunately, none of the current LCIA software have the possibility of including uncertainty of the ecotoxicological characterizations. As data on uncertainties becomes more available across several of the parameters included in LCIAs, we highly recommend LCIA software developers to engage in the current developments of uncertainty estimations of data used, and work towards including such parameters in updated versions of their respective software.

**Future work**  
Several ideas are queued for further investigations based on the present dataset. For example, how much does the endpoint conversions using regression coefficients affect the final value? A robust methodology to include the variability of each endpoint conversion could elucidate how feasible such data conversion really are. Moreover, since data scarcity is an issue, we are commencing an investigation on the the potential to include quantitative structure-activity relationship models, investigate the reliability of these models and evaluate if this data is suitable albeit potentially large uncertainties. Finally, many LCA studies are published despite missing toxicological characterizations for some chemicals in the LCIA step, leading to potentially underestimated toxicological impacts. We aim to investigate the data gaps within a large set of agi-food LCA studies, hopefully filling these gaps and assessing the updated outcomes.

**Conclusions**  
Variability in toxicological effect estimates can be included in calculations of toxicological characterizations of chemicals and inform us on the variability in the calculated ecotoxicological effect factor, but data scarcity is limiting the number of chemicals that can be properly evaluated. Not only the number of effect records per chemical is a limiting factor, but the number of different species and number of effect data per species will affect the possibility to use the suggested nonlinear least squares model utilizing weighted means of species-specific effect estimations.

**Author contributions**  
Conceptualization: PJGH, ON; Data curation: ON; Formal analysis:ON, RH; Investigation: ON; Methodology: ON, PJGH, RH; Software: ON; Project administration: PJGH, ON; Supervision: PJGH; Visualization: ON; Writing - original draft: ON; Writing - review & editing: ON, PJGH, RH.

**Acknowledgements**  
The authors want to thank Andreu Rico, PhD, for input and discussions on ecologically relevant effect criterion selection, and Rachel Foster for her comments to the manuscript.

**Referees**

* David Volz - Editorial Board of Chemosphere
* Bingsheng Zhou - Editorial Board of Chemosphere
* Mikołaj Owsianiak -
* Peter Fankte - Email: [pefan@dtu.dk](mailto:pefan@dtu.dk)
* Leo Posthuma -
* SUggestions?? -

*With the submitted manuscript authors are requested to provide full contact details of six potential reviewers including email addresses. Please suggest potential reviewers for this submission and provide specific reasons for your suggestion in the comments box for each person. Please note that the editorial office may not use your suggestions, but your help is appreciated and may speed up the selection of appropriate reviewers. Of the six potential reviewers, please suggest two members of the Editorial Board (but NOT Editors) of Chemosphere as reviewer candidates. The other suggested reviewers should not be from the same institution as the author, or co-authors/collaborators on any books, articles, reports, papers, or projects of the author. Not more than one should come from the same country as the author. It should also be avoided to suggest referees that are living in a different country but have the same nationality as the author. Please provide institutional email addresses only; if this is not possible then the potential reviewer’s institution should be clearly stated.*

### References

Aurisano, N., Albizzati, P.F., Hauschild, M., Fantke, P., 2019. Extrapolation Factors for Characterizing Freshwater Ecotoxicity Effects. Environmental Toxicology and Chemistry 38, 2568–2582. <https://doi.org/10.1002/etc.4564>

Aurisano, N., Jolliet, O., Chiu, W.A., Judson, R., Jang, S., Unnikrishnan, A., Kosnik, M.B., Fantke, P., 2023. Probabilistic Points of Departure and Reference Doses for Characterizing Human Noncancer and Developmental/Reproductive Effects for 10,145 Chemicals. Environmental Health Perspectives 131, 037016. <https://doi.org/10.1289/EHP11524>

Bohnes, F.A., Hauschild, M.Z., Schlundt, J., Laurent, A., 2019. Life cycle assessments of aquaculture systems: A critical review of reported findings with recommendations for policy and system development. Reviews in Aquaculture 11, 1061–1079.

Brömssen, C. von, Röös, E., 2020. Why statistical testing and confidence intervals should not be used in comparative life cycle assessments based on monte carlo simulations. The International Journal of Life Cycle Assessment 25, 2101–2105.

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., Borges, B., 2023. [Shiny: Web application framework for r](https://shiny.rstudio.com/).

Connors, K.A., Beasley, A., Barron, M.G., Belanger, S.E., Bonnell, M., Brill, J.L., Zwart, D. de, Kienzler, A., Krailler, J., Otter, R., Phillips, J.L., Embry, M.R., 2019. Creation of a Curated Aquatic Toxicology Database: EnviroTox. Environmental Toxicology and Chemistry 38, 1062–1073. <https://doi.org/10.1002/etc.4382>

Der Werf, H.M. van, Knudsen, M.T., Cederberg, C., 2020. Towards better representation of organic agriculture in life cycle assessment. Nature Sustainability 3, 419–425.

Dimitrov, S., Diderich, R., Sobanski, T., Pavlov, T., Chankov, G., Chapkanov, A., Karakolev, Y., Temelkov, S., Vasilev, R., Gerova, K., others, 2016. QSAR toolbox–workflow and major functionalities. SAR and QSAR in Environmental Research 27, 203–219.

Fantke, P., Bijster, M., Guignard, C., Hauschild, M., Huijbregts, M., Jolliet, O., Kounina, A., Magaud, V., Margni, M., McKone, T., Posthuma, L., Rosenbaum, R.K., Meent, D. van de, Zelm, R. van, 2017. USEtox 2.0, documentation version 1. <https://doi.org/10.11581/DTU:00000011>

FAO, 2023. Pesticides use (FAOSTAT, accessed july 2023) [WWW Document].

Gordon, L.J., Bignet, V., Crona, B., Henriksson, P.J., Van Holt, T., Jonell, M., Lindahl, T., Troell, M., Barthel, S., Deutsch, L., others, 2017. Rewiring food systems to enhance human health and biosphere stewardship. Environmental Research Letters 12, 100201.

Hauschild, M.Z., Huijbregts, M.A., 2015. Introducing life cycle impact assessment. Springer.

Heijungs, R., 2021. The average versus marginal debate in LCIA: Paradigm regained. The International Journal of Life Cycle Assessment 26, 22–25. <https://doi.org/10.1007/s11367-020-01835-4>

Henriksson, P.J., Rico, A., Zhang, W., Ahmad-Al-Nahid, S., Newton, R., Phan, L.T., Zhang, Z., Jaithiang, J., Dao, H.M., Phu, T.M., others, 2015. Comparison of asian aquaculture products by use of statistically supported life cycle assessment. Environmental Science & Technology 49, 14176–14183.

Huijbregts, M.A.J., Hellweg, S., Hertwich, E., 2011. Do We Need a Paradigm Shift in Life Cycle Impact Assessment? Environmental Science & Technology 45, 3833–3834. <https://doi.org/10.1021/es200918b>

Huijbregts, M.A., Thissen, U., Guinée, J., Jager, T., Kalf, D., Van de Meent, D., Ragas, A., Sleeswijk, A.W., Reijnders, L., 2000. Priority assessment of toxic substances in life cycle assessment. Part i: Calculation of toxicity potentials for 181 substances with the nested multi-media fate, exposure and effects model USES–LCA. Chemosphere 41, 541–573.

Krewski, D., Acosta Jr, D., Andersen, M., Anderson, H., Bailar III, J.C., Boekelheide, K., Brent, R., Charnley, G., Cheung, V.G., Green Jr, S., others, 2010. Toxicity testing in the 21st century: A vision and a strategy. Journal of Toxicology and Environmental Health, Part B 13, 51–138.

Maggi, F., Tang, F.H.M., Tubiello, F.N., 2023. Agricultural pesticide land budget and river discharge to oceans. Nature Advance online publication. <https://doi.org/10.1038/s41586-023-06296-x>

Malode, S., Prakash, R., Mohanta, J.C., 2023. A life cycle assessment of coal-fired thermal power plants with post-combustion control techniques: An india scenario. Environmental Science and Pollution Research 1–17.

Millard, S.P., 2013. EnvStats: An r package for environmental statistics. Springer.

Nemecek, T., Antón, A., Basset-Mens, C., Gentil-Sergent, C., Renaud-Gentié, C., Melero, C., Naviaux, P., Peña, N., Roux, P., Fantke, P., 2022. Operationalising emission and toxicity modelling of pesticides in LCA: The OLCA-pest project contribution. The International Journal of Life Cycle Assessment 27, 527–542.

Owsianiak, M., Fantke, P., Posthuma, L., Saouter, E., Vijver, M., Backhaus, T., Schlekat, C., Hauschild, M.Z., 2019. Ecotoxicity, in: Frischknecht, R., Jolliet, O. (Eds.), Global Guidance on Environmental Life Cycle Impact Assessment. UNEP/SETAC Life Cycle Initiative, Paris, France, pp. 138–172.

Owsianiak, M., Hauschild, M.Z., Posthuma, L., Saouter, E., Vijver, M.G., Backhaus, T., Douziech, M., Schlekat, T., Fantke, P., 2023. Ecotoxicity characterization of chemicals: Global recommendations and implementation in USEtox. Chemosphere 310, 136807. https://doi.org/<https://doi.org/10.1016/j.chemosphere.2022.136807>

Pennington, D.W., Margni, M., Ammann, C., Jolliet, O., 2005. Multimedia fate and human intake modeling: Spatial versus nonspatial insights for chemical emissions in western europe. Environmental science & technology 39, 1119–1128.

Persson, L., Carney Almroth, B.M., Collins, C.D., Cornell, S., De Wit, C.A., Diamond, M.L., Fantke, P., Hassellöv, M., MacLeod, M., Ryberg, M.W., others, 2022. Outside the safe operating space of the planetary boundary for novel entities. Environmental science & technology 56, 1510–1521.

Posthuma, L., Gils, J. van, Zijp, M.C., Meent, D. van de, Zwart, D. de, 2019. Species sensitivity distributions for use in environmental protection, assessment, and management of aquatic ecosystems for 12 386 chemicals. Environmental Toxicology and Chemistry 38, 905–917. <https://doi.org/10.1002/etc.4373>

Posthuma, L., Suter II, G.W., Traas, T.P., 2001. Species sensitivity distributions in ecotoxicology. CRC press.

R Core Team, 2023. [R: A language and environment for statistical computing](https://www.R-project.org). R Foundation for Statistical Computing, Vienna, Austria.

Sala, S., Biganzoli, F., Mengual, E.S., Saouter, E., 2022. Toxicity impacts in the environmental footprint method: Calculation principles. The International Journal of Life Cycle Assessment 27, 587–602. <https://doi.org/10.1007/s11367-022-02033-0>

Saouter, E., Biganzoli, F., Ceriani, L., Versteeg, D., Crenna, E., Zampori, L., Sala, S., Pant, R., 2018. Environmental footprint: Update of Life Cycle Impact Assessment Methods : Ecotoxicity freshwater, human toxicity cancer, and non-cancer. Publications Office of the European Union, Luxembourg.

Saouter, E., Biganzoli, F., Pant, R., Sala, S., Versteeg, D., 2019. Using REACH for the EU Environmental Footprint: Building a Usable Ecotoxicity Database, Part I. Integrated Environmental Assessment and Management 15, 783–795. <https://doi.org/10.1002/ieam.4168>

Schreinemachers, P., Tipraqsa, P., 2012. Agricultural pesticides and land use intensification in high, middle and low income countries. Food policy 37, 616–626.

Simpson, A.M.-A., Mitch, W.A., 2022. Chlorine and ozone disinfection and disinfection byproducts in postharvest food processing facilities: A review. Critical Reviews in Environmental Science and Technology 52, 1825–1867.

Standardization, I.O. for, 2006. Environmental management: Life cycle assessment; requirements and guidelines. ISO Geneva, Switzerland.

Szöcs, E., Stirling, T., Scott, E.R., Scharmüller, A., Schäfer, R.B., 2020. webchem: An R package to retrieve chemical information from the Web. Journal of Statistical Software 93, 1–17. <https://doi.org/10.18637/jss.v093.i13>

Van Zelm, R., Huijbregts, M.A., De Meent, D. van, 2009a. USES-LCA 2.0—a global nested multi-media fate, exposure, and effects model.

Van Zelm, R., Huijbregts, M.A., Harbers, J.V., Wintersen, A., Struijs, J., Posthuma, L., Van de Meent, D., 2007. Uncertainty in msPAF-based ecotoxicological effect factors for freshwater ecosystems in life cycle impact assessment. Integrated Environmental Assessment and Management: An International Journal 3, 203–210.

Van Zelm, R., Huijbregts, M.A., Posthuma, L., Wintersen, A., Meent, D. van de, 2009b. Pesticide ecotoxicological effect factors and their uncertainties for freshwater ecosystems. The International Journal of Life Cycle Assessment 14, 43–51.

Warren-Hicks, W.J., Hart, A., 2010. Application of uncertainty analysis to ecological risks of pesticides. CRC Press.