

Letter to the Editor

Reply to “Concerns About Reproducibility, Use of the Akaike Information Criterion, and Related Issues in Hoondert et al. 2019” and Focus in Developing QSAR-Based Species Sensitivity Distributions

Renske P.J. Hoondert,^{a,b} Rik Oldenkamp,^b Dick de Zwart,^c Dik van de Meent,^{b,c} and Leo Posthuma^{a,b}

^aRIVM, Centre for Sustainability, Environment and Health, Bilthoven, The Netherlands

^bDepartment of Environmental Sciences, Faculty of Science, Radboud University Nijmegen, Nijmegen, The Netherlands

^cARES, Odijk, The Netherlands

Authors' Response:

Our exploratory study on the quantitative structure–activity relationship (QSAR)-based species sensitivity distributions (SSDs)—which boils down to directly deriving SSDs for untested compounds from those of tested compounds (Hoondert et al. 2019)—has been appreciated by Iwasaki and Hayashi (2020). However, they also voice concerns about the statistics used in the study, particularly concerning the improper use of statistical methods and parameters in selecting appropriate models and the reproducibility of outcomes. Herewith we provide our reply to the comments, hoping to clarify our methods and reemphasize the higher aim of the study, which is to help forward this field to maturity.

ON NEEDS AND WIDER

There is a societal need to handle >350 000 chemicals and their mixtures, protecting environmental quality and human health when possible and restoring them when needed (Wang et al. 2020). In their comments, Iwasaki and Hayashi first mention the derivation of protective standards as a key goal of ecological risk assessment, hinting at the need for QSARs to bridge data gaps in the light of limited ecotoxicity data. To be clear, we argue that the utility of the methods we proposed is wider, with ecotoxicity data serving in established operational methods for environmental protection, life cycle impact assessment, and environmental pollution assessment and management (Hoondert et al. 2019; Posthuma et al. 2019). We also foresee opportunities to support the development of sustainable chemistry methods (“benign by design” chemicals; Blum et al. 2017). Furthermore, looking at, for example, the US Environmental Protection Agency's (2020) ecotoxicity database alone, these methods can be based on nearly a million

ecotoxicity endpoint data, covering tests for >12 000 chemicals, >13 000 species, representing >50 000 references. Wider policy needs and practices can be based on many data.

OUR METHOD

The original article to which Iwasaki and Hayashi's letter is referring can be summarized by its graphical abstract (Figure 1), which captures the societal need; the use of SSDs in environmental protection, assessment, and management; and the QSAR-based SSD idea as an option to explore the ecotoxicity of all chemicals, even when data for most (>350 000 – ≈12 000) are lacking.

STATISTICAL COMMENTS

Supporting the principles and outcomes of our study, the Iwasaki and Hayashi letter voiced concerns about reproducibility and about the model selection procedure followed. Regarding the reproducibility, there is likely a misunderstanding. Where Iwasaki and Hayashi identify alternative models with our data, all with fewer descriptors, we explicitly derived models that should involve an abundance of descriptors instead of solely being derived as statistically the best model. We wanted to show how different descriptors play a role and can be ranked, anticipating that future more precise models would likely be composed of multiple, potentially mechanistically relevant predictors. For this reply, we made some additional calculations and provide the R-script for that as Supplemental Data to support reproduction of the current findings. We think this solves the key issue of reproducibility.

Iwasaki and Hayashi's further concerns relate to the use of statistical entities and their interpretation in model selection, arguing that 1) if the (corrected) Akaike Information Criterion (AIC[c]) is used in determining the “best” model, then the (adjusted) R^2 should not also be used for model selection, and 2) even if the AIC(c) is used correctly to determine the “best” model, final conclusions on the importance of predictors should not be based on this “best” model alone.

This article includes online-only Supplemental Data.
Published online 22 June 2020 in Wiley Online Library
(wileyonlinelibrary.com).
DOI: 10.1002/etc.4737

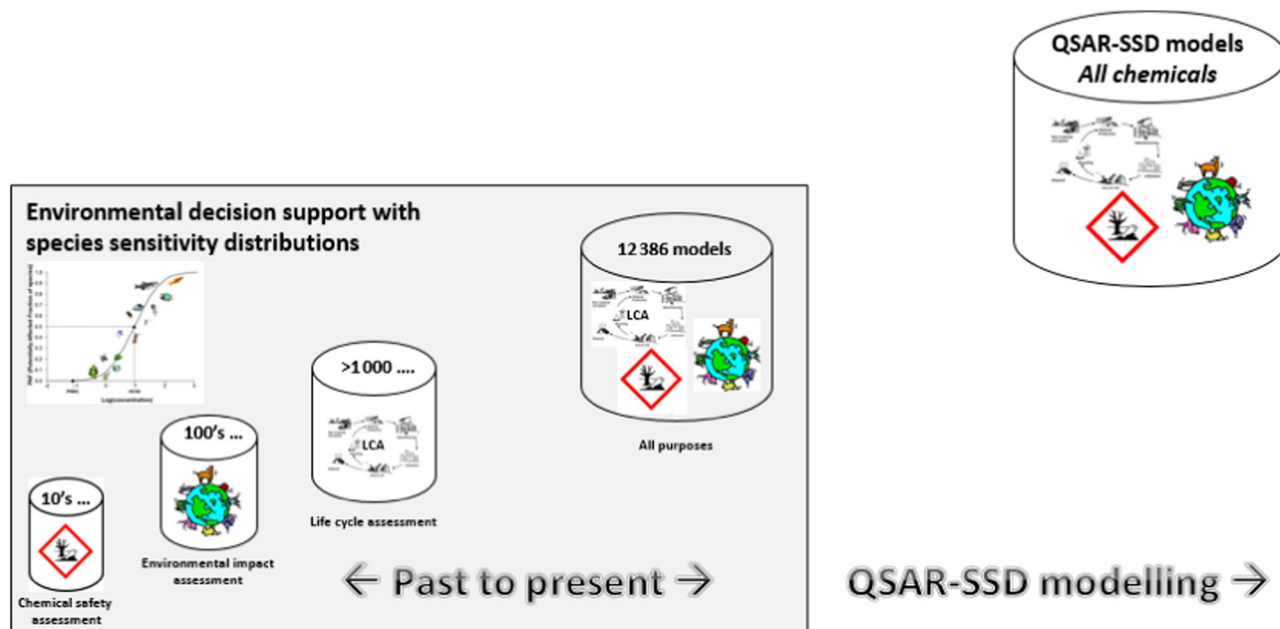


FIGURE 1: Utility of species sensitivity distributions for environmental protection, assessment, and management purposes and numbers of compounds currently covered for various applications, while there are needs to address >350 000 chemicals and their mixtures. Source: graphical abstract of Hoondert et al. (2019). LCA=life cycle assessment; SSD=species sensitivity distribution; QSAR=quantitative structure–activity relationship.

The first point has likely arisen from our text: “Then, the most parsimonious models for μ and σ were selected using the dredge function in R statistics, Ver. 3.5.1, based on the corrected Akaike information criterion (AIC) as well as the adjusted R^2 .” In combination with the misunderstanding of the reproducibility argument, we understand that this might have created ambiguity about the procedure followed and how we valued statistical and other arguments. First, we fully agree that simultaneous use of the AIC and adjusted R^2 in model selection is not good. Because the AIC represents a trade-off between a model's complexity and its maximum likelihood, it already implicitly encompasses a metric of a model's goodness of fit. Second, the Iwasaki and Hayashi letter also points out that AIC values cannot be compared between models developed with different response variables. We followed this principle: the candidate models in our study that we compared were solely “nested” models for the 4 separate outcome models, and selection of our 4 models was done separately and independently.

Within each “nested” model group, we used the dredge function in R (Ver 3.5.1.) to create a list of candidate models with their associated AIC(c). The relative empirical support for each candidate model is represented by the absolute difference (delta) between its own AIC(c) and the minimum of all AIC(c) values. As Iwasaki and Hayashi mention, to distinguish between models with and without empirical support, deltas might be compared to a threshold value, typically 10. As such, if a single candidate model has a delta below that threshold, we should prefer it. In our study, multiple candidate models pertaining to SSD-median effect concentration (EC50)- μ , SSD-EC50- σ had delta values (in the nested approach) below 10 (with the value of 4.44 [and 4.38, respectively] as shown in

Table 1 of Iwasaki and Hayashi). When comparing plots pertaining to all candidate models with delta values below 10 for SSD-EC50- μ as also used in the letter, we observe a similar degree of scattering across all plots (see Supplemental Data). In addition, the potential to predict ecotoxicity for untested compounds (Q^2 or $R^2_{\text{predicted}}$) did not considerably differ among the 12 candidate models shown (0.39–0.431). Delta values for our SSD-no-observed-effect concentration (NOEC)- μ , SSD-NOEC- σ models, however, were much higher (21.07 and 37.07); and although we did not provide delta values originally, we already warned of the limitations of the initial models, especially for the NOEC-based models. We agreed and still agree with Iwasaki and Hayashi that model derivation and selection need further work, especially given the high delta values pertaining to the NOEC models. In addition, we agree that in future correspondence related to this research, we should be more transparent in describing our methodology.

This leaves the point that multiple candidate models can have low delta values and thus statistical support. Here, one may argue that it is illogical to select the model with the lowest AIC(c) because it is not the only one with a reasonable claim to effectively describe the data structure. One way of dealing with multiple plausible models is to base predictions on all of these models combined, that is, through averaging model predictions or regression coefficients (Burnham et al. 2011). Instead, we chose one of the plausible candidate models, based on a complete representation of the descriptors. That original choice was made in line with the higher aim of our modeling exercise and the predictive context in which the selected models are to be applied: exploring optimal ways to fill the void of acute and chronic ecotoxicity data that exists for >340 000 chemicals. As expected, looking ahead, future

models would likely be described by a moderately sized set of well-selected predictors; we thus used (future) utility criteria in presenting outcome models. We thus agree that model selection should be based on good statistical procedures but add to that the need to consider (future) utility.

This brings us to the next point, the descriptors and their relative importance, as also mentioned in the Iwasaki and Hayashi letter. We fully agree that such an assessment would have substantial value. However, the aim of this short communication was primarily to explore the possibility of predicting SSD parameters for substances for which ecotoxicity data are lacking, based initially on some easy-to-obtain molecular descriptors. The working hypothesis put forward in the Iwasaki and Hayashi letter, and by us earlier, would ask for selecting molecular descriptors that make mechanistic-toxicological sense; and those do not necessarily need to be easy to obtain. Furthermore, using molecular descriptors that are not easy to measure as predictors in our models may also increase the risk of introducing a bias in model training because these chemical data may only be available for highly toxic substances (because those trigger the most research). Consequently, applying the model to a (larger) data set may lead to model extrapolation beyond its applicability domain. This may already be the case in our models, in which the distinction between training and test data is based on the number of ecotoxicity data per compound, providing the further basis we mentioned not to overinterpret the models derived so far.

In our study, we so far aimed to pay less attention to the relative importance of specific individual predictors of the currently studied set, but we agree (and indeed discussed) that a next step could encompass this idea, especially given the upcoming attention to safe chemical design purposes (Geiser 2015). We are happy that the Iwasaki and Hayashi letter emphasizes the need for this additional debate, which highlights some teasing aspects of QSAR-based SSDs for important novel purposes.

NEXT STEPS

Our short communication was mainly aimed at exploring a novel idea, that is, the derivation of QSAR-based SSDs for nontested compounds, which is the majority of compounds. We already discussed in the short communication that this novel idea can certainly be developed further, with improved approaches in any step (in collecting/curating raw data, the derivation of μ and σ values, the choice of molecular descriptors, the statistics, the final model selection, and finally responsible use). A logical expansion would consider mechanism-related descriptors that take into account the importance of each individual predictor. In addition, future models could include species traits, yielding SSDs for separate taxonomic groups, because certain molecular descriptors affect

certain traits in particular. However, for now, in the original short communication, we generally focused on applicability rather than (solely) statistical arguments. We went for an approach aiming to predict acute and chronic ecotoxicity for substances lacking these data, initially exploring whether we can fulfill the first question asked: Would it work?

The answer is, yes. The idea of QSAR-based SSDs can be operationalized. This result is the stepping stone for answering next questions, to address pressing societal concerns about the overwhelming numbers of chemicals (Wang et al. 2020).

CONCLUSION

We reiterate that the aim of our short communication was to explore the novel idea of predicting SSD parameters based on molecular descriptors, to get an initial grip on assemblage-level threats of untested chemicals. Overall, the concerns in the Iwasaki and Hayashi letter would not affect the general conclusions of the article, as confirmed by Iwasaki and Hayashi. We do agree that some components of the modeling exercise can be improved in future steps, as stated in the discussion section of our article and in the Iwasaki and Hayashi letter. We therefore highly value the views of Iwasaki and Hayashi in terms of model selection and keep his comments and other important issues in mind to improve model selection in further steps. We look forward to swift developments of this field.

Supplemental Data—The Supplemental Data are available on the Wiley Online Library at <https://doi.org/10.1002/etc.4737>.

REFERENCES

- Blum C, Bunke D, Hungsberg M, Roelofs E, Joas A, Joas R, Blepp M, Stolzenberg HC. 2017. The concept of sustainable chemistry: Key drivers for the transition towards sustainable development. *Sustain Chem Pharm* 5:94–104.
- Burnham KP, Anderson DR, Huyvaert KP. 2011. AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. *Behav Ecol Sociobiol* 65:23–35.
- Geiser K. 2015. *Chemicals Without Harm: Policies for a Sustainable World*. MIT Press, Cambridge, MA, USA.
- Hoondert RP, Oldenkamp R, de Zwart D, van de Meent D, Posthuma L. 2019. QSAR-based estimation of species sensitivity distribution parameters: An exploratory investigation. *Environ Toxicol Chem* 38:2764–2770.
- Iwasaki Y, Hayashi TI. 2020. To the Editor: Concerns about Reproducibility, Use of the Akaike Information Criterion, and Related Issues in Hoondert et al. 2019. *Environ Toxicol Chem* 39:1300–1301.
- Posthuma L, van Gils J, Zijp MC, van de Meent D, de Zwart D. 2019. Species sensitivity distributions for use in environmental protection, assessment, and management of aquatic ecosystems for 12 386 chemicals. *Environ Toxicol Chem* 38:905–917.
- US Environmental Protection Agency. 2020. Ecotoxicity database. Washington, DC. [cited 2020 March 30]. Available from: <https://cfpub.epa.gov/ecotox/stats.cfm?reldate=20200312>
- Wang Z, Walker GW, Muir DCG, Nagatani-Yoshida K. 2020. Toward a global understanding of chemical pollution: A first comprehensive analysis of national and regional chemical inventories. *Environ Sci Technol* 54:2575–2584.