Here is my review of the paper “Machine learning models based on molecular descriptors to predict human and environmental toxicological factors in continental freshwater”. I would like to highlight that I am specialised in ecological modelling (development and analysis of IBM, state space models, ODE based models) and Bayesian statistics. My fields of research are movement ecology, forest ecology and ecotoxicology. I am not a specialist of: chemistry, toxicology, risk assessment, machine learning.

This study aims aims at predicting characterization factors of chemicals based on molecular descriptors using statistical models. Those predictions would complement experimental approaches in order to decrease experimental costs. The intersection of USEtox database comprising CFs and TyPol database comprising molecular descriptors made available the variables and the predictors for 274 chemicals. The authors tested three different models predicting CFs from molecular descriptors, one linear model (PLS) and two machine learning based models (RF and SVM). They also tested whether a clustering step of the chemicals before applying the different statistical models results in better predictive performance. Therefore six models were tested in total. The choice of the “best model” relied on the absolute error between the prediction and the true value. Then missing CFs (for some chemicals of the database those were not available) were predicted with the “best model”. Also, the five variables (ie molecular descriptors) that contributed the most to the predictions were identified. Overall, the clustering step improved the prediction performance for one variable and absolute errors were smaller with the machine learning models than the linear model for both. The “best models” lead to acceptable predictions.

Overall the paper is well written and easy to understand. It is actually a useful study for the ecotoxicology community, since it shows that CFs can be predicted from the molecular descriptors stored in TyPol database with an acceptable error, using a rather easy method (clustering then machine learning regression models, or only machine learning regression models, depending on the predicted variable).

Here is my general comments on the paper sections, a more detailed list of modifications I suggest follows. The introduction is clear and exposes well the context, motivation and interest of the study. The method section is clear enough, except for a few paragraphs. The result section could benefit from changes in the structure and in the choice of figures. Indeed I think that the main results are not clearly displayed, and are therefore rather difficult to get at first. The discussion section is clear, however, it seems that a result, namely the identification of the five most explicative variables in the

models, is not discussed. I get that it can be uneasy to do so - this is well highlighted in the discussion - but I think it might be useful for the readers to have more insight in the authors' opinion on this specific result (note that I am not a specialist in the field of chemistry or toxicology).

Title

I wonder if it is really necessary to specify "in continental freshwater".

Materials and methods

p10 l.24: "Split each cluster", I guess you considered that, in the case of the "global" models, there is one cluster including all the chemicals ? Maybe you could specify, for instance by moving there the phrase "(the whole dataset […] a cluster-then-predict model)" which is currently p11 l.25-26.

p10 l.24: Is there a reason/reference for choosing those percentages of training and test dataset ?

p11 l.17: It seems to me that this paragraph, which ends p12 l.7, is not part of the "comparison procedure" section (2.5). I suggest to start a new section 2.6 Predictions, for instance.

P11 l.24: I do not get how many repetitions you use to compute the 95% prediction interval, by "leave-one-out bootstrap" do you mean that you compute the prediction n times (each time without one of the chemicals), n being the number of chemicals for which there is a CF value in the cluster ?

Results

p12 l.14-19: I suggest to move this paragraph to the Materials and Methods section, after the two first subsections describing the databases.

I also suggest to move Figures 1 and 2 to Supplementary material and add the lines 7 to 9 p14 as a part of the legend, or a comment. I suggest to start p15 l.1 as a first subsection of the Results (a title could be "clustering"), and Figure S2 could be moved to the main text, as the entire paragraph develops on clustering.

Sections 3.2 and 3.3: I found that the chosen structure does not highlight the results enough. I suggest to reorganise those in two sections focusing first on the model comparison and second on the performance and predictions of the "best model". Also, having figure S5 and the equivalent figure for CF_{HT} in the main text would help visualise and support one of the paper statements, namely the "best model" shows good performances.

Similarly, since you have assumed (and it is supported by references) that CF can vary by 2-3 orders of log-magnitude (p11 l.14 and p22 l.14) it would be interesting, again for better visualisation of the results, to highlight that value in Figures 4 and 5. The result stated in the discussion p22 l.17 only appears there, it should be moved to the results section.

In general, it would be useful to write down in the text some quantiles, not only the medians, of the distributions of absolute errors. I would also find interesting to display a table (which could be in the supplementary material) that sums up Figures 4 and 5, with median and quantiles of the absolute error for the 6 models for each CFs and each cluster. I do not find Figure S6 very useful.

Discussion and conclusion

p23 l.2 “the usual ones”: I guess you mean the approaches without the clustering step, I would rather write it like that than “usual”.
p23 l.2 “local”: I do not get what you mean by “local” in that context, could you specify ?
p24 l.10 “a new modelling method”: I would not call the method you describe in the paper “new”.

More generally, I found that the molecular descriptors that were identified as the “most important” are not discussed, although those are highlighted in the results section (p19 l.10-23 and p21 l.17 to p22 l.7) and in two tables in the main text. Similarly, the clustering result is not discussed either. For those two results, I suggest that you try to deepen the interpretation, or you shorten the corresponding paragraphs (and move the tables to Supplementary material) in the results section.

Figures

The colour palette of the boxplot figures is not colour blind accessible, that would be good to change for another colour palette (like the viridis one for instance).
Figures 4 and 5: it would be useful to have the number of replicates (I guess it is the 200 repetitions of the algorithm) in the legend.

Table List of the molecular descriptors: “Number of hydrogen atoms” is mentioned twice (1st and 2nd rows).
There is an issue in numbering, there are two Table 1. It seems there is a mistake in the caption of Table 1 (the second) and 2 “The *most important descriptors* are in the first line of the table” should rather be “The *best model* [...]”.

R script

It would be better to have the comments in English rather than in French. Also it is not very easy to quickly get how it is structured, so if you could separate the different steps of the analysis and put explicit titles, that would be great.