Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC1
https://doi.org/10.5194/hess-2021-618-RC1, 2022

Comment on hess-2021-618

Anonymous Referee #1

Referee comment on "A global assessment of nitrogen concentrations using spatiotemporal random forests" by Razi Sheikholeslami and Jim W. Hall, Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-618-RC1, 2022

Overall I was not impressed by the methodological advancements or the scientific implicaitons of this project. The authors built a standard ML algorithm, random forests, and applied it to a global nitrogen dataset. The portrayal of it as a spatiotemporal model is misleading: It is a standard random forest that uses month and lat/lon as additional predictors, which is perfectly fine but not a novel type of random forest model. The model appears to perform well on held out data, so the authors take that as evidence to then apply it globally and make maps. From those maps and the model itself, the authors pull out generally banal conclusions that have been recorded elsewhere. The 'so what' of the paper from the discussion section is that it could conceivably be used by other stakeholders in applications, but this link felt very light, so I am left skeptical of any pathway to impact.

Specific feedback:

The literature review of ML methods in water quality (Section 1.2) does not powerfully motivate the present study. Why are the “three critical observations” listed interesting and relevant? This section overall feels disconnected from the rest of the paper. The authors also fail to discuss why so few papers attempt to apply their models at a global scale (extrapolation risks) and make it seem that the community just never tried before, which is not the case.

Broadly, the motivation for this work is not clear and compelling. "The primary goal of this study is to introduce a global WQ model that is based on ML approach. (Section 1.3)". There are hundreds of water quality models that are based on ML, as they mention in their previous section. This paper is not really about providing a new dataset then, but a new model. **But their model is relatively off the shelf and does not tell us anything about the systems we do not already know.** Overall building a ML model, because we can, is not a compelling motivation in 2022.

Observational data: Why is the data collection period ceased in 2010? Data continues to be collected, so this seems an arbitrary cutoff removing potentially more data. Second, the stations used to build the model have large geographic disparities the authors do not discuss at length (e.g. abundance of sites in Brazil and Europe). **Sampling bias by location is a huge consideration when applying the map globally. I think this manuscript is an unsupported (by the data validation presented) extrapolation of a model to locations far different that those used to train the model. The authors gloss over this critical consideration when making the main global maps (Figs. 6-8).**

Predictors data: The authors say they started with a list of 27 candidate variables but then reduced it by more than half, to around 13 variables, to "reduce…redundant information" but one of the key advantages of random forests is that they work well with highly correlated variables. What were the other variables considered that were ultimately not included? Were the datasets aggregated over the watershed boundaries corresponding to each sampling location (for variables like precipitation and runoff they need to be)? Land cover is also known to be relevant but only cropland area was included.

Model development: **The novelty of this random forest methodology is greatly overemphasized.** There is research into spatio-temporal random forests, but those are far more advanced than what was applied here, making the title of the paper misleading. Here is an off-the-shelf random forest that anyone taking a Coursera data science course could apply successfully. To clarify, I am okay with the algorithm but troubled by the emphasis on its importance and novelty. Including latitude and longitude as predictors hardly makes this a spatial statistical model. Including month of the year hardly makes this a time series model.

Testing set: I would like to see sites completely held out as well to see how well the model predicts at new locations.

Model evaluation: What is the distribution of R2 values by location? Presumably some locations perform better than others. Also, the metrics are produced on a log-transformed

scale. **What is the mean absolute error or root-mean-squared error in interpretable, mg/L units? A strong performing model in log-log space is quite easy to produce (across domains, not just water quality) so it is important to record performance metrics in the back-transformed data space relevant to decision makers.**

Model evaluation (contd): I would like to see comparison of this model to benchmark models. For example, how does this compare against a simple linear regression? Against a mixed effects regression? Against simply fitting linear trends independently at each site? Not all of these need to be done, but some sort of well selected benchmarks are needed to contextualize model performance.

Model evaluation (contd): How to performance metrics compare with similar nitrogen modeling studies? If this model is the core advancement of the paper, its performance relative to other literature has to be clear and impressive. Unclear at this point if that is the case.

Model interpretation: The variable importance feature is interesting, but I want to see the influence of each variable on the outcome to check they make scientific sense. Otherwise the model could be getting it 'right' for the 'wrong' reasons. Partial dependence plots or the like are one way to plot those dependencies and could provide more interesting scientific findings rather than the surficial relationships presented so far in the paper.

Literature discussion: **Overall it did not seem like the results were sufficiently contextualized in the literature.** This goes for the performance metrics and the identification of increasing/decreasing trends in certain regions. Several of these regions have already been identified as having increasing/decreasing trends so how do these results build off of (or contradict) the prior literature?

Figure 3 is not helpful, perhaps move to SI if authors feel it is relevant.

Figures 4 and 5: What do the observed and predicted look like in original units? If this model and data outputs will ultimately be useful, it has to perform well in the original units. Figure 5 (test data) is more relevant than Figure 4 (training data), so Fig 4 could go to the SI.

Figures 6-8 **I worry considerably about extrapolation, so I do not trust the majority of locations shown.** Also, how about accompanying uncertainty maps?

Figure 8: Adds little not shown elsewhere.

Figure 9: Why do they find time series is more relevant? Is that surprising? Is it interesting that cattle is ranked where it is? The ‘so what?’ is missing here.