Response to Reviewer's Comments

This document contains copies of all comments of the reviewer 2 (*in italicized, blue* text) and our planned effort to address them (in normal, black text). Our proposed manuscript revisions are underlined.

Reviewer 2

The manuscript, titled "A global assessment of nitrogen concentrations using spatiotemporal random forests" by Sheikholeslami and Hall, introduced a machine learning (ML) approach (random forest model - RF) for predicting in-stream nitrogen (NOx-N) concentrations at the global scale. According to the authors, the novelties of this work are (1) its global scale application and (2) the spatio-temporal RF approach proposed in this study. In general, the manuscript was well written. Despite the results (instream NOx-N concentration) look quite well, the are several points regarding the model/approach used in this study that need to be addressed.

Thanks for reviewing our manuscript and providing extensive and valuable suggestions. The constructive comments of this reviewer are highly appreciated. We will address all the comments as described in this rebuttal document.

General comments:

1) Representation of nitrogen (N) lag times from input to riverine N export*: For water quality (e.g., N) modeling, it is expected that there could be significant N accumulated in the soil as biogeochemical legacy and the long travel time within the unsaturated/groundwater zone that could result in a lag time of years to decades between N input and riverine N export signals (e.g., Meals et al., 2010; Van Meter et al., 2017; Chen et al., 2018). It is unclear to me how the proposed RF model could take into account these factors. From my interpretation of the result, the "cumulative month count" variable (Figure 9) somehow could compensate for this kind of effect. However, we should not try to get the right result for the wrong reason.*

Thank you for sharing this important point and for providing the useful references. We certainly agree with reviewer's comment that legacy nutrients stored in soils (and groundwater) can create time lags between changes in nutrient inputs and the response of nutrient outputs. We also confirm reviewer's interpretation that the 'cumulative month of the year' somehow might compensate for legacy effect. An alternative approach would be using other machine learning algorithms such as hybrid long short-term memory-random forest method as proposed by Ahmed et al. (2021). This multi-model deep learning algorithm enables adding memories and lags of the predictors into the model. We will be very happy if other suggestions for how to address this challenge in machine learning-based water quality modelling become the focus of future discussion in the community. In the revised manuscript, we will provide reflections on this critical issue that have to be addressed in future studies.

Nonetheless, we should highlight that there is evidence that the diversity of nutrient recovery trajectories following reductions in nutrient loading suggest that nutrient removal capacity, contemporary nutrient loading, and nutrient-specific transport dynamics are as or more important than nutrient legacy in determining nutrient fluxes in watersheds, particularly when considering anthropogenic factors at large scales (see, e.g., Frei et al., 2021; Frei et al., 2020; Abbott et al., 2018; etc.). We will also elaborate more on this with the supporting literature.

2) Variable importance: *Why are the month of the year and the cumulative month count the most important variables?*

Note that these variables were included in our model to represent 'distance' in the time domain, particularly to capture the temporal dynamics of nitrogen concentrations. Variable importance results shown in Fig. 9 indicate that the most important covariate for predicting monthly nitrogen concentration given the utilized datasets is: time, i.e., cumulative (CM) and/or month of the year (MOY). This reveals that seasonality effect and long-term trends play an important role in prediction accuracy of the random forest algorithm. To address this comment, we will add more details to the relevant discussions in Section 4.3 (High importance factors influencing predictions of nitrogen levels).

I am wondering if the data used in the model has a strong seasonality that makes the variable "month of year" really matter. If this is the case, what is the implication for model application/performance in other areas that have less/no clear seasonality?

This is an excellent concern. Yes, apparently the input data of the model has a considerable seasonality that makes the variable "month of year" important. Our finding is justifiable noting that generally there is a seasonal variability in major factors influencing water quality, such as vegetation, land-use change, hydro-climatic parameters, and farming activities, which strongly influence constituents' concentrations in different seasons (see, e.g., Pejman et al., 2009; Shabalala et al., 2013; Xu et al., 2019; etc.). Therefore, it is expected to observe seasonal variation of water quality in many regions of the world. In the revised manuscript, we will add more discussion in Section 4.3 to highlight seasonal changes in water quality and its main influencing factors.

Regarding the implication for model application in areas that have less/no clear seasonality, we anticipate the proposed model can learn from given data that the month of the year is relevant or not. This is the fundamental ability of the learning algorithms that they can uncover and extract useful information from the input data (e.g., spatial and/or temporal variations).

Is the predictor "cumulative month count" highly important because of an increasing trend in the output variables in many areas (lines 495-497)? If yes, what are the implications from this?

First, we have to clarify that the importance of this variable measured by random forest does not necessarily imply that there is a strong increasing/decreasing trend in target variable (i.e., nitrogen level). In fact, random forest evaluates variable importance by estimating the mean decrease in prediction accuracy before and after randomly permuting the values of a given predictor. Therefore, as we mentioned in our response to your second comment, our factor importance results indicate that the most important covariate for predicting monthly nitrogen concentration given the utilized datasets is: *time*, i.e., cumulative (CM) and/or month of the year (MOY). In other words, the seasonality effect, and long-term trends play an important role in prediction accuracy of the random forest model. But, one cannot say if there is an increasing or decreasing trend in nitrogen level merely based on factor importance results.

Why does "fertilizer application" have a low rank?

We disagree with the reviewer's interpretation that "fertilizer application" is among the less-influential factors. In contrast, as mentioned in Section 4.3, nitrogen fertilizer use is one of the most important factors. In fact, as shown in Fig. 9, the strongly influential predictors are (in rank order): (i) cattle population, (ii) nitrogen fertilizer use, (iii) temperature, (iv) precipitation.

Why is there not much difference in the variables that were ranked 3rd to 15th (Figure 9)?

We believe that for the top 7 important variables the ranking is distinguishable. However, for the rest of predictors, i.e., 8th to 15th, we agree that the difference in variable importance values is not significant. It means that randomly permuting the values of these predictors resulted in quite the same change in prediction accuracy. In other words, the importance of these variables cannot be robustly ranked, even though they are all influential. As mentioned in the manuscript (Section 5), to comprehensively analyze how various factors influence model output variability, a more advanced approach is required. Global sensitivity analysis methods are suitable candidates in this regard.

***3) Spatial unit**: For predicting instream nitrogen concentrations, it is not clear to me why the authors did not use river network (instead of grid cell) as a spatial unit. In 1 grid cell (size of ≈ 55 km2) there could be* *several rivers, so it is unclear if the predicted values are applied for the main or tributary rivers. In addition, with rivers in big basins (e.g., Elbe, Mississippi, Amazon, Mekong River Basins, etc) that are running across multiple grid cells, it would be useful to incorporate the effect of upstream management/catchment characteristics into the model, it is not clear if this was considered in the RF model or not. As I understood from the description, the predictors for instream N concentration prediction currently only cover the properties within the grid cell of interest (no consideration of information from the upstream grid cell for large rivers).*

Thank you, this is a very good point, and we confirm that your interpretation about the spatial resolution of our study is all correct, i.e., the current covariates used for predicting in-stream nitrogen concentrations only cover the properties within the grid cell of interest.

It is widely understood that human activities modify runoff regimes in different spatio-temporal scales and has been proven by several studies, according to long-term observations and/or hydrological modelling experiments (see, e.g., Ren et al., 2002; Zhang and Schilling et al., 2006; Ferguson and Maxwell, 2012). Therefore, we think the impacts of upstream management and human interventions (e.g., land use changes, reservoir operation, river network modification, and building of dams) were, at least implicitly, reflected in runoff data as one of the predictor variables. We'd like to explore the impacts of other human-induced modifications on nitrogen concentrations by inclusion of other variables directly into the model, such as groundwater extraction, irrigation withdrawals, and existence of dams, in future improvements.

Regarding the catchment characteristics, this can be resolved, for example, by adding hydrography data delineating global river networks, though it will presumably add more complexity to the model. In the revision, we will add more variables to the list of predictors, including upstream characteristics, stream proximity (e.g., distance up to the stream) or log-transformed flow accumulation for better capturing spatial characteristics of watersheds. Previous studies have shown that these variables can be key drivers of water quality responses in rivers (see, e.g., Staponites et al., 2017; Lintern et al., 2017; Grabowski et al., 2016; Peterson et al., 2010). Particularly, they reported that accounting for the hydrological flow paths and flow accumulation through the landscape and coupling these processes with specific landscape features can improve model performance. We will also elaborate more on this with the supporting literature.

4) Capacities/limitation section*: I would suggest including a section describing the model capacities and limitations of the model/approach. Although the authors mentioned these points briefly in the conclusion section, however, they could be extended in a separate section.*

We highly appreciate this very important comment. In the revised manuscript, we will add a new Discussion section to better explain caveats/limitations of our approach and will provide possible recommendations for future research.

Specific comments:

Lines 30: "In addition, extensive construction of dams, excessive extraction of groundwater, deforestation, and expanding agricultural land use have altered sedimentary processes, mobilization of salts, and nutrient export to river systems, all of which drive WQ deterioration and groundwater pollution in many parts of the world…". Were these factors considered in the model?

Our model only accounts for expanding agricultural land use in terms of 'cropland area' and 'fertilizer use'. In the revised manuscript, we will also add 'forest fraction' and 'urban fraction' to the list of predictor variables, which will partially address this concern. We think addition of these variables can help us better capture the impact of anthropogenic forces on the global nitrogen variability in surface waters. Moreover, we'd like to explore the impact of other factors such as groundwater extraction and the existence of dams in future model improvements.

Lines 183-184: What is the temporal resolution of the NOxâ-N data and how were they aggregated to monthly timestep?

Most of the GEMStat stations provide monthly measurements. When daily values were reported for a specific month, we simply used the median of the daily measurements for that specific month. We will revise the text (Section 2.1) and add more details on this during revision.

Line 206: Please indicate where the readers could find the list of 27 potential explanatory variables

As mentioned in the paper, the selection of these predictor variables was based a process-informed manner through an extensive literature review (please see Section 2.2 and the references therein). To address this comment, we will also provide the list of 27 potential explanatory variables in Appendix of the revised manuscript.

Line 287: "The second strategy …" The second strategy has not been mentioned before this point

As mentioned in line 287, the second strategy is using geographic information as an auxiliary input to capture the spatial variability of the target variable, for instance, by adding geographic coordinates (Behrens et al., 2018; Meng et al., 2018) or other spatial distances (Li et al., 2011; Wei et al., 2019) into the list of predictors. We will revise this paragraph to improve the clarity of the writing.

Line 291: "Cumulative Month since 1992": how sensitive are the results to the start of the month count? This is a critical point if someone wants to run the model for other periods

Thank you, this is a very good point. Note that this variable has been included in our model to capture the long-term trends. Therefore, this variable can be critical when running our model for other periods that are much longer than our study period (1992-2010) and when the start of the time is far from 1992 as well. We will also discuss this point in the Results to highlight the importance of this factor.

Line 295: "..17 variables" – Please point out the list of 17 variables

Thank you, we will address this comment in the revised manuscript.

Figure 3: I would suggest adding more frames to the "output" panel (as already done in the "input" panel) to reflect that the spatial and temporal properties of outputs

Good point. In the revised manuscript, we will modify Fig. 3.

Lines 418-419: Is there a high correlation between elevation and latitude/longitude?

As mentioned in the manuscript, we have used DEM along with latitude and longitude as predictor variables in our model. Simply speaking, DEM is a discrete representation of the surface of the Earth using points generally placed on a regular grid. For each point its position is known in a chosen reference frame and represented through a chosen coordinate system (horizontal coordinates: geographic (latitude, longitude) or cartographic (East, North); elevation (heights or depths): orthometric with respect to a chosen geoid model or ellipsoidal. In other words, data sets can be either UTM-based, with points on a regular rectangular grid, or lat/long-based, with points at regular intervals on the geographic grid. Because of earth curvature, the two types of data behave differently over large areas. We have not conducted a correlation analysis to investigate the relationship between elevation and lat/lon. We believe this issue is beyond the scope of current study.

Lines 426 – 430: back to table 1, it is not clear whether these data (livestock population) were assumed to be time-invariant or time-variant

For our analysis, we assumed livestock population and wastewater production are constant for the study period mainly due to lack proper information regarding temporal variability of them. We will explain this and revise Table 1 to describe which predictors are time-variant/invariant.

Line 447-448: "This might be partially due to a high correlation between the agricultural fraction of land area and nitrogen fertilizer use" – Why do highly correlated variables have different rankings?

As we mentioned in the manuscript, several studies showed that estimating the importance of predictors by random forest algorithm might be problematic in the presence of correlation between variables. As another example, Toloşi and Lengauer (2011) identified the same issue based on extensive numerical experiments, which they called it "correlation bias". In summary, most of these studies reported two key effects of the correlation on the permutation importance measure: (1) the importance values of the most discriminant correlated variables are not necessarily higher than a less discriminant one, and (2) the permutation importance measure depends on the size of the correlated groups.

Gregorutti et al. (2017) provided theoretical validations for this issue, in a particular statistical framework. Based on Gregorutti et al. (2017)'s study, one possible answer for the reviewer's question on 'why correlated variables exhibit distinct ranking when using permutation-based importance measure', is that when one of the two correlated variables is permuted, the error does not increase that much because of the presence of the other variable, which carries a similar information. The value of the prediction error after permutation is then close to the value of the prediction error without permutation and the importance is small. In addition, it has been asserted that the permutation-based importance measure typically tends to discard most of the correlated variables even if they are discriminants and randomly selects one representative among a set of correlated predictors as the most influential factor (Gregorutti et al., 2017; Bühlmann et al., 2013). We will strengthen relevant discussions on the correlation and variable importance in random forests to better justify our results in Section 4.3 and provide more appropriate references.

Table 2: How were annual time step data (especially fertilizer application) disaggregated to monthly time step?

We assumed that these predictors, i.e., population, cropland area, and synthetic nitrogen fertilizer use, are invariant in month (i.e., only the annual variability is considered), and thus no disaggregation was performed. In other words, for each month of the year, we used the annual value of that year. We will modify Table 1 and revise Section 2.2 to clarify this.

Table 1: Please provide full names for the technical terms (e.g., ANN, DT, MLT,…) in Table 1

We will correct this as suggested.

References

Abbott, B. W., Moatar, F., Gauthier, O., Fovet, O., Antoine, V., & Ragueneau, O. (2018). Trends and seasonality of river nutrients in agricultural catchments: 18 years of weekly citizen science in France. Science of the Total Environment, 624, 845-858.

Ahmed, A. M., Deo, R. C., Feng, Q., Ghahramani, A., Raj, N., Yin, Z., & Yang, L. (2021). Deep learning hybrid model with Boruta-Random forest 8ptimizer algorithm for streamflow forecasting with climate mode indices, rainfall, and periodicity. Journal of Hydrology, 599, 126350.

Behrens, T., Schmidt, K., Viscarra Rossel, R. A., Gries, P., Scholten, T., & MacMillan, R. A. (2018). Spatial modelling with Euclidean distance fields and machine learning. European journal of soil science, 69(5), 757-770.

Bühlmann, P., Rütimann, P., van de Geer, S., & Zhang, C. H. (2013). Correlated variables in regression: clustering and sparse estimation. Journal of Statistical Planning and Inference, 143(11), 1835-1858.

Ferguson, I. M., & Maxwell, R. M. (2012). Human impacts on terrestrial hydrology: climate change versus pumping and irrigation. Environmental Research Letters, 7(4), 044022.

Frei, R. J., Abbott, B. W., Dupas, R., Gu, S., Gruau, G., Thomas, Z., ... & Pinay, G. (2020). Predicting nutrient incontinence in the Anthropocene at watershed scales. Frontiers in Environmental Science, 7, 200.

Frei, R. J., Lawson, G. M., Norris, A. J., Cano, G., Vargas, M. C., Kujanpää, E., ... & Abbott, B. W. (2021). Limited progress in nutrient pollution in the US caused by spatially persistent nutrient sources. PloS one, 16(11), e0258952.

Grabowski, Z. J., Watson, E., & Chang, H. (2016). Using spatially explicit indicators to investigate watershed characteristics and stream temperature relationships. Science of the Total Environment, 551, 376-386.

Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. Statistics and Computing, 27(3), 659-678.

Li, J., Heap, A. D., Potter, A., & Daniell, J. J. (2011). Application of machine learning methods to spatial interpolation of environmental variables. Environmental Modelling & Software, 26(12), 1647-1659.

Lintern, A., Webb, J. A., Ryu, D., Liu, S., Bende-Michl, U., Waters, D., ... & Western, A. W. (2018). Key factors influencing differences in stream water quality across space. Wiley Interdisciplinary Reviews: Water, 5(1), e1260.

Pejman, A. H., Bidhendi, G. R., Karbassi, A. R., Mehrdadi, N., & Bidhendi, M. E. (2009). Evaluation of spatial and seasonal variations in surface water quality using multivariate statistical techniques. International Journal of Environmental Science & Technology, 6(3), 467-476.

Peterson, E. E., Sheldon, F., Darnell, R., Bunn, S. E., & Harch, B. D. (2011). A comparison of spatially explicit landscape representation methods and their relationship to stream condition. Freshwater Biology, 56(3), 590-610.

Ren, L., Wang, M., Li, C., & Zhang, W. (2002). Impacts of human activity on river runoff in the northern area of China. Journal of Hydrology, 261(1-4), 204-217.

Shabalala, A. N., Combrinck, L., & McCrindle, R. (2013). Effect of farming activities on seasonal variation of water quality of Bonsma Dam, KwaZulu-Natal. South African Journal of Science, 109(7), 1-7.

Staponites, L. R., Barták, V., Bílý, M., & Simon, O. P. (2019). Performance of landscape composition metrics for predicting water quality in headwater catchments. Scientific reports, 9(1), 1-10.

Toloşi, L., & Lengauer, T. (2011). Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics, 27(14), 1986-1994.

Wei, J., Huang, W., Li, Z., Xue, W., Peng, Y., Sun, L., & Cribb, M. (2019). Estimating 1-km-resolution PM2. 5 concentrations across China using the space-time random forest approach. Remote Sensing of Environment, 231, 111221.

Xu, G., Li, P., Lu, K., Tantai, Z., Zhang, J., Ren, Z., ... & Cheng, Y. (2019). Seasonal changes in water quality and its main influencing factors in the Dan River basin. Catena, 173, 131-140.

Zhang, Y. K., & Schilling, K. E. (2006). Increasing streamflow and baseflow in Mississippi River since the 1940 s: Effect of land use change. Journal of Hydrology, 324(1-4), 412-422.