Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-423-RC4, 2020

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Seasonal watershed-scale influences on nitrogen concentrations across the Upper Mississippi River Basin" *by* Michael L. Wine et al.

Anonymous Referee #2

Received and published: 29 December 2020

General comments In this paper, Wine et al. developed a modelling approach to predict the seasonal variation of nitrogen concentration [TN] along the river network of the Upper Mississippi River basin (UMRB) using catchment variables and wetland configuration as predictor variables. The model allows predicting future [TN] in the UMRB under different scenarios of wetlands restoration. I found the work very interesting and timely, because nature-base solutions, like wetland restoration, might represent very adequate tools to improve water quality along large watershed and hence, to be considered within the integrated catchment management plans. However, there are some part of the manuscript that should be deeply reviewed by authors. In its current form, methods and results sections are very difficult to follow. I include several comments in

the following sections indicating the paragraphs that I think are messy and hard to understand. In addition, authors should tackle some other technical issues that, from my knowledge in modelling, were not completely well resolved. For instance, it is needed to include anthropogenic pressures (that might be significant for the [TN]) as predictor variables, issues related with the correlation of predictor variables or the lack of indicators of models performance, among some others. However, my major concern and criticism to this manuscript relates with the subjectivity to select variables in the LME while, as far as I have understood, the random forest model was precisely included in the modelling approach to make this selection based on the data. This issue is addressed in depth in my comments to each section, but in summary, authors should support objectively why they eliminate from the LME the most important catchment variable (forest cover) according to the results of the random forest, while the include other variables that were not selected by the machine learning approach. It seems that they eliminate/include variables according to previous knowledge and a priori hypothesis, so if authors are following this via I wonder why they used the random forest before the LME.

Abstract The abstract is clear and informative of what has been done in the work. However, it should be modified after the review and consideration of several of my main comments.

Introduction The introduction is well ordered and is clear in the exposition of the main ideas until reaching the aims of the study. The state of the art and bibliographical references are appropriated although I miss and I recommend "Alvarez-Cabria et al. (2016). Modelling the spatial and seasonal variability of water quality for entire river networks: Relationships with natural and anthropogenic factors. Science of the Total Environment 545–546, 152–162". This paper is related with some of the issues considered in the present work and some parallelism appeared (e.g. determination of main catchment and human variables influencing the seasonal variability of nitrogen concentration in rivers considering the seasonal variability and using machine learning approaches).

Methods Figure 1. It is not clear for me why authors included stream gauges (green triangles)? Are these the sites where [TN] was measured or just sites where the streamflow was measures? Clarify this.

Lines 107-109. Are there any big dams/reservoirs in the UMRB. USA is one of the countries with the largest number of big dams in its rivers and specially, in agriculture areas. Please, include this information. I believe that considering the effects of dams in this work is critical because: 1) Reservoirs have a significant influence in nutrient retention, and hence change of nutrient loadings downstream. 2) Dams operations produce an alteration of the flow regime that can even produce a seasonal inversion when they are dedicated to agriculture, and hence, influencing the seasonal patterns of [TN] loadings downstream (see my next comment regarding flow regime information requirement). If big/dams are present in the UMRB this should be included as a predictor variable of [TN] (e.g. number of big dams upstream, distance from the measurement site to the nearest dam upstream, rate of impoundment discharge related... there are many examples in literature)

Lines 109-111. It is hard to believe that a river network covering a river basin of almost 500,000 km2 shows such a homogenous flow regime. According to what it is stated previously in the study area description, it is true that climatic and other catchment variables do not change much along the catchment. However, I really miss a figure (in the main manuscript or as supplementary dada) showing the actual flow regime in several gauge station located in different rivers of the UMRB. I think that this issue is paramount as the intrannual variability of [TN] is highly related with the flow regime. Moreover, it would be great if author include any flow record of any altered flows by reservoirs upstream.

Line 130: What is "Major River Basin 3 water quality modeling group"? It is hard to understand this if you are not familiarized with the SPARROW database. Please provide better information about this database and the measuring points used, so reader will be able to know what data have authors used in this work. Line 132: Which criteria

have you used to set a minimum of 25 measurements? Which was the minimum number of years? Which was the minimum measurements/years to select the site (or year). The criteria to select/discard sites should be clearly stated. I would really appreciate the provision of further information in Table S1: Number of years with data and mean number of measurements per year.

Line 134-135: The mean vales of [TN] must be provided in the results (as, indeed, they are), so delete this sentence form the methods sections.

Lines 151-152: It is difficult to figure out the variables that authors have used here when referencing to a work that is still under review. Actually, they do this in several occasions (e.g. Line 85: Mengistu et al, In Revision; Leibowitz et al., In Review), So, if I can not access this information, how can I assess if it is valid or not? Consider eliminating the "under review" works or provide a better explanation, or both.

Line 154: I think that using daily discharge (of the day that TN was measured) as a time varying predictor is not a very correct approach. Daily discharge can be a too specific and uninformative. I recommend authors to use 10 to 7-days mean daily flow previous to the [TN] measurement to provide a average value of the discharge condition previous to the measurement and not just in the exact moment of the measurement. Moreover, if you are modelling [TN] in sites with highly variable catchments size (ranging 45 to 52,000 km2) it makes no sense to use the actual value of discharge as a predictor variable, because this value depends on the catchment area and its spatial variability will highly depend on the catchment area. Hence, I strongly recommend to change the models and use normalized daily discharge data. Normalization in usually achieved by diving daily discharges by the mean nannual flow of the whole series. Other option is dividing each daily stream flow by the catchment area. In this regard, it is very likely that the correlation of discharge with [TN] (Table 1 and Table 2) is highly dependent on this. For instance, if [TN] increase downstream, as it is probable given the increase of nitrogen inputs, and discharge also increase with catchment area, the relationship discharge-[TN] might be just and artefact.

Line 166-178: The whole subsection 4.2 is very confusing and difficult to follow. For instance: - the term “three-phased model approach” is quite confusing. In this regard, the third phase in the application of the model but not the creation of the model itshelf. - What is seasonal harmonics? - What does “counterfactual” actually means in the simulation approach? Is this paragraph really needed or does it need to be son long? I mean, in its current form is very confusing while specific details of the terms referred above are provided in subsequent subsections. So please, rewrite the paragraph just as a very brief introduction of the modelling approach followed in the next subsections. Otherwise, the manuscript starts to be is very y confusing.

Line 194: How did authors determine cross-correlation among predictor variables and which threshold have they applied to eliminate variables from the model?

Line 196: If some variables were known to be not important for [TN], why did you included them initially in a nitrogen concentration model? What is really disappointing here is that you eliminated variables from the model although they were good predictors in your model? Please, clarify this. It is not a good practice to eliminate variables if they were important. Contrary, you should explain and discuss why your model selected them as important variables. Otherwise, it seems that authors have used subjective approaches while they were supposed to do the contrary by means of the machine learning approach. It seems that authors include or delete variables based on quite subjective decisions (e.g. lines 194-196, 217-218, 234-236) while the random forest was supposed to make this selection automatically.

Line 222-233: This is confusing. Did authors actually use the ADRE equation in any way along the paper (after reading the whole paper my impression is that this paragraph is just and uninformative and confusing add). I think that all this explanation should be deleted here and just explained and discuss how variables included in the LME might be related to ADRE parameters. The methods section is already too long and very confusing because

After reading section 2.5 and 2.6 and, in accordance with my previous comments related with the seleccion of predictors, I wonder if the application of the initial random forest is needed to comply with aims of the study. I think that the LME model could be just applied with a set of variables selected according to the previous knowledge and literature. Please, provide support to the application of the random forest or just delete this initial step of the modelling approach from the analysis. I sincerely believe that, if it is not completely needed, the work will be much clearer and direct while currently it makes the paper more misleading and contradictory.

Line 240-246: It is not clear how authors included these seasonal functions (harmonics) in the LME models. Further information is needed to clarify their inclusion and parametrization. I think that this issue is better explained in the results section, so I strongly recommend to move the explanation to the results to gain clarity to understand the methodology.

Line 257. I think that the term “counterfactual” in not intuitive and does not actually inform of what to have been done and might be difficult to understand, especially for the non-native english speakers. I would really recommend to change this term throughout the text to facilitate the understanding.

Line 263: What is “GLWD”? Provide the complete abbreviation meaning.

Line 266-267. The modelled conditions are not clear at all since, in these lines, authors presented two future scenarios related with the increase (restoration) of wetlands areas while in lines 258-260 they stated that they are varying the proportion of both, wetlands and the proportion of cultivated areas. I understand that the increase of one of them produce the decrease of the other, as authors stated, but I think that this paragraph is not clear and it seems redundant in some parts. Please, rewrite and explain better what has been done.

Results Results are very messy. In this regard, there are several sentences (as some specified below) that are part of the methods while other are part of the discussion.

This section needs significant improvement to be understandable (and publishable). In my opinion, the current form of the results are impossible to follow. Sentences like: - Line 292. “though this correlation cannot be interpreted independently of other terms in ADRE”. - “a proxy for interannual climate or land-management variability” - “this does not necessarily indicate an inability on their part to influence water quality”. These statements are not part of the results but the discussion. In the results, authors must stick to what are the results that they have obtained from modelling, setting aside other kind of statements, rationales or conjectures. In this way, results are more direct and easy to follow. Please, check this in the whole results section and re-phrase them.

Another general comment and criticism to the results section is that I missed an indicator (e.g. adjusted-R2 or RMSE) to assess the model fitting performance, both in the random forest and in the LME models. Providing this kind of indicators are very valuable to evaluate how well the modelled values adjust to the observations and it is more useful and intuitive (they are commonly provided in many papers) that just providing the graphical solution (e.g. Fig 4, Fig9) or the bias. Moreover, these values help scientist (and watershed management in the future) to evaluate the uncertainty of the projected scenarios of wetland restoration.

Figure 2. Why discharge is presented in mm day-1 while according to Table S2 discharge is provided (within the models) in m3s-1. Is this specific discharge? Explanation of the upper part of the figure (discharge) must be provided in the figure caption because, currently, there is not any reference neither in the text (Line 280) nor in the figure caption.

Lines 280-282, Figure 3/ Line 314, Figure 6. Which criteria have authors followed to segregate watersheds by different catchment areas? It is not explained anywhere why do they have made this distinction. Why did they select catchments <350 km2 and why then they showed results segregated by catchment area? Same for figure 6. This seems random or subjective selection showing partial results. Please, clarify.

Line 299-303. The selected important variables, i.e. those that arose from data and random forest modelling, cannot be eliminated just because they are not supposed to be important (while authors included in the LME other variables that were not selected by random forest). The modelling approach used here supports that catchment area covered by forests is the most important catchment variable and authors have delete it from the LME? In my opinion, this make no sense and is not scientifically acceptable. This issue should be tacked in the discussion section, e.g. why a variable like forest that was not supposed to be important in explaining changes in [TN] is actually the most important catchment variable. It is also possible that this result is related to the correlation between forest and agriculture (more forest brings less agriculture and viceversa?), but the analysis of correlation and the elimination of correlated variables must be done before the development of the random forest model. So there are two options, 1) keep the model as it is regarding the results (without omitting/including variables subjectively) and discuss it or 2) make a previous correlation analysis and eliminate correlated variables (using an objective threshold).

Another impotant issue and a very useful option included in the random forest modelling is the use of partial dependence plots, which indicate the effect of each predictor on the response variable after taking into account the average effect of all other predictors in the model. They provide a useful basis for understanding the relationship between the response and the predictor variable. I strongly recommend authors to analyse these plots and include them, at least, at supplementary material, given that they can be fully explanatory of how each selected variable is affecting [TN].

Line 304. I don’t understand why authors include “total watershed N inputs” as the 3rd most important variable (table 2) while it was not selected within the top 13 important variable selected objectively by random forest.

Line 304. Again, authors are citing a paper under review to contrast their result. This cannot be done so frequently in the same manuscript.

Line 310 and Table 3. This result is poorly explained and in its current form is very confusing. In addition, the caption of Table 3 is not completely informative. - What are the grey shadowed parameter in each equation? - Why do they included variables like cultivated area, watershed area, wetland area that were not selected by random forest. This seems very subjective in the way that authors included the variables that they thought were going to be interesting a priori but were not selected in the variable selection model. So where is the interest of applying a machine learning approach? Very confusing. These results must be clarify.

Line 313-322: This paragraph is very messy. There are part of methods (e.g. line313-316) and other parts that are discussion (e.g. Lines 318-321). This section must be completely re-write and stick to what are results and what should be included in other sections of the document. In its current form is very difficult to follow.

Line 323-329. Again, authors included several variables as fixed terms, in my opinion, in a subjective way, as they just stated that "Several fixed effect terms further improved the model". How much the inclusion of these variables influence the model (a quantitative and objective assessment value is needed)? How it affects AIC? Moreover, it is very likely that correlations exists between variables that were selected by the random forest and those that the authors are including here (e.g. discharge and catchment area). Including, this variables might produce overfitting of the models to the observations. Correlation between each pair of variable can be easily determined using Pearson's correlation coefficients or any similar approach. Please, provide an objective reason why you include these variables. Provide and objective procedure that explain how you specifically select these variables from the complete variables pool (Table S2). Provide and objective quantification of the actual influence of the inclusion of these variables into the model (changes in R2, RMSE, AIC. . .).

Line 352: "implying that denitrification by wetlands is secondary to the reduction in fertilization, which is of primary importance" and lines 355-359. These are not results but explanation/implications of the results; they should be part the discussion, if needed.

Discussion After considering the comments proposed above, discussion must account for the changes of the results and new findings. In addition, authors should re-order the text and review the parts of the results that should be part of the discussion (any explanation of their results). In addition, I personally do not like the order that authors follow to discuss their results are discussed. I think that the discussion and the manuscript would be cleared if this section follows the same order than in previous sections. For instance, in section 4.2 they discuss the counterfactual modelling results. I understand that it is an important part of the papers but I expected a previous discussion of the results of the modelling approach, which were presented at the initial part of the results (only discharge is discussed). I strongly recommend to re-order the discussion to gain clarity.

Line 405-407. I will insist on this issue. I think that this discussion is not scientifically acceptable. It cannot be said that the model is not working just because results do not much with your expectations, hence invalidating part of modelling results and providing other subjective decisions (I am specifically referring to the variables selection). If authors hypothesize that forest and cultivated areas are correlated you must provide an objective assessment of this assumption. If they are correlated it is hence possible to discard the use of one of them in the model (because they are providing the information for the model). But this should be done with all pair of variables.

Line 415-420. This is not a discussion of your results but a divagation about the potential chances related with the use of big data and machine learning. Hence, this paragraph seems more a part of the introduction than a discussion of the results because authors did not state how the advantages of big data and machine learning approach reflects in their results. Moreover, they stated, in the previous paragraph, that their results question the use of big data and machine learning, so it results contradictory.

Lines 430-439. Even though in this paragraph you can perceive the intentions of what authors want to convey, I found that the text is too general, ambiguous, and difficult to relate with specific results, as it is expected in the discussion.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2020-423, 2020.