Review of manuscript “A predictive model for spatio-temporal variability in stream water quality” submitted to HESS (HESS-2019-342).

This manuscript describes a statistical modelling exercise for stream water quality in Victoria, Australia. The manuscript is well written, however, I have some concerns regarding the modelling framework, performance measures, site bias, and results in drought impact. These comments are outlined below, and need to be clarified before publication.

1) Model framework
   While I understand the notion of using catchment spatial variables to represent site-level mean (which is the focus of the published Lintern 18 paper), and using temporal variables to represent deviation from the mean (which is the focus of the published Guo 19 paper), I do not understand equation 6 – why is it necessary to add additional catchment characteristics in the temporal component? Why 2 variables? What’s the implication of this equation for the framework overall? i.e. the framework started with distinct spatial and temporal components, but ended with the temporal component also include spatial variables? Wouldn’t that means the spatial variables are double counting, i.e. does this lead to the model overly focusing on spatial variability while less representing temporal variations? In any case, this need to be explained better in the manuscript.

2) Model performance measures
   a. The manuscript uses long-term mean concentrations frequently in the result and discussion sections (e.g. Figs3-5). My understanding is, based on equation 2, the long-term mean results would be very close to spatial variability, while the temporal component does not have much role in determining the long-term mean:

      $\mathrm{Long-term\ mean\ in\ model\ results\ for\ k\ time\ steps} = \mathrm{C_j} + \frac{\sum \Delta ij}{k}$
      Assuming $\sum \Delta ij$ can be close to 0 as the positive and negative derivations more or less cancel each other out.

      If this is the case, then I’m not sure the long-term mean results are representative for both spatial and temporal variability, and the authors may consider using different result measures to better demonstrate the model’s ability to represent spatial AND temporal variability.

   b. The NSE values for 4 of the 6 constituents are not great. Based on a widely used classification in water quality model performance measures (Moriasi etal 07), the model performance (i.e. NSE values) for these 4 constituents are “unsatisfactory”, while that for TKN is “good”, and EC is “very good”. While it’s perfectly fine to report results even if they are not great, it is questionable to use these 4 poorly performed models for further inference, i.e. change in system response for TSS since drought. Granted, the authors used the long-term mean concentration results for TSS (which have higher NSE values), then it’s back to the previous comment regarding the long-term mean concentration may not adequately represent temporal variability.

3) Site bias

a. The areas of sites are highly diverse, from 5km2 to 16,000 km2. It's reasonable to expect that these different sized sites may be dominated by different processes, e.g. smaller sites may be constituent supply driven, while larger sites may be transport driven. These differences may be translated to different explanatory variables for these sites. But in the model, these sites share the same explanatory variables AND model parameters (ie the betas). The implications needs to be discussed, e.g. if there're more sites with large areas, then the model may bias towards representing large catchments, and the explanatory variables selected does not have strong predictively power for smaller catchments, and thus leading to poor model performance.

b. Data transformation: the authors chose to transform observation data, rather than back-transform modelled data. There are a few issues with transforming observation data: 1) the transformation involves additional parameters (such as lambda, instead of a straight transformation, e.g. logx), thus the "observed" data is in effect a "modelled" data, albeit a simple model. 2) The observation data across sites is transformed using the same parameter value (mean), thus the site bias issue in the comment above also applies. 3) the choice of transformation (log) leads to a decrease in the sensitivity of large values due to the log() function, and increase the sensitivity of small values. Thus, it is unclear to me whether using transformed observation data is any better than back-transforming modelled data. These implications need to be pointed out in the manuscript.

4) Results in drought impacts

a. Assuming the model is appropriate for inference (i.e. have good enough performance measure), a better (more insightful) way to demonstrate the impact of drought could be to show what the parameters (beta) for pre and post drought models are. This is because (I assume) these parameters represent the system behaviours, i.e. how strong different explanatory variables are to predict concentrations.

Other comments:

1) Pg 17, L374: please explain why "out models are very useful in representing and predicting proportional changes in concentrations"?
2) Maybe consider putting supplement tables S5 and S6 in to main text as these are important part of the model.

Reference:

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the ASABE, 50(3), 885-900.