Responses to Reviewer Comments on "A predictive model for spatio-temporal variability in stream water quality" (RC4)

Our proposed revisions are underlined.

General comments

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 –

1.1 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?

Our modelling framework accounts for spatial variation in the parameter of each predictor that has been selected to explain the temporal variability, which were observed in Guo et al. (2019), as well as in Musolff et al. (2015) from a separate dataset. Therefore, the purpose of Eqn. 6 is to explain these spatial variations and thus enabling spatial prediction of those temporal effects according to catchment characteristics. This equation essentially makes the modelling framework fully spatio-temporal (i.e. being able to predict any location at any time-step). The choice of two variables was mainly due to the consideration of controlling model complexity (i.e. number of parameters).

- *Musolff, A., Schmidt, C., Selle, B., and Fleckenstein, J. H.: Catchment controls on solute export, Advances in Water Resources, 86, 133-146, https://doi.org/10.1016/j.advwatres.2015.09.026, 2015.*

1.2 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.

We do not agree with your opinion that using additional spatial variables to explain temporal variability is redundant in our models. We believe that the reviewer is concerned about considering catchment characteristics twice in both Eqn. 3 and Eqn. 6; however, we acknowledge that these two sets of catchment characteristics served contrasting purposes. In Eqn. 6, the two additional catchment characteristics represent spatial variation in the relationships between temporal variability in water quality and its key predictors (e.g. hydro-climatic conditions, vegetation cover). For example, the impacts of streamflow on temporal changes in water quality are stronger at some catchments than at others, and these differences can be explained with additional catchment properties. This contrasts from the purpose of Eqn. 3, where uses a separate set of catchment characteristics to explain the spatial variation in ambient (average) water quality conditions (with more details in Lintern et al., 2018b). Therefore, both sets of spatial predictors serve unique purposes and are necessary components of the models.

To address Comment #1 and improving the clarification of modelling framework, we will:

1) *add brief description of Eqn. 6 in Section 2.1 to better justify the purpose of including this equation. We will also further emphasize this additional modelling capacity (i.e. modelling temporal variability across catchments) that we gained from Eqn.6, apart from the two preceding studies.*
2) *present additional results and discussions on the key drivers for varying temporal relationships across catchments to illustrate the value of this specific model component.*

2. Model performance measures

2.1 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:

Long−term mean in model results for k time steps= $C_j + \frac{\sum \Delta ij}{k}$

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.

Thank you, this is a very good point and we confirm that your interpretation about the spatial and temporal variabilities are all correct. We agree that the existing results presented on model performance are predominantly focused on spatial variability. To improve this, in the revision we will present more results on how the model represent temporal variability in the Results section. This is in line with a number of important additions that we propose to present, to improve a) the presentation of different aspects of model performance and b) the illustration of model utilities that are useful for catchment management. In specific, we propose to include the following topics:

1) *Modelled and observed time-series at selected catchments, to illustrate model capacity to predict trends and changes over time;*
2) *Proportions of spatial and temporal variabilities explained for each constituent, to illustrate the importance of the two variability components for each constituent, and how they can be explained by our models;*
3) *Extending the existing model cross-validation from 5 replicates to 50 replicates, to provide a more comprehensive summary on model sensitivity to calibrated dataset.*

We believe that these additions could add more evidences on the model capacity to represent temporal variability.

2.2 The manuscript 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 longterm mean concentration may not adequately represent temporal variability.

We agree with the reviewer that the NSE achieved in our models is not as high as those recommended in Moriasi et al. (2007). However, this discrepancy should not be a key concern for our models. Firstly, we would like to point out that the water quality models reviewed in Moriasi et al. (2007) were all physically-based, spatially-distributed models (SWAT, HSPT, DRAINMOD-DUFLOW and DRAINMOD-W) which focused on individual catchments within the US, where the key practical implication of modelling is to simulate catchment processes and management activities, and thus to support local catchment management. In contrast, the statistical models that we developed aimed to predict spatio-temporal variabilities over a large Australian region, which has an area of over 200,000 km^2 and more than 100 catchments. The key practical implication was to support higher-level catchment management at a state- or even a national-scale. Due to the different model types, contrasting scales and practical implications between our models and the models reviewed in Moriasi et al. (2007), we are not convinced that the performance standards summarized in Moriasi et al. (2007) are directly transferable to our models. Furthermore, due to the extensive spatial and temporal extents that our models cover and the less focus to support local-scale management activities, it is both more difficult and less necessary for our models to achieve the same performance standards as suggested in Moriasi et al. (2007).

We understand the second part of your comment as questioning: 1) whether the 4 poorer models, TSS, TP, FRP and NOx, are capable to make further inference from, specifically on exploring TSS changes to drought; 2) the validity of using long-term mean concentration of TSS to represent temporal variability, when exploring the drought effects on TSS. Our response to each question is as follows:

1) *We completely agree with you that when a model is not performing well, we should be careful on drawing further inferences. However, we understand such 'further inferences' as to making predictions and/or interpreting parameter values with respect to physical processes. What we presented on the responses of TSS to drought was different to such 'inferences', as this analysis was based on a model validation against three distinct periods which are differently affected by a prolong drought in the region. In this experiment, the focus was not the model performances in an absolute perspective, but instead, the relative performances of different calibration/validation periods. Specifically, Figure 4 focused on how performance deteriorated when calibrating to one sub-period and validating on the other. Similarly, Figure 5 focused on the variation of model performance of the 'full-model' when simulating individual sub-periods of the full data period, so the focus is again on the relative model performance. We believe that exploring these 'changes in model performances' is an informative approach to explore any drought effects especially when the absolute model performances are not optimal.*
2) *We would like to clarify that although this analysis explored changes in water quality across different sub-periods of the full dataset, the focus was any consistent shift across three periods, as opposed to the day-to-day variability of water quality (i.e. as how 'temporal variability' has been defined in our modelling framework). We believe that such cross-period changes can be more clearly summarized with the long-term mean concentrations for each period, as currently presented. <u>As in our responses to Comment #2.1, to better illustrate the model capacity to represent temporal variability, in the revised manuscript we will provide additional results and discussions.</u>*

3. Site bias

3.1 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.

This is an excellent concern. However, we identified a major misunderstanding of our models which we would like to clarify – the statement 'in the model, these sites share the same explanatory variables AND model parameters (i.e. the betas)' is incorrect. This is because that our Bayesian hierarchical models do allow parameters for the temporal predictors across catchments to vary depending on catchment characteristics (Equation 6, which we explained with more details in response to your Comment #1.1). Such variations in temporal parameter sets are capable to account for differences in the key water quality processes across catchments e.g. different roles of surface and sub-surface flows on water quality due to different scales of catchment processes.

Furthermore, in representing these variations of temporal relationships across catchments in our models, we have already considered catchment area as a potential predictor (see Table S1 in the supplementary materials which lists all 50 potential predictors that we considered). However, catchment area has not been identified as a key predictor for variation in these temporal relationships for any constituent, which indicates the less important role that catchment area has on affecting the temporal variability patterns across space.

Our choice of the use of a consistent set of model predictors across all catchments was to ensure that models are able to represent key processes and controls in a large-scale perspective, rather than being dominated by catchment-specific patterns that are difficult to generalize and interpret. For example, if we allow 102 catchments to have different numbers of predictors, it would be extremely difficult to obtain a large-scale understanding on the role of streamflow, as well as to understand how the impacts of streamflow vary across catchments.

To resolve this comment, we propose to improve our model description in Section 2.1 (Spatio-temporal modelling framework), to further emphasize the point that the temporal parameters were allowed to vary across space, which considered potentially different key processes and controls for water quality across the diverse catchment conditions in our catchments. We will also provide some examples to explain how the key processes can vary across catchments e.g. contrasting processes for larger and smaller catchments.

3.2 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.

Transforming data for our modelling was a decision informed by previous phases of the same research project, where we used linear statistical models to identify the key drivers of water quality variability across space and time (Lintern et al., 2018b; Guo et al., 2019). To incorporate the previously obtained

understanding into this study, we used similar linear model structures – which were calibrated with transformed data to satisfy the assumptions of linear modelling or otherwise the untransformed data would be too skewed to work with (see more details in Figure R1 under point 3)). Since the model was calibrated in a transformed scale, we believe that the transformed scale is also most relevant and informative for performance assessments as we presented. We clarify each specific issue you raised as following:

1) *Politely disagree. The log transformation, as referred to as a 'straight' transformation in the comment, is a special case of Box-Cox with lambda=0. Therefore, even a log would still introduce an additional parameter (although = 0), which has fundamentally no difference with a parametric Box-Cox transformation. In a more general sense, parametric transformations (e.g. log, Box-Cox, Log-sinh) have been widely applied and recognized as data pre-processing approaches, instead of a step in the modelling process.*
2) *Using the same parameter value for transformation across all catchments ensured that the results (performance of the calibrated water quality models for each constituent) are in a consistent scale and are thus comparable across catchments. This is an essential requirement to achieve large-scale spatio-temporal modelling capacity as addressed in this paper. Regarding site bias, we have explained in our response to your Comment #3.1 that our Bayesian hierarchical modelling framework can effectively address the concern of site bias, by allowing variation in the temporal parameters to represent potentially different key processes across catchments.*
3) *The whole purpose of data transformation was to reduce the impacts of the extremely high values on model calibration. This is because that those high values often present in extremely low proportions within the data – as illustrated in Figure R1 in which untransformed data were plotted against corresponding quantiles, for each constituent. If those extreme values (right tails in each panel in Figure R1) were left untransformed, they may cause the models to emphasize too much on rare extreme events, and thus largely affect our ability to represent the overall large-scale patterns in water quality.*

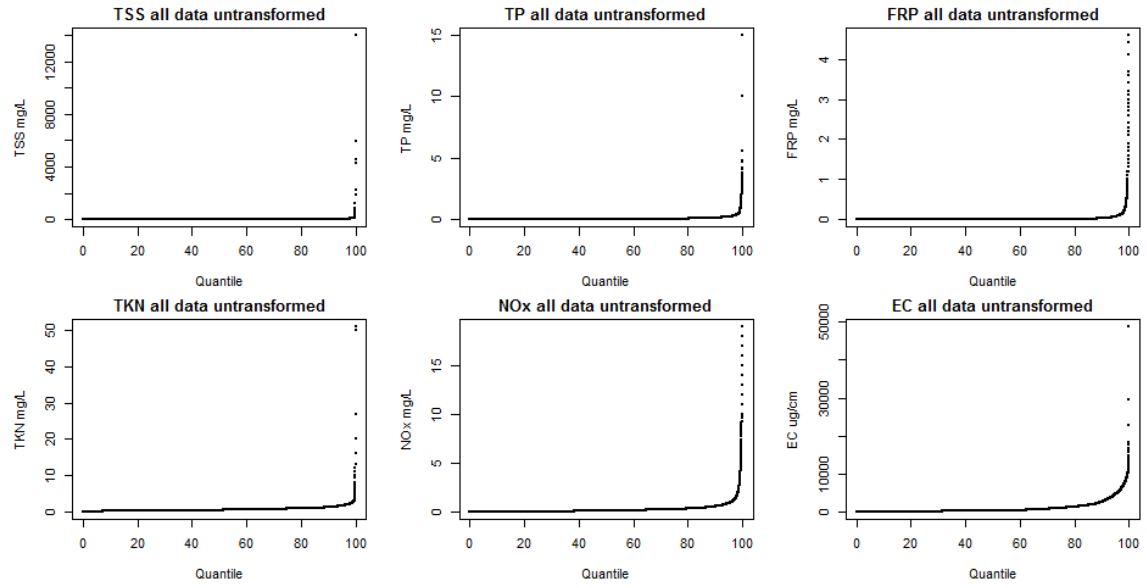

Figure R 1. Untransformed data against their quantiles for each constituent.

To address this comment, we will add more discussions in Section 2.2 (Data collection and processing) to improve the justification of data transformation. We will also add explanations in Section 2.4 (Model performance and sensitivity analyses) on why model performance assessments are presented in a transformed scale.

While we decided to focus this study on the transformed data, we have noted back-transforming modelled data as a possible option, and we would like to explore the differences between these two approaches in future studies.

4. Results in drought impacts

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.

Thanks for sharing the interesting idea. Firstly, we have compared the parameter values for the key spatial and temporal predictors of TSS when the model was calibrated to different periods. The effects of key predictors for spatial variability did not vary much across periods (Figure R2). In contrast, the effects of key predictors for temporal variability showed a clear shift in the role of antecedent flow (prior 7-day flow) across different drought periods (Figure R3). Specifically, the flow effects are mostly positive across catchments before the drought, which shift to mostly negative during the drought; after the drought, the flow effects have mixed directions among different catchments.

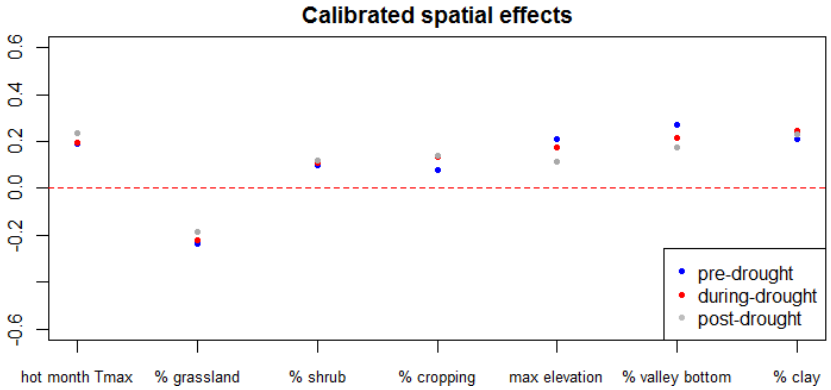

Figure R 2. Effects of the seven key predictors for the spatial variability in TSS across 102 sites, summarized by the posterior mean of the calibrated parameter values for each predictor, to the pre-, during- and post-drought periods (differentiated by colour). The seven key predictors are, from left: hottest month maximum temperature, percentage catchment area as grassland, percentage catchment area as shrub, percentage catchment area as cropping land, maximum catchment elevation, percentage catchment area made up of valley bottoms, and average soil clay content.

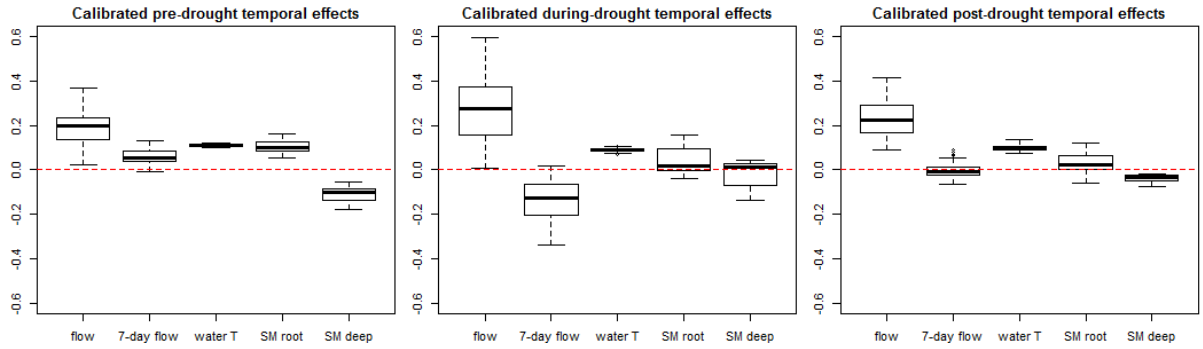

Figure R 3. Effects of the five key predictors for the temporal variability in TSS across 102 sites, summarized by the posterior mean of the calibrated parameter values for each predictor (box shows values across all sites), from left: flow, 7-day antecedent flow, water temperature, root-zone soil moisture and deep soil moisture.

Despite the interesting results and potential discussions, we note that these results are likely not reliable considering the limited performance of the TSS model. This has been explained in our

response to your Comment #2.2, point 1), where we agreed with your comment that we should be careful in making further inferences with a model with limited performance. Therefore, for this drought analysis, we propose to not presenting/discussing further results beyond the existing ones in the manuscript (i.e. relative performance of model over different calibration/validation periods as in current Figures 4 and 5).

Other comments

5. Pg 17, L374: please explain why "out models are very useful in representing and predicting proportional changes in concentrations"?

The Box-Cox transformation which our models were developed with is essentially similar to log transformation, which is widely used in water quality to represent proportional differences in linear space. We will improve clarification of this argument during revision.

6. Maybe consider putting supplement tables S5 and S6 in to main text as these are important part of the model.

Agreed. The key model predictors shown in Tables S5 and S6 are important parts of this model, although they have been identified from our two preceding studies (Lintern et al., 2018b; Guo et al., 2019). To address this comment, we will move these tables to the main text. In addition, since the second column of Table S6 (which summarizes the key factors relating to the spatial variability in temporal effects) are new findings in this study, we will provide more interpretations and discussions on these results.

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.