Responses to Reviewer Comments on "A predictive model for spatio-temporal variability in stream water quality" (RC2)

Our proposed manuscript revisions are underlined.

Context

This paper introduces a Bayesian hierarchical model for spatio-temporal prediction of water quality variables in Australia. After model construction and validation, the results are discussed in terms of influences on prediction accuracy and regarding the influence of a long drought period on average suspended sediment concentrations. The paper concludes with recommendations regarding model improvement.

General comments

Generally, the paper is well written and the methods and results are interesting. However, I have some major concerns regarding (i) the statements drawn from the results, (ii) influences on the simulation accuracy and (iii) the focus of the study. These major points need to be clarified before publication.

Thank you very much for your comprehensive review and identification of key areas of improvement. We provide detailed response to your comments in the subsequent sections.

Focus of the study

1. The study is introduced as a new model for water quality prediction. It is mentioned that the construction of the site-specific model was already published in two preceding papers (Lintern et al., 2018b; Guo et al., 2019). It is not really clear which additional information this paper provides. In the discussion section, there is a long chapter about the influence of a long-term drought to TSS concentrations, which was found as a by-product (?) of the study. The papers ends with conclusions suggesting higher-frequency sampling data, which was not analysed in this study at all. Thus, the study lacks a clear focus and coherent conclusions.

Great points. We acknowledge that the two preceding papers (Lintern et al., 2018b; Guo et al., 2019) focused on identifying the key controls for spatial and temporal variabilities of stream water quality, and understanding the effects of these controls. In contrast, this study presents the integrated model developed based on the previous understanding. Although the model structure was informed by the preceding studies, this study established, for the first time, a spatio-temporal model which is capable to predict across multiple catchments in a regional scale. In addition, in this study we have also developed new understanding on how the temporal drivers of water quality vary spatially, which is a key component of spatio-temporal predictive capacity. To address the comment on the innovations that this study brings, we will revise the Introduction and relevant sections in the Discussion to highlight how this study differs from its preceding works. We will also add more results on how the temporal effects vary spatial as a new understanding obtained from this study.

In addition, to improve the linkage between study objectives and results, we propose to present additional results to highlight several model capabilities, with further discussions on how these can benefit catchment management. These include the following potential topics:

1) Modelled and observed time-series at selected catchments, to illustrate model capacity to predict and interrogate 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 disagree that the effects of long-term drought on TSS is a by-product of the study, but rather consider it as an illustration of model utility – to identify potential changes in water quality processes associated with major catchment changes. The same approach can also be applied to simulate catchments with regions which experienced significant changes in land use and dam development, etc.. and assess corresponding impacts on water quality. To address this comment on the purpose of analyzing drought effects on TSS, we will revise the Discussion section to make this point clearer.

We consider the use of high-frequency monitoring data to be an interesting discussion point, as also reflected in the discussions in Comment #4 from Reviewer 1 and Comment #9 from Reviewer 3 and our responses to those comments. Although we have not analyzed high frequency data in our study (since only a few sites in Victoria have such data), we would like to discuss the potential benefits and additional requirements of using high-frequency data in our modeling framework, and thus to provide useful guidance to future studies. To address this comment on discussing high-frequency sampling data in the conclusion, we will clarify that these are recommendations based on our experiences with the models. We will also thoroughly revise the Conclusion to better align it with the rest of paper.

The influence of LOR on simulation accuracy

2. First of all: What is LOR (Limit of Reporting)? Is it a limit of detection (LOD) or a limit of quantification (LOQ) or something different? Which value was used for the calculation of Nash-Sutcliff (Neff) efficiency if the measurement was below LOR? Zero? Half the LOR? Please clarify.

Our use of the term 'LOR' actually refers to the concept 'detection limit' as defined in the Victorian Water Quality Monitoring Network and State Biological Monitoring Programme (1999), as:

- *'minimum concentration detected for which there is 95% confidence of accuracy and therefore is accurate enough to report. Detection limits are based on a minimum of 10 replicates of a sample or standard of low concentration of the analyte, taken through the whole procedure (including digestion if required by the method).'*

This is different to either of LOD and LOQ, which have been defined as (Armbruster & Pry, 2008):

- *LOD: 'the lowest analyte concentration likely to be reliably distinguished from the LoB and at which detection is feasible. LoB is the highest apparent analyte concentration expected to be found when replicates of a blank sample containing no analyte are tested.'*
- *LOQ: 'the lowest concentration at which the analyte can not only be reliably detected but at which some predefined goals for bias and imprecision are met.'*

Regarding the second part of your question, for the calculation of NSE, when the measurement was below LOR (below LOR values were used only for model evaluation in Section 3.1), the value of half of LOR was used.

To minimize confusion and keep consistency with our monitoring dataset, we will replace the term 'LOR' with 'detection limit' in the revised manuscript. We will clarify the definition of 'detection limit'

where this term is first introduced (L153). We will also add clarification on how the data below detection limit were used when describing the relevant model performance assessments (L213).

- *Australian Water Technologies: Victorian Water Quality Monitoring Network and State Biological Monitoring Programme: Manual of Procedures, 1999.*
- *Armbruster, D. A., and Pry, T.: Limit of blank, limit of detection and limit of quantitation, Clin Biochem Rev, 29 Suppl 1, S49-S52, 2008.*

3. For model construction, the values below LOR were excluded due to statistical reasons and due to the fact that these low concentrations were of less interest. Thus, why were the values below LOR included in model validation at all? Please clarify.

The only place which we considered below-LOR data in model evaluation was to understand the ability of this model (which was calibrated to truncated data) to simulate the full distribution of observations (as justified in Section 2.4), which was not a validation strictly speaking (where independent dataset should be used). Due to the exclusion of below-LOR data for our model calibration, readers may question how much the model performance would be affected by including the below-LOR data. If inclusion of the below-LOR data leads to a good fit, then the models calibrated to above-LOR data is transferable to below-LOR data too.

To address your comment, we will add these discussions to Section 2.4 to better highlight the purpose of this specific model performance evaluation and the fact that subsequent evaluations excluded the below-LOR data.

We acknowledge that all model cross-validations (shown in Section 3.2) were performed without below-LOR data.

4. Later on it is analysed that the fraction of LOR on total measurement values influences model performance, especially the P fractions and TSS. The discussed reasons are mainly methodical/statistical. I think, the effect of LOR on model performance might also be a secondary effect: the parameters with a high proportion of LOR are mainly those with the highest natural concentration variability, since their concentration peaks are event-driven. Thus, monthly grab samples might capture peaks or not. Since some of the catchments are as small as a few km², even the specific time of a day might influence the sampled concentration to a large extent. Thus, the probability of sampling low between-event concentrations is higher for P and TSS than for e.g. Nitrate. Therefore, the low model performance might rather be an effect of the overall lower information content of the samples, which results in models which are based on a lower information content. What do you think?

Thank you for sharing this very interesting point. We understand that you suggest another possible explanation for the influences of high proportions of below-LOR samples on our model performance, that is, constituents with large number of below-LOR samples are often also driven by high streamflow events, which are otherwise insufficiently captured by the monthly monitoring data.

To check whether the high LOR issue is more observed in event-driven constituents we referred to Figure 5 in Guo et al. (2019) (included below), in which the impacts of flow on concentrations across all 102 sites within same dataset were summarized by the 'Q same day' boxes for each constituent (i.e. panel). We focus on the two constituents that are most affected by the high LOR issue (TSS and FRP), and found that TSS is highly event-driven while FRP is relatively less influenced by flow, compared with other constituents. These suggests that the high LOR issue might be related to lack of

sampling high flow events for TSS, but there is no systematic pattern across constituents. We will add some discussions on this point when highlighting the high LOR issue for TSS.

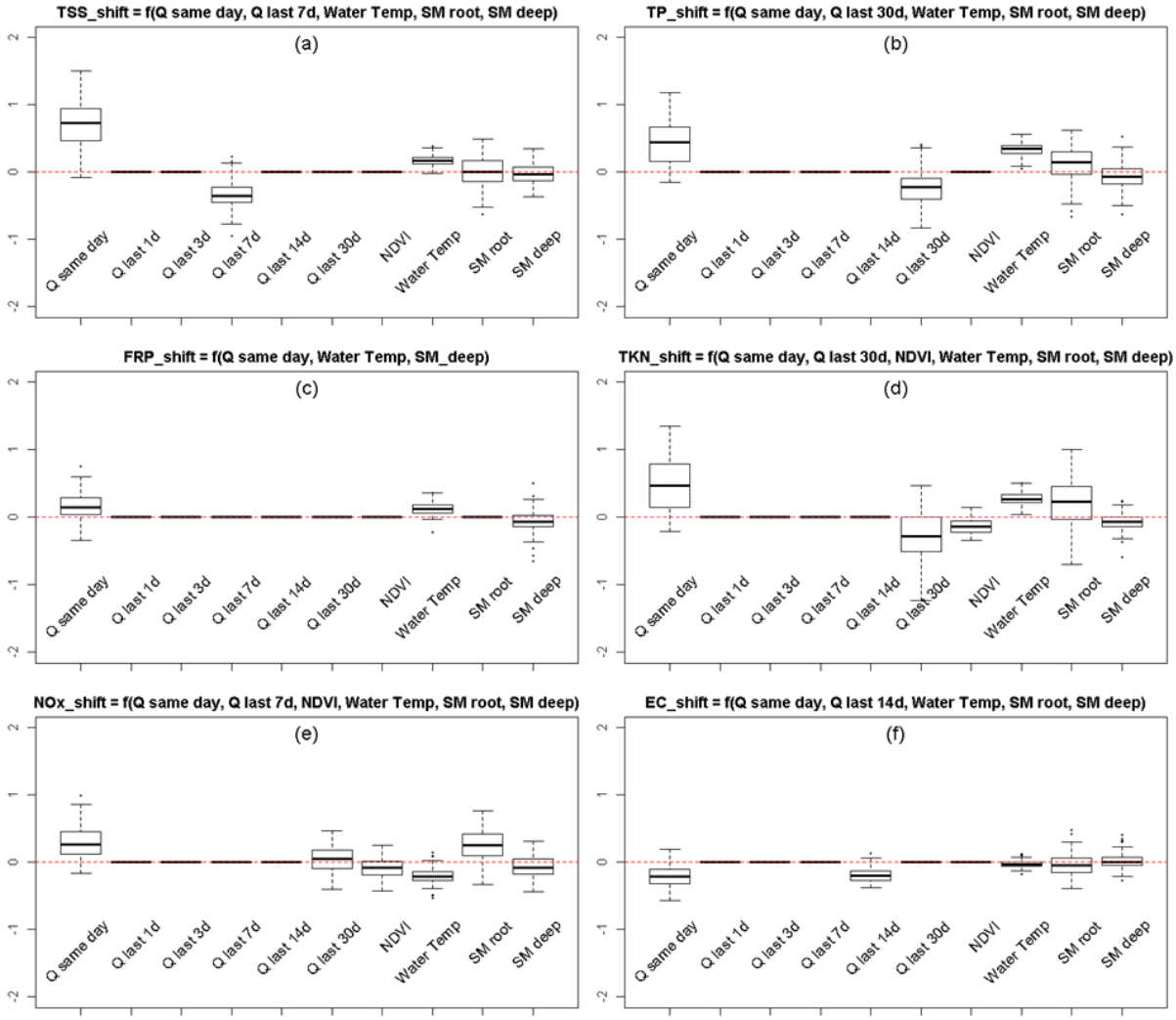

Figure R 1. Figure 5 in Guo et al. (2019): Effects of hydro-climatic predictors on the temporal variability of each constituent across 102 sites, summarized by the posterior mean of the calibrated parameter values for each predictor. Y-axis shows the effect as the number of standard deviations away from site mean. Note that only predictors identified for at least one constituent during the selection of predictors are shown on the X-axis.

Influences of drought on TSS

5. During the modelling process, the authors note, that a long-term drought influenced TSS concentration, which is a really interesting observation. However, I do not understand why a model is required for this analysis. Wouldn't simple statistics (such a t-test or Mann-Whitney-U-Test) have done the same job? I don't see that this is a special result of this model application.

Firstly, the main purpose of the paper is to develop the modelling framework, not examine hypotheses about drought. We use this analysis as an illustration of our model capacity. We agree that simple statistics would indicate trends/changes over time, but the interpretation is limited to only changes in concentrations without further indication on potential causes. Specifically, with simple trend statistics we would be able to identify changes of TSS concentrations during the drought, but not able to tell whether such changes are due to decrease in streamflow or other more complex processes. In contrast, using the models developed in this study, we were not only able to

identify changes in TSS concentrations, but also able to suggest that these systematic changes are not due to changes in any of the key controls of sediments (e.g. streamflow) since drought, but instead, related to a shift in the relationships between sediment concentrations and its key controls (e.g. streamflow) during different periods – this reveals much more understanding compared with simple trend statistics.

To clarify this, we will add brief discussions in Section 4.3 (Potential impacts of long-term drought on water quality dynamics) to compare and contrast our analysis to simple trend analyses, and to highlight the additional understanding obtained through our approach.

Meaning of factors

6. Since the model is a (multidimensional) statistical model, the explaining variables (factors) not necessarily contain process-based meaning for the target water quality parameters. For example, the water temperature is an explaining variable for temporal variability of TSS (Table S6), which is not really clear to me. In L.15-17 it is stated that the paper addresses the key controls (factors) explaining water quality variability, but an in-depth analysis and discussion is missing in the text. I would encourage the authors to even discuss the factors in more detail or to change the focus of the paper.

As in our response to your Comment #1, we acknowledge that the focus of this paper is not to identify the key controls for spatial and temporal variabilities of stream water quality and understand their effects – which have been addressed in the two preceding companion papers (Lintern et al., 2018b; Guo et al., 2019). The effects of key controls on water quality have been presented in detail and discussed extensively in these two preceding papers and are therefore not repeated in this study.

To avoid the confusion which this comment reflects, we will revise the Introduction to better clarify the focus of this study and how this study differs with the preceding ones, along with revision on other parts of the manuscripts.

As also mentioned in response to your Comment #1, this study developed new understanding on how the temporal drivers of water quality vary spatially, which is a key component of spatio-temporal predictive capacity. To highlight this, we will add new results and discussions on the key factors relating to the spatial variability in temporal effects.

Regarding the example that you mentioned (the mechanisms via which water temperature influences the temporal variability of TSS), we have provided extensive discussion on possible mechanisms in Guo et al., 2019, along with those for other key temporal drivers identified. In short, the strong impacts of water temperature may be due to the high correlations between water temperature and air temperature; while higher air temperature (warmer periods) can be further associated with enhanced source and mobilization processes, which potentially lead to: (1) greater soil desiccation and soil erodibility, (2) more intense agricultural activities that occur during warmer periods such as tillage, or (3) lower plant canopy cover in drier and warmer months. Due to the different study focuses, these are not further discussed in this study.

Specific comments

7. 1-2: The title "A predictive model for spatio-temporal variability in stream water quality "suggests a generic model for different sites and different water quality parameters. However, the described model is very site-specific. Thus, I would suggest to change the title to a more site-specific one, probably including the region or similar, including the applied method.

We would like to clarify that the models developed in this study are not completely site-specific, but were integrated space-time models that are capable to predict across 102 sites over a 200,000km^2 region at once. The model structures were informed by previously obtained understanding on both the catchment- and regional-scale water quality variability and their key controls from the two preceding companion papers (Lintern et al., 2018; Guo et al., 2019). This data-driven modeling framework is transferable to any other parts of the world.

Adding locations or study region to the paper title is likely causing misunderstanding that this paper describes a case study of existing modelling approach, which would in turn greatly hamper the communication of key contributions of this study. Therefore, we politely disagree with the reviewer on adding study locations to the paper title. However, to improve clarity, we propose to add the phrase 'data-based' in our title to suggest that an empirical model is introduced:

'A data-based predictive model for spatio-temporal variability in stream water quality'

8. 71: Change "…quality can not…" to "…quality not…"

We will implement this change as suggested.

9. 76: Change "…model built…" to "… model was built…"

We will implement this change as suggested.

10. 76-78. It is stated that the model was constructed and published in two previous papers. Please elaborate on the additional information this paper provides.

As explained in our responses to your Comments #1 and #6, this paper presents the first spatio-temporal model developed over a large geographical region across multiple catchments. We will clarify this better in the Introduction. In addition, we will also adjust the earlier parts of the Introduction to focus more on the knowledge gap relevant to this study (i.e. developing spatio-temporal predictive capacity), instead of those that are relevant to the preceding studies (obtaining new understanding).

As also highlighted previously, this study obtained new understanding on how the key controls of temporal variability of water quality vary spatially, and thus developed spatio-temporal predictive capacity where the two preceding papers have not achieved. New results and discussions on the spatial variability in temporal effects will be added to support the new findings.

11. 79: It is stated, that this study aims at bridging the gap between fully distributed and statistical models. Well, what is this model if not a statistical model? Probably, it was meant to bridge the gap between fully/semi-distributed and lumped models.

Thank you. We meant to say that the model bridges the gap between fully-distributed physically based models (which are driven by equations representing physical processes e.g. SWAT) and data-driven statistical models (which are fully relying on observations e.g. black-box ANN type models). We will adjust the phrase here during revision.

12. 154-156. During the Box-Cox transformation of the data, the high sampling values lose their significance, especially for goodness-of-fit calculations. This effect can be seen after back-transformation (figure S13), which results in low Neff values. Thus, how is the statement "poor water quality conditions…were our primary concerns…" compatible to the fact that the data was transformed?

By saying 'poor water quality conditions (i.e., high constituent concentrations) were our primary concerns to model' (L155), we are referring to higher concentrations with respect to the below-LOR data. Our modelling was not specifically focusing on the representing the extremely high values, but rather focusing on the large-scale patterns of water quality. Such modelling focus required data to be transformed so that the modelling was not overly sensitive to extreme values.

13. 159. Insert a blank between "as each"

We will implement this change as suggested.

14. 186. "… via a Spearman correlation analysis" (note the typo "analyses"). Please add the correlation coefficients and the p-values in the supplement.

We will correct the typo and add the Spearman correlation results in the SI, as suggested.

15. 246. "…in Sect. 4.2." Isn't it section 4.1?

Thank you for identifying this, we will make the correction.

16. 265. Fix "… is also show…"

We will correct this as suggested.

17. 414. Fix "For examples, …"

We will correct this as suggested.

18. 418 Fix "adjscent"

We will correct this typo.

19. 449-451. In the beginning, this paper aims at introducing a model. In this lines, the reader has the impression, that the main aim of this paper is the analysis of drought on TSS concentrations. Please think about the focus of the paper.

As in our responses to your Comments #1, #6 and #10, we will revise the Introduction and relevant sections in the Discussion to better highlight the key study objective. We will also update the Conclusion accordingly to ensure that the paper is coherent and focused.

20. 466-469: "1) collection … in the model". These are not a results/conclusions of this study. Data frequency was not evaluated in this study.

This sentence intends to summarize the key areas of improvement for this modelling framework which have been identified in the Discussion section instead of study results – as seen in the phrase 'to further enhance the performance of the current models, we recommend that future… (L465)'. We have provided summaries of the key results in previous sections of the Conclusion (L453-365).

21. 469-470. "These improvements will be very helpful…" How?

The models that we developed are very useful to provide insights on the overall patterns of water quality variation and potential key controls of these variation, and thus inform the development of mitigation strategies. Therefore, our models are likely more beneficial to support mid- to long-term management, planning and policy making. Our model capacity is likely enhanced by increasing availability of high-frequency monitoring data, since they are likely providing better representation of the temporal variability. However, these data might also have extremely high variability e.g. due to unknown point sources and measurement noises, which brings new challenges for the statistical

modelling framework. Considering these, we have decided to revise this recommendation as an open question on the opportunities and challenges that our modelling framework will face when presented with more high-frequency monitoring data. We will also revise relevant sections in the Discussion accordingly.