Responses to Reviewer Comments on "A predictive model for spatio-temporal variability in stream water quality" (RC1)

Anonymous Referee #1 Received and published: 20 August 2019

Our proposed manuscript revisions are underlined.

This manuscript presents a Bayesian-based approach to analyze spatio-temporal variability in stream water quality. The approach is demonstrated with an application to a large set of monitoring data in Australia. Overall, I think the manuscript is well written and will become a worthwhile contribution to the hydrological community. The proposed method also has the potential of being applied to monitoring data elsewhere. I do have some major and specific comments for the authors, which I hope can help improve the manuscript. I recommend its publication after the following comments are addressed.

Thank you very much for your comprehensive review and recognition of the study contribution. We provide detailed responses to your comments in the subsequent sections.

General comments:

1. On model applications: I recommend the authors to add a separate sub-section to provide some guidelines to potential users of the proposed approach, including at least the computer running time of the model, the required no. of stations and required no. of water-quality samples for running the model, as well as approaches to evaluate if the model does a reasonable job.

We agree with the reviewer and we will add these recommendations for future users of this modelling framework to Section 4.1 (Implications for statistical water quality modelling) of the revised manuscript.

2. On calibration/validation analysis: The authors randomly selected 80% of the sites for calibration and used the remaining 20% for validation, and repeated this validation process for five times for each constituent, in order to evaluate the sensitivity of the model to the monitoring sites. Could you justify the use of five times for each constituent? If this cannot be easily justified, I recommend the authors to increase the replicates from five to a larger number (say 30 or 50). The results may be summarized as boxplots instead of Table 2, which can provide an overall evaluation of the model's ability to capture the dynamics of the different constituents.

We agree that increasing the number of this calibration/validation runs would provide a more comprehensive understanding of model robustness to calibration datasets. We are currently running this cross-validation process with 50 replicates, and we will update the relevant results and discussions in the revised manuscript.

3. On the below-LOR data: The authors argue that the model performance is related to the proportions of below-LOR data. The results appear to support the argument that model works better when the proportion of below-LOR data is low. Can you further prove this? The authors may quantify the proportion of below-LOR data for each monitoring site and conduct a separate analysis for sites of low proportions vs. sites of high proportions (perhaps 50% of sites for each group?) and see if the performance varies significantly between the two groups. This analysis may be implemented for each constituent.

Thank you for the interesting idea. Since our focus is to explain performance variation across constituents, we believe that a more informative analysis to support our argument (below-LOR data impacted model performance) is as follows:

1) *For each of the constituents which has the highest proportions of below-LOR issue (e.g. TSS, FRP and NOx), calibrate the corresponding model by leaving out 10% sites which have the highest proportions of below-LOR data.*
2) *Use a 'control' constituent which are less influenced by the below-LOR issue (e.g. EC) and re-calibrate the model with the same 10% sites removed as in 1).*
3) *Compare the change in model performance from the full model between the focused constituent and the control constituent. If the former shows clear increase in performance whereas performance for the latter remains similar to the full model, then that is supporting our argument that our model performance for constituents with higher proportions of below-LOR data will be affected.*

We will investigate these results and expand the discussing accordingly.

4. On monitoring data: In this pilot application of the proposed approach, water-quality variability is modeled based on monthly monitoring data. First, I think the authors have made a good point that high-temporal-resolution data can further strength the model capacity to explain temporal variability in water quality. Second, I think the approach's ability to reasonably capture that variability based on just monthly monitoring data is a big strength of the proposed approach. After all, a lot of the monitoring records at many locations are based on a monthly sampling scheme. This aspect should be more emphasized. Third, how about high-flow sampling? Many monitoring programs supplement regular sampling with targeted stormflow sampling to capture concentration variability during storm events (e.g., Chanat et al., 2016; Zhang et al., 2017). It is widely acknowledged that sediment and particulate constituents are heavily affected by storms. However, I cannot find any discussion of this aspect in the manuscript. Would you expect the models to be further improved if the monitoring data contain targeted stormflow samples? References: âA˘ c Chanat et al. ´ (2016) (URL: http://dx.doi.org/10.3133/sir20155133) âA˘ c Zhang et al. (2017) (URL: ´ https://doi.org/10.1016/j.jhydrol.2016.12.052)

Thank you very much for sharing these great discussion points. To address this, we propose to add more discussions in Section 4.2 (Implications for water quality monitoring programs) to:

1) *Emphasize that the strength of our model in being able to predict ST variation in monthly data across large region, as many large water quality datasets are composed of monthly samples;*
2) *Discuss a) the common limitation regarding lack of high-flow (event) sampling data, and also b) transferability of the statistical model structure to event-data (where re-calibration is required). In general, monthly monitoring data (like what we used in this study) can typically well represent ambient water quality concentrations across the flow duration curve; however, they would be limited in representing and predicting conditions during events, which becomes a greater issue in estimating total load.*

   We cannot directly compare these two sampling schemes, since event-based data are not available for this particular monitoring dataset. However, we have reasonable confidence in the model capability to capture variations in flow-weighted event mean concentrations, which we obtained from a parallel study of us in the Great Barrier Reef region in Queensland (Australia). Specifically, we applied a similar spatial-temporal statistical modelling framework to an event-based water quality sampling dataset and achieved satisfactory performance. The study is currently under review with Water Resources Research.

5. On key controlling variables: Table S5 and Table S6 may be combined to a single table and moved to the main text. I think this information is critical and deserves to be placed in the main text.

Agreed, we will move Tables S5 and S6 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.

Specific comments:

6. The term "filterable reactive phosphorus (FRP)" may be replaced with "soluble reactive phosphorus (SRP)". I think the latter is more widely used.

Thank you for raising this point, and we agree that SRP is more widely used than FRP in the water quality field. However, the term 'FRP' has been used by the State Government of Victoria where all our water quality data were accessed from (i.e. Victoria Water Measurement Information System, available at: http://data.water.vic.gov.au/*). We would like to keep consistent terminology, and thus to keep the term FRP throughout this manuscript. To avoid confusion, we will clarify the naming convention of FRP and relate it with the more commonly used terminology in the literature (SRP), when FRP is first introduced in the manuscript (L121).*

7. L46: Add a few more references to support the argument "differ significantly".

We will add more recent references to support this argument, e.g.,

- *Chang, H.: Spatial analysis of water quality trends in the Han River basin, South Korea, Water Research, 42, 3285-3304, https://doi.org/10.1016/j.watres.2008.04.006, 2008.*
- *Varanka, S., Hjort, J., and Luoto, M.: Geomorphological factors predict water quality in boreal rivers, Earth Surface Processes and Landforms, 40, 1989-1999, 10.1002/esp.3601, 2015.*

We will also replace 'significantly' with 'substantially' to avoid confusion with 'statistically significant' here (and also for other similar occurrences throughout the paper).

8. L56: Provide some specific examples on "other catchment conditions". One could be antecedent condition, which is heavily discussed in the manuscript. In this regard, Zhang et al. (2017) (URL: https://doi.org/10.1016/j.jhydrol.2016.12.052) provides a study on how antecedent conditions affect the estimation of riverine constituent concentrations. This is also relevant to your discussion at L430.

Thank you for the recommendations. We will elaborate more on the impact of 'other catchment conditions' with the supporting literature.

9. L103-L107: These sentences can be removed. I think the subsection titles are already very clear.

We'd like clarify the paper structure as much as possible for the readers' benefit with these overview sentences. To address this comment while maintain clarity, we propose to move these sentences to start of Section 2 (before Section 2.1) – which is a more suitable place to have an overview of the entire Method section.

10. Figure 1: Use a different color or a larger font for the dots to make them more clear.

We will revise this figure to improve visualization.

11. L130: Add a few more references to support the argument "widely known to influence water quality condition".

We will add more recent references to support this argument, e.g.,

- *Giri, S., and Qiu, Z.: Understanding the relationship of land uses and water quality in Twenty First Century: A review, Journal of Environmental Management, 173, 41-48, https://doi.org/10.1016/j.jenvman.2016.02.029, 2016.*
- *Heathwaite, A. L.: Multiple stressors on water availability at global to catchment scales: understanding human impact on nutrient cycles to protect water quality and water availability in the long term, Freshwater Biology, 55, 241-257, 10.1111/j.1365-2427.2009.02368.x, 2010.*

12. L131: "literature review" is vague. Could you briefly describe how it was conducted?

We will add more details on this process, which involved both an extensive review of published literature (focusing on key controls which affect spatial variability of water quality, see Lintern et al., 2018a) and a consultation with all project partners via a project scoping workshop.

- *Lintern, A., Webb, J. A., Ryu, D., Liu, S., Bende-Michl, U., Waters, D., Leahy, P., Wilson, P., and Western, A. W.: Key factors influencing differences in stream water quality across space, Wiley Interdisciplinary Reviews: Water, 5, e1260, doi:10.1002/wat2.1260, 2018.*

13. L164: I do think one or two references should be provided for "Box-Cox transformation" to help readers. The meaning of the parameter lambda should be also briefly described.

We will add more details on the transformation approach, equation and parameterization with supporting literature.

14. L352: This ranking is roughly consistent with particular constituent vs. dissolved constituent. Any comment in this regard?

We believe that you are referring to our finding that the ranking of the 'categorical' issue most heavily affected FRP (dissolved), followed by TSS (particulate), TP (mostly particulate), NOx (dissolved), TKN (mostly particulate) and EC (dissolved) in a decreasing order. We find it difficult to relate this ranking to the form of constituent (i.e. particulate or dissolved), as there is no distinct pattern of which form of constituent has more 'categorical' issue.

15. L366: The authors list here some processes for N. How about processes for P?

We will expand this discussion onto alternative phosphorus pathways (e.g. P desorption and adsorption) in catchments.

16. L206: What is the "Rhat" value? Please clarify.

Rhat is a summary statistic on the convergence of the Bayesian models implemented in package rstan, which indicates the differences in the estimated model parameters between and within the independent Markov chains (4 chains used in this study, as in L204). Rhat>>1 indicates that the chains have not mixed well (i.e., the between- and within-chain estimates are not consistent) and a value of below 1.1 is often recommended to check convergence (Stan Development Team, 2019). We will add these clarifications in the revised manuscript.

- *Stan Reference Manual Version 2.20: https://mc-stan.org/docs/2_20/reference-manual-2_20.pdf, access: 28/09/2019, 2019.*

Editorial comments:

17. L71: Fix usage of ". . .not only. . .but also. . ." In addition, "limits" should be "limit".
18. L76: The model built. . . –> The model was built. . .
19. Equation 3 and Equation 4: For the betas, consider using subscript instead of dash.
20. L180: "General speaking" –> "Generally speaking"
21. L317: Fix "a results of"
22. L382: Fix "oppourtunities"
23. L417: Fix "droguht"
24. L420: Similarly to –> Similar to Comments on the SM:
25. Supplementary Materials lack of "title-page" information.
26. Table S4: Change "lambda" to its Greek form.

We appreciate your valuable comments and we will address all these in the revised manuscript.