Responses to Reviewer Comments on "A predictive model for spatio-temporal variability in stream water quality" (RC3)

Our proposed manuscript revisions are underlined.

General comments:

1. The study describes a Bayesian statistical model of selected water quality variables in 102 catchments. The model successfully described both the spatial and temporal variability of certain variables, and performed quite well at describing the site-specific means for all variables. Based on the results, the model can serve as a valuable prediction tool in the calibration region (and potentially adapted elsewhere too).

   The main issue with the manuscript is that the otherwise valuable work is presented in an unsuitable (and constantly evolving) context. The title appropriately focuses on the main element of the study, the model and emphasised predictions as the primary field of utilisation. In the Abstract the motivation for the study is summarised as: "To address this [knowledge gap compromising present water quality models], we developed a Bayesian hierarchical statistical model to analyse the spatio-temporal variability in stream water quality across the state of Victoria, Australia." This shifts from predictions to analysis and promises that the model will cover knowledge gaps presumably by revealing so far unknown relations between water quality and its drivers. Interestingly, this objective is not featured in the Introduction. There it reads: "Our approach aims to bridge the gap between fully-distributed water quality models and statistical approaches to provide useful information for catchment managers, especially for largescale water quality assessments." This alters the context again, now the model is meant to be a "missing link" between very detailed (deterministic) models and simple statistical tools and the reason is to serve catchment managers. These context shifts do not help to assess the values of the study and generate expectations that are fulfilled later.

Thank you very much for your comprehensive review and contribution of valuable ideas. We would like to clarify that:

1) *'Revealing unknown relations between water quality and its drivers' has been covered in our previous two papers. Specifically, Lintern et al. (2018b) investigated the key catchment characteristics that are related to spatial variability of catchment water quality; Guo et al. (2019) investigated the key controls for temporal variability of water quality at each catchment.*
2) *The core objective of this study is to develop statistical models that can predict spatial and temporal variabilities of catchment water quality. In achieving this spatio-temporal predictive capacity, we have developed new understanding on how the temporal drivers of water quality vary spatially, which has not been explored in the two preceding studies.*
3) *The main model objective is to improve the predictive power for these variabilities and thus allow catchment managers to better plan/manage water quality changes across both space and time.*

We will thoroughly revise the Introduction (and other parts accordingly) to ensure that the main study objective, research questions and study implications are clearly communicated, and that individual results are mapped to the corresponding research questions. We will also revise the abstract to better capture the full story. In addition, to support point 2) above, we propose to include more results and discussions on how temporal relationships vary spatially.

2. Unfortunately, none of the above alternative contexts is completely followed in the Results and Discussion. The Results consist almost exclusively of performance indicators calculated and plotted in transformed scale. The Discussion focuses on the effects of the drought period on model performance and future development directions without mentioning potential major obstacles and pitfalls (gathering more detailed data and developing more detailed models is an idealistic recipe). The manuscript would greatly benefit from following a clearly defined logical structure, objectives and featuring topics that are truly relevant for the work. Performance indicators should not occupy all the Results section. There is much more to show about the model, especially considering the ideas that show up in the present Introduction and Abstract. The potential topics include:

While confirming that our core study objective is to develop statistical models that are capable to predict spatial and temporal variabilities of catchment water quality, we agree that there are more dimensions to present/discuss on the capability of the models and their practical implications for catchment managers. We appreciate your suggestions on potential topics and we provide responses to individual ones as following:

2.1 Untransformed comparison of measured and modelled time series for selected catchments

Investigating model performance with time series is an excellent idea. We propose to investigate the model capacities to capture trends and changes within water quality time-series, together with discussions on their relevance to catchment managers.

We would like to focus model evaluation at the transformed scale, since this is what the model was calibrated for, and would thus provide the most informative assessment of model performance (as also explained in more details in our response to your Comment #43). We will add these to Section 4.1 to justify the use of transformed data for model evaluation.

2.2 The needs of catchment managers with respect to predictions and how this model fulfils them (If it does so. If not, management should not be emphasised so much).

We believe that our proposed investigations to address your Comment #2.1 (model capacities to capture water quality trends) is a good example of the management utility of our models, because water quality trends and changes are of great interests for catchment management. Our models linked water quality variations over time with the hydro-climatic conditions and vegetation cover of catchments (i.e. controlling temporal variabilities) which means that the model would be able to capture some trends/changes that are 'expected' (e.g. due to droughts or major changes in vegetation cover); on the other hand, catchment managers can use these models to identify 'unexpected' trends (e.g. due to additional discharge points and major land use changes which are not included in our model), and prompt further investigation. These insights are very beneficial to catchment management particularly in regional/national scales, as well as long-term catchment planning and policy making. We will strengthen relevant discussions on the existing results and the proposed additional results (as in response to Comments #2.1, 2.3, 2.4 and 2.6) to better highlight the management implications of our models.

2.3 Key controls and mechanisms governing water quality. What do we learn from this study compared to Guo et al. 2019 and Lintern et al 2018a, 2018b?

As highlighted in our response to your Comment #1, this study does not focus on identifying key controls for spatio-temporal variabilities in water quality, as these have already been extensively

analyzed and discussed in our previous companion studies (Guo et al. 2019 and Lintern et al 2018b). To address this comment, we will revise the Introduction thoroughly to better clarify this different focus of this study with the two preceding studies.

In addition, we note here that understanding and predicting differences in the influence of controls on temporal variability between catchments, as discussed in response to your Comment #1, is a unique contribution of this paper compared with the two preceding papers. This will be better clarified in the revised manuscript.

2.4 The grade of intrinsic randomness (and its compatibility with management), predictability of water quality variables.

We interpret the comment as requesting assessment of natural variability in water quality that is explainable through deterministic predictors. Unexplained variability has been summarized in our model performance evaluation (Section 3.1). To address this comment, we will strengthen the discussion on the potential components of unexplained variability that we have already presented, and further explore the percentages of spatial and temporal variabilities that remain unexplained. These additional results and discussions will better illustrate the model ability to explain water quality changes and the relevant limitations. These results are informative to catchment management, because they highlight how much variability is explainable over space and time, and thus, the modelling error we expect when predicting long-term conditions across space, and when predicting over time at individual sites. We will add relevant discussions to highlight the management implications.

2.5 Model limitations: implicit assumptions, conditionality on the calibration set and the present layout of calibration units (what would happen if the model was calibrated on merged catchments?)

In section Section 4.1, we have presented extensive discussions on the key model limitations due to 1) the linear statistical model structure, 2) the impacts of below-LOR records and 3) data transformation. Furthermore we have also discussed model limitations due to lack of representation of of biogeochemical processes.

We agree that calibration dataset and model transferability are typical limitations of statistical models, which we will add to the above Discussion. On a related note, we also propose to increase the number of cross-validation replicates from the current 5 to 50. We believe that this would give us a more comprehensive summary of model limitation to calibration dataset.

We are unclear on the interpretation of your comment 'what would happen if the model was calibrated on merged catchments' and thus provide two possible interpretations here.

1) *If your comment refers to the differences between a model calibration to individual catchments versus one using all catchments merged as a single one, we are unsure about value of modelling on merged catchments. This is because that we have already used a joint model calibration across all 102 catchments (instead of site-specific calibration) for our Bayesian hierarchical models. The key benefit of the Bayesian hierarchical modelling structure that we applied is its capacity to include varying temporal relationships across catchments, which we identified as a critical consideration when exploring temporal variability of water quality in large regions (as seen in Guo et al. 2019). In contrast, modelling on merged catchments is unable to represent how temporal variability differs across catchments. We believe that this specific comment on the calibration on*

merged catchments can be resolved by improving the description and justification of the Bayesian hierarchical modelling approach in the relevant Method section (Section 2.1).

2) *If your comment is referring to the impact of nested catchments in our models, we would like to clarify that most of the 102 catchments that we used in this study were independent (as seen in Figure 1 in the manuscript). Therefore, our dataset is not suitable to answer this question.*

2.6 Spatial and temporal distributions of validation errors, their relationship with model development alternatives.

We are currently running 50 replicates of cross-validation for the model developed for each constituent. Once completed we will investigate the spatial and temporal distribution of errors from this more comprehensive validation and update the results section accordingly.

Specific comments:

3. Lines 15-16: In my opinion it is not the lack of understanding, but the lack of information. The effects of many key controls on water quality are well understood, albeit in an isolated, idealised context. It is clear, for example what certain polluting sources (like a WWTP effluent, a plot of arable land, etc.) do, how different landcover types affect the transport of pollutants along a specified pathway. The problem with the modelling of stream water quality on the (sub)catchment scale is that numerous key factors and controls act together and in practice there is no hope to get relevant information on all/most of them. That's why detailed and dynamic models fail on all components except those that behave quite simply and are not affected by too many factors. The challenge of modelling is to include the relevant factors AND the necessary information about them. So I would rephrase the sentence to mention that despite the long history of research there are too many key controls and very high complexity in both space and time compared to the available information.

Thank you for the thoughts. We agree that lack of information is a critical issue, because reliable information is the basis for gaining new understanding and/or validating existing understanding. However, we also cannot ignore important limitations in the current understanding of water quality behavior across multiple catchments in large regions, which agrees with what you have summarized – 'The effects of many key controls on water quality are well understood, albeit in an isolated, idealised context'. Certainly, at some catchments we have much better understanding of locally specific water quality mechanisms, which is supported by detailed data and local knowledge (i.e. information). However, this understanding is limited in transferability to other catchments as well as to inform the development of water quality models in other catchments. In addition, current understanding also tends to be focused on characteristics such as land use rather than natural catchment characteristics. These limitations are especially important when the interest is in a large geographical region across multiple catchments. For example, from conceptual understanding we would expect surface flow to enhance transport of sediments, but we have not well understood: a) the relative importance of surface flow effects compared with other key factors of water quality e.g. sub-surface flow and other climatic conditions etc., and how all the key controls interact with each other; and b) the varying extents to which surface flow influences sediment concentration between catchments.

As in the above example, developing such large-scale water quality models across catchments involves the identification of the key explanatory variables at larger scales, which would be ideally developed from more extensive information, but often only limited data exist. Therefore, a key innovation of our two preceding studies is to sift through many potential explanatory variables that we have from conceptual understanding to identify the more important ones for building a parsimonious predictive model at large scales (Lintern et al., 2018b; Guo et al., 2019). As a step forward, this study illustrated

good ability to represent spatio-temporal variability in water quality can be achieved based on understanding developed with limited information, at a regional scale of over 200,000 km^2 and across more than 100 catchments.

Therefore, we suggest that lack of information and lack of understanding should both be discussed as the key limitations to modelling catchment water quality, especially at a regional scale and across catchments (as the focus of this study). To address this comment, we propose to add more elaboration in the Introduction on the tradeoff between having good understanding and at a large scale, as the two critical requirements for modelling water quality at a regional scale. We believe that these revisions will help to clarify the knowledge gap that we address i.e. the need for better modelling capacity at large scales.

4. Line 16: Even if there would be a lack of understanding (which I doubt, see previous comment), how would this issue be addressed by a Bayesian statistical model? Statistical models build on covariance instead of causal relations and therefore are rarely suitable for modelling conditions that are different from the calibration dataset in any significant aspect âA˘T which is the primary objective of most modelling exercises.

Firstly, we believe there is a lack of both information and understanding, as explained in our response to your Comment #3.

We completely agree with you that our modelling approach does not improve our understanding of causality at all, but it still allows us to make better predictions, which is the aim of the paper as we clarified in our response to your Comment #1. Bayesian hierarchical approach enables us to build better empirical models that allow for differences in parameter relationships to exist for individual catchments. This is a key advantage for modelling over large geographical regions across multiple catchments which physically-based models struggle to achieve.

We believe that our proposed updates in the Introduction in response to Comments #1 and #3 would better clarify the key study objectives and provide more evidence to support the knowledge gaps, specifically via:

- *Clarifying the key study objective as to develop statistical models that can predict spatial and temporal variabilities of catchment water quality (Re Comments #1).*
- *Providing more discussion on the tradeoff between having good understanding and at a large scale, as the two critical requirements for modelling water quality in a regional scale (Re Comments #3).*

These changes which will also provide better justification for applying the Bayesian hierarchical model, which helps to address this comment.

Regarding your last comment on modelling different conditions, we believe that it is challenging for all fitted models (including calibrated process-based models) to predict well for conditions that are different from the calibration dataset.

5. Line 20: Please mention how FRP relates to the more commonly known Soluble Reactive Phosphorus (SRP).

FRP (Filterable Reactive Phosphorus) is defined as 'Reactive Phosphorus for a filtered sample to a defined filter size (e.g. RP(<0.45 µm))', which is equivalent to SRP (Soluble Reactive Phosphorus) when the same filter size is referred to (Jarvie et al., 2002).

- *Jarvie, H. P., Withers, J., and Neal, C.: Review of robust measurement of phosphorus in river water: sampling, storage, fractionation and sensitivity, Hydrology and Earth System Sciences, 6, 113-131, 2002.*

We use the term FRP following the terminology being used in the overall research project which this study belongs to. It is also the terminology used in the water quality database which we extracted the study datasets from (i.e. Victoria Water Measurement Information System, available at: http://data.water.vic.gov.au/; the sampling method and terminology definition for this dataset are documented in the Victorian Water Quality Monitoring Network and State Biological Monitoring Programme (1999):

- *Australian Water Technologies: Victorian Water Quality Monitoring Network and State Biological Monitoring Programme: Manual of Procedures, 1999.*

To avoid confusion, we will clarify the naming convention of FRP along with the more commonly used terminology in literature (SRP), when FRP is first introduced in the manuscript (L121).

6. Line 21: The abbreviation of "NOx" is not the best choice, as this is a widely known name of the air pollutant group of gaseous nitrogen oxides. Why not "NOi" or something else?

NOx refers to nitrate-nitrite (NO_{3-} + NO_{2-}) in our study, and this definition has been widely used in water quality research, e.g.,:

- *Bunn, S., Abal, E., Smith, M., Choy, S., Fellows, C., Harch, B., Kennard, M., and Sheldon, F.: Integration of science and monitoring of river ecosystem health to guide investments in catchment protection and rehabilitation, Freshwater Biology, 55, 223-240, 2010.*
- *Eyre, B. D., and Pepperell, P.: A spatially intensive approach to water quality monitoring in the Rous River catchment, NSW, Australia, Journal of Environmental Management, 56, 97-118, https://doi.org/10.1006/jema.1999.0268, 1999.*
- *Bruland, G. L., Hanchey, M. F., and Richardson, C. J.: Effects of agriculture and wetland restoration on hydrology, soils, and water quality of a Carolina bay complex, Wetlands Ecology and Management, 11, 141-156, 2003.*

We prefer to keep the term NOx to maintaining consistency with the overall research project and related papers. NOx is the terminology that has been used in the water quality database which we extracted the study datasets from (i.e. Victoria Water Measurement Information System, available at: http://data.water.vic.gov.au/, the terminology and relevant definitions are provided in Victorian Water Quality Monitoring Network and State Biological Monitoring Programme (1999):

- *Australian Water Technologies: Victorian Water Quality Monitoring Network and State Biological Monitoring Programme: Manual of Procedures, 1999.*

7. Lines 21-22: Yes, the model described variation, but above an improvement of understanding is promised.

As explained in response to your Comments #1 and #2.3, improving understanding is not the key focus of this study, but instead, we focused on developing models to predict spatial and temporal variabilities in stream water quality. We will work throughout the abstract and the manuscript (mainly the Introduction) to improve clarity of the study objective.

8. Lines 29-30: How would a statistical model include those mechanisms that govern non-conservative constituents? Such a development would indeed be a major step forward, but it is definitely not trivial.

In a statistical modelling framework, this could be achieved by considering additional predictors that are related to the key processes that affect the non-conservative constituents and biogeochemical processes (e.g. DO, channel habitat condition, microbial activity in soils etc.) without major changes of the model structure. Another option is to use non-linear structures that attempt to characterize the processes more directly. This sentence within the abstract intends to provide only a brief introduction of potential model improvements. To address this comment, we will add more details to the relevant discussions in Section 4.1 (Implications for statistical water quality modelling).

9. Line 32: High frequency data often reveal phenomena that are typically not parts of models and therefore model performance further declines.

Great point. High-frequency data can be helpful, but only to the point where they do not require much more complicated model structures to account for the fine scale temporal structure, otherwise these higher frequency data will contain temporal variation patterns that are not explainable by the driving data that we have. To avoid confusion, we will delete this statement from the abstract. We will elaborate more on the utility of high-frequency data in our modelling framework in Section 4.2, and provide more comprehensive discussions on the benefit/loss that we can get from using high-frequency data with the current model structure.

10. Line 33: Besides the classical landuse, agricultural activities (ploughing, fertiliser/pesticide application, livestock handling practices, etc.) would need to be known too.

This is an excellent point, which we are also planning to include in future model improvements. However, considering landuse and land management activities at the large-scale that we modelled for requires an extensive amount of good quality datasets that are currently not available. We expect such lack of information to be improved with novel data collection and/or systematic interviewing approaches in the future. To address this comment, we will include land management with some brief examples in this sentence in the abstract and expanding relevant discussions in Section 4.2 (Implications for water quality monitoring programs).

11. Lines 40-42: Unpredictable variability does not preclude management. Robust measures can address issues without having to predict the full dynamics. It is well known that the elimination of pollution sources and artificial hydrological factors improves water quality. If the statement in lines 40-42 was true, water quality management would not exist yet.

This is good point that practical management decisions are often made with low predictive capacity. However, here we were not aiming to criticize such management practices, but to highlight how management would benefit from better understanding and prediction of variabilities. The fact that we are able to manage water quality with limited prediction capacity does not suggest that improving modelling capacities is an unnecessary effort.

To better clarify this, we propose to revise the sentence:

From: 'However, our ability to manage and mitigate water quality impacts is hampered by the variability in water quality both across space and time, and our inability to predict this variability'

To: 'Effective management and mitigation of water quality are key to reduce these impacts. High variability in water quality both across space and time challenges assessment of management options.

Improving modelling frameworks to help predict and interpret this variability provides better tools for such assessments.'

12. Lines 42-46: This is a bit lengthy description of the high variability in both space and time. Please consider compressing.

We agree that this is a long description that might not be necessary for experts in this field. However, considering the broad readership of HESS, we believe that it is necessary to provide all these details. These are particularly helpful for the readers to learn the background and to understanding reasoning behind the spatio-temporal structure that we used to model water quality.

13. Lines 46-51: Briefly, there are allochthonous and autochthonous emissions and both are subject to transport. Please consider compressing.

We will condense this while mentioning the three key processes, which we consider as important background information for the broad readership of HESS as a multi-disciplinary journal (as also explained in our last response).

14. Lines 55-59: This listing is somewhat odd. Emission dynamics are completely missing, others are a bit over-detailed and supported with arbitrary references (is the importance of temperature only known since Robert and Mulholland, 2007?).

Thank you, we will improve the emphasis on emission dynamics, specifically on water quality variation due to changes in land use and land management etc. We will also reduce the discussion on hydro-climatic factors and support it with more appropriate references.

15. Lines 60-62: This sentence contradicts the abstract statement (lines 15-16). Water quality modeling faces high epistemic uncertainty, unpredictable variability stems rather from an information gap than the lack of understanding. And what do you mean here by "larger scales"? And please include why effective policy and mitigation need information on variability.

We will re-phrase the sentence to highlight that current understandings remain largely at a conceptual level and/or are specific to a catchment, as explained in our response to your Comment #3.

Modelling capacity at 'larger scales' refers to the ability to model across multiple catchments over large geographical regions. We will better clarify this by highlighting 'multiple catchments' in this sentence.

Better ability to predict variability in large scales would inform policy and mitigation via multiple pathways, such as: a) informing hot-spots; b) identifying trends/changes in water quality and attribute them to potential causes; c) identifying unexplained variability and thus potential future improvements needed in monitoring and modelling. We will include these discussions to illustrate how management can benefit from better abilities to model water quality variability.

16. Lines 66-69: It would be worth to mention that most statistical models have weak explanatory and predictive power and therefore it is difficult to use them for designing management interventions.

Thank you, we will include this in building up the knowledge gap.

17. Lines 71-72: Please check and fix this sentence, by e.g. deleting "can" or any other way.

We will delete 'can'.

18. Lines 74-80: After mentioning management so many times above, one would expect a brief summary about the requirements of managers against water quality models plus a sentence in the objectives on how the current model would fulfil these.

Good suggestion. We will thoroughly revise the manuscript and make sure that the key requirements for the model to benefit management are mentioned in the Introduction, and further strengthened by relevant sections in the Results and Discussion.

19. Line 103: Please fix "Beyesian".

This will be corrected during revision.

20. Line 112: Please delete "however". Either you describe data processing or not. The present formulation suggest that you don't want to describe it, but later âA˘T reluctantly ˘ âA˘T still do so. ˘

We will delete 'however'.

21. Line 132: Please briefly mention the forms and indicators of landuse considered among the drivers, because these are non-trivial.

As stated in L136, details of all potential predictors are provided in Table S1 in the supplementary materials. There are 50 potential predictors that we included in the predictor selection process, so they are not individually introduced in the main text. In addition, as understanding water quality spatial variability are not key focuses of this study (as in our responses to your Comment #1) we would keep the descriptions of the relevant approach brief, more details have been presented in Lintern et al. (2019b)

22. Line 143: You mean "area-specific streamflow"? Streamflow also has the unit of volume/time.

Yes, streamflow has the unit of volume/time. We will replace the original phrase 'streamflow (mm d^{-1})' here with 'catchment-average runoff (mm d^{-1})' to avoid confusion.

23. Lines 144-149: How did you convert 2D climatic data to soil moisture? This must have included a complete soil hydrological model, but no hints are given in the main text.

The 2D climate dataset was provided by the AWRA project by the Australian Bureau of Meteorology (Frost et al., 2016). It included the average percentage volumetric water contents for the root zone (at 1m depth) and the deep zone (deeper than 1m). We will add these details to better clarify the data information during revision.

- *Frost, A. J., Ramchurn, A., and Smith, A.: The bureau's operational AWRA landscape (AWRA-L) Model, Bureau of Meteorology, 2016.*

24. Lines 156-157: Low flow days often mean the periods of concern with regard to water quality. What was the case here?

Please note what we removed were not low-flow days, but days with zero (no) flow – during which it was impossible to take water quality samples. This is clearly communicated in L156: 'Water quality records corresponding to days with zero flows were also excluded from further analyses'.

25. Lines 162-166: This means that you conditioned the transformation on the dataset. Since the predictive nature of the model is emphasised, please explain the procedure of including new catchments. What to do when the new data suggest a different transformation parameter?

To model new catchments within the study region, we would expect that they follow the same statistical relationships as reflected in our models and thus the transformation parameters (along with other model parameters) to remain the same. However, we still recommend assessment of the statistical properties of the new input datasets (i.e. the key factors controlling spatial and temporal variabilities) and the water quality datasets. The calibrated model can only be applied directly if the statistical properties of the new dataset are similar to those of the calibration dataset.

For new catchments out of the regions, we do not recommend direct application of the calibrated models (including parameter values), since they would best represent the key water quality controls only for the calibrated region. It would be possible to apply this modelling approach in a new region to inform water quality prediction, which however, requires extensive selection of key predictors and model calibration, as what we have addressed with this study and the two preceding ones (Lintern et al., 2018b; Guo et al., 2019).

For either cases, if the new data suggest a transformation parameter that is substantially different to that in our model, then we recommend re-calibration of the model.

We will discuss these more in Section 4.1 regarding future applications of this modelling framework.

26. Lines 172-174: A random forest approach could have been an alternative for the selection process.

This is true, but it is a choice of approach, rather than the only approach. The predictor selection processes were developed in our two previous companion studies (Lintern et al., 2018b; Guo et al., 2019), from which the key spatial and temporal controls were already selected. These models presented in this study were developed using those key controls previously identified, without any additional predictor selection processes. We also believe that this comment would be addressed by our proposed revision to address your Comment #1 and #2.3, which involve a through revision of the Introduction to help better clarifying the focus of this study.

27. Lines 179-183: Aren't these results? Since management is emphasised in the introduction, how would you reflect on the final set of key factors? Climate is close to impossible to manipulate, temperature, soil moisture and streamflow are difficult. Why no direct human factors other than landuse?

These are not results from this study but instead, from our two previous companion studies (Lintern et al., 2018b; Guo et al., 2019). As explained in response to your Comments #1 and #2.3, the two companion studies focused on identifying key factors that influence spatial and temporal variabilities in stream water quality, whereas this study focuses on developing models to predict spatial and temporal variabilities in stream water quality. We have had extensive discussions in the two companion studies on each key control identified and the potential implications for catchment management, which were thus not repeated in this study. As proposed previously (responses to your Comment #1 and #2.3), we believe that a through revision of the Introduction will help clarifying the study focus better and thus addressing this comment too.

28. Lines 194-196: What is the rationale behind the half-normal prior? What is the advantage compared to an exponential? The half-normal suggests that relatively small standard deviations are equally likely, while the exponential prioritises as small std. deviation as possible. Please justify your choice.

When no a-priori knowledge on the distribution of a parameter is available, the prior distribution should be as minimally informative. Gelman (2006) demonstrated that a Gamma prior on precision

among exchangeable units (which we consider as the equivalent of using an exponential prior in this context) is actually highly informative and can skew results. His recommendation was the half-normal uninformative prior distribution for the standard deviation term in a linear Bayesian hierarchical model. We will add these justifications during revision.

- *Gelman, A.: Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper), Bayesian Anal., 1, 515-534, 10.1214/06-BA117A, 2006.*

29. Lines 212-214: This is a rather extreme test, why do you expect the model to describe the below-LOR data, after excluding all of them from the calibration dataset. Would a good fit mean that below-LOR data follow the same rules as above-LOD data do?

We agree that this is an extreme test, but it provides a useful perspective in model performance assessment. 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.

We agree with your interpretation that 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 (i.e. they follow the same rules). We will add this interpretation here to better highlight the utility of this specific performance evaluation.

30. Line 217: The verb "suggested" sounds weird to me here.

Presumably you are questioning the validity of using 'suggest' together with a quantitative measure. To address this, we will replace 'suggested' with 'quantified'.

31. Lines 238-240: FRP is a subset of TP. TP has complicated relations to TSS. The FRPTP relationship is governed by several (fast) biochemical processes simultaneously. Consequently, it is no surprise that FRP is hard to model without considering all these intricate interactions. By the way, a negative NSE suggests that the model entirely failed to capture any of the real dynamics (negative NSE means that a constant model at the mean would perform better).

We agree with your opinions. However, please note that this is the Results section where we refrain from providing extensive discussions. Later in the Discussion section (Specifically 4.1), we have commented on the poor performance of FRP and have specifically discussed the model limitation for representing biochemical processes for FRP.

32. Figures 2-3: It would be great to see some visualisation beyond 1:1 plots in transformed space (of unknown transformation parameters unless one digs them up from elsewhere).

Please note that all transformation parameters have been presented in Tables S3 and S4 in the Supplementary Information (which have been introduced in L168, Section 2.2).

As in response to your Comment #2, we propose to present more results to summarize different aspects of model performance, and also to illustrate model utilities that are useful for catchment management. Specifically, 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.*

33. Lines 268-269: This sentence is not necessary, the section title tells the same.

Thank you, we will delete the redundant sentence.

34. Lines 269-270, 273: Please delete the "Note that . . . in Sect. .3.1." sentence and add "We exclude the FRP model from the analysis due to its poor performance (section 3.1)." into Line 273 after "monitoring sites.".

We will address this in during the manuscript revision.

35. Tables 1-2: These tables are all about calibration indicators, and not the subject of the model. These could be moved to the SI. Why not showing something about the factors? The introduction promised filling some knowledge gaps yet we do not learn about anything except performance indicators (and later the influence of drought on them in table 3).

As explained in response to your Comments #1 and #2.3, the two companion studies focused on identifying key factors that influence spatial and temporal variabilities in stream water quality, whereas this study focuses on developing models to predict spatial and temporal variabilities in stream water quality. We have had extensive discussions in the two companion studies on each key control identified and the potential implications on catchment management, which we would not repeat in this study. As proposed in these responses, we believe that a through revision of the Introduction will help clarify the focus of this study better. In addition, as responded to Comment #1, we will include more results to summarize on how the temporal variations of water quality vary spatially, as this is a new finding that has not been reported in preceding studies.

Also, as in response to your Comment #2, we propose to present more results to summarize different aspects of model performance, and also to illustrate model utilities that are useful for catchment management. These results would be more aligned with the key objective of this study. Specifically, 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.*

36. Line 324: The Results section is over, yet the roles of "key controls", the proportions of "inherent randomness" both remain untold. The primary value of such a model is its information content, which is embodied in the relationships that turn inputs to outputs using the parameters. Model performance indicators are important too, but in a secondary sense: they help to assess the quality of information that can be obtained from the model. Here the reader learns about the model performance in various cases, yet the lesson can't be learnt. What governs the different water quality variables? Are there covariations between the variables? Are certain models similar to others? Are errors clustered in certain situations? Which environmental factors influence the variables, how sensitive are they to the most important one? Etc.

As explained in responses to your Comments #1, #2.3 and #36, the two companion studies focused on identifying key factors that influence spatial and temporal variabilities in stream water quality, whereas this study focuses on developing models to predict spatial and temporal variabilities in stream water quality. In these two papers we have extensively discussed the following topics: the key controls of water quality, there individual roles, interactions and how they can inform management. Therefore, we are not repeating or adding new discussions regarding key controls of spatial and temporal variabilities in water quality. As proposed in these previous responses, we believe that a through revision of the Introduction will help in clarifying this better.

On the other hand, as mentioned in the response to Comment #1, while developing this spatio-temporal model in this study, we have obtained new understanding on how the temporal drivers of water quality vary spatially, which has not been explored in the two preceding studies. To address this, we propose to include more results and discussions on how temporal relationships vary spatially.

Regarding your comment on showing model performance in various cases, we believe this can be addressed with the few additional results that we propose to add to summarize different aspects of model performance (as in response to your Comment #2). Specifically, 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 the calibrated dataset.*

We believe these proposed revisions on the Introduction, Results and Discussion could improve the clarification of the study objective as well as the coherence of the entire paper.

37. Lines 333-334: Would be more positive to start with the opportunities and afterwards with limitations.

Thank you. We will summarize the key model capabilities, contributions and opportunities at the start of Section 4.1.

38. Lines 334-335: Or when the variability is high and explanatory power is weak. Very low FRP values could be much better simulated given that the model knows all the influencing factors and processes.

The LOR issue is the first factor that we identified to influence model performance in Section 4.1. We have discussed the point you raised (i.e. model limitations on representing processes for FRP) in the subsequent discussions from L360.

39. Line 336: This can also be by chance. TKN and EC are “more conservative” than the others, and have much weaker relations to sediment.

Good point. Since this paragraph focuses on the impact of LOR data on model performance, we will add some discussions on how model performance can be affected by the degree of non-conservativeness of different constituents later in Section 4.1 (from L360).

40. Lines 347-359: It is true that transformation increases the distance between distinct values close to the numerical resolution of data, which violates the linearity assumption. But when

you do not transform, linearity is violated by default (as one of the aims of transformation is to reduce nonlinearity). Besides the alternative model structures mentioned, a practical solution is to perturb the data with random small values (small fraction of numerical resolution), which dissolves the discrete bands of the low values without significantly altering the data. This is basically the same as "measurement noise" beyond the resolution of the time-series.

Thank you for the interesting idea, we will include this in the discussion.

We note that one practical issue with this approach would be the requirement of 'data replicates' to reach some convergence of the calibrated model – since different perturbations of the raw data would lead to different calibrated models, particularly where the 'categorical issue' occupy high proportions of the data (e.g. TSS and FRP).

41. Lines 360-361: Yes, this was obvious from the start. That's why the "positioning" of the model study is not optimal. The applied methodology tested whether temporal / regional differences could be replicated by a simple statistical model that lacks any mechanistic background. The exposition of knowledge gaps, management-relevant factors, general predictive power for ungauged catchments create expectations that simply cannot be fulfilled by this model. A lot of mechanistic knowledge is available for these water quality variables, no single bit of this knowledge is reflected by the model structure. A more realistic context would have been to investigate the overarching patterns in this region of Victoria, emphasising that the model only considers emissions only implicitly, through landuse, which in turn assumes similar human activities in the same landuse type. The results are completely in line with previous experiences, more conservative and less sediment-related variables are easier to predict than the others. The model can be a valuable predictive tool, but only in the region of calibration and only for those water quality variables, for which have the model performed acceptably.

Firstly, we'd like to clarify again that this study focuses on developing integrated models to predict spatial and temporal variabilities in stream water quality, which we will better emphasize in the revised manuscript (as detailed in responses to your Comments #1, #2.3 and #36). We will also improve clarification in the Introduction, that the key knowledge gap this study addressed was the lack of statistical modelling approaches that are suitable for large-scale application (as we responded to your Comment #3).

We also propose to present more results on the model capabilities, to strengthen the key study objectives and to better highlight the values of these models to catchment management (see our responses to your Comment #2). Potential topics include:

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 the calibrated dataset.*

We believe that the abovementioned revisions will improve the alignment of the key knowledge gaps, the study objectives and the current results presented.

Furthermore, although our models consist of parsimonious linear relationships between water quality variables and their predictors as opposed to physically-based models governed by more complex equations, our model structures are informed by plausible conceptual relationships between potential predictors and water quality variables. Therefore, they should not be considered as 'simple statistical models that lack any mechanistic background'. The potential predictors of the models were determined following an extensive literature review and consultation with our industrial partners who are all actively working on catchment management (as described in L129-133, the consultation has not been described in the current manuscript but will be added during revision).

These potential predictors then went through an exhaustive predictor selection processes to extract key predictors that explain the most variabilities in water quality, which were detailed in our two companion studies (Lintern et al., 2018b; Guo et al., 2019). In the two studies we have also reviewed extensive literature to confirm the plausibility of those results from a process perspective. Considering all these process-informed decisions in model development, together with a rather comprehensive Bayesian hierarchical structure that we applied over a large region (>200,000 km^2) across multiple catchments, we identified a clear misunderstanding here to consider these models as 'simple statistical models that lack any mechanistic background'. For the same reasons, we also politely disagree the comment that 'A lot of mechanistic knowledge is available for these water quality variables, no single bit of this knowledge is reflected by the model structure'. To avoid further misunderstanding on this, we will highlight the use of process-based evidences when introducing the predictor selection process in Method.

This study is the first time that water quality in the study region (and even world-wide) has been modelled over such a large geographical extent with a statistical approach. We believe that even though the performances of these models are 'as expected', the modelling experiences and understanding on capability of large-scale statistical water quality models will provide very useful contribution to existing studies, as these have been predominantly focused on physical models that operate at catchment scales. We would also add these points to the Discussion to highlight the study contributions.

42. Lines 364-369: Making the model more detailed can potentially lead to a dead end. Non-linear statistical model structures may perform a bit better, but need more data for a meaningful calibration and still often lack the mechanistic background, and are much more complicated numerically. Adding descriptions of different mechanisms to the model either moves it towards a deterministic direction, which is a wrong way for this spatial and temporal scale because data will anyway appear to be at least partly random due to the lack of information on all relevant drivers, or leads to a stochasticdynamic model, which is extremely complicated and difficult to calibrate.

Thank you for sharing the valuable opinions. Within a statistical modelling framework, a most feasible option would be to include additional predictors that are related to the key processes that affect the non-conservative constituents (e.g. DO, channel habitat condition etc.). Alternatively, non-linear structures can also be used to characterize the processes more directly. However, as discussed in our response to Comment #3, we also need to be aware of the trade-off between the complexity of model for detailed process representing versus the spatial of scale that the model is capable to present. To address this comment, we will add some examples to illustrate potential improvement of this modelling of biogeochemical processes within a statistical modelling framework, while also noting this trade-off between model scale and complexity.

43. Lines 372-373: If this was an issue, why don't we learn about the "real-world" (=nontransformed) model accuracy earlier? The NSE values and the figures are all in transformed space, so it is difficult to judge what these mean for the practice.

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 will add these justifications to Section 4.1.

Our transformed models focus more on proportional errors instead of absolute errors, since the latter is less important at high concentrations in practice. We will better clarify this around Line 375-377 along with the revision proposed for the comment below.

44. Lines 375-377: I don't understand this example. Completely usual floods often bring much more sediments in almost pristine mountain catchments. Why would such an event be an alarm for management?

Agreed, we will amend this explanation for proportional changes.

45. Lines 377-379: How? This should have been the main topic if the logical line of the Introduction was followed. How strong is the predictive power of the calibrated models considering practical needs? Are they suitable for real forecasting either for the far future or for shorter periods during operative management?

As in the response to your Comment #2, we propose to present more results to summarize different aspects of model performance, and also how these model capabilities can be useful for catchment management, as well as planning and policy making. Specifically, 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 this comment can be addressed once we have illustrated the value of our models to catchment management with these additional results and relevant discussion.

46. 384-386: The references are "too new" for this statement in its present general form. Commercial solutions for online monitoring with <10 minute resolution is available for turbidity (proxy for TSS), temperature, EC, chlorophyll, dissolved oxygen since at least 20 years. Nutrient sensors are indeed newer, yet they are often not sensitive enough to yield meaningful data in surface waters (unless they are heavily polluted).

Good point, but we also acknowledge that these sensors are only made more accessible (i.e. cheaper with improved mass production) recently, for more wide application in practices. To address this comment, we will add more well-established literature to the list.

47. Lines 386-388: How would you apply remote sensing in stream networks? Except for larger rivers (and of course, lakes and reservoirs), these water surfaces are difficult to analyse because the number of "clean" pixels without any terrestrial or littoral influence is very low or even zero.

Although available information extractable from remote-sensing are limited only to larger rivers, which will certainly involve further development before operational uses. However, current remote sensing data could still provide valuable information which allows us to augment the temporal resolution of existing monthly data. For example, it is possible to make spatial inferences once a network structure is developed from the larger rivers. In addition, for small streams we might also be able to extract information from available drone-based monitoring data with much higher resolutions (e.g. cm scale pixels). To address this, we propose to add more details on potential approaches to use remote sensing data in our modelling framework in this discussion (Section 4.2)

48. Lines 390-391: There may be better (older, original) references for this. This is known since at least 30 years.

We will add more appropriate literature to the list.

49. Lines 391-397: Please remove, this is too case-specific.

Yes, these are all examples from Australia. However, we believe that this is a useful quick summary of data availability, as well as highlighting opportunities in some major regions for future water quality modelling in Australia. Due to the study region, we expect this study to attract many Australian readers who might find such information particularly useful. We believe that this comment could be addressed by broaden the examples to more international examples.

50. Lines 398-399: This is the exact reason why models fail despite the rather solid understanding of mechanisms (and this is a data or information gap and not a knowledge gap). Relevant, representative, and accurate data on such activities is close to impossible to obtain, even for smaller regions or shorter periods. Therefore, the temporal and spatial variability of these contribute to apparent "inherent randomness" and undescribed variance (the difference of NSE from 1) and weaken the predictive power of models. At the moment the solution to this issue remains an open question even for the past/present, not to mention the potentially changing practices of the future.

We completely agree and appreciate your concerns on the challenges with obtaining good information, while also acknowledge that much of these 'ideal' information are not currently available, especially at the modelling scale that we focused on (i.e. regional, across multiple catchments). We agree that this lack of information is not a trivial point and is thus worth highlighting here as a priority for future monitoring; and it is very likely that such lack of information can only be achieved in the future with novel data collection approaches. To address this comment, we will include more examples on land use and land management activities that are relevant to water quality, and the need of monitoring these potentially via improved data collection and surveying approaches.

On the other hand, as in our response to your Comment #3, we believe that modelling capacity is limited by the lack of both information and understanding. We would like to highlight that current understanding of water quality remains largely at a conceptual level when focusing at a catchment scale. In contrast, understanding is still highly limited to enable large-scale water quality modelling, regarding e.g., the relative importance of key controls for water quality, how these key controls interact and how they vary across space and time etc. While information is limited, it is still possible to advance these understandings to build better modelling capacities, as we illustrated and considered as a key novelty of this study. As also mentioned in the response, there is a trade-off between having good understanding and being representative for large regions. We will add relevant discussions on this as well.

51. Lines 422-423: Direct livestock input may increase concentrations during drought.

Thank you, we will add this point to the discussion with supporting references.

52. Lines 438-443: As the results of this study showed, this would be a hard job without implementing at least a few mechanistic features in the model. However, more features would require more data, potentially beyond the scope of the presented dataset.

The existing dataset would be useful to reveal many aspects of the proposed analyses. One way to conduct this is to assess the rainfall and streamflow time-series at individual catchments to identify specific periods of droughts (which tends to vary across catchments, see Saft et al. (2015)). We could then assess how the strengths and directions of statistical relationships between water quality and its key controls change over droughts. This analysis has not been performed as part of the study due to the different focus. However, we will briefly discuss possible approaches to address this comment.

- *Saft, M., Western, A. W., Zhang, L., Peel, M. C., and Potter, N. J.: The influence of multiyear drought on the annual rainfall-runoff relationship: An Australian perspective, Water Resources Research, 51, 2444-2463, doi:10.1002/2014WR015348, 2015.*

53. Lines 455-457: This is a crucially important sentence. I would add explicitly that the model is not only bound to the period, but also to the region for which calibration took place.

Agreed, we will add this in the revised manuscript.

54. Supplementary material: Figures could be structured better graphically. When 4x4 panel units are to be seen, please structure the figure so that the units get obvious. Please indicate the contents in the subfigure title. Print Box-Cox or log-sinh transformation parameters on figures or in the caption, because without knowing the strength of transformation it is difficult to judge the quality of fit.

Thank you for the suggestions. We will implement these in the revised manuscript.