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

Editor comments published on 22 Oct 2019

Thank you for responding to the four reviews of your manuscript. I appreciate the serious manner of how you have provided answers to comments and suggest that you revise the manuscript accordingly.

Nevertheless, a few aspects deserve more attention than what you have proposed. I list them below and recommend that you pay them due attention during the revision of the manuscript as well.

Thank you for your decision. We have thoroughly revised the manuscript as proposed in our previous responses to the four reviewers. We also carefully considered your additional comments and have revised the manuscript accordingly. We provide specific responses to each of your comment as below (with specific revisions shown in underlined text).

Transformation of the data:

1. Several reviewer comments questioned aspects of how using the transformed concentration values (R2: comment 12, R3: comment 2.1, 32, Rev. 3, comment 3.2). Your arguments to avoid comparing observations and simulations in the original space is not fully convincing: you argue that it was best to evaluate model performance in the transformed space because (e.g., response to Rev. 2, comments 2.1, 43) it was most informative and because absolute errors were less important in practice.

    1.1 First, you don't provide an argument WHY it should be most informative in the transformed space. Actually, inspection of Fig. S13, reveals obvious model biases (even for the site-specific average concentrations, if I interpret the figure caption correctly). A careful look at Fig. 2 and 3 show similar deficiencies (e.g. systematic underestimation of high concentrations for TSS, TP, FRP, and NOx. However, these deviations are much less conspicuous than in the transformed space. This holds especially true because some of the chosen transformations are very non-linear making it very difficult to have a sense for the actual meaning of the transformed values. Additionally, inspection of Fig. S13 for EC suggests that there might be two populations of catchments: one population is very well represented by the model (close to the 1:1 line), while the second is definitely off. This can hardly be seen with the transformed data. Do the catchment being off share some commonality?

    We agree with you that the untransformed plot can better help us to understand absolute model errors so we should discuss some results and implications of this. However, we also acknowledge that since the model was developed in a transformed space, performance evaluations in the transformed space would allow us to best explore a wide range of factors that can influence model performance (e.g. the LOR issue – now referred to as the 'detection-limit issue' in the revised manuscript, the limitation in simulating non-conservative constituents, and any changes in model performance across different monitoring sites and periods used for model calibration).

    To resolve this comment, we first improved the justifications in Section 2.2 (Model performance and sensitivity analyses) on why model performance assessments are presented in a transformed scale.

    - *L297: Since the model was calibrated in a Box-Cox transformation scale (see justification in Section 2.1.2), the Box-Cox transformation scale was used for model evaluation to enable a clear investigation on the influences of a wide range of factors that can influence model performance.'*

We also moved Fig. S13 to the main text to better clarify the back-transformed model performance – which becomes Fig. 5 and placed after the transformed model performance is shown in Fig. 4. Along with the figure we have added corresponding explanations on how the model performance is limited by back-transformation, as:

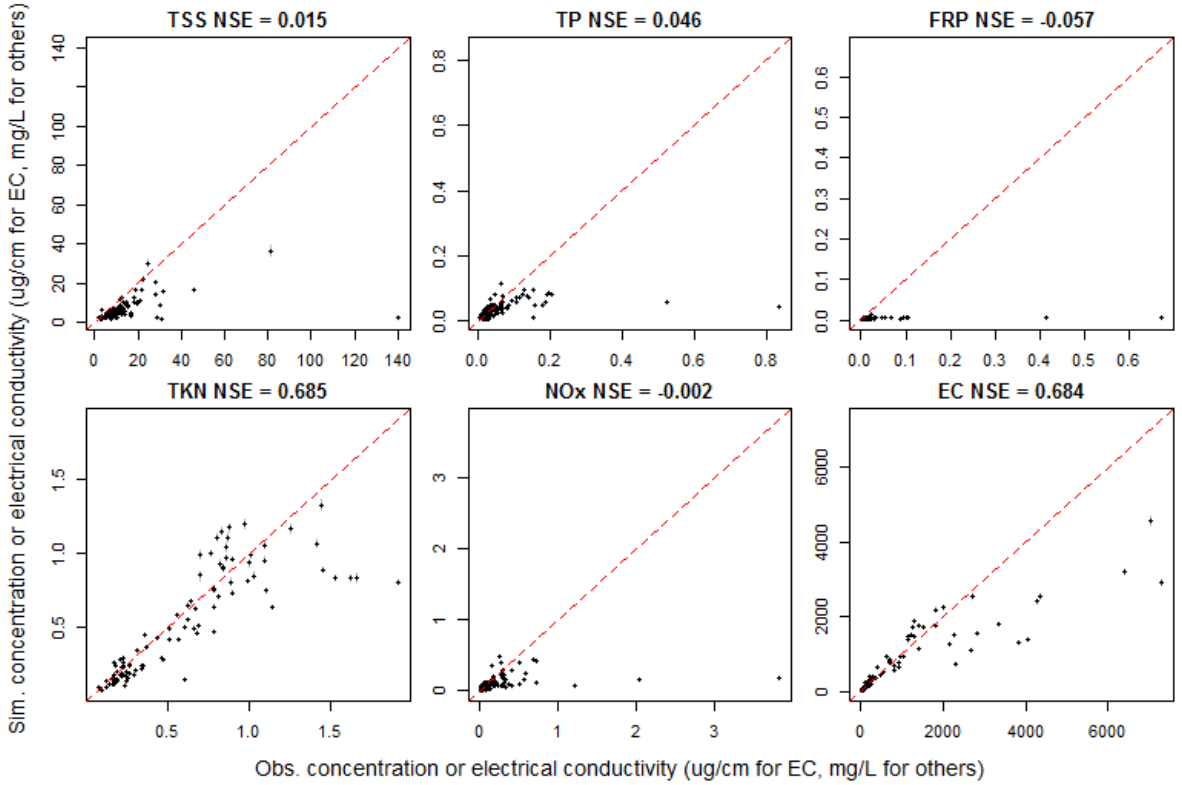


Figure 1. Back-transformation of the model simulations to the measurement scale emphasizes lack of fit for the highest concentrations, illustrated by simulated against observed site-level mean concentrations of each constituent in a back-transformed scale. The 95% lower and upper bounds of all posterior simulations shown in vertical grey lines. The NSE for each constituent is also shown and red dash lines show the 1:1 lines.

- *L421: 'At the back-transformed scale, the model shows greater biases for sites with higher concentrations (approximately the highest 10% sites for each constituent) (Fig. 5). This is not surprising as the model was fitted to a Box-Cox transformed space that reduces focus on high values and increases the focused on low values. This compromised its ability to represent sites with unusually high concentrations. The implications of the model having higher predictive capacity in the transformed scale is further discussed in Section. 4.1.'*

1.2 Second, the relevance of absolute errors is probably very context-specific. In some situations, practitioners do care about high concentrations and model uncertainty was important to them.

We agree with you that absolute values and high concentrations can be important to practitioners in many cases. We have added discussions the following discussions in Section 4.1 to clarify this:

- *L577: 'As previously noted, our model was developed in a Box-Cox transformed scale to ensure the validity of the statistical assumptions (see details on data transformation in Sect. 2.1.2), which shows limited performance for high constituent concentrations when simulations are back-transformed to the measurement scale (Figs. 4 and 5). However, our model approximately represents proportional changes in water quality, which can thus help managers to understand proportional changes to inform practical catchment management.*

Footnote: All Box-Cox transformation parameters for water quality constituents are approximately 0 (Table S4), which means that the transformations are similar to a log transformation.'

Regarding your concern on uncertainty, we have added the 95% uncertainty bounds to the back-transformed model performance shown in Figure 5 (previously as Figure S13), as well as the all relevant plots showing model performance in the transformed space (Figures 4, 6, 7 and 8). Note that for clear visualization, we did not add the uncertainty bands in Figure 2 which compares all model simulations with corresponding observations. The model uncertainty is generally very small, and no clear pattern of heteroscedasticity is observed.

1.3 It is important to note that a systematic model deficiency (e.g., under or overestimation in a certain concentration range) is not alleviated by transforming the data. However, it allows for better fulfilling distributional assumption for making statistical inference. Therefore, the transformations make sense. However, to proper and transparently present the model performance and the effects of transformations, more information needs to be provided (as also suggested by the reviewers):

Thank you, we respond to each of your specific suggestion as below.

1.3.1 Provide information on how you have determined the optimal log-sinh and Box-Cox parameters (L. 161, 164). What was the optimality criteria and how did you assess optimality (manual calibration, visual inspection of quantile plots etc.)?

We have added the equation of each transformation (Eqs. 7 and 8), with further details on approaches to determine the transformation parameters in Section 2.1.2:

- *L216: 'The GA package in R (Luca Scrucca, 2019) was used to identify the log-sinh transformation parameters (a and b) for each spatial explanatory variable that minimized the data skewness (i.e. symmetry is maximized) across all 102 catchments.'*
- *L223: 'For each variable, the optimal Box-Cox transformation parameter λ was identified using the car R package and a maximum likelihood-like approach. We first identified the optimal Box-Cox parameter λ using the data at each site (i.e. 21-year time-series). The averaged λ across all sites was then used to transform the data across all catchments together. This transformation approach ensured that all sites used a consistent transformation parameter.'*

In addition, we also summarized the assessment of the quality of transformation, as:

- *L227: 'All transformation parameters used are summarized in Tables S3 and S4 in the Supplementary Material. The transformation process has greatly improved the data symmetry and thus suitability for use in a linear model (the quality of the transformations was assessed via visual inspection in Lintern et al., 2018b; Guo et al., 2019; and summarized in Figures S2, S4 and S6 in the Supplementary Material).'*

1.3.2 Provide information on the distribution per site and constituent in Tab. S4. You may also consider to plot the respective distributions in the SI.

We have specified the between-site variation by the standard deviation of the lambda in Table S4.

We have added Figures S1-S6 in the Supplementary Materials for the distributions of (a) the raw data and (b) the transformed data for each of the six water quality constituents, all 50 potential spatial predictors and all 19 potential temporal predictors.

1.3.3 Complement Fig. 2 – 5 with the regression lines between observations and simulations and provide the slope estimates (including uncertainty).

Thank you for the suggestions. Firstly, we have added the 95% uncertainty bounds to the back-transformed model performance shown in Figure 5 (previously Figure S13), as well as

the all relevant plots showing model performance in a transformed space (Figures 4, 6, 7 and 8, although note that Figure 2 was not imposed by uncertainty bounds for clear visualization). The model uncertainty is generally very small with no visible patterns of heteroscedasticity.

However, we are unsure about the value added by having regression lines between observations and simulations in these plots, because:

1) *These regression lines can potentially affect the visualization of the 1:1 lines which are currently shown – which we believe are sufficient, more relevant and more direct visual aids for assessing over-/under-estimation of the model we developed.*
2) *These regression lines are also not directly related to our models. Different to what these regression lines show (i.e. relationship between observed and simulated concentrations), our models described relationships between WQ constituent concentrations with catchment landscape characteristics and temporal hydro-climatic and vegetation conditions. Adding regression lines between simulations and observations may thus confuse readers when interpreting the model evaluations.*

1.3.4 Clarify whether Fig. 3 and Fig. S13 correspond to the same data.

We have moved Fig. S13 to the main text (now as Fig. 5, after presenting the transformed results, as now in Fig. 4). These back-transformed results are introduced as:

- *L422: 'At the back-transformed scale, the model shows greater biases for sites with higher concentrations (approximately the highest 10% sites for each constituent) (Fig. 5). This is not surprising as the model was fitted to a Box-Cox transformed space that reduces focus on high values and increases the focused on low values. This compromised its ability to represent sites with unusually high concentrations. The implications of the model having higher predictive capacity in the transformed scale is further discussed in Section. 4.1.'*

1.3.5 Include one figure in the main text comparing observations and model results in backtransformed form. This could be Fig. S13 or a time series that you have mentioned several times (e.g., response 2.1 to Rev. 4).

This is addressed in our response to your last Comment #1.3.4

Further editorial comments:

2. L. 27: You focus here on improving the model fit for low concentrations. However, Fig. 2 and 4 suggest that the model is deficient in the low and the high concentration ranges. These systematic deviations should be addressed. If my interpretation was wrong, please provide a convincing argument why to put emphasis on the low concentrations. The argument mentioned above about the practical relevance that was less for high concentrations is not convincing. This very much depends on the actual context and some practitioners may be much more interested in high concentrations. Note that L. 154 – 155 would support this view as well.

Thank you for raising this issue. We acknowledge that the transformation issue has limited model capacity to predict absolute values for high concentrations. However, this is less of a concern is the model focuses on predicting proportional changes (e.g. as presented in a Box-Cox transformed scale) and when the interest in large-scale patterns instead of individual catchments. To address your comment and better highlighting the scope of this model, we first revised the abstract and the main text to highlight the model limitation on simulating absolute values. In the abstract we added:

- *L22 (abstract): 'The model is best used to predict proportional changes in water quality in a Box-Cox transformed scale, but can have substantial bias if used to predict absolute values for high concentrations.'*

We also revised the discussion on the implication of the transformation impacts on model performance to better clarify the limitations and recommended model usage for management as:

- *L577: 'As previously noted, our model was developed in a Box-Cox transformed scale to ensure the validity of the statistical assumptions (see details on data transformation in Sect. 2.1.2), which shows limited performance for high constituent concentrations when simulations are back-transformed to the measurement scale (Figs. 4 and 5). However, our model approximately represents proportional changes in water quality, which can thus help managers to understand proportional changes to inform practical catchment management.'*

We have also revised the justification to remove the below LOR (now referred to as the detection limit (DL) data) in the Method section. Specifically, we removed an inaccurate statement that the below-LOR data were removed from analyses because our model focused on the high concentrations:

- *L207: '…This was because the uncertainty in values below the DL would be amplified after transformation, which would influence the subsequent model fitting. Furthermore, those undetectable low concentrations were of less interest for management purposes. Water quality records corresponding to days with zero flows were also excluded from further analyses.'*

3. The data presented in the main text (e.g., Fig. 3 – 5) refer to site-specific mean concentrations across space. Of course, Fig. 4 and 5 represent such mean concentrations for different periods. But there is no information on how well temporal dynamics are captured at shorter time scales. Strengthening this temporal aspect as you mention several times is important.

We have added evaluations of the model capacity to represent temporal variability by adding the following results and interpretations:

- *Fig. 3, which shows the proportions of spatial and temporal variability within total observed variability, as well as the model performance in explaining each component of variability. These results indicate that the model performs much better in capturing spatial variability compared to the temporal variability.*

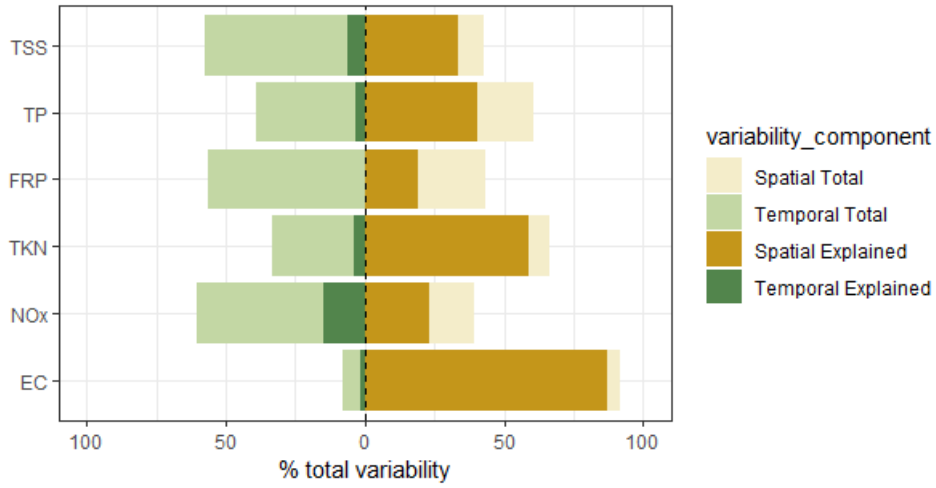


Figure 3. Observed spatial and temporal variabilities as proportions of the total variability (total width of each bar, 100%). The dashed line differentiates temporal variability (left side) with spatial variability (right side), and the darker colours highlight the proportions of spatial and temporal variabilities that are explainable by the model. All values were estimated in Box-Cox transformed space.

- *Fig. 6, which shows the simulated and observed temporal variability for each constituent, at the catchment where the model performs the best. These results further illustrated that the model largely underestimated temporal variability across all constituents, but is generally capable to represent long-term trend (except for FRP).*

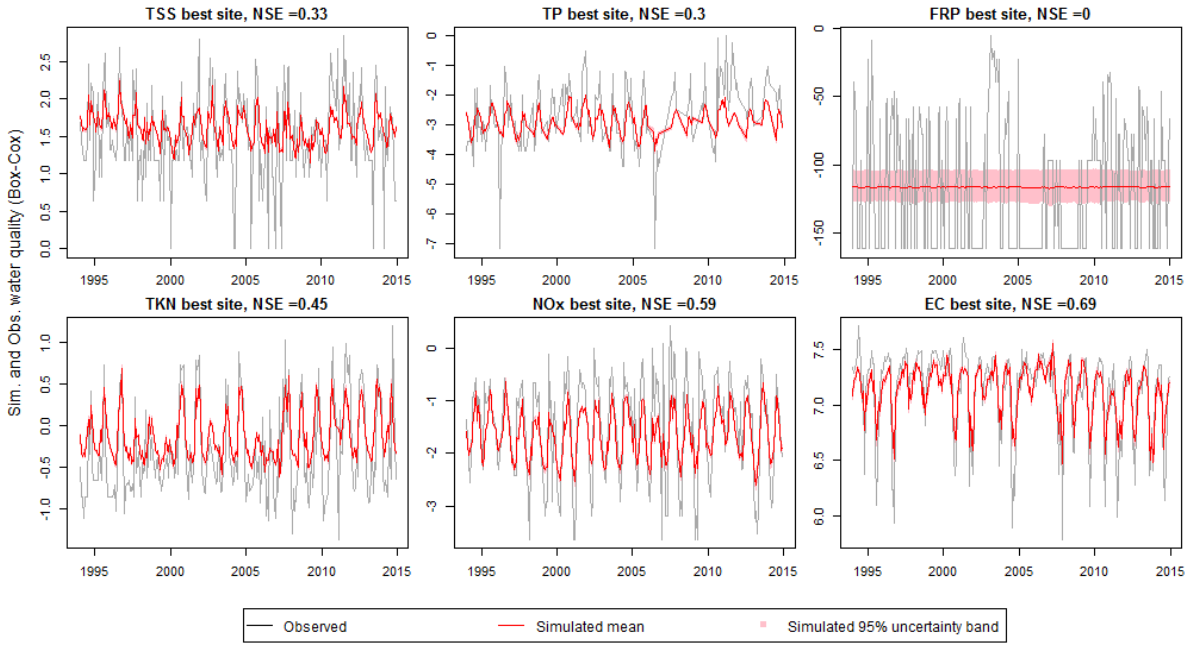


Figure 6. Model fit of the within-site (temporal) water quality variability, illustrated with the observed and simulated time-series for the best-performing site for each constituent. All values are presented in Box-Cox transformed space. The NSE for each constituent is also shown. The red line indicates the corresponding mean of all posterior simulations, while the pink bands show the corresponding 95% lower and upper bounds (only visible for FRP).

- *Table 4, which summarizes the proportions of observed positive and negative water quality trends that are recognized by the model. These results add further evidences to Fig. 6 on model capability to represent long-term trend.*

Table 4. Model ability to capture observed water quality trends across all monitoring sites for each constituent. The percentages of sites where observed positive and negative trends are captured by the model are presented separately. Values in brackets indicate numbers of sites where corresponding positive or negative trends are observed. For detailed estimation of these percentages please refer to Section 2.2.

Constituent	% positive trends captured	% negative trends captured
TSS	33.3 (12)	85.0 (20)
TP	82.1 (28)	16.7 (12)
FRP	47.1 (17)	55.6 (9)
TKN	81.1 (37)	40.0 (10)
NO_x	68.6 (35)	66.7 (27)
EC	82.6 (23)	77.3 (22)

2 In this context, I am not fully convinced of your argument not to discuss in some more details how the model simulates the drought effects (see Fig. R3). If you consider the results solid in Fig. 4 and 5 enough to be presented in the manuscript you have also to demonstrate what makes the difference in the parameters for different periods. This is simply reporting your findings. It is subsequently fair enough to critically mention that an over-interpretation isn't warranted because of model deficiencies.

We agree with you and have added Fig. 9 and related discussions in Section 4.3, where the drought impacts on sediments concentrations and the potential mechanisms are discussed:

- *L611: 'A further analysis of the calibrated model parameters for pre-, during and post-drought periods suggest that the effects of key spatial predictors do not vary much across periods (Figure S14). In contrast, the effects of key temporal predictors highlight a clear shift in the role of antecedent flow (prior 7-day flow) across different time periods (Figure 9). Specifically, the*

antecedent flow effects are mostly positive across catchments before the drought, and shift to mostly negative during the drought. After the drought, the antecedent flow effects have mixed directions among different catchments.'

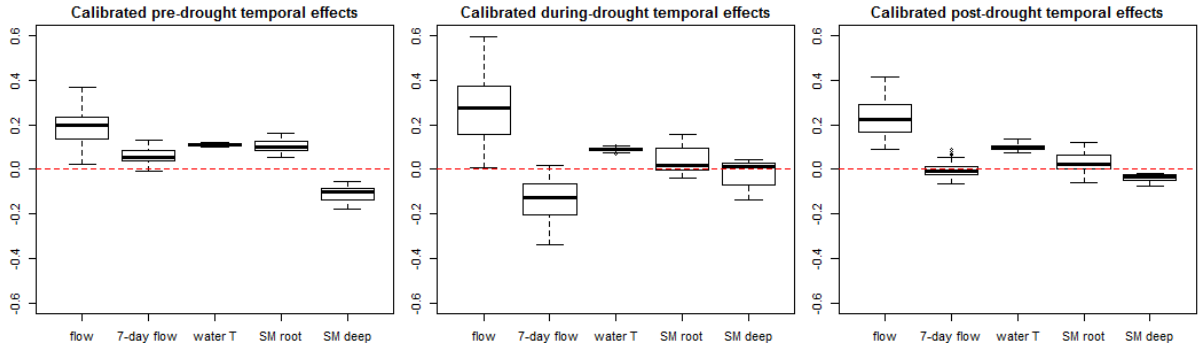


Figure 9. Effects of the five key predictors for the temporal variability in TSS across 102 sites, summarized by the posterior mean of the calibrated parameter values for each predictor (box shows values across all sites), from left: flow, 7-day antecedent flow, water temperature, root-zone soil moisture and deep soil moisture.

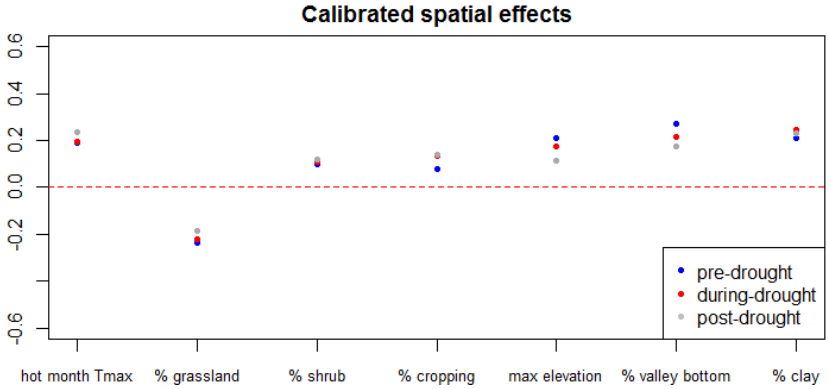


Figure S14. Effects of the seven key predictors for the spatial variability in TSS across 102 sites, summarized by the posterior mean of the calibrated parameter values for each predictor, to the pre-, during- and post-drought periods (differentiated by colour). The seven key predictors are, from left: hottest month maximum temperature, percentage catchment area as grassland, percentage catchment area as shrub, percentage catchment area as cropping land, maximum catchment elevation, percentage catchment area made up of valley bottoms, and average soil clay content.

After presenting these results, we have also added acknowledgement on the model deficiencies and thus recommended specific care in interpretation, as:

- *L616: 'Considering the limited performance of the TSS model (i.e. substantial under-estimation of temporal variability in Section 3.1), these changing relationships suggested in the calibrated parameters might be unreliable. However, this should not affect the reliability of the observed change in TSS since the drought (Section 3.3), which was based on the systematic differences of model fitting between different periods, revealing a broad-scale patterns across the state on the drought influences.'*

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

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 (our specific manuscript revisions are underlined).

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 have added recommendations for future implementations of our modelling framework to Section 4.1 (Implications for statistical water quality modelling) of the revised manuscript. Note that we have not included the model run time because it is highly dependent on the amount of data and the number of model predictors used – which have been determined/selected in our two preceding papers. Due to the highly specific nature of the run time and the multiple modelling procedures involved, we do not see model run time as a useful information to report:

- *L583: 'For future implementations, the established model structure and parameterization would be best suited to within the study region. Before performing new simulations (e.g. for new monitoring sites or for current study sites over a different time-period), the statistical properties of the new input datasets should be checked to ensure that they are similar to the calibration datasets. To model new catchments outside of the study region, a re-calibration of the model is required. This would involve extensive selection of key predictors and model calibration, much as performed in this study and the two preceding ones (Lintern et al., 2018b; Guo et al., 2019). A sufficiently long record length (e.g. 20 years) is ideal for such modelling, as it ensures a reasonable understanding of the temporal variability to be obtained.'*

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.

This is an excellent idea. To provide a more comprehensive understanding of model robustness to calibration datasets, we have increased the number of cross-validation replicates from the current 5 to 50, each of which used 80% monitoring sites for model calibration and the other 20% for validation. The results are summarized in Table 5.

Table 5. Comparison of model performances (as NSE) of the full model (Column 2) and the 50 partial models (Columns 3 to 5) with each calibrated to 80% randomly selected monitoring sites. Columns 3 to 5 summarize the mean, minimum and maximum NSE values across the 50 runs, where for each constituent, the top row showing calibration performance and the bottom row showing the validation performance (i.e. at the 20% sites that were not used for calibration).

Constituent	Full model	50 CV mean	50 CV min	50 CV max
TSS	0.225	0.413	0.376	0.439
		0.382	0.292	0.513
TP	0.433	0.461	0.427	0.501
		0.411	0.151	0.575
FRP	-1.92	0.168	0.067	0.232
		0.129	-0.078	0.272
TKN	0.658	0.654	0.622	0.670
		0.622	0.468	0.691
NO_x	0.216	0.453	0.414	0.489
		0.397	0.258	0.563
EC	0.907	0.893	0.882	0.903
		0.875	0.809	0.924

The results are introduced as:

- *L467: 'The calibration and validation results for the 50 partial models are summarized in Table 5 along with the performance of the full model calibrated to all 102 sites (see Figs. S6 and S7 in the Supplementary Material for detailed comparison of model residuals of the partial calibration/validation). Across constituents, the calibration performance of the full model was comparable with the 50 partial models. In addition, model performance is highly consistent between corresponding calibration and validation, with most differences in NSEs less than 0.1. These suggest that the spatio-temporal model performance is highly robust and unaffected by the choice of calibration sites.'*

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. However, during the revision, we have identified more factors that contributing to weak model performances other than the below-LOR data issue (see additional results presented in Section 3.2 and the enhanced Discussions in Section 4.1). Considering all the interacting factors, we decided to condense the discussions related to the below-LOR data and not to presenting additional 'proof' for the impacts of the below-LOR data – this could thus allow more space for additional discussions on the other influencing factors.

We have introduced three potential limiting factors on model performance in Sect 4.1 as:

- *L549: 'Despite the opportunities highlighted above, the model's performance also suggests some current limitations of the modelling framework in the following situations:*
  1) *High within-site temporal variability. ….*

2) *Presence of high proportions of below-DL data. ….*
3) *Non-conservativeness of constituent. ….'*

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: Chanat et al. (2016) (URL: http://dx.doi.org/10.3133/sir20155133); 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. Regarding your first point, we have extended the discussion in Sect. 4.2 on utilizing high-temporal-resolution data by considering potential challenges in using these data as well:

- *L592: 'The current spatio-temporal model extracts water quality temporal variability from monthly data. Utilizing data with higher temporal resolution may further strengthen the model capacity to explain temporal variability, especially by capturing more information on water quality dynamics during flow events. This may be possible into the future; however, current high-frequency water quality sensors (Bende-Michl and Hairsine, 2010;Outram et al., 2014;Lannergård et al., 2019;Pellerin et al., 2016) still have very high resourcing requirements that limits widespread deployment in operational networks.'*

We also added the following to the conclusion:

- *L686: 'The inclusion of high-frequency water quality sampling data may also extend the model's ability to represent temporal variability. However, high-frequency water quality data are also typically highly variable with large noise. Therefore, the implication of such data for the spatio-temporal modelling framework remains an open question, which needs further investigation in future applications of this modeling framework.'*

To address your second point, we have emphasized in Section 4.1 the strength of our model in being able to predict spatio-temporal variation in monthly data across a large region, despite the relatively low requirement of input data:

- *L527: 'In this study, we developed the first process-informed statistical model that is capable of explaining a reasonable proportion of water quality variability for a large spatial area of over 130,000km2. Although the calibration data have relatively low sampling frequency (i.e. monthly), our model generally performs satisfactorily in explaining the total variability in water quality. This demonstrates the effectiveness of the Bayesian hierarchical modelling framework in predicting spatio-temporal variability in water quality across large scales. The Bayesian hierarchical model is: a) more advantageous than other simpler statistical water quality models with its more comprehensive and process-informed approach, and capacity to represent varying temporal*

relationships across large-scale regions; b) less demanding for input data compared with those required by fully-distributed, processes-based models.'

Regarding your third point, we added discussions in Section 4.2 on the current limitations with using only monthly data on capturing event conditions:

- *L593: 'Utilizing data with higher temporal resolution may further strengthen the model capacity to explain temporal variability, especially by capturing more information on water quality dynamics during flow events.'*

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 have merged part of Tables S5 and S6 related to the key spatial and temporal predictors of the model to Table 1. Since these are results reported in the two preceding studies (Lintern et al. 2018 and Guo et al. 2019b) which were used for model development in this study, these results are presented in Section 2.1.3 under the Method section of the main text.

Table 1. Key factors affecting the spatial and temporal variability for each of six constituents, as identified in Lintern et al. (2018) and Guo et al. (2019b), respectively.

Constituent	Key factors that affect spatial variability	Key factors that affect
TSS	Hottest month maximum temperature Percentage area covered by grass Percentage area covered by shrub Percentage cropping area Maximum elevation Dam storage Percentage clay area	Same-day streamflow 7-day antecedent streamflow Water temperature Soil moisture root Soil moisture deep
TP	Erosivity Percentage area covered by grass Percentage area covered by shrub Percentage area made up of roads Percentage cropping area Average soil TP content	Same-day streamflow 30-day antecedent streamflow Water temperature Soil moisture root Soil moisture deep
FRP	Percentage area covered by shrub Percentage cropping area Catchment area Average soil TP content Mean channel slope	Same-day streamflow Water temperature Soil moisture deep
TKN	Percentage clay area Warmest quarter mean temperature Coldest quarter rainfall Percentage cropping area Percentage pasture area Average soil TP content	Same-day streamflow 30-day antecedent streamflow NDVI Water temperature Soil moisture root Soil moisture deep
NO_x	Annual radiation Warm quarter rainfall Hottest month maximum temperature Average soil TP content Mean channel slope	Same-day streamflow 30-day antecedent streamflow NDVI Water temperature Soil moisture root Soil moisture deep
EC	Annual radiation Annual rainfall Wettest quarter rain Hottest month maximum temperature Percentage agriculture area Percentage cropping area Percentage area covered by shrub Average soil TN content	Same-day streamflow 14-day antecedent streamflow Water temperature Soil moisture root Soil moisture deep

In addition, the second column of Table S6 (which summarizes the key factors relating to the spatial variability in temporal effects) have not been reported in any preceding studies. Therefore, these results are further enhanced and presented in Table 2 in Section 3.1, under the Results section.

Table 2. The key catchment landscape characteristics that are related to the varying relationships of water quality and same-day streamflow across space, which were selected as the two predictors for the streamflow effect in our model. The corresponding Spearman's correlation (ρ, at $p<0.05$) between the effect of streamflow and each catchment characteristic is presented.

Constituent	Key factors that affect spatial variability in temporal effects	Spearman's ρ ($p<0.05$)
TSS	Annual rainfall	0.722
	Hottest month maximum temperature	-0.575
TP	Annual rainfall	0.695
	Percentage area used for cropping	-0.556
FRP	Percentage agriculture area	0.392
	Percentage area underlain by mixed igneous bedrock	0.314
TKN	Annual rainfall	0.713
	Hottest month maximum temperature	-0.618
NO_x	Total storage capacity of dams in catchment	-0.493
	Mean soil TN content	0.458
EC	Percentage area covered by grassland	-0.347
	Percentage area covered by woodland	-0.317

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 have clarified 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 in Section 2.1.2:*

- *L165: 'Note that in the sampling protocol, FRP is defined as 'Reactive Phosphorus for a filtered sample to a defined filter size (e.g. RP(<0.45 µm))', which is equivalent to the more widely-used terminology, SRP i.e. Soluble Reactive Phosphorus (Jarvie et al., 2002).'*

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

We added more recent references to support this argument, as:

- *L45: 'Water quality conditions also typically differ substantially across locations (Meybeck and Helmer, 1989;Chang, 2008;Varanka et al., 2015;Lintern et al., 2018a).'*
  Added references:
  - *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 have also replaced the term 'significantly' with 'substantially' to avoid confusion with 'statistically significant' here. We have also checked throughout the manuscript to correct misuses of the term 'significant',

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 have improved the clarification of this discussion and have deleted the phrase ‘other catchment conditions’ as part of this, the updated discussion is as:

- *L56: ‘At the same time, temporal shifts in water quality are also influenced by changes in pollutant sources, such as land use and land management including urbanization, agriculture and vegetation clearing (Ren et al., 2003;Smith et al., 2013;Ouyang et al., 2010). In addition, water quality can also vary in time with variations in the mobilization and delivery processes, which are largely driven by the hydro-climatic conditions at a catchment, such as streamflow (Ahearn et al., 2004;Mellander et al., 2015;Sharpley et al., 2002;Zhang and Ball, 2017), the timing and magnitude of rainfall events (Fraser et al., 1999;Miller et al., 2014) and temperature (Bailey and Ahmadi, 2014).’*

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

We would like to clarify the paper structure as much as possible for the readers’ benefit with these overview sentences. To address this comment while maintain clarity, we have moved these sentences to the start of Section 2 before Section 2.1. We believe that this is a more suitable place to have an overview of the entire Method section:

- *L110: ‘We first discuss the process used to develop the integrated spatio-temporal model (Section 2.1). Sections 2.1.1 and 2.1.2 introduces the statistical modelling framework and the data used for model development, respectively. The approaches to determine model structure was then introduced, which include the choice of key predictors (Section 2.1.3) and the calibration for model parameters (Section 2.1.4). Finally, the approaches to evaluate model performance and robustness are described in Section 2.2.’*

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

We have revised this figure to improve visualization.

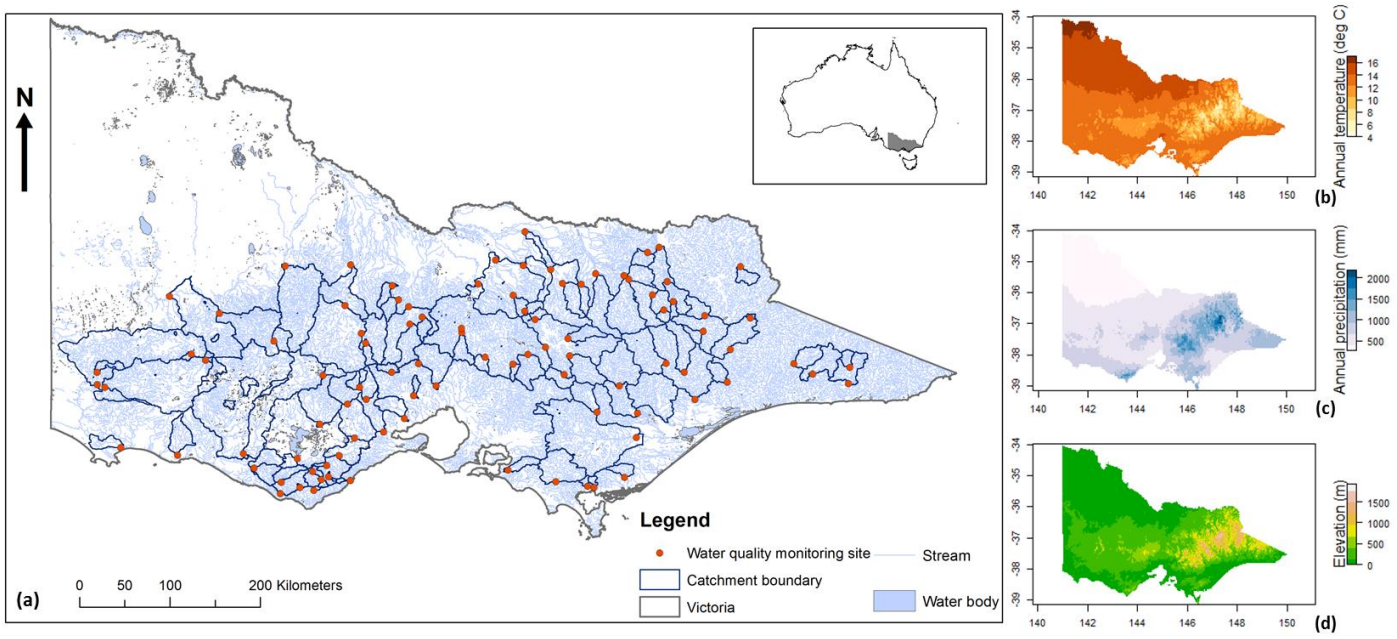


Figure 1. Map of (a) the 102 selected water quality monitoring sites and their catchment boundaries, with inserts showing the location of the state of Victoria within Australia; (b) annual average temperature and (c) annual precipitation and (d) elevation across Victoria.

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

We would like to clarify that both the literature review and the process to identify the potential spatial predictors for the model have been performed as part of a preceding study (Lintern et al., 2018a), where a more comprehensive reference list was presented, and we thus prefer not to repeat the details in this paper. To better clarify this, we have revised the following text in Section 2.1.2 to highlight that the preceding effort in literature review:

- *L173: 'To compile a dataset for the potential spatial explanatory variables (i.e. predictors to explain spatial variability in water quality), a comprehensive literature review was conducted (Lintern et al., 2018a), which summarized the key catchment landscape characterisitics that are widely known to influence water quality. Further, as part of Lintern et al. (2018b), fifty potential explanatory catchment characteristics were selected, which included catchment land use, land cover, topographic, climatic, geological, lithological and hydrological catchment characteristics.'*

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

We believe that our response to your last comment (#11) have resolved this concern too.

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 have added the key literature and the equation of the Box-Cox transformation, along with explanation of the transformation parameter, λ, as:

- *L220: 'all observed constituent concentrations and temporal explanatory variables were Box-Cox transformed (Box and Cox, 1964) (Eq. 8).*

$$y_{Box-Cox} = \begin{cases} \frac{y_{Raw}^{\lambda}-1}{\lambda}, & for\ \lambda \neq 0 \\ logy, & for\ \lambda = 0 \end{cases} \quad (8)$$

For each variable, the optimal Box-Cox transformation parameter λ was identified using the car R package and a maximum likelihood-like approach.'

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

During revision we have identified several other factors that potentially impact model performance and have added relevant discussions (as detailed in responses to your Comment #3). We have now removed discussions on the 'categorical issue' which is relatively less important.

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

This list has been extended to include phosphorus pathways in catchments, as:

- *L572: 'To better capture changes in reactive constituents, the model may require greater consideration of and more extensive spatial and temporal data to represent bio-geochemical processes. Examples include improvements on the process representation for nitrogen cycling and the desorption and adsorption of phosphorus (Granger et al., 2010;Smyth et al., 2013;Tian and Zhou, 2007).*

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). To clarify this we have added the following:

- *L288: 'In each model run there were four independent Markov chains. A total of 20,000 iterations were used for each chain. Convergence of the chains was ensured by checking the Rhat value (Sturtz et al., 2005), which is a summary statistic on the convergence of the Bayesian models from the four Markov chains used in model calibration (Stan Development Team, 2018). Specifically, an Rhat value much greater than 1 indicates that the independent Markov chains have not been mixed well, and a value of below 1.1 is recommended (Stan Development Team, 2018).'*

Editorial comments:

17. L71: Fix usage of ". . .not only. . .but also. . ." In addition, "limits" should be "limit".

Due to the substantial revision of the Introduction, this sentence has been revised as:

- *L75: 'Due to the limited understanding of large-scale water quality patterns, we currently lack the capacity to model spatio-temporal variabilities in water quality at larger scales across multiple catchments. This hinders our ability to inform the development of effective policy and mitigation strategies over large regions.'*

18. L76: The model built. . . –> The model was built. . .

Due to the substantial revision of the Introduction, this sentence has been revised as:

- *L95: 'To obtain the necessary understanding of process drivers required to develop this model, two preceding studies were conducted on the same dataset to identify the key drivers for the spatial and temporal variability of water quality, respectively (Lintern et al., 2018b; Guo et al., 2019).'*

19. Equation 3 and Equation 4: For the betas, consider using subscript instead of dash.

We have improved the clarify if the Beta terms in all of Equations 3, 4, 5 and 6. Consider that we have already used three sets of subscripts (n, i and j), we decided to only remove the dash, but keeping the T or S (that were previously after the dash) as normal-sized text. These equations are revised as:

$$\overline{C}_j = intC + \beta S_1 \times S_{1,j} + \beta S_2 \times S_{2,j} + \cdots + \beta S_m \times S_{m,j} \quad (3)$$

$$\Delta_{ij} = \beta T_{1,j} \times T_{1,ij} + \cdots + \beta T_{n,j} \times T_{n,ij} \quad (4)$$

$$\beta T_{N,j} \sim N(\mu\beta T_{N,j}, \sigma\beta T), for\ N\ in\ 1, 2, \ldots n \quad (5)$$

$$\mu\beta T_{N,j} = int\beta T_N + \beta ST_{N1} \times ST_{N1,j} + \beta ST_{N2} \times ST_{N2,j} \quad (6)$$

20. L180: "General speaking" –> "Generally speaking"

We have revised this as suggested (now L244).

21. L317: Fix "a results of"

We have revised this as suggested (now L513).

22. L382: Fix "oppourtunities"

Due to the substantial revision of the Discussion, this sentence has been revised as:

- *L592: 'The current spatio-temporal model extracts water quality temporal variability from monthly data. Utilizing data with higher temporal resolution may further strengthen the model capacity to explain temporal variability, especially by capturing more information on water quality dynamics during flow events.'*

23. L417: Fix "droguht"

We have revised this as suggested (now L631).

24. L420: Similarly to –> Similar to Comments on the SM:

We have revised this as suggested (now L637).

25. Supplementary Materials lack of "title-page" information.

The manuscript preparation guidelines of HESS suggested not to have title page for the Supplementary Materials, as:

"Supplements will receive a title page added during the publication process including title ("Supplement of"), authors, and the correspondence email. Therefore, please avoid providing this information in the supplement."

https://www.hydrology-and-earth-system-sciences.net/for_authors/manuscript_preparation.html

26. Table S4: Change "lambda" to its Greek form.

We have revised this as suggested.

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

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, with our specific manuscript revisions are shown in underlined text.

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 thoroughly revised the Abstract, Introduction and Conclusion to improve clarification of the knowledge gaps and the corresponding study objectives, and how this study differs from its preceding works. Some key revisions include:

- *L11 (Abstract): 'Our current capacity to model stream water quality is limited particularly at large spatial scales across multiple catchments. To address this, we developed a Bayesian hierarchical statistical model to simulate the spatio-temporal variability in stream water quality across the state of Victoria, Australia. The model was developed using monthly water quality monitoring data over 21 years, across 102 catchments, which span over 130,000 km2.'*

- *L75 (Introduction): 'Due to the limited understanding of large-scale water quality patterns, we currently lack the capacity to model spatio-temporal variabilities in water quality at large scales across multiple catchments. This hinders our ability to inform the development of effective policy and mitigation strategies over large regions. ... Modelling the spatio-temporal variability simultaneously remains challenging over long time periods and large regions.'*
- *L90 (Introduction): 'Accordingly, this research attempts to bridge the gap between fully-distributed physically-based water quality models and data-driven statistical approaches. We aim to develop a process-informed, data-driven model to predict spatio-temporal changes in stream water quality over a large region consisting of multiple catchments. Specifically, this model was established using long-term (21 years) stream water quality observations across 102 catchments in Australia, with an aggregate catchment area of 130,000 km2. To obtain the necessary understanding of process drivers required to develop this model, two preceding studies were conducted on the same dataset to identify the key drivers for the spatial and temporal variability of water quality, respectively (Lintern et al., 2018b; Guo et al., 2019). The aim of this study is to develop an integrated spatio-temporal model using the previously-identified spatial and temporal predictors, and to then assess the performance of this model.'*
- *L666 (Conclusion): 'This study aims to address the current lack of water quality models that operate at large scales across multiple catchments. To achieve this, we used long-term stream water quality data collected from 102 sites in south-eastern Australia, and developed a Bayesian hierarchical statistical model to simulate the spatio-temporal variabilities in six key water quality constituents: TSS, TP, FRP, TKN, NOx and EC. The choice of model predictors was guided by previous studies on the same dataset (Lintern et al., 2018b; Guo et al., 2019).'*

We added further results on how the temporal effects vary spatially, which have not been reported in preceding studies (Table 2).

Table 2. The key catchment landscape characteristics that are related to the varying relationships of water quality and same-day streamflow across space, which were selected as the two predictors for the streamflow effect in our model. The corresponding Spearman's correlation (ρ, at $p<0.05$) between the effect of streamflow and each catchment characteristic is presented.

Constituent	Key factors that affect spatial variability in temporal effects	Spearman's ρ ($p<0.05$)
TSS	Annual rainfall	0.722
	Hottest month maximum temperature	-0.575
TP	Annual rainfall	0.695
	Percentage area used for cropping	-0.556
FRP	Percentage agriculture area	0.392
	Percentage area underlain by mixed igneous bedrock	0.314
TKN	Annual rainfall	0.713
	Hottest month maximum temperature	-0.618
NO_x	Total storage capacity of dams in catchment	-0.493
	Mean soil TN content	0.458
EC	Percentage area covered by grassland	-0.347
	Percentage area covered by woodland	-0.317

In addition, to improve the linkage between study objectives and results, we also presented additional results to highlight several model capabilities and discussed on how these can benefit catchment management. The following specific results have been presented:

- *Fig. 3, which shows the proportions of spatial and temporal variability within total observed variability, as well as the model performance in explaining each component of variability. These results indicate that the model performs much better in capturing spatial variability compared to the temporal variability.*

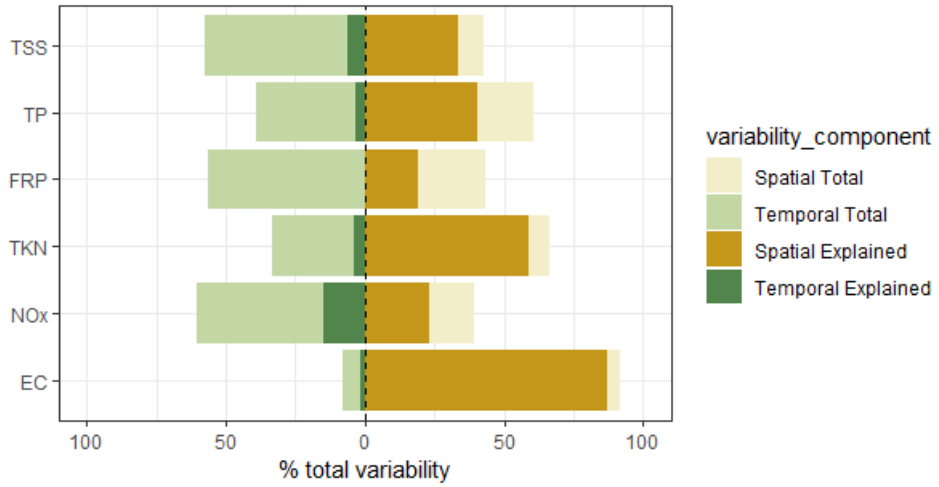


Figure 3. Observed spatial and temporal variabilities as proportions of the total variability (total width of each bar, 100%). The dashed line differentiates temporal variability (left side) with spatial variability (right side), and the darker colours highlight the proportions of spatial and temporal variabilities that are explainable by the model. All values were estimated in Box-Cox transformed space.

- *Fig. 6, which shows simulated and observed temporal variability at the catchment where the model performs the best at, for each constituent. These results further illustrated that the model largely underestimated temporal variability across all constituents, but is generally capable to represent long-term trend.*

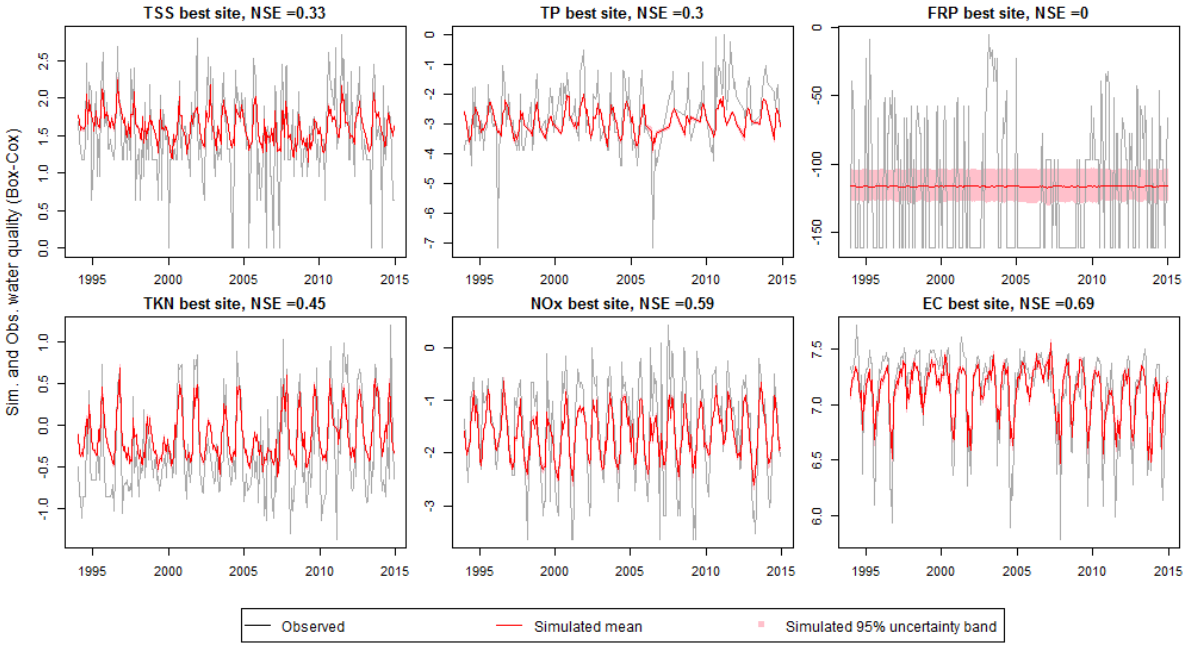


Figure 6. Model fit of the within-site (temporal) water quality variability, illustrated with the observed and simulated time-series for the best-performing site for each constituent. All values are presented in Box-Cox transformed space. The NSE for each constituent is also shown. The red line indicates the corresponding mean

of all posterior simulations, while the pink bands show the corresponding 95% lower and upper bounds (only visible for FRP).

- *Table 4, which summarizes the proportions of observed positive and negative water quality trends that are recognized by the model, which adds further evidence to Fig. 6 on model capability to represent long-term trend.*

Table 4. Model ability to capture observed water quality trends across all monitoring sites for each constituent. The percentages of sites where observed positive and negative trends are captured by the model are presented separately. Values in brackets indicate numbers of sites where corresponding positive or negative trends are observed. For detailed estimation of these percentages please refer to Sect. 2.2.

Constituent	% positive trends captured	% negative trends captured
TSS	33.3 (12)	85.0 (20)
TP	82.1 (28)	16.7 (12)
FRP	47.1 (17)	55.6 (9)
TKN	81.1 (37)	40.0 (10)
NO_x	68.6 (35)	66.7 (27)
EC	82.6 (23)	77.3 (22)

To highlight the practical value of these results, we have also added discussions on specific management utilities that the model could benefit, as:

- *L535: 'From a practical perspective, this model has the potential to contribute to a number of management activities including catchment planning, management and policy-making activities, specifically:*
  1) *The spatial predictive capacity can be used to identify pollution hot-spots and the catchment conditions that are likely causes of high concentrations. This can be used to help identify target catchment(s) to prioritize future water quality monitoring and management (Figs. 4 and 5);*
  2) *Further to 1), since water quality has been linked with catchment characteristics in this model, it can also be used to assess potential impacts of alternative options of land use and land cover change, as well as potential effects of climate change, on ambient water quality conditions;*
  3) *The model's temporal predictive capacity can identify changes in water quality due to changes in hydro-climatic conditions and vegetation cover, and thus enabling attribution of detected trends. On the other hand, any 'unexpected' trends can be identified to prompt further investigation to identify causes (Figure 6 and Table 4). The model could also be used for assessing the impacts of long-term catchment changes on water quality (Figures 7 and 8).'*

As highlighted in 3) above, 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 first revised the abstract to remove the specific focus on this drought analysis but instead focusing on more general illustration of model utility:

- *L20: 'Across constituents, the model generally captures over half of the observed spatial variability; temporal variability remains largely unexplained across all catchments, while long-term trends are well captured. The model is best used to predict proportional*

changes in water quality in a Box-Cox transformed scale, but can have substantial bias if used to predict absolute values for high concentrations. This model can assist catchment management by (1) identifying hot-spots and hot moments for waterway pollution; (2) predicting effects of catchment changes on water quality e.g. urbanization or forestation; and (3) identifying and explaining major water quality trends and changes. Further model improvements should focus on: (1) alternative statistical model structures to improve fitting for truncated data, where a large amount of data below the detection-limit; and (2) better representation of non-conservative constituents (e.g. FRP) by accounting for important biogeochemical processes.'

The description of the two cross-validation experiments in Section 2.1.2 was expanded to better clarify the purpose of this analysis:

- *L328: 'Additional evaluations of model sensitivity were conducted with calibration and validation on subsets of the full data (Section. 3.3), to understand model transferability and stability:*
    1. *Model sensitivity to the monitoring sites used for calibration. We randomly selected 80% of the sites for calibration and used the remaining 20% for validation, and repeated this validation process 50 times. We compared all calibration and validation performances of these 'partial models' were compared with each other, as well as with the performance of the full model, to obtain a comprehensive evaluation of the sensitivity of model performance to calibration sites.*
    2. *Model sensitivity to calibration data period. Since the study region was greatly influenced by a prolonged drought from 1997 to 2009 – known as the Millennium Drought (van Dijk et al., 2013), we also investigated model robustness for before, during and after this drought period. Specifically, we calibrated the model to each pre-, during- and post-drought period (1994-1996, 1997-2009 and 2010-2014, respectively) with model validation on the remaining data. For example, when calibrating to the pre-drought period (1997-2009), validation was performed on the merged during and post-drought period (1994-1996 plus 2010-2014). The corresponding calibration and validation performances were compared with each other as well as against that of the full model, to identify potential impacts of the drought on model robustness. '*

We also added discussions referring to results on the drought effects on TSS (Figs. 7 and 8) in Section 4.1 to emphasize the link to management:

- *L544: 'The model's temporal predictive capacity can identify changes in water quality due to changes in hydro-climatic conditions and vegetation cover, and thus enabling attribution of detected trends. On the other hand, any 'unexpected' trends can be identified to prompt further investigation to identify causes (Figure 6 and Table 4). The model could also be used for assessing the impacts of long-term catchment changes on water quality (Figures 7 and 8).'*

Last but not least, to address your concern on discussion high-frequency sampling data in the Conclusion, we have added clarification that these are recommendations based on our model evaluations.

- *L680: 'Based on the above model evaluations, we discussed potential ways to further enhance the model performance. … Regarding data availability, the current models could potentially benefit from improved monitoring of changes in land use intensity and*

management to be able to include these drivers in the model. The inclusion of high-frequency water quality sampling data may also extend the model's ability to represent temporal variability. However, high-frequency water quality data are also typically highly variable with large noise. Therefore, the implication of such data for the spatio-temporal modelling framework remains an open question, which needs further investigation in future applications of this modeling framework.'

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.'*

To minimize confusion and keep consistency with our monitoring dataset, we replaced the term 'LOR' with 'detection limit' in the revised manuscript. We added the following clarification for 'detection limit' where this term is first introduced, as:

- *L205: 'We also removed any values below the detection limit (DL), which was defined as the 'minimum concentration detected for which there is 95% confidence of accuracy and therefore is accurate enough to report' in the monitoring protocols for this dataset (Australian Water Technologies, 1999). This was because the uncertainty in values below DL would be amplified after transformation, which would largely influence in the subsequent model fitting. Furthermore, those undetectable low concentrations were of less interest for management purposes. Water quality records corresponding to days with zero flows were also excluded from further analyses.'*

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 clarify this, we added details on how the data below detection limit were used when describing the relevant model performance assessments:

- *L301: 'Firstly, the simulations from the fitted model and the corresponding observed concentrations were compared at 102 sites altogether to understand how the overall spatio-temporal variabilities were captured. For each constituent, this evaluation was performed with: 1) these above-DL data to focus only on data used for calibration (as*

detailed in Section. 2.1.2); 2) the full dataset including the below-DL data (set to half of the DL of the specific constituent), to understand how well the model represents the full distribution of constituent concentrations.'

References:

- *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 added these discussions to Section 2.2 to better highlight the purpose of this specific model performance evaluation:

- *L301: 'Firstly, the simulations from the fitted model and the corresponding observed concentrations were compared at 102 sites altogether to understand how the overall spatio-temporal variabilities were captured. For each constituent, this evaluation was performed with: 1) these above-DL data to focus only on data used for calibration (as detailed in Section. 2.1.2); 2) the full dataset including the below-DL data (set to half of the DL of the specific constituent), to understand how well the model represents the full distribution of constituent concentrations. A good model performance when including the below-DL data would suggest that the calibrated model is transferable to below-DL data too.'*

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.

We agree that a large amount of below-DL data may reflect a limitation of grab sampling to capture temporal variability for more event-driven constituents (e.g. TSS). However, considering the substantial increase of the paper length after implementing additional results during revision, we have refrained from further discussing the causes of below-DL data but have focused on their impact on model performance and potential improvements:

- *L558: 'The full datasets for the three poorly modelled constituents (FRP, TSS and NOx) all have higher proportions of data below the detection limit (38.2% 17.3% and 15% of all data, respectively) compared with other constituents. As illustrated in Fig. 2, for each of these constituents, removal of below-DL data before model calibration had created clear a truncation on the left-hand side of the distribution. This substantially increases the degrees of skewness and discontinuity of the data, essentially violating the assumption of normally distributed residuals and thus limiting model performance. The model capacity to handle truncated data might be improved by model fitting approaches explicitly designed for this issue. For example, Wang and Robertson (2011) and Zhao et al. (2016) illustrated an approach to resolving the discontinuity of the likelihood estimation in model fitting to data with presence of a lower bound such as zero rainfall values.*

We have also added discussion on how the model could potentially be enhanced with better representation of events, as:

- *L592: 'The current spatio-temporal model extracts water quality temporal variability from monthly data. Utilizing data with higher temporal resolution may further strengthen the model capacity to explain temporal variability, especially by capturing more information on water quality dynamics during flow events.'*

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. To clarify this we have added links to this analyses when discussing the practical utilities of the model:

- *L544: 'The model's temporal predictive capacity can identify changes in water quality due to changes in hydro-climatic conditions and vegetation cover, and thus enabling attribution of detected trends. On the other hand, any 'unexpected' trends can be identified to prompt further investigation to identify causes (Figure 6 and Table 4). The model could also be used for assessing the impacts of long-term catchment changes on water quality (Figures 7 and 8).'*

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 suggest 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:

- *'L653: Our findings provide extra dimensions to what would be offered by simple trend analyses using approaches such as Mann Kendall test or Sen's slope (e.g. Smith et al., 1987;Chang, 2008;Hirsch et al., 1991;Bouza-Deaño et al., 2008). Those approaches are only capable of indicating direction and magnitude of observed trends. In contrast, our model was able to attribute the consistent upward shift in TSS concentration to change in relationships between water quality and its key driving factors since the start of drought.'*

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 have revised 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. Key revisions include:

- *L11 (Abstract): 'Our current capacity to model stream water quality is limited particularly at large spatial scales across multiple catchments. To address this, we developed a Bayesian hierarchical statistical model to simulate the spatio-temporal variability in stream water quality across the state of Victoria, Australia. The model was developed using monthly water quality monitoring data over 21 years, across 102 catchments, which span over 130,000 km2.'*
- *L75 (Introduction): 'Due to the limited understanding of large-scale water quality patterns, we currently lack the capacity to model spatio-temporal variabilities in water quality at large scales across multiple catchments. This hinders our ability to inform the development of effective policy and mitigation strategies over large regions. ... Modelling the spatio-*

temporal variability simultaneously remains challenging over long time periods and large regions.'

- *L90 (Introduction): 'Accordingly, this research attempts to bridge the gap between fully-distributed physically-based water quality models and data-driven statistical approaches. We aim to develop a process-informed, data-driven model to predict spatio-temporal changes in stream water quality over a large region consisting of multiple catchments. Specifically, this model was established using long-term (21 years) stream water quality observations across 102 catchments in Australia, with an aggregate catchment area of 130,000 km2. To obtain the necessary understanding of process drivers required to develop this model, two preceding studies were conducted on the same dataset to identify the key drivers for the spatial and temporal variability of water quality, respectively (Lintern et al., 2018b; Guo et al., 2019). The aim of this study is to develop an integrated spatio-temporal model using the previously-identified spatial and temporal predictors, and to then assess the performance of this model.'*
- *L666 (Conclusion): 'This study aims to address the current lack of water quality models that operate at large scales across multiple catchments. To achieve this, we used long-term stream water quality data collected from 102 sites in south-eastern Australia, and developed a Bayesian hierarchical statistical model to simulate the spatio-temporal variabilities in six key water quality constituents: TSS, TP, FRP, TKN, NOx and EC. The choice of model predictors was guided by previous studies on the same dataset (Lintern et al., 2018b; Guo et al., 2019).'*

We have added clarification that the key spatial and temporal predictors for water quality are discussed in more details in the two preceding studies, as:

- *L346: 'The key controls of the spatial and temporal variations in water quality have been identified in our two preceding studies (Lintern et al. 2018b, Guo et al. 2019) and briefly summarized in Section 2.1.3. and are thus not discussed here.'*

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 have added new results and discussions on the key factors relating to the spatial variability in temporal effects (Table 2).

Table 2. The key catchment landscape characteristics that are related to the varying relationships of water quality and same-day streamflow across space, which were selected as the two predictors for the streamflow effect in our model. The corresponding Spearman's correlation (R, at $p<0.05$) between the effect of streamflow and each catchment characteristic is presented.

Constituent	Key factors that affect spatial variability in temporal effects	Spearman's R ($p<0.05$)
TSS	Annual rainfall	0.722
	Hottest month maximum temperature	-0.575
TP	Annual rainfall	0.695
	Percentage area used for cropping	-0.556
FRP	Percentage agriculture area	0.392
	Percentage area underlain by mixed igneous bedrock	0.314
TKN	Annual rainfall	0.713
	Hottest month maximum temperature	-0.618
NO_x	Total storage capacity of dams in catchment	-0.493
	Mean soil TN content	0.458
EC	Percentage area covered by grassland	-0.347
	Percentage area covered by woodland	-0.317

- *L360: 'TSS, TP and TKN show consistent patterns of the spatial variation in the effects of streamflow on water quality, which are strongly driven by the differences in average rainfall conditions across catchments. Specifically, streamflow generally has a larger effect on water quality in catchments with higher average annual rainfall. Since the streamflow effects are positive for the majority of catchments (as shown in Figure S5), these correlations indicate that for the same increase in transformed streamflow, a greater increase in transformed concentrations of TSS, TP and TKN will occur at a catchment with higher annual average rainfall. Given that the Box-Cox lambda values (Table S4) are close to zero, the transformation is log-like and hence changes in transformed flow and concentration approximately correspond to proportional changes in the real values of flow and concentration. In contrast, for FRP, NOx and EC, the spatial patterns of streamflow effects are specific to each constituent.'*

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 130,000km^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 have added the phrase 'data-based' in our title to suggest that an empirical model is introduced. The revised title is:

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

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

Due to the substantial revision of the Introduction, this sentence has been revised as:

- *L75: 'Due to the limited understanding of large-scale water quality patterns, we currently lack the capacity to model spatio-temporal variabilities in water quality at larger scales across multiple catchments. This hinders our ability to inform the development of effective policy and mitigation strategies over large regions.'*

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

Due to the substantial revision of the Introduction, this sentence has been revised as:

- *L95: 'To obtain the necessary understanding of process drivers required to develop this model, two preceding studies were conducted on the same dataset to identify the key drivers for the*

spatial and temporal variability of water quality, respectively (Lintern et al., 2018b; Guo et al., 2019).'

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 have clarified this better in the Introduction. In addition, we have also adjusted 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 in these responses, 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 (Table 2 and relevant discussions, as detailed in responses on Comment #6) have been 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 have adjusted this phrase as:

- *L90: 'Accordingly, this research attempts to bridge the gap between fully-distributed physically-based water quality models and data-driven statistical approaches.'*

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?

We have removed the statement 'poor water quality conditions (i.e., high constituent concentrations) were our primary concerns to model' since this was not accurately reflecting the key consideration for us to removing below-LOR data (now referred to as below-detection-limit data in the revised manuscript). The revised discussion is as:

- *L205: 'We also removed any values below the detection limit (DL), which was defined as the 'minimum concentration detected for which there is 95% confidence of accuracy and therefore is accurate enough to report' in the monitoring protocols for this dataset (Australian Water Technologies, 1999). This was because the uncertainty in values below DL would be amplified after transformation, which would largely influence in the subsequent model fitting. Furthermore, those undetectable low concentrations were of less interest for management purposes. Water quality records corresponding to days with zero flows were also excluded from further analyses.'*

13. 159. Insert a blank between “as each”

We have revised this as suggested (now in L213).

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

We have corrected the typo and have added the Spearman’s correlation values (ρ) and the significance level ($p<0.05$) in the corresponding results i.e. Table 2.

15. 246. “…in Sect. 4.2.” Isn’t it section 4.1?

We confirm this and have corrected the mistake (now in L385).

16. 265. Fix “… is also show…”

We have revised this as suggested (now caption to Fig. 4).

17. 414. Fix “For examples, …“

We have revised this as suggested (now in L631).

18. 418 Fix “adjscent”

We have revised this as suggested (now in L635).

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 Comment #1 and #6, we have:

1) Thoroughly revised the Abstract, Introduction and Conclusion to improve clarification of the knowledge gaps and the corresponding study objectives, and how this study differs from its preceding works. We also added results on how the temporal effects vary spatially, which have not been reported in preceding studies.

- *L11 (Abstract): ‘Our current capacity to model stream water quality is limited particularly at large spatial scales across multiple catchments. To address this, we developed a Bayesian hierarchical statistical model to simulate the spatio-temporal variability in stream water quality across the state of Victoria, Australia. The model was developed using monthly water quality monitoring data over 21 years, across 102 catchments, which span over 130,000 km2.’*
- *L75 (Introduction): ‘Due to the limited understanding of large-scale water quality patterns, we currently lack the capacity to model spatio-temporal variabilities in water quality at large scales across multiple catchments. This hinders our ability to inform the development of effective policy and mitigation strategies over large regions. ... Modelling the spatio-temporal variability simultaneously remains challenging over long time periods and large regions.’*
- *L90 (Introduction): ‘Accordingly, this research attempts to bridge the gap between fully-distributed physically-based water quality models and data-driven statistical approaches. We aim to develop a process-informed, data-driven model to predict spatio-temporal changes in stream water quality over a large region consisting of multiple catchments. Specifically, this model was established using long-term (21 years) stream water quality observations across 102 catchments in Australia, with an aggregate catchment area of 130,000 km2. To obtain*

the necessary understanding of process drivers required to develop this model, two preceding studies were conducted on the same dataset to identify the key drivers for the spatial and temporal variability of water quality, respectively (Lintern et al., 2018b; Guo et al., 2019). The aim of this study is to develop an integrated spatio-temporal model using the previously-identified spatial and temporal predictors, and to then assess the performance of this model.'

- *L666 (Conclusion): 'This study aims to address the current lack of water quality models that operate at large scales across multiple catchments. To achieve this, we used long-term stream water quality data collected from 102 sites in south-eastern Australia, and developed a Bayesian hierarchical statistical model to simulate the spatio-temporal variabilities in six key water quality constituents: TSS, TP, FRP, TKN, NOx and EC. The choice of model predictors was guided by previous studies on the same dataset (Lintern et al., 2018b; Guo et al., 2019).'*

2) *<u>Improved the linkage between study objectives and results, via additional results to highlight several model capabilities and discussed on how these can benefit catchment management, specifically:</u>*

- *Table 2, which summarizes the key landscape characteristics that are relevant to the variation in the strengths of the temporal predictors of water quality across space.*
- *Fig. 3, which shows the proportions of spatial and temporal variability within total observed variability, as well as the model performance in explaining each component of variability. These results indicate that the model performs much better in capturing spatial variability compared to the temporal variability.*
- *Fig. 6, which shows simulated and observed temporal variability at the catchment where the model performs the best at, for each constituent. These results further illustrated that the model largely underestimated temporal variability across all constituents, but is generally capable to represent long-term trend.*
- *Table 4, which summarizes the proportions of observed positive and negative water quality trends that are recognized by the model, which adds further evidence to Fig. 6 on model capability to represent long-term trend.*

3) *<u>Added discussions on specific management utilities that the model could benefit, as:</u>*

- *L535: From a practical perspective, this model has the potential to contribute to a number of management activities including catchment planning, management and policy-making activities, specifically:*
  1) *The spatial predictive capacity can be used to identify pollution hot-spots and the catchment conditions that are likely causes of high concentrations. This can be used to help identify target catchment(s) to prioritize future water quality monitoring and management (Figs. 4 and 5);*
  2) *Further to 1), since water quality has been linked with catchment characteristics in this model, it can also be used to assess potential impacts of alternative options of land use and land cover change, as well as potential effects of climate change, on ambient water quality conditions;*
  3) *The model's temporal predictive capacity can identify changes in water quality due to changes in hydro-climatic conditions and vegetation cover, and thus enabling attribution of detected trends. On the other hand, any 'unexpected' trends can be identified to prompt further investigation to identify causes (Figure 6 and Table 4). The model could also be used for assessing the impacts of long-term catchment changes on water quality (Figures 7 and 8).'*

To address this comment on the purpose of analyzing drought effects on TSS, we first revised the abstract to remove the specific focus on this analysis but instead focusing on more general illustration of model utility:

- *L20: 'Across constituents, the model generally captures over half of the observed spatial variability; temporal variability remains largely unexplained across all catchments, while long-term trends are well captured. The model is best used to predict proportional changes in water quality in a Box-Cox transformed scale, but can have substantial bias if used to predict absolute values for high concentrations. This model can assist catchment management by (1) identifying hot-spots and hot moments for waterway pollution; (2) predicting effects of catchment changes on water quality e.g. urbanization or forestation; and (3) identifying and explaining major water quality trends and changes. Further model improvements should focus on: (1) alternative statistical model structures to improve fitting for truncated data, where a large amount of data below the detection-limit; and (2) better representation of non-conservative constituents (e.g. FRP) by accounting for important biogeochemical processes.'*

We also expanded the description of the two cross-validation experiments in Section 2.2 to better clarify the purpose of these analyses:

- *L328: 'Additional evaluations of model sensitivity were conducted with calibration and validation on subsets of the full data (Section. 3.3), to understand model transferability and stability:*
  1) *Model sensitivity to the monitoring sites used for calibration. We randomly selected 80% of the sites for calibration and used the remaining 20% for validation, and repeated this validation process 50 times. We compared all calibration and validation performances of these 'partial models' were compared with each other, as well as with the performance of the full model, to obtain a comprehensive evaluation of the sensitivity of model performance to calibration sites.*
  2) *Model sensitivity to calibration data period. Since the study region was greatly influenced by a prolonged drought from 1997 to 2009 – known as the Millennium Drought (van Dijk et al., 2013), we also investigated model robustness for before, during and after this drought period. Specifically, we calibrated the model to each pre-, during- and post-drought period (1994-1996, 1997-2009 and 2010-2014, respectively) with model validation on the remaining data. For example, when calibrating to the pre-drought period (1997-2009), validation was performed on the merged during and post-drought period (1994-1996 plus 2010-2014). The corresponding calibration and validation performances were compared with each other as well as against that of the full model, to identify potential impacts of the drought on model robustness.'*

In addition, we added discussions referring to results on the drought effects on TSS (Figs. 7 and 8) in Section 4.1 to highlight this analysis as a way that the model can be used to inform management:

- *L544: 'The model's temporal predictive capacity can identify changes in water quality due to changes in hydro-climatic conditions and vegetation cover, and thus enabling attribution of detected trends. On the other hand, any 'unexpected' trends can be identified to prompt further investigation to identify causes (Figure 6 and Table 4). The model could also be used for assessing the impacts of long-term catchment changes on water quality (Figures 7 and 8).'*

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. To clarify this, we have added clarification that these are recommendations based on our model evaluations.

- *L680: ‘Based on the above model evaluations, we discussed potential ways to further enhance the model performance. …Regarding data availability, the current models could potentially benefit from improved monitoring of changes in land use intensity and management to be able to include these drivers in the model. The inclusion of high-frequency water quality sampling data may also extend the model’s ability to represent temporal variability.’*

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 Conclusion accordingly.

- *L684: ‘Regarding data availability, the current models could potentially benefit from improved monitoring of changes in land use intensity and management to be able to include these drivers in the model. The inclusion of high-frequency water quality sampling data may also extend the model’s ability to represent temporal variability. However, high-frequency water quality data are also typically highly variable with large noise. Therefore, the implication of such data for the spatio-temporal modelling framework remains an open question, which needs further investigation in future applications of this modeling framework.’*

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

Our 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 a statistical model that can predict spatial and temporal variabilities of catchment water quality. The model development was informed by the understanding obtained from the two preceding studies.*
3) *The key practical implication of this model is to allow catchment managers to better assess/plan/manage water quality changes across both space and time.*

To address the comment on the key focus of the study, we thoroughly revised the Abstract, Introduction and Conclusion to better clarify the knowledge gaps and study objectives, and how this study differs from its preceding works, as:

- *L11 (Abstract): 'Our current capacity to model stream water quality is limited particularly at large spatial scales across multiple catchments. To address this, we developed a Bayesian hierarchical statistical model to simulate the spatio-temporal variability in stream water quality across the state of Victoria, Australia. The model was developed using monthly water*

quality monitoring data over 21 years, across 102 catchments, which span over 130,000 km2.'

- *L75 (Introduction): 'Due to the limited understanding of large-scale water quality patterns, we currently lack the capacity to model spatio-temporal variabilities in water quality at large scales across multiple catchments. This hinders our ability to inform the development of effective policy and mitigation strategies over large regions. ... Modelling the spatio-temporal variability simultaneously remains challenging over long time periods and large regions.'*
- *L90 (Introduction): 'Accordingly, this research attempts to bridge the gap between fully-distributed physically-based water quality models and data-driven statistical approaches. We aim to develop a process-informed, data-driven model to predict spatio-temporal changes in stream water quality over a large region consisting of multiple catchments. Specifically, this model was established using long-term (21 years) stream water quality observations across 102 catchments in Australia, with an aggregate catchment area of 130,000 km2. To obtain the necessary understanding of process drivers required to develop this model, two preceding studies were conducted on the same dataset to identify the key drivers for the spatial and temporal variability of water quality, respectively (Lintern et al., 2018b; Guo et al., 2019). The aim of this study is to develop an integrated spatio-temporal model using the previously-identified spatial and temporal predictors, and to then assess the performance of this model.'*
- *L666 (Conclusion): 'This study aims to address the current lack of water quality models that operate at large scales across multiple catchments. To achieve this, we used long-term stream water quality data collected from 102 sites in south-eastern Australia, and developed a Bayesian hierarchical statistical model to simulate the spatio-temporal variabilities in six key water quality constituents: TSS, TP, FRP, TKN, NOx and EC. The choice of model predictors was guided by previous studies on the same dataset (Lintern et al., 2018b; Guo et al., 2019).'*

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 a statistical model 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 respond to individual ones as following:

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

To address this comment, we have presented additional results on the time-series for selected catchments at the modelling (transformed) scale, and the un-transformed model performance in two separate figures. In this way we can closely assess a) the model ability to simulate temporal variability and b) any impacts of transformation on model performance:

- *Fig. 6, which shows simulated and observed temporal variability at the catchment where the model performs the best at, for each constituent. These results further illustrated that the model largely underestimated temporal variability across all constituents, but is generally capable to represent long-term trend.*

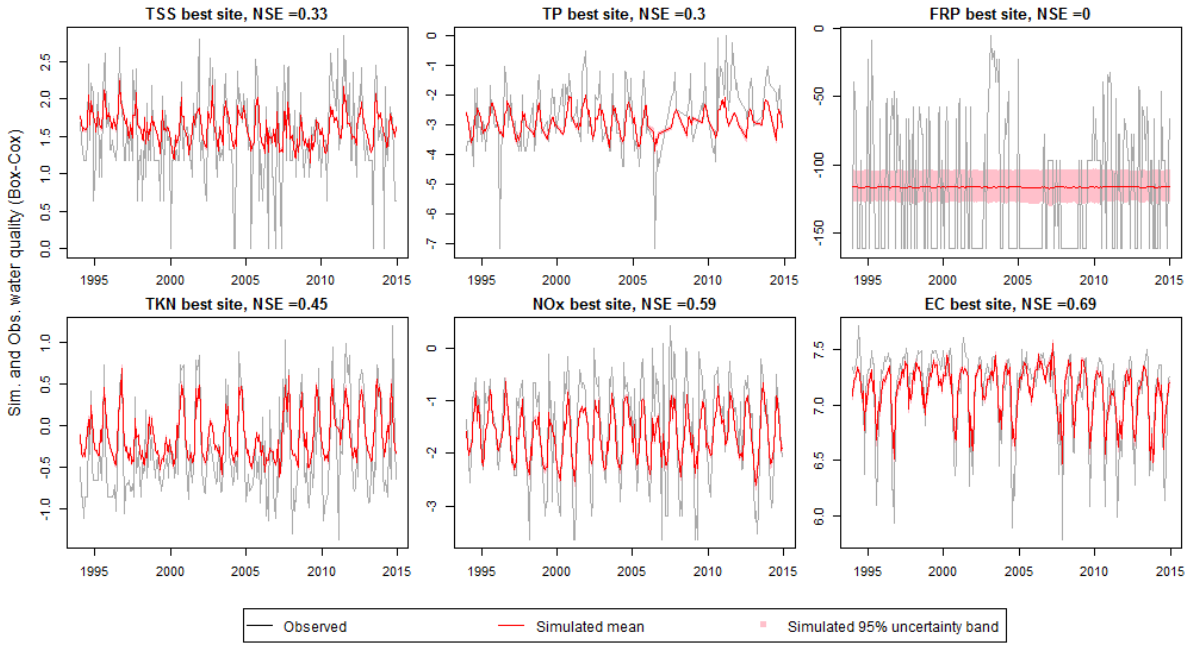


Figure 6. Model fit of the within-site (temporal) water quality variability, illustrated with the observed and simulated time-series for the best-performing site for each constituent. All values are presented in Box-Cox transformed space. The NSE for each constituent is also shown. The red line indicates the corresponding mean of all posterior simulations, while the pink bands show the corresponding 95% lower and upper bounds (only visible for FRP).

- *Fig. 5, which shows the modelled and observed site-mean concentrations in a back-transformed scale. This is presented after Fig. 4, which shows the simulated against observed site-mean concentrations in Box-Cox transformed scale, and thus highlighting any impact that data transformation has on model performance.*

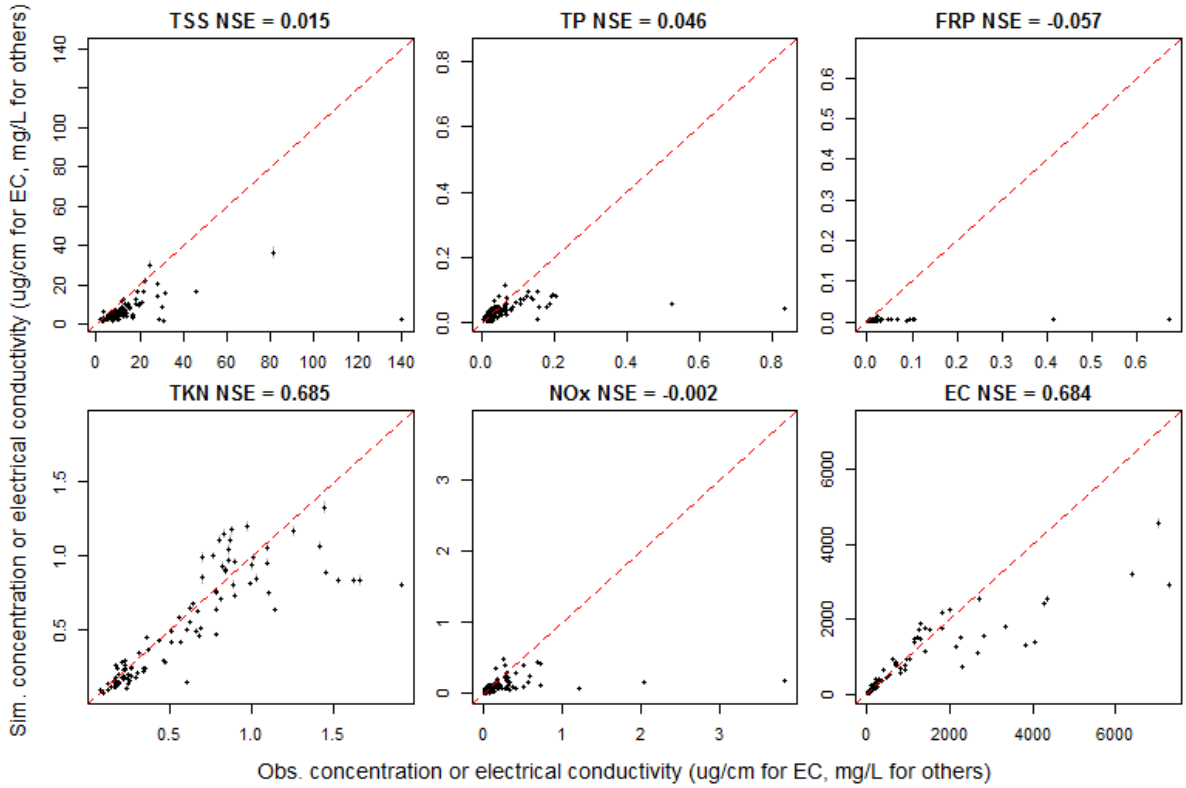


Figure 5. Back-transformation of the model simulations to the measurement scale emphasizes lack of fit for the highest concentrations, illustrated by simulated against observed site-level mean concentrations of each constituent in a back-transformed scale. The 95% lower and upper bounds of all posterior simulations shown in vertical grey lines. The NSE for each constituent is also shown and red dash lines show the 1:1 lines.

We have added discussions on the back-transformation as:

- *L422: 'At the back-transformed scale, the model shows greater biases for sites with higher concentrations (approximately the highest 10% sites for each constituent) (Fig. 5). This is not surprising as the model was fitted to a Box-Cox transformed space that reduces focus on high values and increases the focused on low values. This compromised its ability to represent sites with unusually high concentrations. The implications of the model having higher predictive capacity in the transformed scale is further discussed in Section. 4.1.'*

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 presented additional results to illustrate the model capability to capture temporal variability and trends in water quality (Figure 6 and Table 4), which is a good example of the management utility of our models, because water quality trends and changes are of great interests for catchment management:

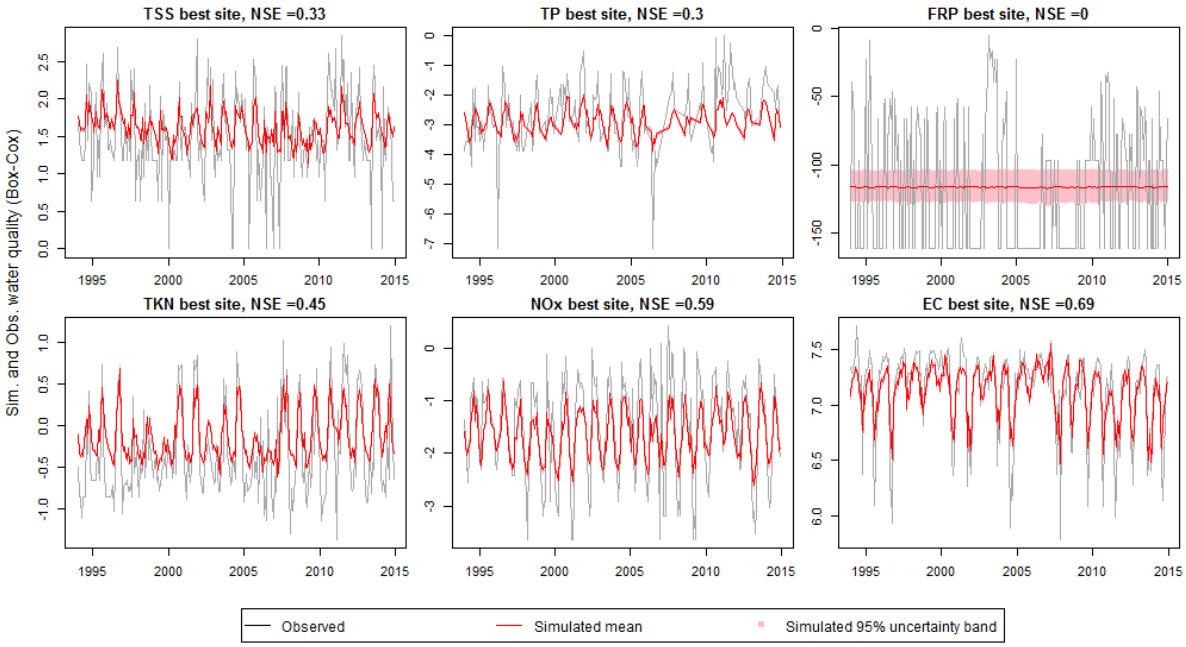


Figure 6. Model fit of the within-site (temporal) water quality variability, illustrated with the observed and simulated time-series for the best-performing site for each constituent. All values are presented in Box-Cox transformed space. The NSE for each constituent is also shown. The red line indicates the corresponding mean of all posterior simulations, while the pink bands show the corresponding 95% lower and upper bounds (only visible for FRP).

Table 4. Model ability to capture observed water quality trends across all monitoring sites for each constituent. The percentages of sites where observed positive and negative trends are captured by the model are presented separately. Values in brackets indicate numbers of sites where corresponding positive or negative trends are observed. For detailed estimation of these percentages please refer to Sect. 2.2.

Constituent	% positive trends captured	% negative trends captured
TSS	33.3 (12)	85.0 (20)
TP	82.1 (28)	16.7 (12)
FRP	47.1 (17)	55.6 (9)
TKN	81.1 (37)	40.0 (10)
NO_x	68.6 (35)	66.7 (27)
EC	82.6 (23)	77.3 (22)

More generally, we also added discussions to summarize how different aspects of model capacity can benefit various catchment management purposes in Section 4.1, as:

- *L535: 'From a practical perspective, this model has the potential to contribute to a number of management activities including catchment planning, management and policy-making activities, specifically:*
  1) *The spatial predictive capacity can be used to identify pollution hot-spots and the catchment conditions that are likely causes of high concentrations. This can be used to help identify target catchment(s) to prioritize future water quality monitoring and management (Figs. 4 and 5);*
  2) *Further to 1), since water quality has been linked with catchment characteristics in this model, it can also be used to assess potential impacts of alternative options of land use and land cover change, as well as potential effects of climate change, on ambient water quality conditions;*
  3) *The model's temporal predictive capacity can identify changes in water quality due to changes in hydro-climatic conditions and vegetation cover, and thus enabling attribution of detected trends. On the other hand, any 'unexpected' trends can be*

identified to prompt further investigation to identify causes (Figure 6 and Table 4). The model could also be used for assessing the impacts of long-term catchment changes on water quality (Figures 7 and 8).'

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 first revised the Introduction and Conclusion thoroughly to better clarify this different focus of this study with the two preceding studies:

- *L90 (Introduction): 'Accordingly, this research attempts to bridge the gap between fully-distributed physically-based water quality models and data-driven statistical approaches. We aim to develop a process-informed, data-driven model to predict spatio-temporal changes in stream water quality over a large region consisting of multiple catchments. Specifically, this model was established using long-term (21 years) stream water quality observations across 102 catchments in Australia, with an aggregate catchment area of 130,000 km2. To obtain the necessary understanding of process drivers required to develop this model, two preceding studies were conducted on the same dataset to identify the key drivers for the spatial and temporal variability of water quality, respectively (Lintern et al., 2018b; Guo et al., 2019). The aim of this study is to develop an integrated spatio-temporal model using the previously-identified spatial and temporal predictors, and to then assess the performance of this model.'*
- *L666 (Conclusion): 'This study aims to address the current lack of water quality models that operate at large scales across multiple catchments. To achieve this, we used long-term stream water quality data collected from 102 sites in south-eastern Australia, and developed a Bayesian hierarchical statistical model to simulate the spatio-temporal variabilities in six key water quality constituents: TSS, TP, FRP, TKN, NOx and EC. The choice of model predictors was guided by previous studies on the same dataset (Lintern et al., 2018b; Guo et al., 2019).'*

To highlight the new understandings obtained in this study, we also added further results on how the effects of the key temporal factors vary spatially. This is a critical understanding to develop the spatio-temporal predictive power of the model, and have not been reported in preceding studies (Table 2).

Table 2. The key catchment landscape characteristics that are related to the varying relationships of water quality and same-day streamflow across space, which were selected as the two predictors for the streamflow effect in our model. The corresponding Spearman's correlation (ρ, at $p<0.05$) between the effect of streamflow and each catchment characteristic is presented.

Constituent	Key factors that affect spatial variability in temporal effects	Spearman's ρ ($p<0.05$)
TSS	Annual rainfall	0.722
	Hottest month maximum temperature	-0.575
TP	Annual rainfall	0.695
	Percentage area used for cropping	-0.556
FRP	Percentage agriculture area	0.392
	Percentage area underlain by mixed igneous bedrock	0.314
TKN	Annual rainfall	0.713
	Hottest month maximum temperature	-0.618
NO_x	Total storage capacity of dams in catchment	-0.493
	Mean soil TN content	0.458

EC	Percentage area covered by grassland	-0.347
	Percentage area covered by woodland	-0.317

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

To address this comment, we added Fig. 3, which shows the proportions of spatial and temporal variability within total observed variability, as well as the model performance in explaining each component of variability.

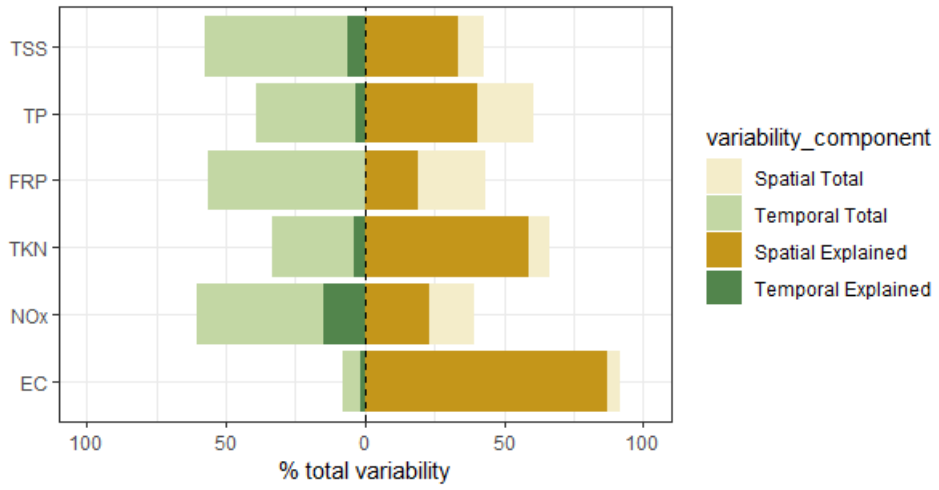


Figure 3. Observed spatial and temporal variabilities as proportions of the total variability (total width of each bar, 100%). The dashed line differentiates temporal variability (left side) with spatial variability (right side), and the darker colours highlight the proportions of spatial and temporal variabilities that are explainable by the model. All values were estimated in Box-Cox transformed space.

This additional result leads to better understanding of the model capability in representing spatial and temporal variabilities, which we added as:

- *L409: 'The explained variability (darker colours) shows that, across all catchments, temporal variability is much more difficult to model compared with spatial variability. It also appears that a substantial part of the model's overall performance is driven by its ability to capture spatial variability in ambient water quality conditions. For example, the models for TSS, FRP and NOx show poorer overall performance (Fig. 2, with NSE values of 0.225, -1.92 and 0.216, respectively)), because the total variability for each of these is dominated by temporal variability (57.4%, 56.6%, 60.5%, respectively), which largely remains unexplained by the model (Fig. 3). In contrast, the EC model shows a very good fit with 90.7% total variability explained – 91.8% of the total observed variability is due to spatial variability, of which 94.7% is explained by the model. Therefore, although EC the model can only explain a small portion of temporal variability (20% out of 8.2% of total variability), the overall model performance remains high.'*

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?)

We have thoroughly revised Section 4.1 to better highlight current limitations in model performance and the potential causes for those, as:

- *L549: 'Despite the opportunities highlighted above, the model's performance also suggests some current limitations of the modelling framework in the following situations:*

1) *High within-site temporal variability. In Section 3.2 we have identified a general lack of predictive power for temporal variability. The potential impacts of high temporal variability on model performance is particularly evident for results of TSS, NOx and FRP in Fig. 3. Since our model has already included hydro-climatic conditions and vegetation cover to explain temporal variability, the unexplained temporal variability is likely due to other uncaptured temporal drivers. These could be: changes in land use and land management, bio-geochemical processes, or transit time of water through catchments.*
2) *Presence of high proportions of below-DL data. The full datasets for the three poorly modelled constituents (FRP, TSS and NOx) all have higher proportions of data below the detection limit (38.2% 17.3% and 15% of all data, respectively) compared with other constituents. As illustrated in Fig. 2, for each of these constituents, removal of below-DL data before model calibration had created clear a truncation on the left-hand side of the distribution. This substantially increases the degrees of skewness and discontinuity of the data, essentially violating the assumption of normally distributed residuals and thus limiting model performance. The model capacity to handle truncated data might be improved by model fitting approaches explicitly designed for this issue. For example, Wang and Robertson (2011) and Zhao et al. (2016) illustrated an approach to resolving the discontinuity of the likelihood estimation in model fitting to data with presence of a lower bound such as zero rainfall values.*
3) *Non-conservativeness constituents. The results indicate that the reactivity of the constituent is broadly associated with performance, which suggest that bio-geochemical processes (e.g. phosphorus cycling, nitrification/de-nitrification) can make water quality dynamics more difficult for the model to capture. To better capture changes in reactive constituents, the model may require greater consideration of and more extensive spatial and temporal data to represent bio-geochemical processes. Examples include improvements on the process representation for nitrogen cycling and the desorption and adsorption of phosphorus (Granger et al., 2010;Smyth et al., 2013;Tian and Zhou, 2007).'*

In addition, we also discussed the implication of transformation on the model predictivity, as:

- *L577: 'As previously noted, our model was developed in a Box-Cox transformed scale to ensure the validity of the statistical assumptions (see details on data transformation in Sect. 2.1.2), which shows limited performance for high constituent concentrations when simulations are back-transformed to the measurement scale (Figs. 4 and 5). However, our model approximately represents proportional changes in water quality, which can thus help managers to understand proportional changes to inform practical catchment management.'*

Limitations to be noted for future implementations are also discussed as:

- *L583: 'For future implementations, the established model structure and parameterization would be best suited to within the study region. Before performing new simulations (e.g. for new monitoring sites or for current study sites over a different time-period), the statistical properties of the new input datasets should be checked to ensure that they are similar to the calibration datasets. To model new catchments outside of the study region, a re-calibration of the model is required. This would involve extensive selection of key predictors and model calibration, much as performed in this study and the two preceding ones (Lintern et al., 2018b; Guo et al., 2019). A sufficiently long record length (e.g. 20 years) is ideal for such modelling, as it ensures a reasonable understanding of the temporal variability to be obtained.'*

To better summarize the model sensitivity to calibration data, we have increased the number of cross-validation replicates from the current 5 to 50, each of which used 80% monitoring sites for model calibration and the other 20% for validation. The results are summarized in Table 5.

Table 5. Comparison of model performances (as NSE) of the full model (Column 2) and the 50 partial models (Columns 3 to 5) with each calibrated to 80% randomly selected monitoring sites. Columns 3 to 5 summarize the mean, minimum and maximum NSE values across the 50 runs, where for each constituent, the top row showing calibration performance and the bottom row showing the validation performance (i.e. at the 20% sites that were not used for calibration).

Constituent	Full model	50 CV mean	50 CV min	50 CV max
TSS	0.225	0.413	0.376	0.439
		0.382	0.292	0.513
TP	0.433	0.461	0.427	0.501
		0.411	0.151	0.575
FRP	-1.92	0.168	0.067	0.232
		0.129	-0.078	0.272
TKN	0.658	0.654	0.622	0.670
		0.622	0.468	0.691
NO_x	0.216	0.453	0.414	0.489
		0.397	0.258	0.563
EC	0.907	0.893	0.882	0.903
		0.875	0.809	0.924

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 2.1.1):
   - *L119: 'A key strength of applying the hierarchical model structure to analyze spatio-temporal variability is that this structure enables the key controls of temporal variability in water quality to vary across locations (Webb and King, 2009;Borsuk et al., 2001). This variability has been found to be important in other study regions where the (temporal) solute export regime varies with catchment characteristics such as climate and land use (Musolff et al., 2015;Poor and McDonnell, 2007).'*

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 have added Figures S6 and S7 in the Supplementary Material for detailed comparison of model residuals of the partial calibration/validation (in which only 80% study sites were used for calibration and this process was repeated for 50 times, see manuscript Section 2.2). Due to the large spatial-temporal extent of our model, we acknowledge that these validation results would provide limited implications on model development alternatives (e.g. how to improve modelling for specific catchments/regions). We have provided ideas future investigations focusing on the large-scale drought impact on water quality in Section 4.3 of the manuscript (L659 and onwards).

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 130,000 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 multiple catchments (which is the focus of this study). To address this comment, we have firstly revised the Abstract to remove the emphasize on the lack of understanding but instead, focusing on our study objective which is to improve modelling capacity:

- *L11: 'Our current capacity to model stream water quality is limited particularly at large spatial scales across multiple catchments. To address this, we developed a Bayesian hierarchical statistical model to simulate the spatio-temporal variability in stream water quality across the state of Victoria, Australia.'*

We also revised discussions 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.

- *L63 (Introduction): 'As abovementioned, we have good understanding of the key controls for variations in water quality, albeit in an isolated, idealized context. We still lack a sound understanding of how relationships between specific landscape characteristics and water quality can shift with influences from other landscape characteristics, and how the drivers of temporal variability in water quality can interact and vary across large spatial scales (Musolff et al., 2015;Lintern et al., 2018a;Ali et al., 2017). In contrast, current detailed understanding have been primarily based on field studies at small scales with detailed information on specific temporal drivers ranging from hydrologic conditions to detailed management decisions such as fertilizer rates and application timing (Smith et al., 2013;Poudel et al., 2013;Adams et al., 2014). While operational weather observation netowrks, stream gauging networks and remote sensing can provide some of this information, developing a large-scale understanding of water quality patterns across catchments would ideally also involve an extensive suite of management information that substantially exceeds what is currently available.'*
- *L75 (Introduction): 'Due to the limited understanding of large-scale water quality patterns, we currently lack the capacity to model spatio-temporal variabilities in water quality at larger scales across multiple catchments. This hinders our ability to inform the development of effective policy and mitigation strategies over large regions.'*
- *L90 (Introduction): 'Accordingly, this research attempts to bridge the gap between fully-distributed physically-based water quality models and data-driven statistical approaches. We aim to develop a process-informed, data-driven model to predict spatio-temporal changes in stream water quality over a large region consisting of multiple catchments. Specifically, this model was established using long-term (21 years) stream water quality observations across 102 catchments in Australia, with an aggregate catchment area of 130,000 km2. To obtain the necessary understanding of process drivers required to develop this model, two preceding studies were conducted on the same dataset to identify the key drivers for the spatial and temporal variability of water quality, respectively (Lintern et al., 2018b; Guo et al., 2019). The*

aim of this study is to develop an integrated spatio-temporal model using the previously-identified spatial and temporal predictors, and to then assess the performance of this model.'

To complete this discussion, we have also highlighted the capability of our model in enabling regional-scale modelling with limited information in Discussion Section 4.1.

- *L527: 'In this study, we developed the first process-informed statistical model that is capable of explaining a reasonable proportion of water quality variability for a large spatial area of over 130,000km2. Although the calibration data have relatively low sampling frequency (i.e. monthly), our model generally performs satisfactorily in explaining the total variability in water quality. This demonstrates the effectiveness of the Bayesian hierarchical modelling framework in predicting spatio-temporal variability in water quality across large scales. The Bayesian hierarchical model is: a) more advantageous than other simpler statistical water quality models with its more comprehensive and process-informed approach, and capacity to represent varying temporal relationships across large-scale regions; b) less demanding for input data compared with those required by fully-distributed, processes-based models.'*

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 of 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, 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 this comment is addressed along with our responses to your 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).*
- *Improving justification of the knowledge gaps that we lack modelling capacity of water quality for large-scale and across multiple catchments, for which two critical limitations are the lack in both understanding and information (Re Comments #3).*

The above-mentioned manuscript revisions can help to address this comment too by strengthening the justification for applying the Bayesian hierarchical model.

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).

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 have clarified 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 in Section 2.1.2:

- *L165: 'Note that in the sampling protocol, FRP is defined as 'Reactive Phosphorus for a filtered sample to a defined filter size (e.g. RP(<0.45 µm))', which is equivalent to the more widely-used terminology, SRP i.e. Soluble Reactive Phosphorus (Jarvie et al., 2002).'*

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.

To address this comment, we thoroughly revised the Abstract, Introduction and Conclusion to improve clarification of the knowledge gaps and the corresponding study objectives, and how this study links to/differs from its preceding works. Key revisions include:

- *L11 (Abstract): 'Our current capacity to model stream water quality is limited particularly at large spatial scales across multiple catchments. To address this, we developed a Bayesian hierarchical statistical model to simulate the spatio-temporal variability in stream water quality across the state of Victoria, Australia. The model was developed using monthly water quality monitoring data over 21 years, across 102 catchments, which span over 130,000 km2.'*
- *L75 (Introduction): 'Due to the limited understanding of large-scale water quality patterns, we currently lack the capacity to model spatio-temporal variabilities in water quality at larger scales across multiple catchments. This hinders our ability to inform the development of effective policy and mitigation strategies over large regions. ... Modelling the spatio-temporal variability simultaneously remains challenging over long time periods and large regions.'*
- *L90 (Introduction): 'Accordingly, this research attempts to bridge the gap between fully-distributed physically-based water quality models and data-driven statistical approaches. We aim to develop a process-informed, data-driven model to predict spatio-temporal changes in stream water quality over a large region consisting of multiple catchments. Specifically, this model was established using long-term (21 years) stream water quality observations across 102 catchments in Australia, with an aggregate catchment area of 130,000 km2. To obtain the necessary understanding of process drivers required to develop this model, two preceding studies were conducted on the same dataset to identify the key drivers for the spatial and temporal variability of water quality, respectively (Lintern et al., 2018b; Guo et al., 2019). The aim of this study is to develop an integrated spatio-temporal model using the previously-identified spatial and temporal predictors, and to then assess the performance of this model.'*
- *L666 (Conclusion): 'This study aims to address the current lack of water quality models that operate at large scales across multiple catchments. To achieve this, we used long-term stream water quality data collected from 102 sites in south-eastern Australia, and developed a Bayesian hierarchical statistical model to simulate the spatio-temporal variabilities in six key water quality constituents: TSS, TP, FRP, TKN, NOx and EC. The choice of model predictors was guided by previous studies on the same dataset (Lintern et al., 2018b; Guo et al., 2019).'*

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.

Please note that this sentence within the abstract intends to provide only a brief introduction of potential model improvements. We have added some details to proposed strategies in Section 4.1 (Implications for statistical water quality modelling) as:

- *L572: 'To better capture changes in reactive constituents, the model may require greater consideration of and more extensive spatial and temporal data to represent bio-geochemical processes. Examples include improvements on the process representation for nitrogen cycling and the desorption and adsorption of phosphorus (Granger et al., 2010;Smyth et al., 2013;Tian and Zhou, 2007).'*

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. The impact of higher temporal resolution of input data on model performance needs more in-depth discussion. To avoid confusion, we have deleted this statement from the abstract.

We have added more discussion In Section 4.2 on the opportunities/challenges associated with using high-frequency data in our modelling framework:

- *L593: 'Utilizing data with higher temporal resolution may further strengthen the model capacity to explain temporal variability, especially by capturing more information on water quality dynamics during flow events. This may be possible into the future; however, current high-frequency water quality sensors (Bende-Michl and Hairsine, 2010;Outram et al., 2014;Lannergård et al., 2019;Pellerin et al., 2016) still have very high resourcing requirements that limits widespread deployment in operational networks.'*

We also added an open question to the end of the Conclusion on value of using high-frequency data in our modelling framework:

- *L686: 'The inclusion of high-frequency water quality sampling data may also extend the model's ability to represent temporal variability. However, high-frequency water quality data are also typically highly variable with large noise. Therefore, the implication of such data for the spatio-temporal modelling framework remains an open question, which needs further investigation in future applications of this modeling framework.'*

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 our model considers would require 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. This requires further discussions to be clearly communicated. Therefore, we have deleted this statement from the abstract to avoid confusion.

To address this comment, we have included land management with some brief examples in the relevant discussions in Section 4.2 (Implications for water quality monitoring programs):

- *L598: 'Furthermore, changes in land use and management over time are currently not considered here as predictors of temporal variability in water quality, which include but not limit to land clearing, urbanization, tillage, fertiliser application and irrigation. This is due to a complete lack, or inconsistency of available data. However, changes in land use/land management practices can occur over short time periods, which can lead to increases in pollutant sources and changes to runoff generation processes (e.g. Tang et al., 2005;DeFries and Eshleman, 2004;Smith et al., 2013). Therefore, our modelling framework can potentially be improved by having additional monitoring data on the temporal patterns of land use/land management to better capture their impacts on water quality.'*

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 have revised this sentence as:

- *L38: 'Reducing these impacts requires effective management and mitigation of poor water quality; however, high variability in water quality both across space and time reduces our ability to accurately assess the status of water quality and to develop effective management strategies. Thus, improved modelling frameworks to predict and interpret this variability would be useful for water quality management (Chang, 2008;Ai et al., 2015;Zhou et al., 2012).'*

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 have condensed these descriptions while mentioning the three key processes, which we consider as important background information for the broad readership of HESS as a multi-disciplinary journal:

- *L47: 'These variabilities in stream water quality are driven by three key mechanisms: (1) source, which defines the total amount of constituents being available in a catchment; (2) mobilization, which detaches constituents (both in particulate and dissolved forms) from their sources via processes such as erosion and biogeochemical processing; and (3) delivery of mobilized constituents from catchments to receiving waters via multiple hydrologic pathways including surface and subsurface flow (Granger et al., 2010).'*

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?).

Agreed. To address this comment, we have improved the emphasis on emission dynamics, specifically on water quality variation due to changes in land use and land management etc., with condensed discussions on individual hydro-climatic factors as following:

- *L56: 'At the same time, temporal shifts in water quality are also influenced by changes in pollutant sources, such as land use and land management including urbanization, agriculture*

and vegetation clearing (Ren et al., 2003;Smith et al., 2013;Ouyang et al., 2010). In addition, water quality can also vary in time with variations in the mobilization and delivery processes, which are largely driven by the hydro-climatic conditions at a catchment, such as streamflow (Ahearn et al., 2004;Mellander et al., 2015;Sharpley et al., 2002;Zhang and Ball, 2017), the timing and magnitude of rainfall events (Fraser et al., 1999;Miller et al., 2014) and temperature (Bailey and Ahmadi, 2014).'

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.

Firstly, we believe there is a lack of both information and understanding, as explained in our response to your Comment #3. To reflect this, we have revised this sentence that you referred to, 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:

- *L63: 'As abovementioned, we have good understanding of the key controls for variations in water quality, albeit in an isolated, idealized context. We still lack a sound understanding of how relationships between specific landscape characteristics and water quality can shift with influences from other landscape characteristics, and how the drivers of temporal variability in water quality can interact and vary across large spatial scales (Musolff et al., 2015;Lintern et al., 2018a;Ali et al., 2017). In contrast, current detailed understanding have been primarily based on field studies at small scales with detailed information on specific temporal drivers ranging from hydrologic conditions to detailed management decisions such as fertilizer rates and application timing (Smith et al., 2013;Poudel et al., 2013;Adams et al., 2014). While operational weather observation netowrks, stream gauging networks and remote sensing can provide some of this information, developing a large-scale understanding of water quality patterns across catchments would ideally also involve an extensive suite of management information that substantially exceeds what is currently available.'*

Modelling capacity at 'larger scales' refers to the ability to model across multiple catchments over large geographical regions. We have better clarified this by highlighting 'multiple catchments' in the discussions following this sentence:

- *L75: 'Due to the limited understanding of large-scale water quality patterns, we currently lack the capacity to model spatio-temporal variabilities in water quality at larger scales across multiple catchments.'*

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 have briefly discussed these in the Introduction, as:

- *L105: 'The model can potentially provide useful information for large-scale catchment management, assessment and policy making, such as testing major changes in land use patterns, informing pollution hot-spots, as well as identification and attribution of water quality trends and changes over time.'*

In Discussion Section 4.1, we further detailed how improved abilities to model water quality variability can benefit management, with references to the results presented:

- *L535: 'From a practical perspective, this model has the potential to contribute to a number of management activities including catchment planning, management and policy-making activities, specifically:*
  1) *The spatial predictive capacity can be used to identify pollution hot-spots and the catchment conditions that are likely causes of high concentrations. This can be used to help identify target catchment(s) to prioritize future water quality monitoring and management (Figs. 4 and 5);*
  2) *Further to 1), since water quality has been linked with catchment characteristics in this model, it can also be used to assess potential impacts of alternative options of land use and land cover change, as well as potential effects of climate change, on ambient water quality conditions;*
  3) *The model's temporal predictive capacity can identify changes in water quality due to changes in hydro-climatic conditions and vegetation cover, and thus enabling attribution of detected trends. On the other hand, any 'unexpected' trends can be identified to prompt further investigation to identify causes (Figure 6 and Table 4). The model could also be used for assessing the impacts of long-term catchment changes on water quality (Figures 7 and 8).'*

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 for highlighting this great point, however, this is not relevant to our key study objectives (developing modelling capacities for large regions across multiple catchments) so we decided to not to divert the flow with this discussion.

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

Due to the substantial revision of the Introduction, this sentence has been revised as:

- *L75: 'Due to the limited understanding of large-scale water quality patterns, we currently lack the capacity to model spatio-temporal variabilities in water quality at larger scales across multiple catchments. This hinders our ability to inform the development of effective policy and mitigation strategies over large regions.'*

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 have briefly discussed these in the Introduction, as:

- *L97: 'The aim of this study is to develop an integrated spatio-temporal model using the previously-identified spatial and temporal predictors, and to then assess the performance of this model. …The model can potentially provide useful information for large-scale catchment management, assessment and policy making, such as testing major changes in land use patterns, informing pollution hot-spots, as well as identification and attribution of water quality trends and changes over time.'*

19. Line 103: Please fix "Beyesian".

This paragraph has been moved to the start of Section 2 and revised as:

- *L110: 'We first discuss the process used to develop the integrated spatio-temporal model (Section 2.1). Sections 2.1.1 and 2.1.2 introduces the statistical modelling framework and the data used for model development, respectively. The approaches to determine model structure was then introduced, which include the choice of key predictors (Section 2.1.3) and the calibration for model parameters (Section 2.1.4). Fianlly, the approaches to evaluate model performance and robustness are described in Section 2.2.'*

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 reluctantly still do so.

We have deleted 'however' (now in L152).

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

We have provided details of all potential predictors 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, since improving understanding water quality spatial variability is not the key focus of this study (as in our responses to your Comments #1 and #2.3), we prefer to keep the descriptions of the relevant approach brief, and referring the readers to Lintern et al. (2019b) for more detailed. We have revised relevant sentences to better clarify this, as:

- *L173: 'To compile a dataset for the potential spatial explanatory variables (i.e. predictors to explain spatial variability in water quality), a comprehensive literature review was conducted (Lintern et al., 2018a), which summarized the key catchment landscape characterisitics that are widely known to influence water quality. Furthre, as part of Lintern et al. (2018b), fifty potential explanatory catchment characteristics were selected, which included catchment land use, land cover, topographic, climatic, geological, lithological and hydrological catchment characteristics. These variables were derived using datasets obtained from Geoscience Australia (2004, 2011), the Bureau of Meteorology (2012), the Bureau of Rural Sciences (2010), Department of Environment Land Water and Planning Victoria (2016) and the Terrestrial Ecosystem Research Network (2016) (see Table S1 in the Supplementary Material for detailed variable names and data sources).'*

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 have replaced the original phrase 'streamflow (mm d^{-1})' here with 'runoff depth (mm d^{-1})' (L188) 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 have added details of this dataset as:

- *L195: '…soil moisture for the root-zone and the deep-zone (averaged volumetric content for shallower and deeper than 1m, respectively).'*

Reference:

- *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 L210 (L156 in the previous version): '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:

- *L583: 'For future implementations, the established model structure and parameterization would be best suited to within the study region. Before performing new simulations (e.g. for new monitoring sites or for current study sites over a different time-period), the statistical properties of the new input datasets should be checked to ensure that they are similar to the calibration datasets. To model new catchments outside of the study region, a re-calibration of the model is required. This would involve extensive selection of key predictors and model calibration, much as performed in this study and the two preceding ones (Lintern et al., 2018b; Guo et al., 2019). A sufficiently long record length (e.g. 20 years) is ideal for such modelling, as it ensures a reasonable understanding of the temporal variability to be obtained.'*

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 model predictors are informed by our two previous companion studies (Lintern et al., 2018b; Guo et al., 2019), from which the key spatial and temporal controls have been selected. Therefore, the model development in this study did not involve additional prediction selection processes. We believe that this comment can

be addressed by our revisions in response to your Comment #1 and #2.3, which have better clarified the focus of this study and its differences to the two preceding studies:

- *L90 (Introduction): 'Accordingly, this research attempts to bridge the gap between fully-distributed physically-based water quality models and data-driven statistical approaches. We aim to develop a process-informed, data-driven model to predict spatio-temporal changes in stream water quality over a large region consisting of multiple catchments. Specifically, this model was established using long-term (21 years) stream water quality observations across 102 catchments in Australia, with an aggregate catchment area of 130,000 km2. To obtain the necessary understanding of process drivers required to develop this model, two preceding studies were conducted on the same dataset to identify the key drivers for the spatial and temporal variability of water quality, respectively (Lintern et al., 2018b; Guo et al., 2019). The aim of this study is to develop an integrated spatio-temporal model using the previously-identified spatial and temporal predictors, and to then assess the performance of this model.'*
- *L666 (Conclusion): 'This study aims to address the current lack of water quality models that operate at large scales across multiple catchments. To achieve this, we used long-term stream water quality data collected from 102 sites in south-eastern Australia, and developed a Bayesian hierarchical statistical model to simulate the spatio-temporal variabilities in six key water quality constituents: TSS, TP, FRP, TKN, NOx and EC. The choice of model predictors was guided by previous studies on the same dataset (Lintern et al., 2018b; Guo et al., 2019).'*

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. We also believe that this comment can be addressed by our revisions in response to your Comment #1 and #2.3 which have better clarified the focus of this study and its differences to the two preceding studies:

- *L90 (Introduction): 'Accordingly, this research attempts to bridge the gap between fully-distributed physically-based water quality models and data-driven statistical approaches. We aim to develop a process-informed, data-driven model to predict spatio-temporal changes in stream water quality over a large region consisting of multiple catchments. Specifically, this model was established using long-term (21 years) stream water quality observations across 102 catchments in Australia, with an aggregate catchment area of 130,000 km2. To obtain the necessary understanding of process drivers required to develop this model, two preceding studies were conducted on the same dataset to identify the key drivers for the spatial and temporal variability of water quality, respectively (Lintern et al., 2018b; Guo et al., 2019). The aim of this study is to develop an integrated spatio-temporal model using the previously-identified spatial and temporal predictors, and to then assess the performance of this model.'*
- *L666 (Conclusion): 'This study aims to address the current lack of water quality models that operate at large scales across multiple catchments. To achieve this, we used long-term stream water quality data collected from 102 sites in south-eastern Australia, and developed*

a Bayesian hierarchical statistical model to simulate the spatio-temporal variabilities in six key water quality constituents: TSS, TP, FRP, TKN, NOx and EC. The choice of model predictors was guided by previous studies on the same dataset (Lintern et al., 2018b; Guo et al., 2019).'

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 have added more justification for the choice of this priors as:

- *L284: 'The hyper-parameters were further assumed to be drawn from minimally informative prior distributions, following recommendations in Gelman (2006) and Stan Development Team (2019): for all the hyper-parameter means, a normal prior distribution of N(0,5) was used; for all the hyper-parameter standard deviations, a half-normal prior distribution of N(0,10) was used, which was truncated to only positive values.'*

Reference:

- *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(now referred to as below-detection-limit or below-DL data in the revised manuscript) 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 added this interpretation to better highlight the utility of this specific performance evaluation:

- *L301: 'The simulations from the fitted model and the corresponding observed concentrations were compared at 102 sites altogether to understand how the overall spatio-temporal variabilities were captured. For each constituent, this evaluation was performed with: 1) these above-DL data to focus only on data used for calibration (as detailed in Section. 2.1.2); 2) the full dataset including the below-DL data (set to half of the DL of the specific constituent), to understand how well the model represents the full distribution of constituent concentrations. A good model performance when including the below-DL data would suggest that the calibrated model is transferable to below-DL data too.'*

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' (now in L310).

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 added comments on the poor performance of FRP and have specifically discussed several factors that might contribute to this model limitation:

- *L548: 'Despite the opportunities highlighted above, the model's performance also suggests some current limitations of the modelling framework in the following situations:*
  1) *High within-site temporal variability….*
  2) *Presence of high proportions of below-DL data..*
  3) *Non-conservativeness constituents…*

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 L227, Section 2.1.2).

Also, as in response to your Comment #2, we added more results to summarize different aspects of model performance, and also to illustrate model utilities that are useful for catchment management. Some key additions are:

1) *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 (Fig. 3);*

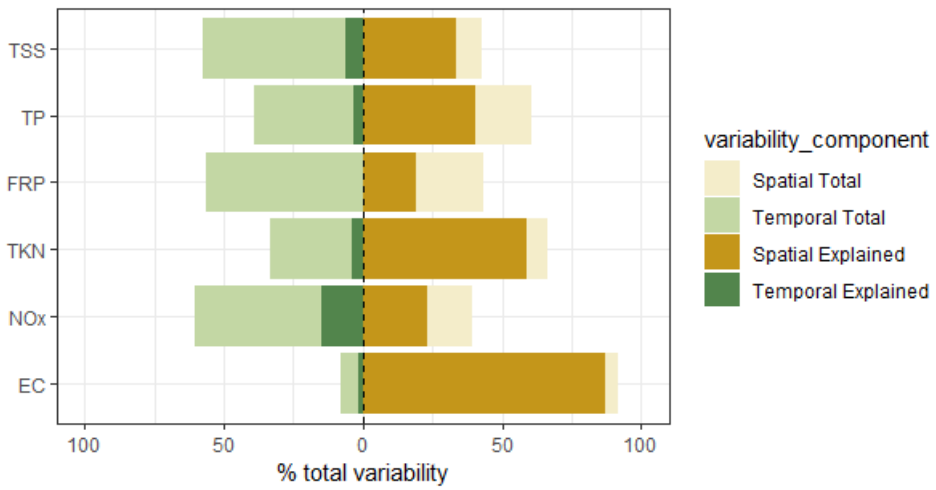


Figure 3. Observed spatial and temporal variabilities as proportions of the total variability (total width of each bar, 100%). The dashed line differentiates temporal variability (left side) with spatial variability (right side), and the

darker colours highlight the proportions of spatial and temporal variabilities that are explainable by the model. All values were estimated in Box-Cox transformed space.

2) *Modelled and observed time-series at selected catchments, to illustrate model capacity to predict trends and changes over time (Fig. 6);*

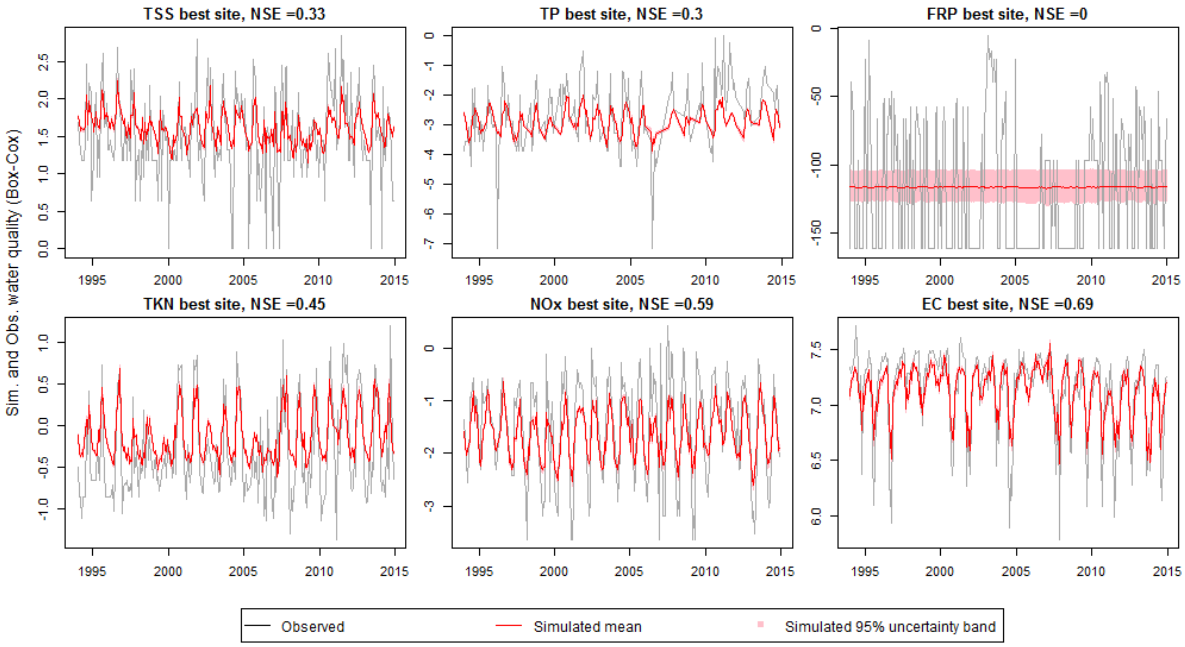


Figure 6. Model fit of the within-site (temporal) water quality variability, illustrated with the observed and simulated time-series for the best-performing site for each constituent. All values are presented in Box-Cox transformed space. The NSE for each constituent is also shown. The red line indicates the corresponding mean of all posterior simulations, while the pink bands show the corresponding 95% lower and upper bounds (only visible for FRP).

3) *Extending the existing model cross-validation from 5 replicates to 50 replicates (each calibrated with 80% monitoring sites and validated on the remaining 20% sites), to provide a more comprehensive summary on model sensitivity to calibrated dataset.*

Table 5. Comparison of model performances (as NSE) of the full model (Column 2) and the 50 partial models (Columns 3 to 5) with each calibrated to 80% randomly selected monitoring sites. Columns 3 to 5 summarize the mean, minimum and maximum NSE values across the 50 runs, where for each constituent, the top row showing calibration performance and the bottom row showing the validation performance (i.e. at the 20% sites that were not used for calibration).

Constituent	Full model	50 CV mean	50 CV min	50 CV max
TSS	0.225	0.413	0.376	0.439
		0.382	0.292	0.513
TP	0.433	0.461	0.427	0.501
		0.411	0.151	0.575
FRP	-1.92	0.168	0.067	0.232
		0.129	-0.078	0.272
TKN	0.658	0.654	0.622	0.670
		0.622	0.468	0.691
NO_x	0.216	0.453	0.414	0.489
		0.397	0.258	0.563
EC	0.907	0.893	0.882	0.903

	0.875	0.809	0.924

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

Thank you, we have deleted 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 have revised the sentences as suggested (now in L466).

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. Therefore, Tables 1 and 2 are necessary to illustrate key capabilities of this model. 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 mentioned in these responses, we have thoroughly revised the Introduction and Conclusion to better clarify the focus of this study.

- *L90 (Introduction): 'Accordingly, this research attempts to bridge the gap between fully-distributed physically-based water quality models and data-driven statistical approaches. We aim to develop a process-informed, data-driven model to predict spatio-temporal changes in stream water quality over a large region consisting of multiple catchments. Specifically, this model was established using long-term (21 years) stream water quality observations across 102 catchments in Australia, with an aggregate catchment area of 130,000 km2. To obtain the necessary understanding of process drivers required to develop this model, two preceding studies were conducted on the same dataset to identify the key drivers for the spatial and temporal variability of water quality, respectively (Lintern et al., 2018b; Guo et al., 2019). The aim of this study is to develop an integrated spatio-temporal model using the previously-identified spatial and temporal predictors, and to then assess the performance of this model.'*
- *L666 (Conclusion): 'This study aims to address the current lack of water quality models that operate at large scales across multiple catchments. To achieve this, we used long-term stream water quality data collected from 102 sites in south-eastern Australia, and developed a Bayesian hierarchical statistical model to simulate the spatio-temporal variabilities in six key water quality constituents: TSS, TP, FRP, TKN, NOx and EC. The choice of model predictors was guided by previous studies on the same dataset (Lintern et al., 2018b; Guo et al., 2019).'*

In addition, as responded to Comment #1, we have included new results to summarize on how the temporal variations of water quality vary spatially (Table 2), as this is a new finding that has not been reported in preceding studies.

Table 2. The key catchment landscape characteristics that are related to the varying relationships of water quality and same-day streamflow across space, which were selected as the two predictors for the streamflow effect in our model. The corresponding Spearman's correlation (ρ, at $p<0.05$) between the effect of streamflow and each catchment characteristic is presented.

Constituent	Key factors that affect spatial variability in temporal effects	Spearman's ρ ($p<0.05$)
TSS	Annual rainfall	0.722
	Hottest month maximum temperature	-0.575
TP	Annual rainfall	0.695
	Percentage area used for cropping	-0.556
FRP	Percentage agriculture area	0.392
	Percentage area underlain by mixed igneous bedrock	0.314
TKN	Annual rainfall	0.713
	Hottest month maximum temperature	-0.618
NO_x	Total storage capacity of dams in catchment	-0.493
	Mean soil TN content	0.458
EC	Percentage area covered by grassland	-0.347
	Percentage area covered by woodland	-0.317

Furthermore, as in response to your Comment #2 and #34, we added more results to summarize different aspects of model performance, and also to illustrate model utilities that are useful for catchment management. We believe that these results would be more aligned with the key objective of this study. Key additions are:

1) *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 (Fig. 3);*
2) *Modelled and observed time-series at selected catchments, to illustrate model capacity to predict trends and changes over time (Fig. 6);*
3) *Extending the existing model cross-validation from 5 replicates to 50 replicates (each calibrated with 80% monitoring sites and validated on the remaining 20% sites), to provide a more comprehensive summary on model sensitivity to calibrated dataset (Table 5).*

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 mentioned in these previous responses, we have thoroughly revised the Introduction and Conclusion to better clarify the focus of this study.

- *L90 (Introduction): 'Accordingly, this research attempts to bridge the gap between fully-distributed physically-based water quality models and data-driven statistical approaches. We aim to develop a process-informed, data-driven model to predict spatio-temporal changes in*

stream water quality over a large region consisting of multiple catchments. Specifically, this model was established using long-term (21 years) stream water quality observations across 102 catchments in Australia, with an aggregate catchment area of 130,000 km2. To obtain the necessary understanding of process drivers required to develop this model, two preceding studies were conducted on the same dataset to identify the key drivers for the spatial and temporal variability of water quality, respectively (Lintern et al., 2018b; Guo et al., 2019). The aim of this study is to develop an integrated spatio-temporal model using the previously-identified spatial and temporal predictors, and to then assess the performance of this model.'

- *L666 (Conclusion): 'This study aims to address the current lack of water quality models that operate at large scales across multiple catchments. To achieve this, we used long-term stream water quality data collected from 102 sites in south-eastern Australia, and developed a Bayesian hierarchical statistical model to simulate the spatio-temporal variabilities in six key water quality constituents: TSS, TP, FRP, TKN, NOx and EC. The choice of model predictors was guided by previous studies on the same dataset (Lintern et al., 2018b; Guo et al., 2019).'*

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 have added Table 2 to show how the temporal effects vary spatially.

Table 2. The key catchment landscape characteristics that are related to the varying relationships of water quality and same-day streamflow across space, which were selected as the two predictors for the streamflow effect in our model. The corresponding Spearman's correlation (ρ, at $p<0.05$) between the effect of streamflow and each catchment characteristic is presented.

Constituent	Key factors that affect spatial variability in temporal effects	Spearman's ρ ($p<0.05$)
TSS	Annual rainfall	0.722
	Hottest month maximum temperature	-0.575
TP	Annual rainfall	0.695
	Percentage area used for cropping	-0.556
FRP	Percentage agriculture area	0.392
	Percentage area underlain by mixed igneous bedrock	0.314
TKN	Annual rainfall	0.713
	Hottest month maximum temperature	-0.618
NO_x	Total storage capacity of dams in catchment	-0.493
	Mean soil TN content	0.458
EC	Percentage area covered by grassland	-0.347
	Percentage area covered by woodland	-0.317

We also presented additional results to summarize different aspects of model performance. Specifically:

1) *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 (Fig. 3);*
2) *Modelled and observed time-series at selected catchments, to illustrate model capacity to predict trends and changes over time (Fig. 6);*
3) *Extending the existing model cross-validation from 5 replicates to 50 replicates (each calibrated with 80% monitoring sites and validated on the remaining 20% sites), to provide a more comprehensive summary on model sensitivity to calibrated dataset (Table 5).*

We believe these 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 have added a summary of the key model capabilities, contributions and opportunities at the start of Section 4.1.

- *L527: 'In this study, we developed the first process-informed statistical model that is capable of explaining a reasonable proportion of water quality variability for a large spatial area of over 130,000km2. Although the calibration data have relatively low sampling frequency (i.e. monthly), our model generally performs satisfactorily in explaining the total variability in water quality. This demonstrates the effectiveness of the Bayesian hierarchical modelling framework in predicting spatio-temporal variability in water quality across large scales. The Bayesian hierarchical model is: a) more advantageous than other simpler statistical water quality models with its more comprehensive and process-informed approach, and capacity to represent varying temporal relationships across large-scale regions; b) less demanding for input data compared with those required by fully-distributed, processes-based models.'*

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.

We have thoroughly revised Section 4.1 to discuss potential contributors to impact model performance in a more structured manner:

- *L549: 'Despite the opportunities highlighted above, the model's performance also suggests some current limitations of the modelling framework in the following situations:*
  1) *High within-site temporal variability….*
  2) *Presence of high proportions of below-DL data..*
  3) *Non-conservativeness constituents…*

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. As mentioned in our response to your last comment, we have thoroughly revised Section 4.1 to discuss potential contributors to impact model performance in a more structured manner. We believe that this comment can also be addressed by these changes:

- *L549: 'Despite the opportunities highlighted above, the model's performance also suggests some current limitations of the modelling framework in the following situations:*
  1) *High within-site temporal variability….*
  2) *Presence of high proportions of below-DL data..*
  3) *Non-conservativeness constituents…*

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. However, during revision of Section 4.1 we have identified several more important factors which influence model performance, so we decided to delete this discussion on the categorical behavior of data in lower concentrations.

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 have revised throughout the manuscript to highlight (as detailed in responses to your Comments #1, #2.3 and #36). For the Abstract and Introduction, we improved emphasizing that the key knowledge gap that this study addressed was the lack of statistical modelling approaches that are suitable for large-scale application (as we responded to your Comment #3):

- *L11 (Abstract): 'Our current capacity to model stream water quality is limited particularly at large spatial scales across multiple catchments. To address this, we developed a Bayesian hierarchical statistical model to simulate the spatio-temporal variability in stream water quality across the state of Victoria, Australia. The model was developed using monthly water quality monitoring data over 21 years, across 102 catchments, which span over 130,000 km2.'*
- *L75 (Introduction): 'Due to the limited understanding of large-scale water quality patterns, we currently lack the capacity to model spatio-temporal variabilities in water quality at larger scales across multiple catchments. This hinders our ability to inform the development of effective policy and mitigation strategies over large regions. ... Modelling the spatio-temporal variability simultaneously remains challenging over long time periods and large regions.'*
- *L90 (Introduction): 'Accordingly, this research attempts to bridge the gap between fully-distributed physically-based water quality models and data-driven statistical approaches. We aim to develop a process-informed, data-driven model to predict spatio-temporal changes in stream water quality over a large region consisting of multiple catchments. Specifically, this model was established using long-term (21 years) stream water quality observations across 102 catchments in Australia, with an aggregate catchment area of 130,000 km2. To obtain the necessary understanding of process drivers required to develop this model, two preceding studies were conducted on the same dataset to identify the key drivers for the spatial and temporal variability of water quality, respectively (Lintern et al., 2018b; Guo et al., 2019). The aim of this study is to develop an integrated spatio-temporal model using the previously-identified spatial and temporal predictors, and to then assess the performance of this model.'*

As mentioned in our previous responses, we also presented additional results to summarize different aspects of model performance. Specifically:

1) *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 (Fig. 3);*
2) *Modelled and observed time-series at selected catchments, to illustrate model capacity to predict trends and changes over time (Fig. 6);*
3) *Extending the existing model cross-validation from 5 replicates to 50 replicates (each calibrated with 80% monitoring sites and validated on the remaining 20% sites), to provide a more comprehensive summary on model sensitivity to calibrated dataset (Table 5).*

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, we politely disagree with you comment that our models are 'simple statistical models that lack any mechanistic background' and 'a lot of mechanistic knowledge is available for these water quality variables, no single bit of this knowledge is reflected by the model structure'. 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. We have better clarified the consultation in the revised manuscript to highlight the use of process-based evidences in determining the model structure:

- *L173: 'To compile a dataset for the potential spatial explanatory variables (i.e. predictors to explain spatial variability in water quality), a comprehensive literature review was conducted (Lintern et al., 2018a), which summarized the key catchment landscape characterisitics that are widely known to influence water quality. Furthre, as part of Lintern et al. (2018b), fifty potential explanatory catchment characteristics were selected, which included catchment land use, land cover, topographic, climatic, geological, lithological and hydrological catchment characteristics. These variables were derived using datasets obtained from Geoscience Australia (2004, 2011), the Bureau of Meteorology (2012), the Bureau of Rural Sciences (2010), Department of Environment Land Water and Planning Victoria (2016) and the Terrestrial Ecosystem Research Network (2016) (see Table S1 in the Supplementary Material for detailed variable names and data sources).'*

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 have added some discussions in Section 4.1 to highlight the significance and contributions that this study brings.

- *L527: 'In this study, we developed the first process-informed statistical model that is capable of explaining a reasonable proportion of water quality variability for a large spatial area of over 130,000km2. Although the calibration data have relatively low sampling frequency (i.e.*

monthly), our model generally performs satisfactorily in explaining the total variability in water quality. This demonstrates the effectiveness of the Bayesian hierarchical modelling framework in predicting spatio-temporal variability in water quality across large scales. The Bayesian hierarchical model is: a) more advantageous than other simpler statistical water quality models with its more comprehensive and process-informed approach, and capacity to represent varying temporal relationships across large-scale regions; b) less demanding for input data compared with those required by fully-distributed, processes-based models.'

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. However, this study has not been focused on these improvements so we prefer to keep this discussion brief. To address this comment, we added some examples to illustrate potential improvement of this modelling of biogeochemical processes within a statistical modelling framework:

- *L572: 'To better capture changes in reactive constituents, the model may require greater consideration of and more extensive spatial and temporal data to represent bio-geochemical processes. Examples include improvements on the process representation for nitrogen cycling and the desorption and adsorption of phosphorus (Granger et al., 2010;Smyth et al., 2013;Tian and Zhou, 2007).'*

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.

We agree with you that the untransformed plot can better help us to understand absolute model errors so we should discuss some results and implications of this. However, we also acknowledge that this is only one factor that potentially limited the model performance, and if all model performance evaluations are presented in the back-transformed space they could potentially mask the effects of all other potential factors that affect model performance (e.g. the LOR issue – now referred to as the 'detection-limit issue' in the revised manuscript, the limitation in simulating non-conservative constituents, and any changes in model performance across different monitoring sites and periods used for model calibration).

To resolve this comment, we moved Fig. S13 to the main text to better clarifying the back-transformed model performance – which becomes Fig. 5 and placed after the transformed model performance is shown in Fig. 4. We also added corresponding explanations on how the model performance is limited by back-transformation, as:

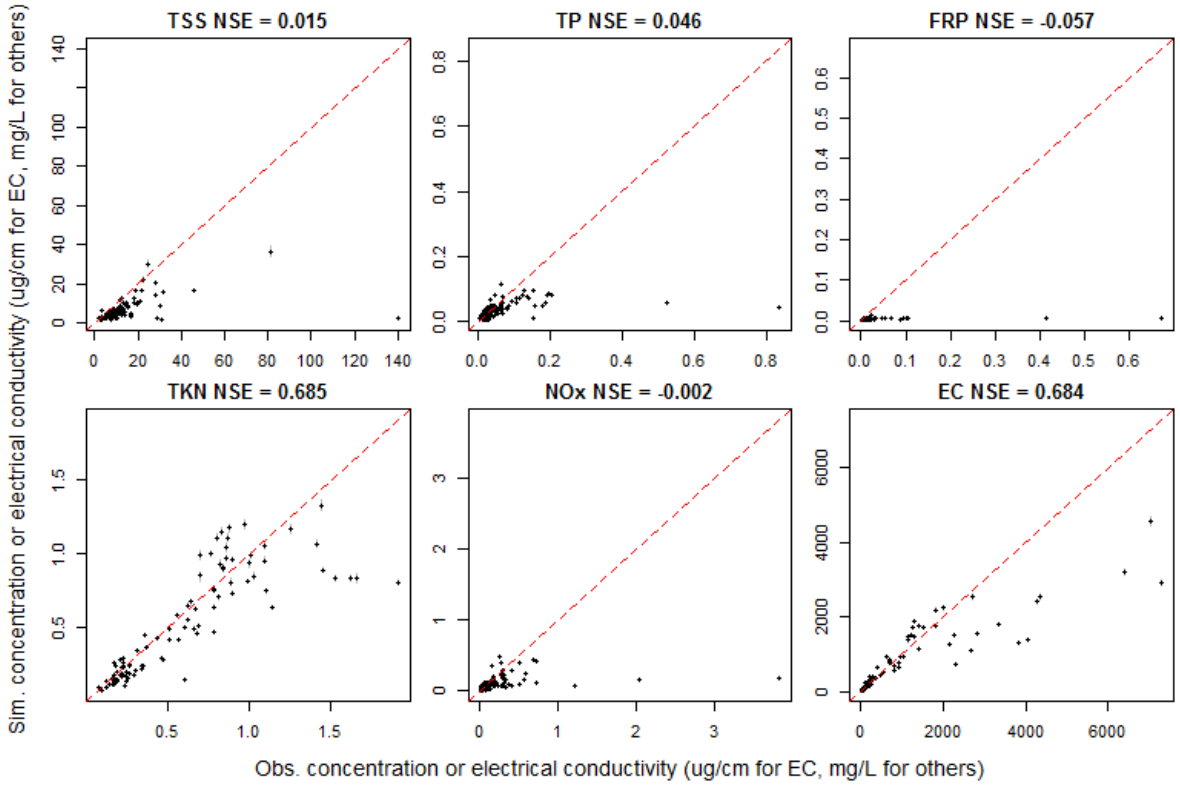


Figure 1. Back-transformation of the model simulations to the measurement scale emphasizes lack of fit for the highest concentrations, illustrated by simulated against observed site-level mean concentrations of each constituent in a back-transformed scale. The 95% lower and upper bounds of all posterior simulations shown in vertical grey lines. The NSE for each constituent is also shown and red dash lines show the 1:1 lines.

- *L422: 'At the back-transformed scale, the model shows greater biases for sites with higher concentrations (approximately the highest 10% sites for each constituent) (Fig. 5). This is not surprising as the model was fitted to a Box-Cox transformed space that reduces focus on high values and increases the focused on low values. This compromised its ability to represent sites with unusually high concentrations. The implications of the model having higher predictive capacity in the transformed scale is further discussed in Section. 4.1.'*

We also improved clarification on our model limitation in simulating absolute values, and recommended potential utilities for this transformed model in Section 4.1:

- *L577: 'As previously noted, our model was developed in a Box-Cox transformed scale to ensure the validity of the statistical assumptions (see details on data transformation in Sect. 2.1.2), which shows limited performance for high constituent concentrations when simulations are back-transformed to the measurement scale (Figs. 4 and 5). However, our model approximately represents proportional changes in water quality, which can thus help managers to understand proportional changes to inform practical catchment management.*

  Footnote: All Box-Cox transformation parameters for water quality constituents are approximately 0 (Table S4), which means that the transformations are similar to a log transformation.'

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 have removed this example and revised this discussion as:

- *L577: 'As previously noted, our model was developed in a Box-Cox transformed scale to ensure the validity of the statistical assumptions (see details on data transformation in Sect. 2.1.2), which shows limited performance for high constituent concentrations when simulations are back-transformed to the measurement scale (Figs. 4 and 5). However, our model approximately represents proportional changes in water quality, which can thus help managers to understand proportional changes to inform practical catchment management.*

  Footnote: All Box-Cox transformation parameters for water quality constituents are approximately 0 (Table S4), which means that the transformations are similar to a log transformation.'

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?

Firstly, we have added the following results to assess different aspects of model performance:

1) *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 (Fig. 3);*
2) *Modelled and observed time-series at selected catchments, to illustrate model capacity to predict trends and changes over time (Fig. 6);*
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 (Table 5).*

We believe that these additional results can better illustrated the value of our models to catchment management in a general sense. We have added discussion on how these model capabilities can benefit management in Section 4.1, as:

- *L535: 'From a practical perspective, this model has the potential to contribute to a number of management activities including catchment planning, management and policy-making activities, specifically:*
  1) *The spatial predictive capacity can be used to identify pollution hot-spots and the catchment conditions that are likely causes of high concentrations. This can be used to help identify target catchment(s) to prioritize future water quality monitoring and management (Figs. 4 and 5);*
  2) *Further to 1), since water quality has been linked with catchment characteristics in this model, it can also be used to assess potential impacts of alternative options of land use and land cover change, as well as potential effects of climate change, on ambient water quality conditions;*
  3) *The model's temporal predictive capacity can identify changes in water quality due to changes in hydro-climatic conditions and vegetation cover, and thus enabling attribution of detected trends. On the other hand, any 'unexpected' trends can be identified to prompt further investigation to identify causes (Figure 6 and Table 4). The model could also be used for assessing the impacts of long-term catchment changes on water quality (Figures 7 and 8).'*

Regarding the specific value that the transformed model added to management, we have explained this with reference to a log-transformed model which has often been used in assessing water quality changes in management, as:

- *L577: 'As previously noted, our model was developed in a Box-Cox transformed scale to ensure the validity of the statistical assumptions (see details on data transformation in Sect. 2.1.2), which shows limited performance for high constituent concentrations when simulations are back-transformed to the measurement scale (Figs. 4 and 5). However, our model approximately represents proportional changes in water quality, which can thus help managers to understand proportional changes to inform practical catchment management.*

  Footnote: All Box-Cox transformation parameters for water quality constituents are approximately 0 (Table S4), which means that the transformations are similar to a log transformation.'

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. We had performed a Web of Science search with key word 'high frequency water quality monitoring', and found that the majority of published research only appeared in this decade. Since this discussion focuses on utilizing high-frequency monitoring data for research (rather than developing high-frequency monitoring techniques), we prefer to retain the most relevant literature which have utilized such high-frequency monitoring data. To improve clarification, we have revised the sentence as:

- *L592: 'Utilizing data with higher temporal resolution may further strengthen the model capacity to explain temporal variability, especially by capturing more information on water quality dynamics during flow events. This may be possible into the future; however, current high-frequency water quality sensors (Bende-Michl and Hairsine, 2010;Outram et al., 2014;Lannergård et al., 2019;Pellerin et al., 2016) still have very high resourcing requirements that limits widespread deployment in operational networks.'*

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.

We agree that 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). However, these potential improvements are not the focus of this study and thus have not been assessed. Considering the growing length of the manuscript during revision, we have removed this discussion.

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

We appreciate and agree with your suggestion, however, this potential improvement (surrogate modelling) are not the focus of this study and thus have not been assessed. Considering the growing length of the manuscript during revision, we have removed this discussion.

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

Thank you for the suggestion. We have removed these 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 have included 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:

- *L598: 'Furthermore, changes in land use and management over time are currently not considered here as predictors of temporal variability in water quality, which include but not limit to land clearing, urbanization, tillage, fertiliser application and irrigation. This is due to a complete lack, or inconsistency of available data. However, changes in land use/land management practices can occur over short time periods, which can lead to increases in pollutant sources and changes to runoff generation processes (e.g. Tang et al., 2005;DeFries and Eshleman, 2004;Smith et al., 2013). Therefore, our modelling framework can potentially be improved by having additional monitoring data on the temporal patterns of land use/land management to better capture their impacts on water quality.'*

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

Thank you, we have expanded the discussion as:

- *L637: 'Similar to sediments, the impact of droughts on stream nutrient and salt concentrations have also commonly been understood as responses to reduced runoff generation and streamflow. In catchments with no significant point-source pollution, nutrient concentrations typically decreased during droughts (Mosley, 2015) with less nutrient leaching and overland flow, but may also increase due to increasing livestock inputs at more local scales (Caruso, 2002).'*

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. However, this potential improvement (long-term water quality trend analysis) are not the focus of this study and needs to be properly assessed in future studies. Considering the growing length of the manuscript during revision, we have removed this discussion.

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 have revised the sentence as:

- *L675: '… the spatio-temporal model can predict water quality in non-monitored locations under similar conditions to the historical period and the calibration catchments that we investigated.'*

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 so much for the suggestions. However, due to the substantial increase of replicates for partial calibration and validation (from 5 times to 50 times, see revised Section 2.2), these figures are now replaced with Figures S6 and S7 which better summarizes performance across 50 replicates. Please note that we have summarized all transformation parameters in Table S4.

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

Our manuscript revisions are underlined.

General comments

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

1. Model framework
   While I understand the notion of using catchment spatial variables to represent site-level mean (which is the focus of the published Lintern 18 paper), and using temporal variables to represent deviation from the mean (which is the focus of the published Guo 19 paper), I do not understand equation 6 –

1.1 Why is it necessary to add additional catchment characteristics in the temporal component? Why 2 variables? What's the implication of this equation for the framework overall? i.e. the framework started with distinct spatial and temporal components, but ended with the temporal component also include spatial variables?

Our modelling framework accounts for spatial variation in the parameter of each predictor that has been selected to explain the temporal variability, which were observed in Guo et al. (2019), as well as in Musolff et al. (2015) and Poor and McDonnell (2007) from separate datasets. Therefore, the purpose of Eqn. 6 is to explain these spatial variations and thus enabling spatial prediction of those temporal effects according to catchment characteristics. This equation essentially makes the modelling framework fully spatio-temporal (i.e. being able to predict any location at any time-step). The choice of two variables was mainly due to the consideration of controlling model complexity (i.e. number of parameters).

- *Musolff, A., Schmidt, C., Selle, B., and Fleckenstein, J. H.: Catchment controls on solute export, Advances in Water Resources, 86, 133-146, https://doi.org/10.1016/j.advwatres.2015.09.026, 2015.*
- *Poor, C. J., and McDonnell, J. J.: The effects of land use on stream nitrate dynamics, Journal of Hydrology, 332, 54-68, https://doi.org/10.1016/j.jhydrol.2006.06.022, 2007.*

1.2 Wouldn't that means the spatial variables are double counting, i.e. does this lead to the model overly focusing on spatial variability while less representing temporal variations? In any case, this need to be explained better in the manuscript.

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

set of catchment characteristics to explain the spatial variation in ambient (average) water quality conditions (with more details in Lintern et al., 2018b). Therefore, both sets of spatial predictors serve unique purposes and are necessary components of the models.

To address both questions raised in Comment #1 and improving the clarification of modelling framework, we have:

1) *Added the following description when introducing the overall modelling framework to highlight the key model advantage of representing variable water quality dynamics across catchments:*

   - *L119: 'A key strength of applying the hierarchical model structure to analyze spatio-temporal variability is that this structure enables the key controls of temporal variability in water quality to vary across locations (Webb and King, 2009;Borsuk et al., 2001). This variability has been found to be important in other study regions where the (temporal) solute export regime varies with catchment characteristics such as climate and land use (Musolff et al., 2015;Poor and McDonnell, 2007). '*

2) *Added detailed description of Eqn. 6 in Section 2.1 to better justify the purpose of including this equation. We also further emphasized this additional modelling capacity (i.e. modelling temporal variability across catchments) that we gained from Eqn.6, apart from the two preceding studies:*

   - *L136: 'The selection of key spatial and temporal predictors for the model has been performed in our two preceding studies (Lintern et al., 2018b; Guo et al., 2019) and is briefly described in Section 2.1.3. Eq. 1 to 4 enable the model to separately represent the spatial and temporal variability in water quality; however, there is still a further step required to make the model fully spatio-temporal (i.e. being able to predict over both time and location). Specifically, in Guo et al. (2019), clear spatial variation was observed in the relationships between water quality and its key temporal predictors (i.e. in the $\beta T_{N,j}$ in Eq. 4). To be able to model multiple catchments across a large spatial area simultaneously, we must account for differences in these temporal influences across sites. To do this, the effect of each temporal variable at site j ($\beta T_{N,j}$ with N in 1,2, … n) is drawn from a distribution with a mean of $\mu\beta T_{N,j}$ (Eq. 5), which is then modelled with a linear combination of two additional chatchment characteristics, $ST_{N1,j}$ and $ST_{N2,j}$ (Eq. 6). Details of the selection for these two additional predictors are presented in Section 2.1.3.'*

3) *Presented additional results and discussions on the key drivers for varying temporal relationships across catchments to illustrate the value of this specific model component (Table 2).*

Table 2. The key catchment landscape characteristics that are related to the varying relationships of water quality and same-day streamflow across space, which were selected as the two predictors for the streamflow effect in our model. The corresponding Spearman's correlation (ρ, at p<0.05) between the effect of streamflow and each catchment characteristic is presented.

Constituent	Key factors that affect spatial variability in temporal effects	Spearman's ρ (p<0.05)
TSS	Annual rainfall	0.722
	Hottest month maximum temperature	-0.575
TP	Annual rainfall	0.695
	Percentage area used for cropping	-0.556
FRP	Percentage agriculture area	0.392
	Percentage area underlain by mixed igneous bedrock	0.314
TKN	Annual rainfall	0.713
	Hottest month maximum temperature	-0.618
NO_x	Total storage capacity of dams in catchment	-0.493

	Mean soil TN content	0.458
EC	Percentage area covered by grassland	-0.347
	Percentage area covered by woodland	-0.317

2. Model performance measures

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

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

Assuming $\sum \Delta ij$ can be close to 0 as the positive and negative derivations more or less cancel each other out.

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

Thank you, this is a very good point and we confirm that your interpretation about the spatial and temporal variabilities are all correct. We agree that the existing results presented on model performance are predominantly focused on spatial variability. To improve this, in the revision we will present more results on how the model represent temporal variability in the Results section. Specifically:

1) 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 (Fig. 3);

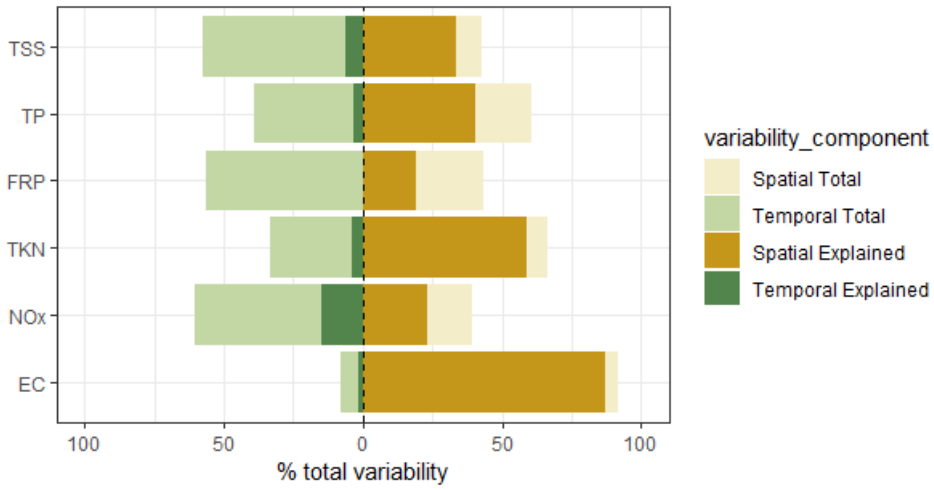


Figure 3. Observed spatial and temporal variabilities as proportions of the total variability (total width of each bar, 100%). The dashed line differentiates temporal variability (left side) with spatial variability (right side), and the darker colours highlight the proportions of spatial and temporal variabilities that are explainable by the model. All values were estimated in Box-Cox transformed space.

2) Modelled and observed time-series at selected catchments, to illustrate model capacity to predict trends and changes over time (Fig. 6);

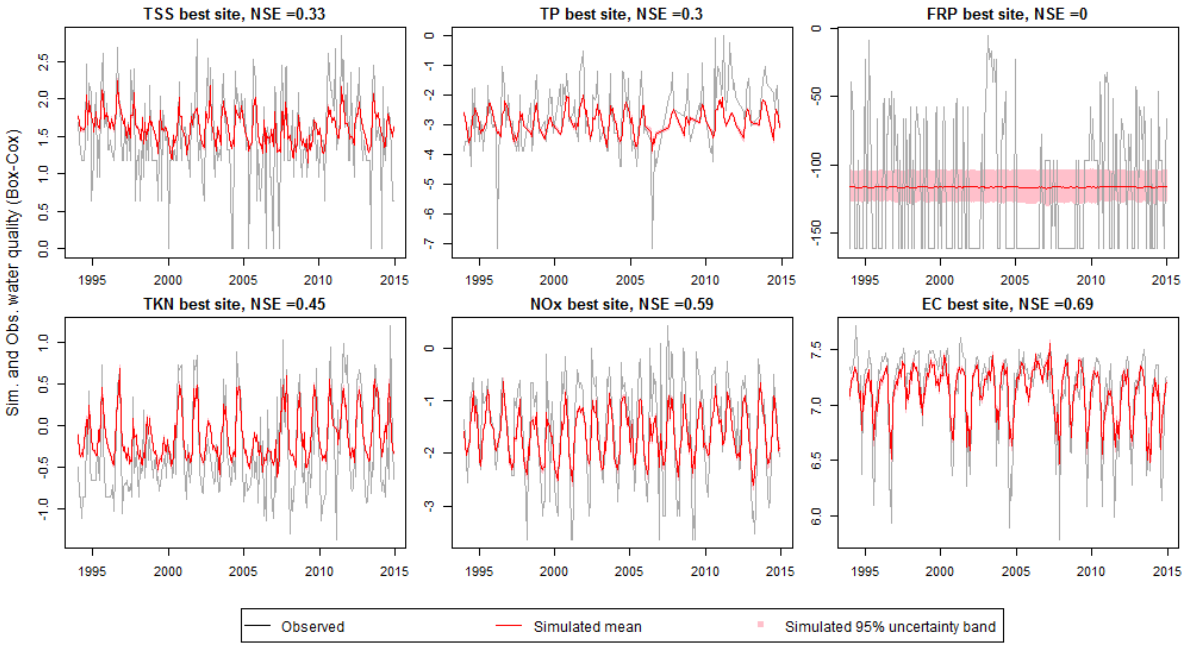


Figure 6. Model fit of the within-site (temporal) water quality variability, illustrated with the observed and simulated time-series for the best-performing site for each constituent. All values are presented in Box-Cox transformed space. The NSE for each constituent is also shown. The red line indicates the corresponding mean of all posterior simulations, while the pink bands show the corresponding 95% lower and upper bounds (only visible for FRP).

3) Summary of model capacity to detect water quality trends at all catchments (Table 4);

Table 4. Model ability to capture observed water quality trends across all monitoring sites for each constituent. The percentages of sites where observed positive and negative trends are captured by the model are presented separately. Values in brackets indicate numbers of sites where corresponding positive or negative trends are observed. For detailed estimation of these percentages please refer to Sect. 2.2.

Constituent	% positive trends captured	% negative trends captured
TSS	33.3 (12)	85.0 (20)
TP	82.1 (28)	16.7 (12)
FRP	47.1 (17)	55.6 (9)
TKN	81.1 (37)	40.0 (10)
NO_x	68.6 (35)	66.7 (27)
EC	82.6 (23)	77.3 (22)

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

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

We agree with the reviewer that the NSE achieved in our models is not as high as those recommended in Moriasi et al. (2007). However, this discrepancy should not be a key concern for

our models. Firstly, we would like to point out that the water quality models reviewed in Moriasi et al. (2007) were all physically-based, spatially-distributed models (SWAT, HSPT, DRAINMOD-DUFLOW and DRAINMOD-W) which focused on individual catchments within the US, where the key practical implication of modelling is to simulate catchment processes and management activities, and thus to support local catchment management. In contrast, the statistical models that we developed aimed to predict spatio-temporal variabilities over a large Australian region, which has an area of over 130,000 km^2 and more than 100 catchments. The key practical implication was to support higher-level catchment management at a state- or even a national-scale. Due to the different model types, contrasting scales and practical implications between our models and the models reviewed in Moriasi et al. (2007), we are not convinced that the performance standards summarized in Moriasi et al. (2007) are directly transferable to our models. Furthermore, due to the extensive spatial and temporal extents that our models cover and the less focus to support local-scale management activities, it is both more difficult and less necessary for our models to achieve the same performance standards as suggested in Moriasi et al. (2007).

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

1) *We completely agree with you that when a model is not performing well, we should be careful on drawing further inferences. However, we understand such 'further inferences' as to making predictions and/or interpreting parameter values with respect to physical processes. What we presented on the responses of TSS to drought was different to such 'inferences', as this analysis was based on a model validation against three distinct periods which are differently affected by a prolong drought in the region. In this experiment, the focus was not the model performances in an absolute perspective, but instead, the relative performances of different calibration/validation periods. Specifically, in Section 3.3, Figure 7 focused on how performance deteriorated when calibrating to one sub-period and validating on the other. Similarly, Figure 8 focused on the variation of model performance of the 'full-model' when simulating individual sub-periods of the full data period, so the focus is again on the relative model performance. We believe that exploring these 'changes in model performances' is an informative approach to explore any drought effects especially when the absolute model performances are not optimal. <u>To highlight this we added to the discussion:</u>*
   - *L616: 'Considering the limited performance of the TSS model (i.e. substantial under-estimation of temporal variability in Section 3.1) … calibrated parameters might be unreliable. However, this should not affect the reliability of the observed change in TSS since the drought (Section 3.3), which was based on the systematic differences of model fitting between different periods, revealing a broad-scale patterns across the state on the drought influences.'*
2) *We would like to clarify that although this analysis explored changes in water quality across different sub-periods of the full dataset, the focus was any consistent shift across three periods, as opposed to the day-to-day variability of water quality (i.e. as how 'temporal variability' has been defined in our modelling framework). We believe that such cross-period changes can be more clearly summarized with the long-term mean concentrations for each period, as currently presented. <u>Regarding model ability to represent temporal variability, we believed that this is now well addressed by the additional results we presented, as detailed in our responses to your Comment #2.1.</u>*

3. Site bias

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

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

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

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

To resolve this comment, we have firstly improved our model description in Section 2.1.1 (Spatio-temporal modelling framework), to further emphasize the point that the temporal parameters were allowed to vary across space, which considered potentially different key processes and controls for water quality across the diverse catchment conditions in our catchments. We have also provided some examples to explain how the key processes can vary across catchments:

- *L119: 'A key strength of applying the hierarchical model structure to analyze spatio-temporal variability is that this structure enables the key controls of temporal variability in water quality to vary across locations (Webb and King, 2009;Borsuk et al., 2001). This variability has been found to be important in other study regions where the (temporal) solute export regime varies with catchment characteristics such as climate and land use (Musolff et al., 2015;Poor and McDonnell, 2007). '*

3.2 Data transformation: the authors chose to transform observation data, rather than back-transform modelled data. There are a few issues with transforming observation data: 1) the transformation involves additional parameters (such as lambda, instead of a straight transformation, e.g. logx), thus the "observed" data is in effect a "modelled" data, albeit a

simple model. 2) The observation data across sites is transformed using the same parameter value (mean), thus the site bias issue in the comment above also applies. 3) the choice of transformation (log) leads to a decrease in the sensitivity of large values due to the log() function, and increase the sensitivity of small values. Thus, it is unclear to me whether using transformed observation data is any better than back-transforming modelled data. These implications need to be pointed out in the manuscript.

Transforming data for our modelling was a decision informed by previous phases of the same research project, where we used linear statistical models to identify the key drivers of water quality variability across space and time (Lintern et al., 2018b; Guo et al., 2019). To incorporate the previously obtained understanding into this study, we used similar linear model structures – which were calibrated with transformed data to satisfy the assumptions of linear modelling or otherwise the untransformed data would be too skewed to work with (see more details in Figure R1 below under point 3)).

We agree with you that the back-transformed plot can better help us to understand absolute model errors so we should discuss some results and implications of this. However, we also acknowledge that since the model was developed in a transformed space, performance evaluations in the transformed space would allow us to best explore a wide range of factors that can influence model performance (e.g. the LOR issue – now referred to as the 'detection-limit issue' in the revised manuscript, the limitation in simulating non-conservative constituents, and any changes in model performance across different monitoring sites and periods used for model calibration).

To resolve this comment, we first added justifications in Section 2.2 (Model performance and sensitivity analyses) on why model performance assessments are presented in a transformed scale.

- *L297: Since the model was calibrated in a Box-Cox transformation scale (see justification in Section 2.1.2), the Box-Cox transformation scale was used for model evaluation to enable a clear investigation on the influences of a wide range of factors that can influence model performance.'*

We also moved Fig. S13 to the main text to better clarifying the back-transformed model performance – which becomes Fig. 5 and placed after the transformed model performance is shown in Fig. 4. We also added corresponding explanations on how the model performance is limited by back-transformation, as:

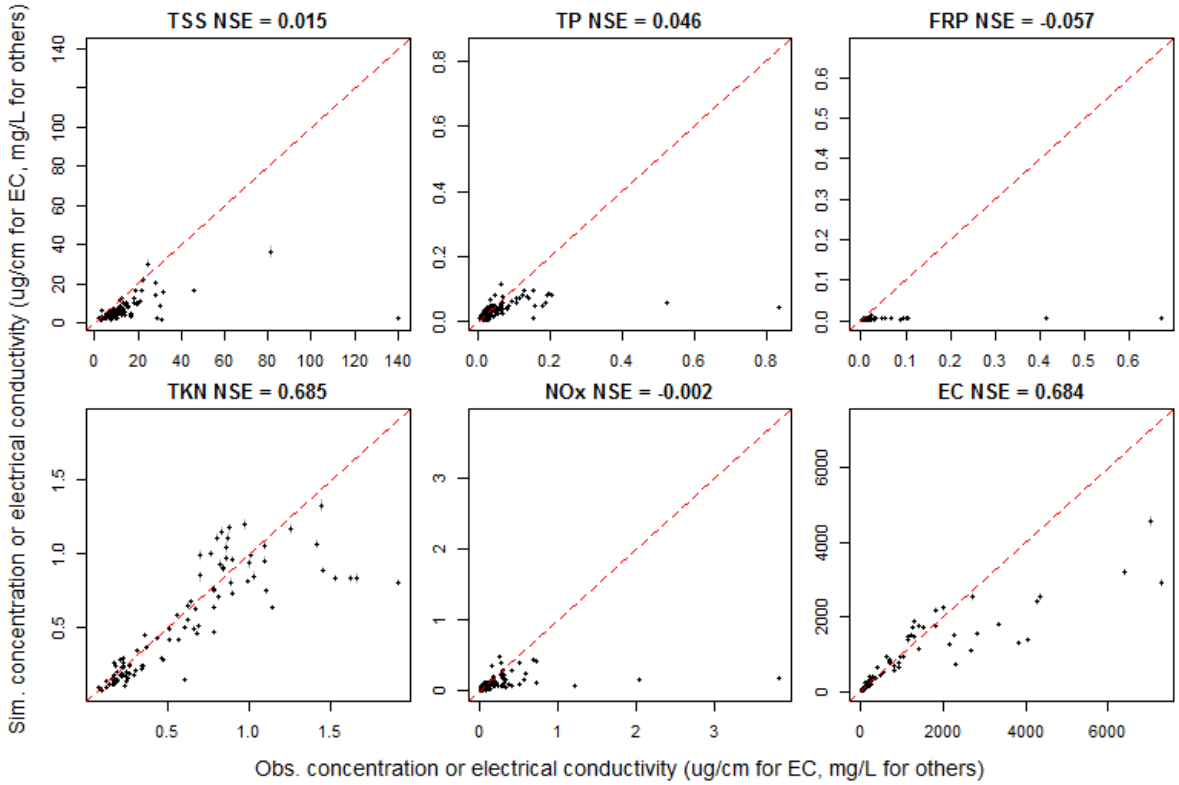


Figure 1. Back-transformation of the model simulations to the measurement scale emphasizes influences of unusually high concentrations and thus heavily affects model fitting, illustrated by simulated against observed site-level mean concentrations of each constituent in a back-transformed scale. The 95% lower and upper bounds of all posterior simulations shown in vertical grey lines.

- *L422: 'At the back-transformed scale, the model shows greater biases for sites with higher concentrations (approximately the highest 10% sites for each constituent) (Fig. 5). This is not surprising as the model was fitted to a Box-Cox transformed space that reduces focus on high values and increases the focused on low values. This compromised its ability to represent sites with unusually high concentrations. The implications of the model having higher predictive capacity in the transformed scale is further discussed in Section. 4.1.'*

In response to the specific issues that you raised on transformation:

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

low proportions within the data. We have highlighted this by the additional Figure S1 in Supplementary Material, in which untransformed data were plotted against corresponding quantiles, for each constituent. If those extreme values (right tails in each panel in Figure S1) were left untransformed, they may cause the models to emphasize too much on rare extreme events, and thus largely affect our ability to represent the overall large-scale patterns in water quality.

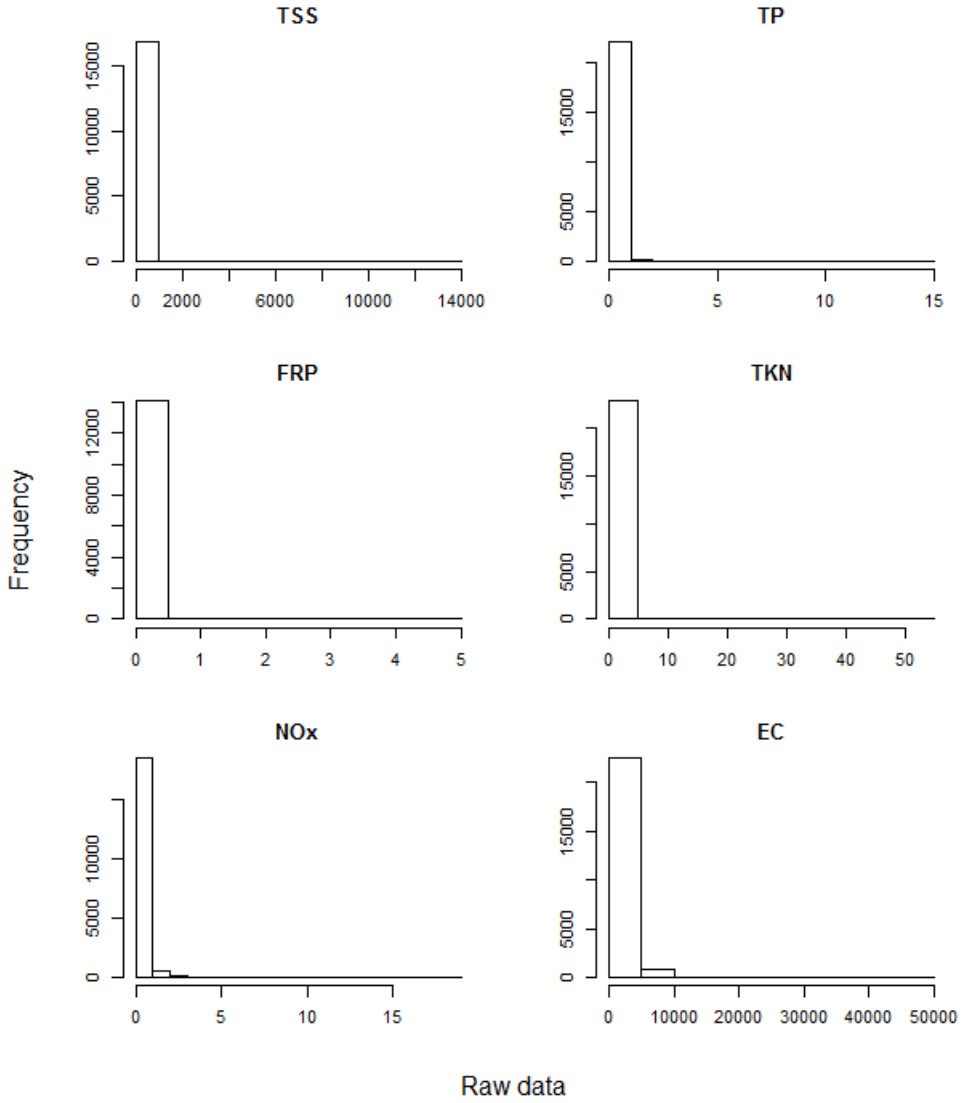


Figure S1. Distribution of the raw water quality data across all catchments. Each panel shows one constituent with only the above-DL data.

To better justify the need of data transformation, we have further strengthened the improvement that the transformation made on the data in Section 2.1.2:

- *L228: 'The transformation process has greatly improved the data symmetry and thus suitability to be used for use in a linear model (the quality of the transformations was assessed via visual inspection in Lintern et al., 2018b; Guo et al., 2019; and summarized in Figures S2, S4 and S6 in the Supplementary Material).'*

4. Results in drought impacts

Assuming the model is appropriate for inference (i.e. have good enough performance measure), a better (more insightful) way to demonstrate the impact of drought could be to show what the

parameters (beta) for pre and post drought models are. This is because (I assume) these parameters represent the system behaviours, i.e. how strong different explanatory variables are to predict concentrations.

Thanks for sharing the interesting idea. Firstly, we have compared the parameter values for the key spatial and temporal predictors of TSS when the model was calibrated to different periods. The effects of key predictors for spatial variability did not vary much across periods (Figure S14 as shown below). In contrast, the effects of key predictors for temporal variability showed a clear shift in the role of antecedent flow (prior 7-day flow) across different drought periods (Figure 9 as shown below). Specifically, the flow effects are mostly positive across catchments before the drought, which shift to mostly negative during the drought; after the drought, the flow effects have mixed directions among different catchments. We added these results when specifically discussing the drought impacts in Section 4.3, as:

- *L611: 'A further analysis of the calibrated model parameters for pre-, during and post-drought periods suggest that the effects of key spatial predictors do not vary much across periods (Figure S14). In contrast, the effects of key temporal predictors highlight a clear shift in the role of antecedent flow (prior 7-day flow) across different time periods (Figure 9). Specifically, the antecedent flow effects are mostly positive across catchments before the drought, and shift to mostly negative during the drought. After the drought, the antecedent flow effects have mixed directions among different catchments .'*

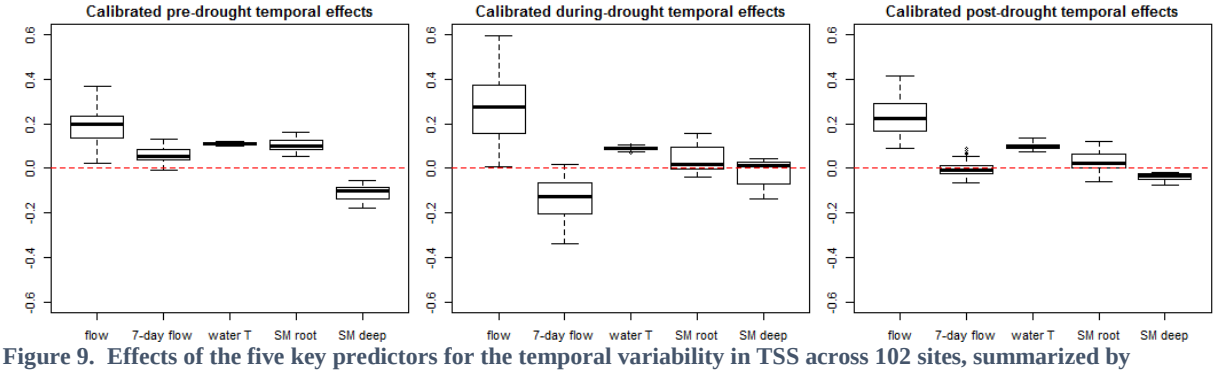


Figure 9. Effects of the five key predictors for the temporal variability in TSS across 102 sites, summarized by the posterior mean of the calibrated parameter values for each predictor (box shows values across all sites), from left: flow, 7-day antecedent flow, water temperature, root-zone soil moisture and deep soil moisture.

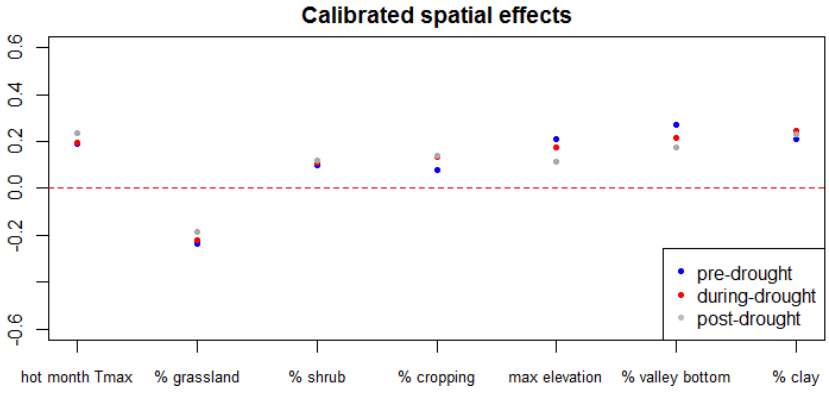


Figure S14. Effects of the seven key predictors for the spatial variability in TSS across 102 sites, summarized by the posterior mean of the calibrated parameter values for each predictor, to the pre-, during- and post-drought periods (differentiated by colour). The seven key predictors are, from left: hottest month maximum temperature, percentage catchment area as grassland, percentage catchment area as shrub, percentage catchment area as cropping land, maximum catchment elevation, percentage catchment area made up of valley bottoms, and average soil clay content.

After presenting these results, we have also added acknowledgement on the model deficiencies and thus recommended specific care in interpretation, as:

- *L616: 'Considering the limited performance of the TSS model (i.e. substantial under-estimation of temporal variability in Section 3.1), these changing relationships suggested in the calibrated parameters might be unreliable. However, this should not affect the reliability of the observed change in TSS since the drought (Section 3.3), which was based on the systematic differences of model fitting between different periods, revealing a broad-scale patterns across the state on the drought influences.'*

Other comments

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

The Box-Cox transformation which our models were developed with is essentially similar to log transformation, which is widely used in water quality to represent proportional differences in linear space. We have improved the clarity of the relevant discussions in Section 4.1 as:

- *L580: 'However, our model approximately represents proportional changes in water quality, which can thus help managers to understand proportional changes to inform practical catchment management.*

  Footnote: All Box-Cox transformation parameters for water quality constituents are approximately 0 (Table S4), which means that the transformations are similar to a log transformation.'

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

Agreed. We have merged part of Tables S5 and S6 related to the key spatial and temporal predictors of the model to Table 1. Since these are results reported in the two preceding studies (Lintern et al. 2018 and Guo et al. 2019b) which were used for model development in this study, this information is presented in Section 2.1.3 under the Method section of the main text.

Table 1. Key factors affecting the spatial and temporal variability for each of six constituents, as identified in Lintern et al. (2018) and Guo et al. (2019b), respectively.

Constituent	Key factors that affect spatial variability	Key factors that affect
TSS	Hottest month maximum temperature Percentage area covered by grass Percentage area covered by shrub Percentage cropping area Maximum elevation Dam storage Percentage clay area	Same-day streamflow 7-day antecedent streamflow Water temperature Soil moisture root Soil moisture deep
TP	Erosivity Percentage area covered by grass Percentage area covered by shrub Percentage area made up of roads Percentage cropping area Average soil TP content	Same-day streamflow 30-day antecedent streamflow Water temperature Soil moisture root Soil moisture deep
FRP	Percentage area covered by shrub Percentage cropping area Catchment area Average soil TP content Mean channel slope	Same-day streamflow Water temperature Soil moisture deep
TKN	Percentage clay area Warmest quarter mean temperature Coldest quarter rainfall	Same-day streamflow 30-day antecedent streamflow NDVI

	Percentage cropping area Percentage pasture area Average soil TP content	Water temperature Soil moisture root Soil moisture deep
NO_x	Annual radiation Warm quarter rainfall Hottest month maximum temperature Average soil TP content Mean channel slope	Same-day streamflow 30-day antecedent streamflow NDVI Water temperature Soil moisture root Soil moisture deep
EC	Annual radiation Annual rainfall Wettest quarter rain Hottest month maximum temperature Percentage agriculture area Percentage cropping area Percentage area covered by shrub Average soil TN content	Same-day streamflow 14-day antecedent streamflow Water temperature Soil moisture root Soil moisture deep

In addition, the second column of Table S6 (which summarizes the key factors relating to the spatial variability in temporal effects) have not been presented before. Therefore, these results are further enhanced and presented in Table 2 in Section 3.1, under the Results section.

Table 2. The key catchment landscape characteristics that are related to the varying relationships of water quality and same-day streamflow across space, which were selected as the two predictors for the streamflow effect in our model. The corresponding Spearman's correlation (ρ, at $p<0.05$) between the effect of streamflow and each catchment characteristic is presented.

Constituent	Key factors that affect spatial variability in temporal effects	Spearman's ρ ($p<0.05$)
TSS	Annual rainfall	0.722
	Hottest month maximum temperature	-0.575
TP	Annual rainfall	0.695
	Percentage area used for cropping	-0.556
FRP	Percentage agriculture area	0.392
	Percentage area underlain by mixed igneous bedrock	0.314
TKN	Annual rainfall	0.713
	Hottest month maximum temperature	-0.618
NO_x	Total storage capacity of dams in catchment	-0.493
	Mean soil TN content	0.458
EC	Percentage area covered by grassland	-0.347
	Percentage area covered by woodland	-0.317

Reference

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

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

Danlu Guo[1], Anna Lintern[1,2], J. Angus Webb[1], Dongryeol Ryu[1], Ulrike Bende-Michl[3], Shuci Liu[1], Andrew William Western[1]

[1] Department of Infrastructure Engineering, The University of Melbourne, Parkville, VIC Australia;

[2] Department of Civil Engineering, Monash University, Clayton, VIC Australia

[3] Bureau of Meteorology, Parkes, ACT, 2600.

Corresponding author's email-: danlu.guo@unimelb.edu.au

Abstract

~~Degraded water quality in rivers and streams can have large economic, societal and ecological impacts. Stream water quality can be highly variable both over space and time. To develop effective management strategies for riverine water quality, it is critical to be able to predict these spatio-temporal variabilities. However, our~~Our current capacity to model stream water quality is limited particularly at large spatial scales across multiple catchments. To address this, we developed a Bayesian hierarchical statistical model to simulate the spatio-temporal variability in stream water quality across the state of Victoria, Australia. The model was developed using monthly water quality monitoring data ~~collected at 102 sites~~ over 21 years, across 102 catchments, which span over 130,000 km^2. The modelling focused on six key water quality constituents: total suspended solids (TSS), total phosphorus (TP), filterable reactive phosphorus (FRP), total Kjeldahl nitrogen (TKN), nitrate-nitrite (NO_x), and electrical conductivity (EC). ~~Among~~ The model structure was informed by knowledge of the ~~six~~key factors driving water quality variation, which had been identified in two preceding studies using the same dataset. Apart from FRP, which is largely unexplainable, the model explains 21.6% (NO_x) to 90.7% (EC) of total spatio-temporal variability in water quality. Across constituents, the ~~models explained varying proportions of variation in water quality. EC was the most predictable constituent (88.6% variability explained) and FRP had the lowest predictive performance (19.9%~~model generally captures over half of the observed spatial variability~~ explained). The models were validated for multiple sets of calibration/validation sites and showed robust performance. Temporal validation revealed a systematic change in the TSS model performance~~; temporal variability remains largely unexplained across ~~most~~all catchments~~ since an extended drought period in the study region, highlighting potential shifts in TSS dynamics over the drought.~~, while long-term trends are well captured. The model is best used to predict proportional changes in water quality in a Box-Cox transformed scale, but can have substantial bias if used to predict absolute values for high concentrations. This model can assist catchment management by (1) identifying hot-spots and hot moments for waterway pollution; (2) predicting effects of catchment changes on water quality e.g. urbanization or forestation; and (3) identifying and explaining major water quality trends and changes. Further model improvements ~~in model performance need to~~should focus on: (1) alternative statistical model structures to improve fitting for ~~the low concentration~~truncated data, ~~especially records~~

for constituents where a large amount of data below the detection limit; and (2) better representation of non-conservative constituents (e.g. FRP) by accounting for important biogeochemical processes. ~~We also recommend future improvements in water quality monitoring programs which can potentially enhance the model capacity, via: 1) improving the monitoring and assimilation of high-frequency water quality data; and 2) improving the availability of data to capture land use and management changes over time.~~

Keywords

stream water quality; spatio-temporal variability; sediments; nutrients; statistical modeling; Bayesian hierarchical model

1. Introduction

Deteriorating water quality in aquatic systems such as rivers and streams can have significant environmental, economic and social ramifications (e.g. Whitworth et al., 2012;Vörösmarty et al., 2010;Qin et al., 2010;Kingsford et al., 2011). ~~However, our ability to manage~~Reducing these impacts requires effective management and ~~mitigate~~mitigation of poor water quality ~~impacts is hampered by the~~; however, high variability in water quality both across space and time~~,~~ reduces our ability to accurately assess the status of water quality and ~~our inability~~to develop effective management strategies. Thus, improved modelling frameworks to predict and interpret this variability would be useful for water quality management (Chang, 2008;~~Bengraïne and Marhaba, 2003;~~Ai et al., 2015;Zhou et al., 2012).

Water quality conditions can vary across individual events, as well as at daily, seasonal and inter-annual scales at an individual location (Arheimer and Lidén, 2000; Kirchner et al., 2004; Larned et al., 2004; Pellerin et al., 2012; Saraceno et al., 2009). Water quality conditions also typically differ ~~significantly~~substantially across locations ~~(Meybeck and Helmer, 1989).~~(Meybeck and Helmer, 1989;Chang, 2008;Varanka et al., 2015;Lintern et al., 2018a). These variabilities in stream water quality are driven by three key mechanisms: (1) ~~the~~ source ~~of constituents~~, which defines the total amount of constituents being available in a catchment; (2) ~~the~~ mobilization ~~of~~, which detaches constituents (both in particulate and dissolved forms~~, which detaches constituents~~) from their sources via processes such as erosion and ~~–~~biogeochemical processing; and (3) ~~the~~ delivery of mobilized constituents from catchments to receiving waters via multiple hydrologic pathways including surface and subsurface flow (Granger et al., 2010).

Spatial variability in stream water quality is driven by ~~natural catchment characteristics (e.g., climate, geology, soil type, topography and hydrology) as well as by human activities within catchments (e.g., land use and management, vegetation cover etc.), all of which control the extent and magnitude of the three key mechanisms described above (Lintern et al., 2018a). At the same time, temporal shifts in water quality are influenced by changes in climatic, hydrological and other catchment conditions, such as temperature (Roberts and Mulholland, 2007), the timing and magnitude of rainfall events (Fraser et al., 1999), runoff generation and streamflow (Ahearn et al., 2004; Mellander et al., 2015; Sharpley et al.,~~

~~2002), and vegetation cover changes over time (Kaushal et al., 2014;Ouyang et al., 2010)~~human activities within catchments (e.g., land use and management, vegetation cover etc.) (Lintern et al., 2018a;Carey and Migliaccio, 2009;Giri and Qiu, 2016;Heathwaite, 2010), along with natural catchment characteristics such as climate, geology, soil type, topography and hydrology (Hrachowitz et al., 2016;Poulsen et al., 2006;Sueker et al., 2001;Onderka et al., 2012). At the same time, temporal shifts in water quality are also influenced by changes in pollutant sources, such as land use and land management including urbanization, agriculture and vegetation clearing (Ren et al., 2003;Smith et al., 2013;Ouyang et al., 2010). In addition, water quality can also vary in time with variations in the mobilization and delivery processes, which are largely driven by the hydro-climatic conditions at a catchment, such as streamflow (Ahearn et al., 2004;Mellander et al., 2015;Sharpley et al., 2002;Zhang and Ball, 2017), the timing and magnitude of rainfall events (Fraser et al., 1999;Miller et al., 2014) and temperature (Bailey and Ahmadi, 2014).

~~Despite understanding of the basic mechanisms~~As abovementioned, we have good understanding of the key controls for variations in water quality, albeit in an isolated, idealized context. We still lack a sound understanding of how relationships between specific landscape characteristics and water quality can shift with influences from other landscape characteristics, and how the drivers of temporal variability in water quality can interact and vary across large spatial scales (Musolff et al., 2015;Lintern et al., 2018a;Ali et al., 2017). In contrast, current detailed understanding have been primarily based on field studies at small scales with detailed information on specific temporal drivers ranging from hydrologic conditions to detailed management decisions such as fertilizer rates and application timing (Smith et al., 2013;Poudel et al., 2013;Adams et al., 2014). While operational weather observation networks, stream gauging networks and remote sensing can provide some of this information, developing a large-scale understanding of water quality patterns across catchments would ideally also involve an extensive suite of management information that substantially exceeds what is currently ~~lack the ability~~available.

Due to the limited understanding of large-scale water quality patterns, we currently lack the capacity to model ~~these~~ spatio-temporal variabilities in water quality at ~~larger~~large scales across multiple catchments. This hinders our ability to inform the development of effective policy and mitigation

strategies~~. Conceptual~~ over large regions. Specifically, conceptual or physically-based water quality models are typically limited by the simplification of physical processes such as flow pathways (Hrachowitz et al., 2016). Furthermore, practical implementation of these models can be also limited by the intensive data requirements ~~of data and~~for calibration ~~effort~~and validation, particularly for large regions with ~~varying~~ highly heterogeneous catchment conditions (Fu et al., 2018;Abbaspour et al., 2015). In contrast, ~~whilst most~~when performed over large geographical regions, statistical water quality models are ~~easier to implement, they~~ generally more capable of simulating water quality variability while requiring less detailed information and thus effort for implementation. However, existing statistical models often focus only on either the spatial variation of time-averaged water quality conditions (Tramblay et al., 2010;Ai et al., 2015)~~,~~ or the temporal variation at individual locations (Kisi and Parmar, 2016;Kurunç et al., 2005;Parmar and Bhardwaj, 2015)~~. Consequently, it remains challenging to address~~, which often limits their value as practical management tools. Modelling the spatio-temporal variability ~~simulaneously~~simultaneously remains challenging over long time periods and large regions. ~~This lack of integrated modelling of both spatial and temporal variability in water quality can not only limits our understanding of the key factors that affect water quality dynamics over both of these dimensions. It also hinders our ability to predict future water quality changes in un-monitored locations.~~

~~The aim of~~Accordingly, this research ~~is~~ attempts to bridge the gap between fully-distributed physically-based water quality models and data-driven statistical approaches. We aim to develop a process-informed, data-driven model to predict spatio-temporal changes in stream water quality~~. This~~ over a large region consisting of multiple catchments. Specifically, this model was established using long-term (21 years) stream water quality observations across 102 catchments in ~~the state of Victoria, Australia. The model built on~~ Australia, with an aggregate catchment area of 130,000 km^2. To obtain the necessary understanding of process drivers required to develop this model, two ~~previous~~preceding studies were conducted on the same dataset ~~that identified~~ to identify the key drivers for ~~water quality~~ the spatial and temporal ~~variabilities~~variability of water quality, respectively (Lintern et al., 2018b; Guo et al., 2019). ~~Our approach aims to bridge the gap between fully distributed water quality models and statistical approaches to~~The aim of this study is to develop an integrated spatio-temporal model using the

previously-identified spatial and temporal predictors, and to then assess the performance of this model. Spatio-temporal variability of water quality was modelled using a novel Bayesian hierarchical approach which can jointly account for both variability components, including accounting for varying temporal water quality dynamics between catchments. This modelling approach also has relatively low requirement for input data, which keeps the modelling detail commensurate with the level of data availability. During the model development, we also obtained additional understanding on the patterns of spatial variations in the effects of each temporal predictor. The model can potentially provide useful information for ~~catchment managers, especially for large scale water quality assessments~~large-scale catchment management, assessment and policy making, such as testing major changes in land use patterns, informing pollution hot-spots, as well as identification and attribution of water quality trends and changes over time.

2. Method

~~2~~We first discuss the process used to develop the integrated spatio-temporal model (Section 2.1). Sections 2.1.1 and 2.1.2 introduces the statistical modelling framework and the data used for model development, respectively. The approaches to determine model structure was then introduced, which include the choice of key predictors (Section 2.1.3) and the calibration for model parameters (Section 2.1.4). Finally, the approaches to evaluate model performance and robustness are described in Section 2.2.

2.1 Model development

2.1.1 Spatio-temporal modelling framework

A Bayesian hierarchical approach was used to model the spatio-temporal variability in stream water quality. The Bayesian approach enables the inherent natural stochasticity of water quality to be incorporated into the model (Clark, 2005)~~, and~~. A key strength of applying the hierarchical model structure to analyze spatio-temporal variability is that this structure enables the key controls of temporal variability in water quality to vary across locations ~~(Webb and King, 2009;Borsuk et al., 2001)~~(Webb and King, 2009;Borsuk et al., 2001). This variability has been found to be important in other study regions where the (temporal) solute export regime varies with catchment characteristics such as climate

and land use (Musolff et al., 2015;Poor and McDonnell, 2007).

The structure of ~~this~~the Bayesian hierarchical model is presented below in Eq. 1 to 6. Eq. ~~The~~1 formulates the transformed constituent concentration ~~of a constituent~~ (see ~~Sect.~~Section 2.1.2 for justification) at time i and site j (C_{ij}) ~~is assumed to be~~ as a normally ~~distributed~~distribution with a mean μ_{ij} and standard deviation σ representing inherent randomness ~~(Eq. 1).~~

$$\boldsymbol{C_{ij} \sim N(\mu_{ij}, \sigma)} \quad (1)$$

To represent spatio-temporal variability, μ_{ij} is modelled as the sum of the site-level mean constituent concentration ($\bar{C}_j$) and the deviation from that mean at time i (Δ_{ij}) (Eq. 2).

$$\boldsymbol{\mu_{ij} = \bar{C}_j + \Delta_{ij}} \quad (2)$$

To describe spatial variability, the site-level mean concentration at site j ($\bar{C}_j$) is modelled as a linear function of a global intercept (~~int~~$intC$), and the sum of ~~the~~ m catchment characteristics $S_{1,j}$ to $S_{m,j}$ (e.g. land use, topography) weighted by their relative contributions to spatial ~~varaibility (β_S_1~~variability (βS_1 to ~~β_S_m~~βS_m) (Eq. 3).

$$\boldsymbol{\bar{C}_j = int + \beta\_S_1 intC + \beta S_1 \times S_{1,j} + + \beta\_S_2 + \beta S_2 \times S_{2,j} + \cdots + \beta\_S_m \beta S_m \times S_{m,j}} \quad (3)$$

The temporal variability, represented by the deviation from the mean (Δ_{ij}), is a linear combination of n temporal variables, $T_{1,ij}$ to $T_{n,ij}$ (e.g., climate condition, streamflow, vegetation cover) (Eq. 4), at time i and site j.

$$\boldsymbol{\Delta_{ij} = \beta\_T_{1,j} \times T_{1,ij} \beta T_{1,j} \times T_{1,ij} + \cdots + \beta\_T_{n,j} \times T_{n,ij} \beta T_{n,j} \times T_{n,ij}} \quad (4)$$

The selection of key spatial and temporal predictors for the model has been performed in our two preceding studies (Lintern et al., 2018b; Guo et al., 2019) and is briefly described in Section 2.1.3. Eq. ~~To~~1 to 4 enable the model to separately represent the spatial and temporal variability in water quality; however, there is still a further step required to make the model fully spatio-temporal (i.e. being able to predict over both time and location). Specifically, in Guo et al. (2019), clear spatial variation was observed in the relationships between water quality and its key temporal predictors (i.e. in the $\beta T_{N,j}$ in Eq. 4). To be able to model multiple catchments across a large spatial area simultaneously, we must account for differences in these temporal influences across sites. To do this, the effect of each temporal

variable at site j (~~$\beta\_T_{N,j}$~~$\beta T_{N,j}$ with N in 1,2, … n) is drawn from a distribution with a mean of ~~$N_{\beta\_T_{Nj}}$~~$\mu\beta T_{N,j}$ (Eq. 5), which is then modelled with a linear combination of two additional chatchment characteristics, ~~$S_{TN1,j}$~~$ST_{N1,j}$ and ~~$S_{TN2,j}$~~$ST_{N2,j}$ (Eq. 6). Details of the selection for these two additional predictors are presented in Section 2.1.3.

$$\beta\_T_{Nj}\beta T_{N,j} \sim N\left(\mu_{\beta\_T_{Nj}}, \sigma_{\beta\_T_N}\right), N\left(\mu\beta T_{N,j}, \sigma\beta T\right), \textbf{for } N \textbf{ in } 1, 2, \ldots n \quad (5)$$

$$N_{\beta\_T_{Nj}} = int_{\beta\_TN} + \beta\_S_{TN1} \times S_{TN1,j} + \beta\_S_{TN2} \times S_{TN2,j}\mu\beta T_{N,j} = int\beta T_N + \beta ST_{N1} \times ST_{N1,j} + \beta ST_{N2} \times ST_{N2,j} \quad (6)$$

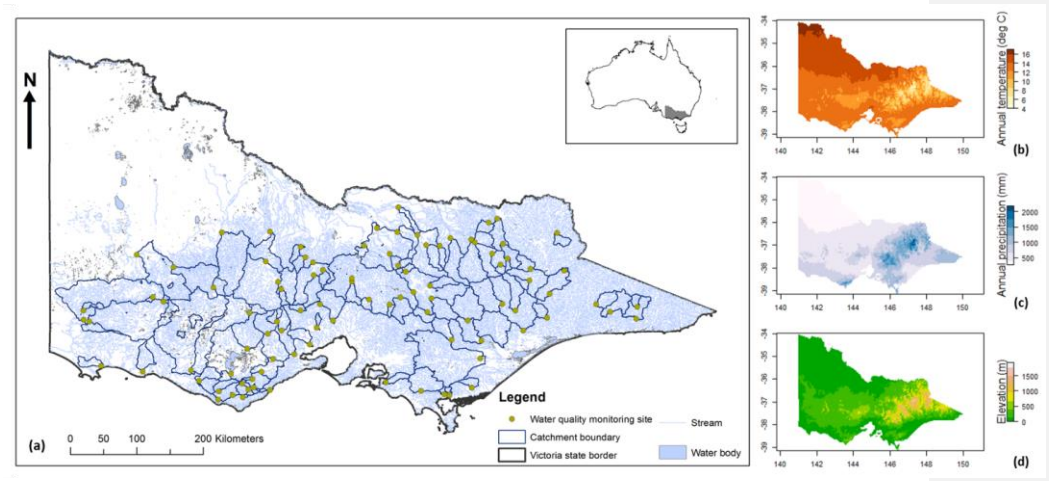
Section 2.2 introduces the data used to develop these Beyesian hierarchical models. Section 2.3 describes how the detailed model strucutre was determined, including the choice of key predictors for the spatial variability (i.e., catchment characteristics $S_{1,j}$ to $S_{m,j}$) and temporal variability (i.e. $T_{1,tj}$ to $T_{n,tj}$ and $S_{TN1,j}$ and $S_{TN2,j}$), and all their corresponding coefficient values. The approaches to evaluate model performance and robustness are described in Sect. 2.4.

2.2 Data collection and processing

2.1.2 Data collection and processing

The Bayesian hierarchical ~~models were~~model was developed with 21~~-year~~ years of monthly stream water quality observations at 102 catchments in the state of Victoria, Australia~~.~~ (aggregate catchment area > 130,000 km^2). The collection and processing of the data are detailed in previous publications that worked with the same dataset (Lintern et al., 2018b; Guo et al., 2019). Briefly~~, however~~, stream water quality data were extracted from the Victorian Water Measurement Information System (Department of Environment Land Water and Planning (DELWP) Victoria, 2016b), which contains monthly grab samples of water quality at approximately 400 sites across Victoria. Water quality data sampled between 1994 and 2014 at 102 sites were used to develop the model (Fig. 1). ~~This was because these~~These sites and ~~this~~ time period were chosen because they provided the longest consistent period of continuous records over the greatest number of monitoring sites. The catchments corresponding to these water quality monitoring sites were delineated using the Geofabric tool (Bureau of Meteorology, 2012), and have areas ranging from 5 km^2 to 16,000 km^2. The water quality parameters of interest were: total suspended solids (TSS), total phosphorus (TP), filterable reactive phosphorus (FRP), total Kjeldahl

nitrogen (TKN), nitrate-nitrite (NO_x) and electrical conductivity (EC). These parameters represent sediments, nutrients and salts, which are some of the key concerns for water quality managers in Australia and around the world. These water quality ~~data~~samples were ~~sampled~~collected following standard DELWP protocols (Australian Water Technologies, 1999) and analysed in National Association of Testing Authorities accredited laboratories~~.~~. Note that in the sampling protocol, FRP is defined as '*Reactive Phosphorus for a filtered sample to a defined filter size (e.g. RP(<0.45 µm))*', which is equivalent to the more widely-used terminology, SRP i.e. Soluble Reactive Phosphorus (Jarvie et al., 2002).



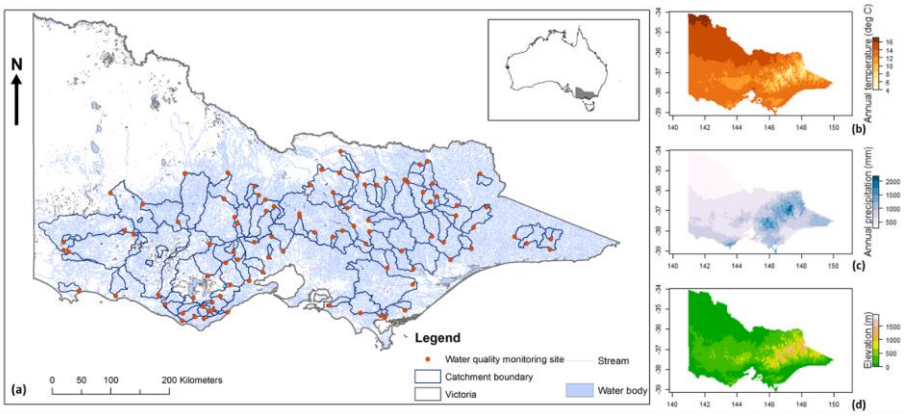


Figure 1. Map of (a) the 102 selected water quality monitoring sites and their catchment boundaries, with ~~insert~~inserts showing the location of the state of Victoria within Australia; (b) annual average temperature and (c) annual precipitation and (d) elevation across Victoria.

~~We selected~~To compile a dataset for the potential spatial explanatory variables (i.e. predictors to explain spatial variability~~) based on~~ in water quality), a comprehensive literature review was conducted (Lintern et al., 2018a), which summarized the key catchment landscape characterisitics that are widely known to influence water quality ~~condition (~~. Further, as part of Lintern et al~~., 2018a). Fifty~~. (2018b), fifty potential explanatory catchment characteristics were selected ~~based on a literature review. These~~, which included catchment land use, land cover, topographic, climatic, geological, lithological and hydrological catchment characteristics. These variables were derived using datasets obtained from Geoscience Australia (2004, 2011), ~~–~~the Bureau of Meteorology (2012), the Bureau of Rural Sciences (2010), Department of Environment Land Water and Planning Victoria (2016) and the Terrestrial Ecosystem Research Network (2016) (see Table S1 in the Supplementary Material for detailed variable names and data sources). We used a static set of land use data from 2005-2006 to represent the entire study period, as a preliminary analysis ~~of land use data~~ between 1996 and 2011 suggested less than 1% changes in the key land uses in these catchments (i.e. agricultural, grazing, conservation).

~~Temporal~~Nineteen potential temporal explanatory variables were included. Firstly, data of discharge (originally in ML d^{-1}) and water temperature (°C) corresponding to the same timestamps for water quality observations were also extracted for each monitoring site over the study period (Department of Environment Land Water and Planning Victoria, 2016). Discharge was converted to ~~streamflow~~runoff depth (mm d^{-1}) for each catchment, ~~which allowed us to also calculate~~and the average streamflows over 1, 3, 7, 14 and 30 days preceding the water quality sampling dates~~.~~ were calculated. In addition, we extracted gridded ~~climate data~~dataset from the Australian Water Availability Project (AWAP) (Frost et al., 2016;Raupach et al., 2009, 2012) ~~and the normalized difference vegetation index (NDVI) data~~ and Australian Water Resources Assessment Landscape (AWRA-L) model (Frost et al., 2016). These datasets were used to calculate catchment averaged values of daily average temperature (°C), daily rainfall (mm), antecedent rainfall (1, 3, 7, 14 and 30 days preceding sampling), dry spell (> 0.1mm rainfall) length in the antecedent 14 days, daily actual evapotranspiration (ET) (mm), as well as soil moisture for the root-zone and the deep-zone (averaged volumetric content for shallower and deeper

than 1m, respectively). In addition, catchment averaged monthly NDVI data were extracted from Advanced Very High Resolution Radiometer (AVHRR) Product (Eidenshink, 1992) and Moderate Resolution Imaging Spectroradiometer MOD13A3 ~~(NASA LP DAAC, 2017;Eidenshink, 1992) were also extracted to calculate the catchment average daily rainfall (mm), daily evapotranspiration (ET) (mm), daily average temperature (°C), daily root zone (shallower than 1m) and deep (deeper than 1m) soil moisture, as well as monthly~ (NASA LP DAAC, 2017). A summary of these datasets of temporal variables and their corresponding sources are in Table S2 in the Supplementary Material and details are provided in Guo et al. 2019. ~~NDVI. A summary of these data and their sources is in Table S2 in the Supplementary Material.~~

The raw input data were filtered and transformed to increase the data reliability, continuity and symmetry, making them more suitable for use in the linear spatio-temporal model structure (~~Eqs~~Eq. 3, 4 and 6). For the filtering process, we first removed all water quality records with flags ~~of~~indicating quality issues~~ and~~. We also removed any values below the ~~limits of reporting (LOR).~~detection limit (DL), which was defined as the '*minimum concentration detected for which there is 95% confidence of accuracy and therefore is accurate enough to report*' in the monitoring protocols for this dataset (Australian Water Technologies, 1999). This was because ~~that~~the uncertainty ~~of~~in values below ~~LOR may amplify~~the DL would be amplified after ~~the~~ transformation, ~~posing large~~which would influence ~~in~~ the subsequent model fitting. Furthermore, those undetectable low concentrations were of less interest~~, poor water quality conditions (i.e., high constituent concentrations) were our primarily concerns to model.~~ for management purposes. Water quality records corresponding to days with zero flows were also excluded from further analyses.

~~For the~~The transformation process~~, we transformed the data of~~ was performed for each of the spatial catchment characteristics, temporal explanatory variables, as well ~~as each~~as each water quality constituent to improve the symmetry of individual distributions. The log-sinh transformation (Wang et al., 2012) (Eq. 7) was used for all catchment characteristics, due to its ability to resolve the presence of zero values in several of the catchment characteristics (e.g., percentage area of ~~different types of~~individual land ~~use~~uses). The ~~best~~*GA* package in R (Luca Scrucca, 2019) was used to identify the log-sinh transformation ~~parameter was determined~~parameters (*a* and *b*) for each spatial explanatory variable

that minimized the data skewness (i.e. symmetry is maximized) across all 102 catchments.

$$y_{log-sinh} = \frac{1}{b}\log(\sinh[a + by_{raw}]) \quad (7)$$

In addition, all observed constituent concentrations and temporal explanatory variables were Box-Cox transformed~~. For each variable, i.e., 21-year time series data across all 102 sites, we first identified the optimal Box-Cox parameter at each site λ, and then the averaged λ across all sites to determine the final λ used to transform a respective variable. This ensured a consistent transformation for each variable across all sites. All log-sinh and Box-Cox transformation parameters used are summarized in Table S3 and S4 in the Supplementary Material.~~ (Box and Cox, 1964) (Eq. 8).

~~**2.3 Model fitting**~~

$$~~Based~~y_{Box-Cox} = \begin{cases} \frac{y_{Raw}^{\lambda}-1}{\lambda}, & for\ \lambda \neq 0 \\ logy, & for\ \lambda = 0 \end{cases} \quad (8)$$

For each variable, the optimal Box-Cox transformation parameter λ was identified using the *car* R package and a maximum likelihood-like approach. We first identified the optimal Box-Cox parameter λ using the data at each site (i.e. 21-year time-series). The averaged λ across all sites was then used to transform the data across all catchments together. This transformation approach ensured that all sites used a consistent transformation parameter. All transformation parameters used are summarized in Tables S3 and S4 in the Supplementary Material. The transformation process has greatly improved the data symmetry and thus suitability for use in a linear model (the quality of the transformations was assessed via visual inspection in Lintern et al., 2018b; Guo et al., 2019; and summarized in Figures S2, S4 and S6 in the Supplementary Material).

2.1.3 Selection of key model predictors

Key predictors for the model were selected in a process-informed and data-driven manner based on ~~the general spatio-temporal modelling structure (Eqs. 2 to 6), we~~our two preceding studies (Lintern et al., 2018b; Guo et al., 2019). Lintern et al. (2018b) identified the best spatial predictors (S_1 to S_m in Eq. 3) ~~and~~for the model, while the best temporal predictors across all sites (T_1 to T_n in Eq. 4) have been identified in ~~two sequential studies (Lintern et al., 2018b;~~ Guo et al., (2019). ~~The~~In both studies, the best predictors were selected using an exhaustive search approach (May et al., 2011;Saft et al., 2016),

which considered ~~a large number of potential predictors and~~ all possible combinations of ~~these~~the potential predictors introduced earlier in this section. This selection approach required firstly fitting an individual model to ~~each~~all possible candidate predictor ~~set~~sets, and then comparing all fitted models to select a single best set of predictors. Alternative models were evaluated based on the Akaike Information Criterion (AIC) (Akaike, 1974) and Bayesian Information Criterion (BIC) (Schwarz, 1978) to ensure optimal balance between model performance and complexity.

The ~~key factors identified for~~best predictors to explain the spatial and temporal variabilities in each constituent are listed in ~~Tables S5 and S6 in the Supplementary Materials. General~~Table 1. Generally speaking, the key factors controlling the spatial variability in river water quality were land-use and long-term climate conditions (Lintern et al., 2018b). Temporal variability was mainly explained by temporal changes in streamflow conditions, water temperature and soil moisture (Guo et al., 2019). ~~We further modelled the spatial variation in each of these temporal relationships (β_T_1 to β_T_n in~~The potential mechanisms via which these key drivers influence water quality are discussed in details in these two previous studies~~ Eq. 4) with two spatial characteristics, S_{TN1} and S_{TN2} (Eq. 6), where a higher number of predictors was not used to avoid over fitting. We found S_{TN1} and S_{TN2} via a Spearman correlation analyses (p<0.05) between the fitted parameter values of each temporal predictor variable (β_T_1 to β_T_n) and potential spatial explanatory variables as mentioned in Sect. 2.2. S_{TN1} and S_{TN2} were selected as the two catchment characteristics which had the highest correlations with the fitted parameter values of each temporal predictor, which were also summarized in Table S6 in the Supplementary Material~~.

Table 1. Key factors affecting the spatial and temporal variability for each of six constituents, as identified in Lintern et al. (2018) and Guo et al. (2019b), respectively.

Whilst the previous studies (Lintern et al. 2018b, Guo et al. 2019) identified the predictors for spatial and temporal variability respectively, they did not provide guidance on the predictors for spatial variability in the relationships between drivers of temporal variability and temporal water quality response (i.e. βT in Eq 4). As such, the final step of the predictor selection process to develop the combined spatio-temporal model was to identify the key catchment characteristics that affect spatial variability in the hydroclimatic parameters driving temporal changers in water quality (βT_1 to βT_n in Eq. 4, also right column in Table 1). This is achieved by selecting two spatial characteristics that are most

closely related to the coefficient for each temporal predictor (ST_{N1} and ST_{N2}, Eq. 6) across all sites, where only two spatial characteristics were used to avoid over-fitting. Selection of these two spatial characteristics were based on a Spearman correlation analysis between the fitted parameter values of each temporal predictor variable and the fifty potential spatial explanatory variables (as mentioned earlier in this section), following three steps:

1. from the 50 candidate spatial predictors, the one with the highest Spearman correlation with βT_N is selected as ST_{N1}, provided the correlation is statistically significant ($p<0.05$);
2. the subset of remaining spatial predictors with spearman correlation with $ST_{N1} < 0.7$ is found; and
3. from this subset, the spatial predictor with the highest spearman correlation with βT_N is selected as ST_{N2}, provided the correlation has $p<0.05$;

Steps 2 and 3 intended to avoid cross-correlations between ST_{N1} and ST_{N2}. The selected spatial characteristics that influence the temporal relationships in our model are presented and interpreted in Section 3.1. Note that the entire process to select ST_{N1} and ST_{N2} was performed with the fitted parameters for each predictor of the temporal variability obtained from Guo et al. (2019).

2.1.4 Model calibration

After identifying the spatial and temporal predictors for each constituent, as well as the spatial characteristics which affect the strengths of each temporal predictor, the Bayesian hierarchical spatio-temporal model was fitted for each constituent across all monitoring sites. simultaneously. To achieve this, we used the R package *rstan* (Stan Development Team, 2018), which enabled both the sampling of parameter values from posterior distributions with Markov chain Monte Carlo (MCMC) and model evaluation. Constituent standard deviation (σ) was assumed to be drawn from a ~~prior of~~ minimally informative ~~distribution of~~ prior half-normal of $N(0,10)$ ~~that was~~distribution truncated to only positive values (Gelman, 2006; Stan Development Team, 2018). The regression coefficient of each spatial predictor (~~$\beta\_S_1, \beta\_S_2, \ldots, \beta\_S_m$ in Eq. 3) was assumed to be drawn from an independent hyper-parameter normal distribution with mean of β_S and standard deviation of σ_S. The site-level regression coefficients of the temporal predictors ($\beta\_T_{1,j}, \beta\_T_{2,j}, \ldots, \beta\_T_{n,j}$ in Eq. 4, respectively) were sampled from the corresponding hyper-parameter normal distribution with means of $\mu.\beta\_T_1, \mu.\beta\_T_2, \ldots, \mu.\beta\_T_n$ and~~

~~standard deviations of $\sigma.\beta\_T_1$, $\sigma.\beta\_T_2$, …, $\sigma.\beta\_T_n$. The hyper-parameters were further assumed to be drawn from minimally informative normal distributions with $N(0,5)$ (for all the means) and minimally informative half-normal distribution of $N(0,10)$ that was truncated to only positive values (for all the standard deviations).~~βS_1, βS_2, …, βS_m in Eq. 3) was independently drawn from hyper-parameter distributions of $N(\mu\beta S_M, \sigma\beta S_M)$. The site-level regression coefficients of the temporal predictors ($\beta T_{1,j}$, $\beta T_{2,j}$, …, $\beta T_{n,j}$ in Eq. 4, respectively) were sampled from the corresponding hyper-parameter distribution of $N(\mu\beta T_N, \sigma\beta T_N)$. The hyper-parameters were further assumed to be drawn from minimally informative prior distributions, following recommendations in Gelman (2006) and Stan Development Team (2019): for all the hyper-parameter means, a normal prior distribution of $N(0,5)$ was used; for all the hyper-parameter standard deviations, a half-normal prior distribution of $N(0,10)$ was used, which was truncated to only positive values. In each model run there were four independent Markov chains. A total of 20,000 iterations were used for each chain. Convergence of the chains was ~~checked using~~ensured by checking the *Rhat* value (Sturtz et al., 2005)~~.~~, which is a summary statistic on the convergence of the Bayesian models from the four Markov chains used in model calibration (Stan Development Team, 2018). Specifically, an *Rhat* value much greater than 1 indicates that the independent Markov chains have not been mixed well, and a value of below 1.1 is recommended (Stan Development Team, 2018).

2.4~~2~~ Model performance evaluation and sensitivity analyses

~~The performance of the fitted model for each constituent was first evaluated by comparing the simulated and observed concentrations.~~ Performance evaluation of the model was undertaken on several aspects of the model results (Section. 3.2). Since the model was calibrated in a Box-Cox transformation scale (see justification in Section 2.1.2), the Box-Cox transformation scale was used for model evaluation to enable a clear investigation on the influences of a wide range of factors that can influence model performance. Detailed performance evaluations include:

1. *Ability to capture total spatio-temporal variability.* Firstly, the simulations from the fitted model and the corresponding observed concentrations were compared at 102 sites altogether to understand how the ~~full~~overall spatio-temporal ~~variabilities~~variabilities were captured ~~(Sect. 3.1). As explained in Sect. 2.2, the model calibration for~~. For each constituent, this evaluation

was performed with~~ only the above LOR data. Therefore, model performance was first evaluated with only~~: 1) these above-~~LOR~~DL data to focus only on data~~. Performance was then evaluated with~~ used for calibration (as detailed in Section. 2.1.2); and 2) the full dataset including the below-~~LOR~~DL data~~;~~ (set to half of the DL of the specific constituent), to understand how well the model ~~capacity to simulate~~represents the full distribution of constituent ~~concentration. In addition, the~~concentrations. A good model performance ~~for capturing spatial differences was assessed by comparing the simulated and observed long-term mean concentration at each site. The~~when including the below-DL data would suggest that the calibrated model is transferable to below-DL data too. All performance assessments were based on both visual inspection of model fitting as well as the Nash-Sutcliffe efficiency (NSE), which ~~suggested~~quantified the proportion of variability that ~~can be~~was explained by the ~~models~~model (Nash and Sutcliffe, 1970).

2. *Proportions of spatial and temporal variability explained*. This involved a decomposition of the total observed variability using Eq. 2., into proportions contributed by spatial variability (variations in all site-mean concentrations from the grand average of site-mean concentrations) and temporal variability (variations in all concentrations from the corresponding site-mean concentrations). The corresponding modelled values were then used to calculate NSE for each variability component of each constituent.

3. *Ability to capture variation in ambient conditions across space, and temporal variation (including trends) across multiple catchments*. These were evaluated by a) comparing all simulated and observed site-averaged long-term mean concentrations; and b) comparing the simulated and observed time-series and long-term trends at representative sites. Further to a), performance was also evaluated on a real measurement scale by first back-transforming all modelled sample concentrations, calculating the back-transformed site-level means and then compared those to the corresponding observations. A further analysis to b) was also performed by comparing the estimated Sen's slope (Akritas et al., 1995) for the observations and simulations at all sites, and then computing the percentage of sites where the observed trends as indicated by the Sen's slope have been correctly represented by the model.

Additional evaluations of model sensitivity were conducted with calibration and validation on subsets of the full data (~~Sect~~Section. 3.~~2). Firstly,~~3), to understand ~~the~~model transferability and stability:

1. *Model sensitivity ~~of the model~~ to the monitoring sites ~~included~~used for calibration~~, we~~*. We randomly selected 80% of the sites for calibration and used the remaining 20% for validation, and repeated this validation process ~~for five~~50 times~~ for each constituent. The~~. We compared all calibration and validation ~~performance was compared to~~performances of these 'partial models' with each other, as well as with the performance of the full model~~.~~, to obtain a comprehensive evaluation of the sensitivity of model performance to calibration sites.

2. ~~We also evaluated the model~~ *Model sensitivity to ~~the periods of~~ calibration data period*. Since the study region was greatly influenced by a prolonged drought from 1997 to 2009 ~~–~~ known as the Millennium Drought~~,~~ (van Dijk et al., 2013), we ~~focused on analysing the impact of~~ also investigated model robustness for before, during and after this drought period. Specifically, we calibrated the model ~~for each constituent~~ to each pre-, during- and post-drought ~~periods~~period (1994-1996, 1997-2009 and 2010-2014, respectively) ~~and then validated the~~ with model validation on the remaining ~~period which was not used for calibration.~~data. For example, when calibrating to the pre-drought period (~~1994-1996~~1997-2009), validation was performed on ~~both~~ the merged during and post-drought ~~data (1997~~ period (1994-1996 plus 2010-2014). ~~Each~~The corresponding calibration and validation ~~performance was~~performances were compared with each other as well as against that of the full model, to identify potential impacts of the drought on model robustness.

3. Results

3.1 Spatial variation in the impact of temporal factors

The key controls of the spatial and temporal variations in water quality have been identified in our two preceding studies (Lintern et al. 2018b, Guo et al. 2019) and briefly summarized in Section 2.1.3. and are thus not discussed here. As also detailed in Section 2.1.3, to achieve full spatio-temporal predictive capacity, the model developed in this study considers the spatial variation in the strength of each temporal predictor by using two additioal catchment spatial characteristics ($ST_{N1,j}$ and $ST_{N2,j}$ in Eq. 6).

on the Spearman's correlations. Here we focus on the most important temporal predictor for each constituent, streamflow, where Table 2 shows the two spatial characteristics identified that are most closely related to the spatial variation of the effects of impact of streamflow on water quality. The full list of the selected key catchment characteristics for all temporal predictors of each constituent is summarized in Table S5 and visualized in Figure S4.

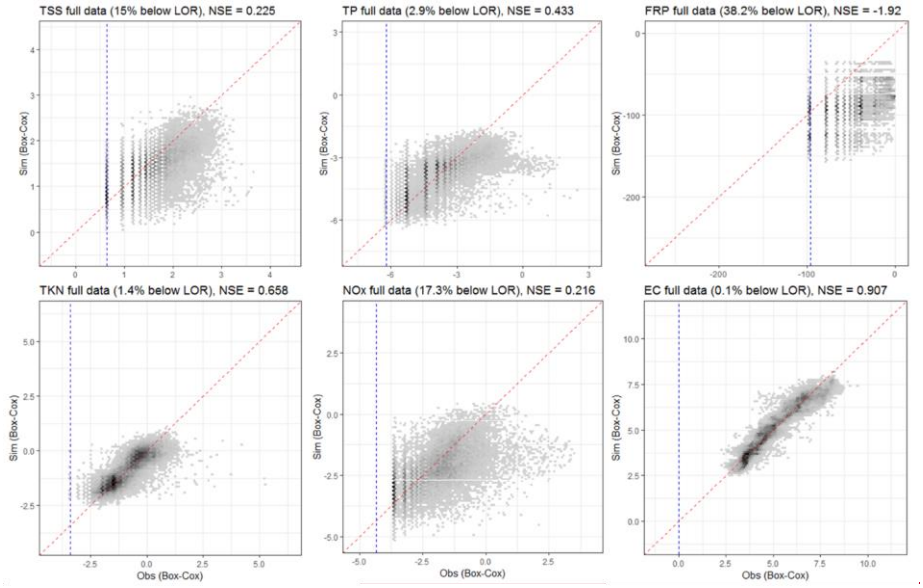
Table 2. The key catchment landscape characteristics that are related to the varying relationships of water quality and same-day streamflow across space, which were selected as the two predictors for the streamflow effect in our model. The corresponding Spearman's correlation (ρ at $p<0.05$) between the effect of streamflow and each catchment characteristic is presented.

TSS, TP and TKN show consistent patterns of the spatial variation in the effects of streamflow on water quality, which are strongly driven by the differences in average rainfall conditions across catchments. Specifically, streamflow generally has a larger effect on water quality in catchments with higher average annual rainfall. Since the streamflow effects are positive for the majority of catchments (as shown in Figure S5), these correlations indicate that for the same increase in transformed streamflow, a greater increase in transformed concentrations of TSS, TP and TKN will occur at a catchment with higher annual average rainfall. Given that the Box-Cox lambda values (Table S4) are close to zero, the transformation is log-like and hence changes in transformed flow and concentration approximately correspond to proportional changes in the real values of flow and concentration. In contrast, for FRP, NO_x and EC, the spatial patterns of streamflow effects are specific to each constituent. This difference in the model results between TSS, TP and TKN against the other constituents might be related to the distinct transport pathways of particulate and dissolved constituents. The former is predominantly related to surface flow and thus relies heavily on rainfall contribution. Dissolved constituents are likely transported along the subsurface pathway. Apart from streamflow, the spatial patterns in other key temporal drivers of water quality (e.g. antecedent streamflow, soil moisture etc.) are less consistent across different constituents (Figure S4).

3.2 Model performance evaluation

The spatio-temporal water quality models show varying performances ~~among~~between the constituents. When assessed with only the above-~~LOR~~DL data (Fig. 2), the best performing models are those for EC and TKN, which capture 90.7% ~~–~~and 65.8% of the total observed spatio-temporal variability. The

modelling ~~power~~performance is lowest for FRP ~~(~~, NO_x and TSS, with NSE ~~=~~values of -1.92~~), which might be related to the large number of FRP records below the LOR (38%). Similar to FRP, poorer model performance is also observed for NO_x and TSS, with NSE values of~~ , 0.216 and 0.225, ~~where the proportion of below-LOR samples were 17.3% and 15%,~~ respectively. When evaluated against the entire dataset (i.e., including both below- and above ~~LOR~~DL data), the models explain 19.9% (FRP) to 88.6% (EC) of spatio-temporal variability (Table ~~1~~3). Model performances for FRP, NO_x and TSS improve notably compared with the previous evaluation ~~on~~of above-~~LOR~~DL data~~. However, FRP, NO_x and TSS~~, however, they remain as the three constituents that are most difficult to predict. We further discuss the possible factors influencing their model performance in ~~Sect.~~Section 4.~~2~~1.



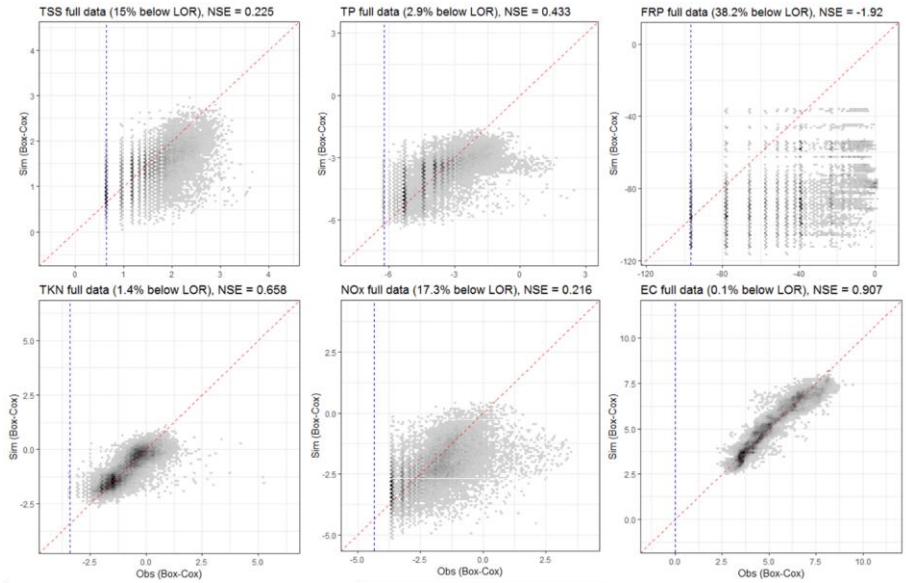


Figure 2. Performance of the spatio-temporal models for each of the six constituents, represented by the simulated ~~and observed~~median concentrations and corresponding observations of above-~~LOR~~DL records across all 102 calibration sites, in Box-Cox transformed space. Darker regions represent denser distribution of simulation and observation points. Dashed red lines show the 1:1 lines whereas dashed blue lines show the ~~LOR~~DL levels. For each constituent, the percentage of data below the ~~LOR~~DL and the model performance (NSE) are also specified.

Table 3. Comparison of model performance for all records and only the above-LOR records for each constituent.

~~When simulating~~The model performance to predict spatial and temporal variability is summarized in Figure 3, which compares the observed and explainable variability for each of the spatial and temporal components (detailed in Section 2.1.4). Regarding the observed variability (lighter colours), EC is strongly dominated by spatial variability (91.8%), highlighting that within-site variation in water quality is minimal compared to between-site variation. To a lesser extent, spatial variability also contributes to major proportions of total variability for TP and TKN (60.8% and 66.6%, respectively). TSS, FRP and NO_x are more influenced by temporal variability (57.4%, 56.6%, 60.5%, respectively).

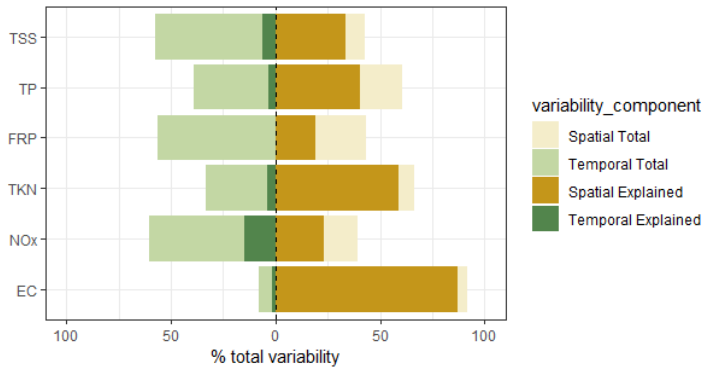
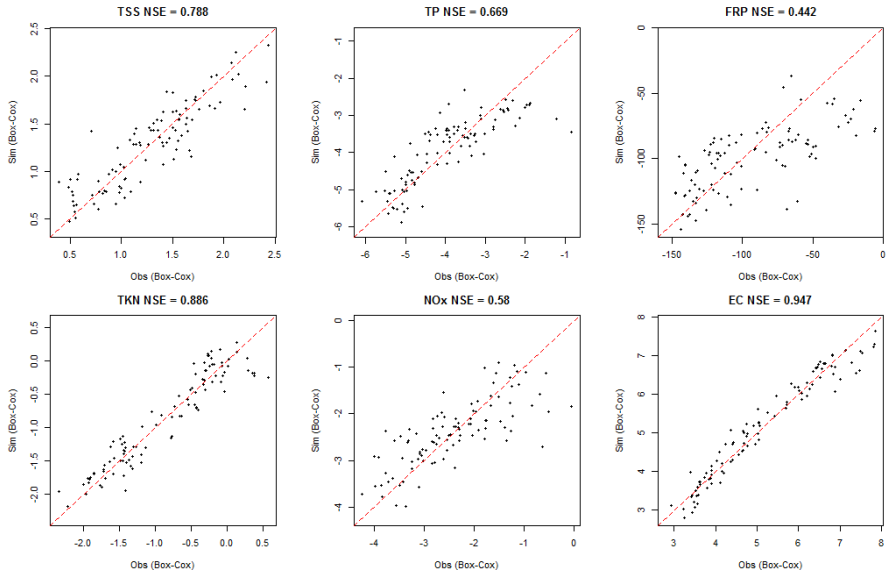


Figure 3. Observed spatial and temporal variabilities as proportions of the total variability (total width of each bar, 100%). The dashed line differentiates temporal variability (left side) with spatial variability (right side), and the darker colours highlight the proportions of spatial and temporal variabilities that are explainable by the model. All values were estimated in Box-Cox transformed space.

The explained variability (darker colours) shows that, across all catchments, temporal variability is much more difficult to model compared with spatial variability. It also appears that a substantial part of the model's overall performance is driven by its ability to capture spatial variability in ambient water quality conditions. For example, the models for TSS, FRP and NO_x show poorer overall performance (Fig. 2, with NSE values of 0.225, -1.92 and 0.216, respectively)), because the total variability for each of these is dominated by temporal variability (57.4%, 56.6%, 60.5%, respectively), which largely remains unexplained by the model (Fig. 3). In contrast, the EC model shows a very good fit with 90.7% of total variability explained – 91.8% of the total observed variability is due to spatial variability, of which 94.7% is explained by the model. Therefore, although the EC model can only explain a small portion of temporal variability (20% out of 8.2% of total variability), the overall model performance remains high.

As highlighted in Fig. 3, the model has good capacity to capture spatial variability in water quality. This is further evaluated in Fig. 4 by comparing the simulated and observed site-level mean concentrations~~, the spatio-temporal models generally show good abilities to capture variability across sites for all constituents (Fig. 3)~~. The highest model performance is for EC ~~(explaining 94.7% of spatial variability)~~ and lowest performance is for FRP (explaining 94.7% and 44.2% spatial variability~~). The relative abilities of models~~, respectively). At the back-transformed scale, the model shows greater biases for sites with higher concentrations (approximately the highest 10% sites for each

constituent) (Fig. 5). This is not surprising as the model was fitted to a Box-Cox transformed space that reduces focus on high values and increases the focused on low values. This compromised its ability to represent ~~spatial variability in different constituents are generally consistent with their capacities to capture spatio-temporal variability (Table 1).~~sites with unusually high concentrations. The implications of the model having higher predictive capacity in the transformed scale is further discussed in Section. 4.1.



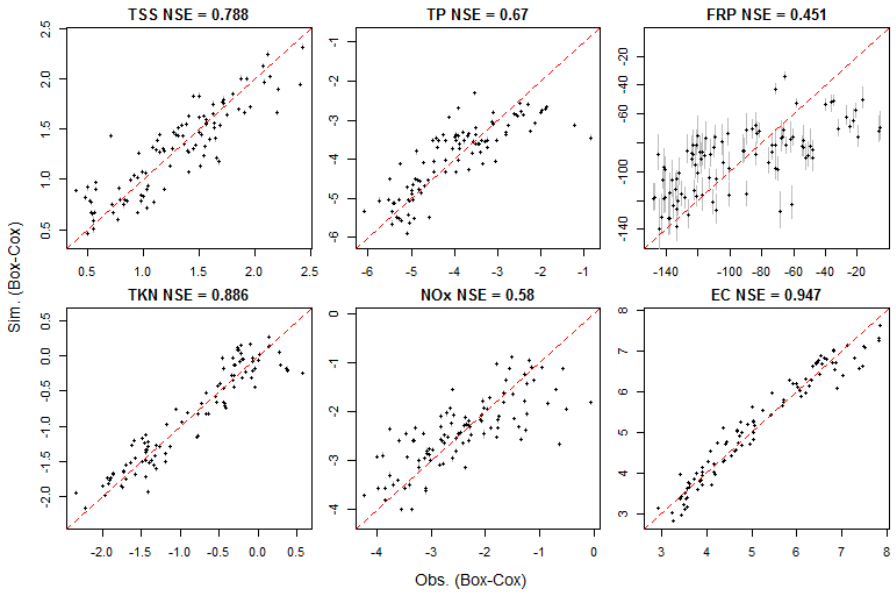


Figure 4. Model fit for site-level mean concentration at the 102 calibration sites~~,~~ for ~~the selected~~ six constituents, with the 95% lower and upper bounds of posterior simulations shown in vertical grey lines. All simulations and observations are presented in in Box-Cox transformed space. The NSE for each constituent is also ~~show~~shown and ~~dashed~~ red dash lines show the 1:1 lines.

~~3.2~~

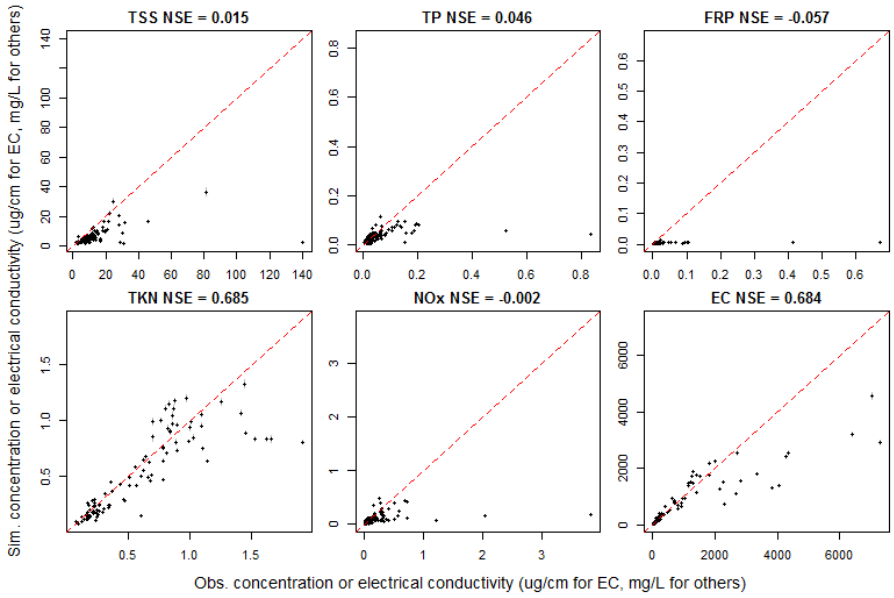


Figure 5. Back-transformation of the model simulations to the measurement scale emphasizes lack of fit for the highest concentrations, illustrated by simulated against observed site-level mean concentrations of each constituent in a back-transformed scale. The 95% lower and upper bounds of all posterior simulations shown in vertical grey lines. The NSE for each constituent is also shown and red dash lines show the 1:1 lines.

As also noted in Fig. 3, the ability of the spatio-temporal model to explain temporal variability remains relatively limited. This is further explored in Fig. 6, where the observed and simulated time-series are presented for one monitoring site for each constituent, at which the model performance (NSE) was the highest. These results show that even for catchments where the model has the highest ability to capture temporal variability, the model consistently underestimated temporal variability for all constituents.

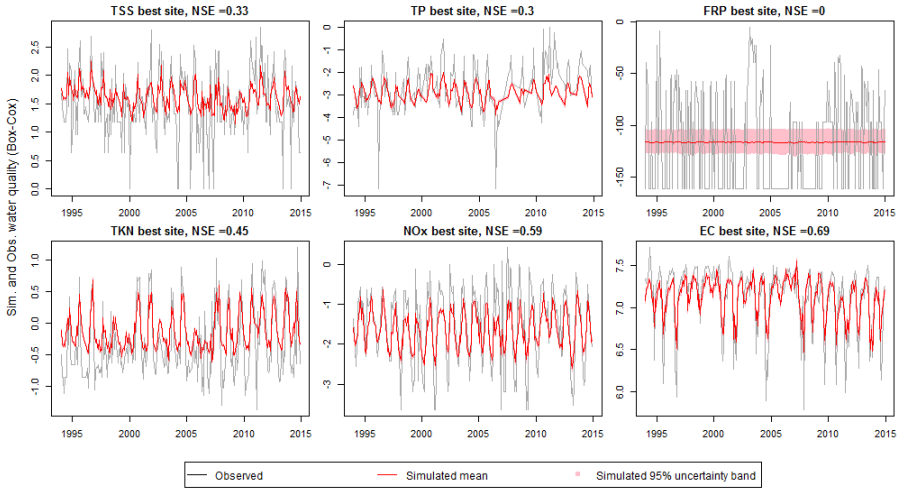


Figure 6. Model fit of the within-site (temporal) water quality variability, illustrated with the observed and simulated time-series for the best-performing site for each constituent. All values are presented in Box-Cox transformed space. The NSE for each constituent is also shown. The red line indicates the corresponding mean of all posterior simulations, while the pink bands show the corresponding 95% lower and upper bounds (only visible for FRP).

Fig. 6 also illustrates that, although the model shows substantial underestimation of temporal variability within site, long-term temporal trends in the time-series are well captured at the best sites (except for FRP). Table 4 summarizes the ability of the model to capture observed trends across all 102 catchments for each constituent. In general, the model is able to capture observed trends in most sites for NO_x and EC and for both positve and negative trends. For TP and TKN, positive trends are well captured while for TSS the negative trends are better captured.

Table 4. Model ability to capture observed water quality trends across all monitoring sites for each constituent. The percentages of sites where observed positive and negative trends are captured by the model are presented separately. Values in brackets indicate numbers of sites where corresponding positive or negative trends are observed. For detailed estimation of these percentages please refer to Sect. 2.2.

3.3 Model sensitivity ~~to calibration sites and periods~~analyses

~~This section presents model sensitivity to different calibration sites and periods of record (as detailed in Sect. 2.4). Note that in these evaluations, the FRP model is not a focus due to the poor model performance observed in Sect. 3.1.~~

We first compare the performance of each spatio-temporal model fitted with the full dataset with those obtained from the ~~five~~50 corresponding "partial" models that were calibrated to only 80% of the monitoring sites. Note that in this comparison, the FRP model was not assessed due to its poor performance (Section 3.2). The calibration and validation results for the 50 partial models are summarized in Table 5 along with the performance of the full model calibrated to all 102 sites (see Figs. S6 and S7 in the Supplementary Material for detailed comparison of model residuals of the partial calibration/validation). Across constituents, the calibration ~~performances obtained from~~performance of the full ~~dataset are~~model was comparable with the ~~five~~50 partial models ~~calibrated with 80% of the sites (calibration dataset).~~ In addition, ~~each pair of calibration and validation~~ model performance is highly consistent~~. In either comparison, the~~ between corresponding calibration and validation, with most differences in ~~NSE are within 0.1 (Table 2, see Figs. S1 to S6 in the Supplementary Material for detailed fitting plots for the partial calibration/validation).~~ NSEs less than 0.1. These suggest that the spatio-temporal model performance ~~are~~is highly robust and ~~remain~~ unaffected by the choice of calibration sites.

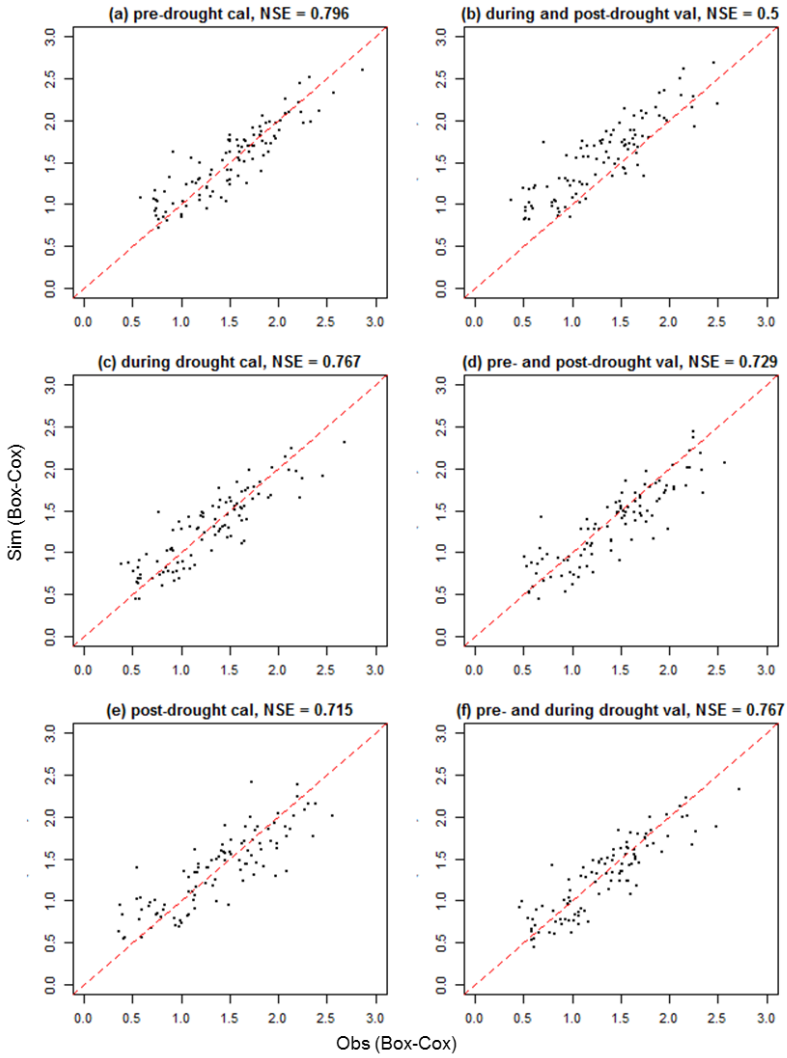
Table 5. Comparison of model performances (as NSE) of the full model (Column 2) and the ~~five~~50 partial models (Columns 3 to ~~7~~5) with each calibrated to 80% randomly selected monitoring sites. ~~In~~ Columns 3 to ~~7~~5 summarize the mean, minimum and maximum NSE values across the 50 runs, where for each constituent, the top row showing calibration performance and the bottom row showing the validation performance (i.e. at the 20% sites that were not used for calibration).

The performance of the full model for each constituent is also compared with that of the three models calibrated to the pre-, during and post-drought periods. In general, we observe consistent performance for each constituent, across calibrations to the three periods of contrasting hydrological conditions (Table ~~3~~6, see Figs. ~~S7~~S8 to ~~S12~~S13 in the Supplementary Material for detailed model fittings). One notable common pattern is that the performance for calibration and validation is more consistent ~~for~~during the drought period than either the pre- and post-drought periods. However, this is most likely explained by relative sizes of the calibration data sets, which are 3, 13 and 5 years for the pre-, during and post-drought periods respectively.

Of all constituents (excluding FRP), TSS shows greater differences in model performances across periods – especially when comparing the pre-drought calibration with its validation~~. Fig. 4 shows the corresponding TSS model fit as represented by the~~ for the site-level mean concentrations ~~for the three~~

~~calibration/validation datasets.~~(Fig. 7). Notably, when calibrated to the pre-drought period and validated on both the during- and post-drought periods, the validated model over-estimates ~~a majority~~most of the data (Fig. ~~4~~7 (b)); and when calibrated to the during-drought period, ~~it~~the validated model slightly under-estimates pre- and post-drought period TSS (Fig. ~~4~~7 (d)).

Table 6. Comparison of model performances (as NSE) of the full model and the three models that were calibrated to the pre-drought (1994-1996), drought (1997-2009) and the post-drought (2010-2014) periods. For each of the models, the calibration performance is shown on the top row and the validation performance (i.e. over the periods that were not used for calibration) is shown on the bottom row. See Section 2.1.4 for details of the calibration and validation approach.


(a) pre-drought cal, NSE = 0.796
(b) during and post-drought val, NSE = 0.5
(c) during drought cal, NSE = 0.767
(d) pre- and post-drought val, NSE = 0.729
(e) post-drought cal, NSE = 0.715
(f) pre- and during drought val, NSE = 0.767
Sim (Box-Cox)
Obs (Box-Cox)
0.0
0.5
1.0
1.5
2.0
2.5
3.0

605

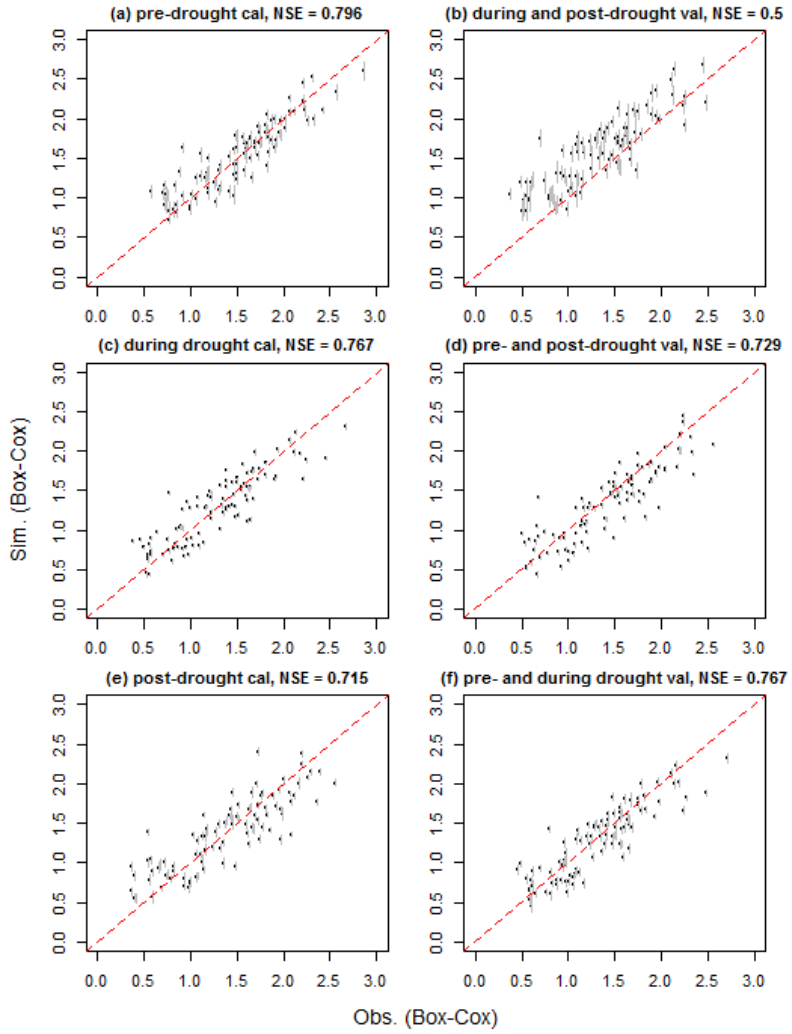


Figure 7. Comparison of the TSS model performance, as the simulated against observed site-level mean concentrations in Box-Cox transformed space. The left column shows calibration performance for the model calibrated to the pre-drought (1994-1996), drought (1997-2009) and the post-drought (2010-2014) periods, respectively; the right column shows the corresponding validation performance for each period. ~~**See Sect. 2.4 for details of the calibration and validation approach.**~~**The 95% lower and upper bounds of simulations shown in vertical grey lines and red dash lines show the 1:1 lines.**

The potential impacts of drought on TSS dynamics are further illustrated with the performance of the ~~full~~ spatio-temporal model (calibrated to the full dataset with all sites and all data from 1994 to 2014)

over the pre-, during and post-drought periods (Fig. ~~5~~8). Both the during- and post-drought periods have consistently good performances, while the model underestimates ~~the majority of~~most sites for the pre-drought period. This is consistent with Fig. ~~4~~7 in suggesting a systematic decrease in TSS concentration since the drought began. The better performance of the full model during and after drought (Fig. ~~5~~8) can be a ~~results~~result of the calibration period of the full spatio-temporal model – between 1994 and 2014 – which was dominated by the during- and post-drought periods~~;~~ ~~consequently, the full spatio-temporal model can be largely defined by observed TSS dynamics during and after the drought~~.

In summary, Figs~~. 4~~ 7 and ~~5~~8 together with Figs. S13-S17 suggest that ~~since~~whilst model performance for most constituents are not affected by the ~~drought, TSS concentrations experienced a large-scale downward shift compared to~~ hydrological periods used for calibration and validation, the ~~pre-drought~~calibration period~~, under otherwise identical spatial and temporal conditions. Such a shift indicates changes in the relationships between TSS and its key spatial and temporal controls since the start of the drought.~~ did have notable impact on TSS. Some possible causes are ~~further~~ discussed in ~~Sect.~~Section 4.3.

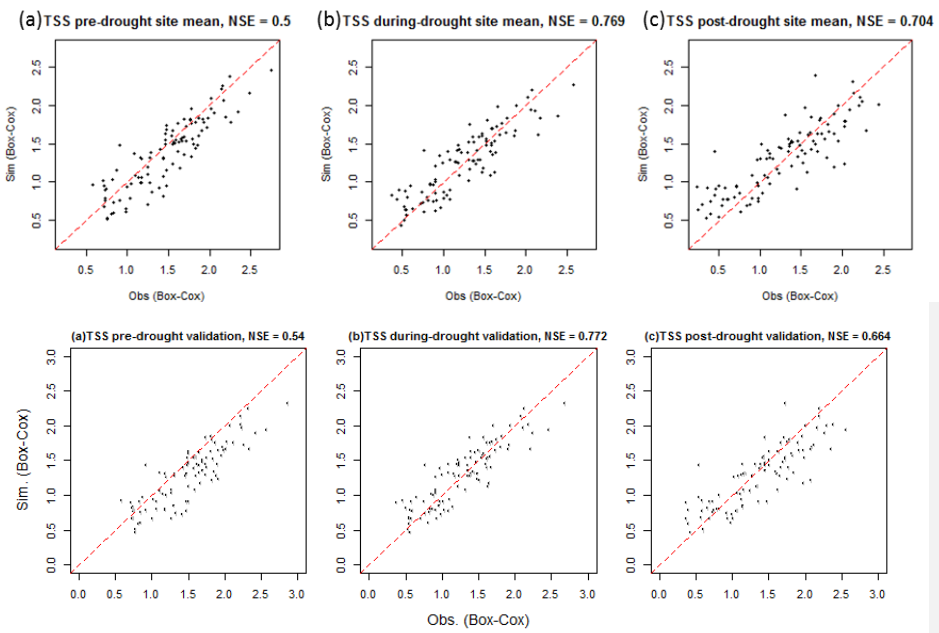


Figure 8. Comparison of the performance of the full spatio-temporal TSS model calibrated to all data across a) pre-drought (1994-1996), b) during drought (1997-2009) and c) post-drought (2010-2014) periods, as represented by the simulated against observed site-level mean concentrations in Box-Cox transformed space. The 95% lower and upper bounds of simulations shown in vertical grey lines and red dash lines show the 1:1 lines.

4. Discussion

4.1 Implications for statistical water quality modelling

~~Our~~ In this study, we developed the first process-informed statistical model that is capable of explaining a reasonable proportion of water quality variability for a large spatial ~~-temporal~~ area of over 130,000km^2. Although the calibration data have relatively low sampling frequency (i.e. monthly), our model generally performs satisfactorily in explaining the total variability in water quality. This demonstrates the effectiveness of the Bayesian hierarchical modelling framework in predicting spatio-temporal variability in water quality across large scales. The Bayesian hierarchical model is: a) more advantageous than other simpler statistical water quality models ~~are able~~with its more comprehensive and process-informed approach, and capacity to ~~capture the majority of observed variability across the 102 sampling locations in Victoria (Sect. 3.~~represent varying temporal relationships across large-scale regions; b) less demanding for input data compared with those required by fully-distributed, processes-based models.

From a practical perspective, this model has the potential to contribute to a number of management activities including catchment planning, management and policy-making activities, specifically:

1) The spatial predictive capacity can be used to identify pollution hot-spots and the catchment conditions that are likely causes of high concentrations. This can be used to help identify target catchment(s) to prioritize future water quality monitoring and management (Figs. 4 and 5);

2) Further to 1~~); the model performances~~), since water quality has been linked with catchment characteristics in this model, it can also ~~allow us to explore~~be used to assess potential impacts of alternative options of land use and land cover change, as well as potential effects of climate change, on ambient water quality conditions;

3) The model's temporal predictive capacity can identify changes in water quality due to changes in hydro-climatic conditions and vegetation cover, and thus enabling attribution of detected trends. On the other hand, any 'unexpected' trends can be identified to prompt further investigation to identify causes (Figure 6 and Table 4). The model could also be used for assessing the impacts of long-term catchment changes on water quality (Figures 7 and 8).

Despite the opportunities highlighted above, the model's performance also suggests some current limitations of the modelling framework~~. The greatest limiting factor for model performance seems to be when~~ in the following situations:

1) *High within-site temporal variability.* In Section 3.2 we have identified a general lack of predictive power for temporal variability. The potential impacts of high temporal variability on model performance is particularly evident for results of TSS, NO_x and FRP in Fig. 3. Since our model has already included hydro-climatic conditions and vegetation cover to explain temporal variability, the unexplained temporal variability is likely due to other uncaptured temporal drivers. These could be: changes in land use and land management, bio-geochemical processes, or transit time of water through catchments.

Presence of *high proportions of* ~~LOR data are present. As shown in Fig. 2 and Table 1, model performance is best for TKN and EC, where proportions of~~ *below* ~~LOR records are low. For~~ *DL data.* The full datasets for the three poorly modelled constituents ~~where the LOR records occupy greater proportions of the entire dataset, we observe poorer model performances (e.g. FRP).~~ (FRP, TSS and NO_x) all have higher proportions of data below the detection limit (38.2% 17.3% and 15% of all data, respectively) compared with other constituents. As illustrated in Fig. 2, ~~the FRP~~for each of these constituents, removal of below-DL data before model calibration ~~data have a~~had created clear ~~left~~ a truncation ~~pattern resulted from removing a large proportion of below LOR data.~~ on the left-hand side of the distribution. This substantially ~~increased~~increases the degrees of skewness and discontinuity of the data, essentially violating the assumption of ~~linear modelling of continuous data,~~normally distributed residuals and thus limiting ~~the~~model performance ~~of the spatio-temporal model. It is worth noting that in this study, since we modelled spatial and temporal variabilities in an integrated manner, the model may compensate representation of the individual components of spatial and temporal variability to improve~~. The model capacity to handle truncated data might be improved by model fitting ~~to the overall variability during calibration. Consequently, in this spatio-temporal modelling framework, large presence of below LOR data can limit the accurate representation of both variability components.~~

~~1)~~2) ~~Figure 2 highlights another possible influence on model performance, which is a combination of our inability to analyse low concentrations and the limited resolution of these low concentration measurements due to heavy transformation in data processing. This is evidenced by visually inspecting the fittings which show distinct "categorical" behaviour for low concentrations for some constituents. This "categorical" issue impacts the six constituents to different extents, ranking from the strongest as: FRP, TSS, TP, NO_x, TKN and EC – a ranking that is broadly aligned with the degree of lacking model performance for these constituents. Similar to the below LOR records, when these categorical values are present in large proportions of the full records (e.g. TSS and FRP), they can also violate the linear model assumptions and cause performance deterioration. This issue could be overcome by alternative~~

[1]All Box-Cox transformation parameters for water quality constituents are approximately 0 (Table S4), which means that the transformations are similar to a log transformation.

~~model structures that explicitly account for truncated data.~~approaches explicitly designed for this issue. For example, Wang and Robertson (2011) and Zhao et al. (2016) illustrated an approach to resolving the discontinuity of the likelihood estimation in ~~modelling~~model fitting to data with presence of a lower bound such as zero rainfall values~~, which can be potentially extended to improve fitting for the categorical levels at low concentrations.~~.

~~In addition, our current models are empirical relationships which are likely unable to represent complex biogeochemical processes. For example, performances for FRP and NO_x might be limited because: 1) the linear model structure can over-simplify constituent dynamics due to biogeochemical processes that are often highly non-linear; 2) the model may not include parameters that can adequately represent relevant biogeochemical processes (due to the lack of these data). To better capture changes in reactive constituents, greater consideration of and data representing biogeochemical processes may be required to address nutrient cycling including denitrification, ammonification and mineralisation (Granger et al., 2010). Therefore, possible ways to improve the statistical modelling of non-conservative constituents are: 1) alternative non-linear statistical model structures; or 2) inclusion of parameters to better represent biogeochemical processes.~~

3) ~~Lastly, it is worth noting that our results are presented in the transformed scale for which the spatio-temporal models were~~*Non-conservativeness of constituents.* The results indicate that the reactivity of the constituent is broadly associated with performance, which suggest that bio-geochemical processes (e.g. phosphorus cycling, nitrification/de-nitrification) can make water quality dynamics more difficult for the model to capture. To better capture changes in reactive constituents, the model may require greater consideration of and more extensive spatial and temporal data to represent bio-geochemical processes. Examples include improvements on the process representation for nitrogen cycling and the desorption and adsorption of phosphorus (Granger et al., 2010;Smyth et al., 2013;Tian and Zhou, 2007).

As previously noted, our model was developed ~~and~~ in a Box-Cox transformed scale to ensure the validity of the statistical assumptions ~~hold~~ (see details on data transformation in Sect. 2.1.2~~). Model~~), which

[1]All Box-Cox transformation parameters for water quality constituents are approximately 0 (Table S4), which means that the transformations are similar to a log transformation.

shows limited performance ~~is heavily affected~~for high constituent concentrations when ~~simulated model outputs~~simulations are back-transformed to the measurement scale (~~see~~ Figs. ~~S13 in the Supplementary Information).~~4 and 5). However, our ~~models are very useful in representing and predicting~~model approximately represents proportional changes in ~~concentrations~~water quality[1], which ~~adds important information for assessing and managing catchment water quality. For example, an increase of 1 mg L^{-1} in suspended solids would be alarming in pristine streams and/or periods of good water quality, while having much less impact on highly polluted conditions. The transformed models developed in this study can~~can thus help managers to understand ~~these~~ proportional changes to ~~identify critical locations~~inform practical catchment management.

For future implementations, the established model structure and parameterization would be best suited

[1]All Box-Cox transformation parameters for water quality constituents are approximately 0 (Table S4), which means that the transformations are similar to a log transformation.

to within the study region. Before performing new simulations (e.g. for new monitoring sites or for current study sites over a different time-period), the statistical properties of the new input datasets should be checked to ensure that they are similar to the calibration datasets. To model new catchments outside of the study region, a re-calibration of the model is required. This would involve extensive selection of key predictors and ~~periods of key water quality concerns.~~ model calibration, much as performed in this study and the two preceding ones (Lintern et al., 2018b; Guo et al., 2019). A sufficiently long record length (e.g. 20 years) is ideal for such modelling, as it ensures a reasonable understanding of the temporal variability to be obtained.

4.2 Implications for water quality monitoring programs

~~Within the~~The current spatio-temporal ~~models,~~model extracts water quality temporal variability ~~is based on~~from monthly ~~monitoring~~ data. ~~This suggests potential oppourtunities to~~Utilizing data with higher temporal resolution may further strengthen the model capacity to explain temporal variability, especially by ~~utilizing data with higher temporal resolution. This approach can~~capturing more information on water quality dynamics during flow events. This may be ~~supported by recent developments that significantly improved the accessibility of~~possible into the future; however, current high-~~-~~frequency water quality ~~monitoring data~~sensors (Bende-Michl and Hairsine, 2010;Outram et al., 2014;Lannergård et al., 2019;Pellerin et al., 2016)~~. Another potential development is to use remote sensing data to augment low frequency sampled data with higher frequency remotely sensed estimates e.g. for sediments and nutrients (Glasgow et al., 2004;Ritchie et al., 2003). Alternatively, where high frequency data are lacking for the target constituent, high frequency proxy data could also be utilized to enhance the understanding obtained from low frequency samples. For example, turbidity can be used as surrogate for sediments and nutrients (Schilling et al., 2017;Robertson et al., 2018;Lannergård et al., 2019). Currently, continuous turbidity data are available from Australia state agencies, such as the Victorian Water Quality Monitoring Network database (Department of Environment Land Water and Planning Victoria, 2016) and the NSW Water information database (WaterNSW, 2018), and collated at national level in the Bureau of Meteorology's Water Data Online portal (Bureau of Meteorology, 2019). These datasets may have great potential to enhance the temporal resolutions of records for other key water quality~~

Formatted: Footer, Line spacing: single, Tab stops: Not at 7.62 cm + 15.24 cm

~~constituents (e.g. nutrients and sediments).~~

~~Changes~~ still have very high resourcing requirements that limits widespread deployment in operational networks.

Furthermore, changes in land use and management over time ~~(e.g. tillage, fertiliser application, irrigation)~~ are currently not considered here as predictors of ~~water quality~~ temporal variability~~.~~ in water quality, which include but not limit to land clearing, urbanization, tillage, fertiliser application and irrigation. This is due to a complete lack ~~of availability and/~~, or inconsistency of available data. However, changes in land use/land management practices can occur over short time periods, which~~-~~ can lead to increases in pollutant sources and changes to runoff generation processes (e.g. Tang et al., 2005;DeFries and Eshleman, 2004;Smith et al., 2013). Therefore, ~~model performance~~our modelling framework can potentially be ~~further~~ improved by ~~increased capacities in the~~ having additional monitoring ~~of~~data on the temporal patterns of land use/land management~~.~~ to better capture their impacts on water quality.

4.3 Potential impacts of long-term drought on water quality dynamics

Results of model calibration and validation to different time periods suggest a systematic decrease in TSS concentrations ~~since~~during and after the prolonged drought, in comparison with the pre-drought period under the same spatial and temporal conditions. Such a shift is not observed for any other five constituents ~~analysed~~analyzed (nutrients and salts) (~~Sect.~~Section 3.~~2).~~3).

A further analysis of the calibrated model parameters for pre-, during and post-drought periods suggest that the effects of key spatial predictors do not vary much across periods (Figure S14). In contrast, the effects of key temporal predictors highlight a clear shift in the role of antecedent flow (prior 7-day flow) across different time periods (Figure 9). Specifically, the antecedent flow effects are mostly positive across catchments before the drought, and shift to mostly negative during the drought. After the drought, the antecedent flow effects have mixed directions among different catchments. Considering the limited performance of the TSS model (i.e. substantial under-estimation of temporal variability in Section 3.1), these changing relationships suggested in the calibrated parameters might be unreliable. However, this should not affect the reliability of the observed change in TSS since the drought (Section 3.3), which

Formatted: Footer, Line spacing: single, Tab stops: Not at 7.62 cm + 15.24 cm

was based on the systematic differences of model fitting between different periods, revealing a broad-scale patterns across the state on the drought influences.

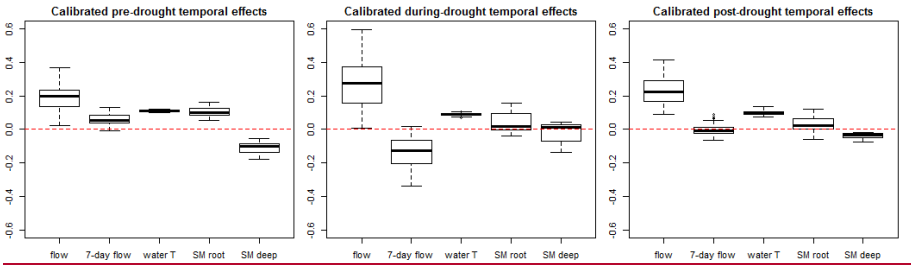


Figure 9. Effects of the five key predictors for the temporal variability in TSS across 102 sites, summarized by the posterior mean of the calibrated parameter values for each predictor (box shows values across all sites), from left: flow, 7-day antecedent flow, water temperature, root-zone soil moisture and deep soil moisture.

In the literature, impacts of the Millennium Drought on the hydrology and runoff regimes of south-eastern Australia are well understood (van Dijk et al., 2013;Leblanc et al., 2012;Saft et al., 2015). However, less is known about how this ~~significant~~major and prolonged drought event has impacted water quality (Bond et al., 2008). Previous studies on other drought events around the world mainly focused on changes in water quality as responses to the reduced streamflow during drought. For ~~examples~~example, reduction in sediment levels ~~have~~during drought has been reported ~~during drought, due~~and attributed to lower erosion from the contributing catchment ~~and~~, together with lower rates of solid transport associated with reduced flows (Murdoch et al., 2000;Caruso, 2002). At a more local scale, increasing sediment concentrations during ~~drogulht~~drought have also been observed in streams ~~adjscent~~adjacent to land with high densities of livestock and bushland, which both constantly contribute to sediment load during drought, leading to elevated concentrations ~~due to~~with lower dilution rate (Caruso, 2002). ~~Similarly~~Similar to sediments, the impact of droughts on stream nutrient and salt concentrations ~~were~~have also commonly been understood as responses to reduced runoff generation and streamflow. ~~Nutrient concentrations typically decrease during droughts in~~In catchments with no significant point-source pollution, nutrient concentrations typically decreased during droughts (Mosley, 2015)~~, as~~ with less nutrient leaching and overland flow ~~are reduced~~, but may also increase due to increasing livestock inputs at more local scales (Caruso, 2002). In contrast, catchments with significant point-source pollution generally experience water quality deterioration during drought due to reduced

Formatted: Footer, Line spacing: single, Tab stops: Not at 7.62 cm + 15.24 cm

dilution (van Vliet and Zwolsman, 2008;Mosley, 2015). For salinity, concentration often increases during drought with reduced dilution and increased evaporation (Caruso, 2002)~~,~~. This is particularly evident for catchments that are more influenced by saline groundwater input ~~where drought can increase~~as the relative contribution of ~~saline~~ groundwater ~~input~~increased during drought (Costelloe et al., 2005).

~~However~~In contrast to these previous studies, our findings ~~highlight other~~suggest additional possible pathways ~~on how~~along which drought can affect stream water quality~~. The results suggest~~ , that ~~the~~ prolonged drought ~~induced changes in sediment dynamics i.e. changes of~~might have altered the relationships between sediments and its predictors (Figs. 4~~7~~ and ~~5~~8). In contrast to sediments, our ~~models for~~ model suggests no clear shifts in the dynamics of nutrients and salts ~~maintain consistent performance for different drought and non-drought periods, suggesting no clear shifts in dynamics. A~~ in a regional scale. Our findings are in line with a few previous studies ~~have also~~which reported temporal changes in the concentration-discharge relationships for sediments and nutrients~~. Specifically, these relationships changed from~~, specifically, when comparing high- ~~to~~and low-flow conditions (Zhang, 2018;Moatar et al., 2017), as well as ~~from~~ drought ~~to the~~and recovery period (Burt et al., 2015)~~. However, effects of extended multi-year droughts on the~~ . Our findings provide extra dimensions to what would be offered by simple trend analyses using approaches such as Mann Kendall test or Sen's slope (e.g. Smith et al., 1987;Chang, 2008;Hirsch et al., 1991;Bouza-Deaño et al., 2008). Those approaches are only capable of indicating direction and magnitude of observed trends. In contrast, our model was able to attribute the consistent upward shift in TSS concentration~~ discharge relationships are less explored. Furthermore, there is also a lack of comprehensive assessments on the~~ to change ~~of~~in relationships between water quality and ~~other relevant controls (e.g. water temperature, land cover etc.) during extended drought over large geographical regions. Our findings highlight great oppourtunities to use this dataset to further investigate the impacts of prolonged droughts on water quality dynamics, especially the changes in relationships between TSS and each of~~ its key ~~controls across multiple catchments.~~driving factors since the start of drought.

In addition, we also acknowledge that our ability to represent the pre- and post-drought conditions in this study may be limited by the record length, since only 2 years of pre-drought and 4 years of post-

Formatted: Footer, Line spacing: single, Tab stops: Not at 7.62 cm + 15.24 cm

drought data were available. Once longer records build up, they will enable us to update our understanding of the impact of this prolonged drought. We would be also able to conduct more sophisticated investigations, such as comparing the impacts of long-term droughts versus individual dry and wet years~~. Addressing these research questions are particularly important in a changing climate that will be characterized by lower streamflows~~ and ~~possibly a shift towards more intermittent flows in many parts of the world~~events (e.g. Saft et al., 2015;~~Chiew~~Outram et al., 2014;~~Ukkola~~Burt et al., 2015).

5. Conclusions

~~Using~~This study aims to address the current lack of water quality models that operate at large scales across multiple catchments. To achieve this, we used long-term stream water quality data collected from 102 sites in south-eastern Australia, ~~we~~and developed a Bayesian hierarchical statistical model to ~~analyse~~simulate the spatio-temporal variabilities in six key water quality constituents: TSS, TP, FRP, TKN, NO_x and EC. ~~The~~ The choice of model predictors was guided by previous studies on the same dataset (Lintern et al., 2018b; Guo et al., 2019). The model generally well captures the spatio-temporal ~~models are capable of predicting future water quality~~variability in water quality, where spatial variability between catchments is much better represented than temporal variability. The model is best used to predict proportional changes in water quality in a Box-Cox transformed scale, and can have substantial bias if used to predict absolute values for high concentrations. Cross-validation shows that the spatio-temporal model can predict water quality in non-monitored locations under similar conditions to the historical period and the calibration catchments that we investigated. ~~A notable shift in TSS dynamics is observed since the extended drought in the study region, which highlights~~This can assist management by (1) identifying hot-spots and key temporal periods for waterway pollution; (2) testing effects of catchment changes e.g. urbanization or afforestation; and (3) identifying and attributing major water quality trends and changes.

Based on the above model evaluations, we discussed potential ~~oppourtunities for further research to better understand the impact of this significant drought event on water quality.~~

~~Despite the promising~~ways to further enhance the model performance~~ of these models, the results also illustrate areas of further improvement, both in the modelling framework but also in the monitoring of~~

Formatted: Footer, Line spacing: single, Tab stops: Not at 7.62 cm + 15.24 cm

~~water quality.~~. In improving the modelling framework, alternative statistical approaches could be considered to reduce the impact of below detection limit ~~and low concentration~~ data on model performance. In addition, the models could be extended to ~~take into account~~consider some key ~~biogeochemical~~bio-geochemical processes to better ~~represent spatial-temporal variability~~ dynamics in non-conservative constituents (e.g., FRP or NO_x). ~~To further enhance the performance of the current models, we recommend that future water quality monitoring programs be enhanced with: 1) collection and assimilation of high frequency sampling data to enhance the temporal resolution of water quality data; and 2) more frequent~~Regarding data availability, the current models could potentially benefit from improved monitoring of changes in land use intensity and management to be able to include these ~~parameters in the model. These improvements will be very helpful to operational catchment management and mitigation~~drivers in the model. The inclusion of high-frequency water quality sampling data may also extend the model's ability to represent temporal variability. However, high-frequency water quality data are also typically highly variable with large noise. Therefore, the implication of such data for the spatio-temporal modelling framework remains an open question, which needs further investigation in future applications of this modeling framework.

Data availability

All data used in this study were extracted from public domain. All stream water quality data were extracted from the Victorian Water Measurement Information System (via http://data.water.vic.gov.au/, provided by the Department of Environment Land Water and Planning Victoria). The catchments corresponding to these water quality monitoring sites were delineated using the Geofabric tool provided by the Bureau of Meteorology, via ftp://ftp.bom.gov.au/anon/home/geofabric/. We have listed the sources of all other data for the spatial and temporal predictors of our models in Tables S1 and S2 in the Supplementary Materials.

Author contribution

All authors contributed to the conceptualization the models and the design of methodology. A. Lintern and S. Liu contributed to the data curatiom. D.Guo carried out the formal analyses, visualization and validation. J.A. Webb, D. Ryu, U. Bende-Michl and A.W. Western contributed to the funding

Formatted: Footer, Line spacing: single, Tab stops: Not at 7.62 cm + 15.24 cm

acquisition. D. Guo, A. Lintern, J.A. Webb, D. Ryu, S. Liu and A.W. Western contributed to the investigation. D. Guo carried out project administration and coding to run the experiments. J.A. Webb, D. Ryu, and A.W. Western contributed to the supervision. D.Guo prepared the manuscript with contributions from all co-authors.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgement

The Australian Research Council ~~and~~, the Victorian Environment Protection Authority, the Victorian Department of Environment, Land Water and Planning, the Australian Bureau of Meteorology and the Queensland Department of Natural Resources, Mines and Energy provided funding for this project through the linkage program (LP140100495). The authors would also like to thank Matthew Johnson, Louise Sullivan, Hannah Sleeth and Jie Jian, for their assistance in the compilation and analysis of data. All water quality data used for this project can be found on: Water Measurement Information System (http://data.water.vic.gov.au/monitoring.htm). Sources of other data are provided in Tables S1 and S2 of the Supplementary Materials.

Formatted: Footer, Line spacing: single, Tab stops: Not at 7.62 cm + 15.24 cm

References

Abbaspour, K. C., Rouholahnejad, E., Vaghefi, S., Srinivasan, R., Yang, H., and Kløve, B.: A continental-scale hydrology and water quality model for Europe: Calibration and uncertainty of a high-resolution large-scale SWAT model, Journal of Hydrology, 524, 733-752, https://doi.org/10.1016/j.jhydrol.2015.03.027, 2015.

Adams, R., Arafat, Y., Eate, V., Grace, M. R., Saffarpour, S., Weatherley, A. J., and Western, A. W.: A catchment study of sources and sinks of nutrients and sediments in south-east Australia, Journal of Hydrology, 515, 166-179, https://doi.org/10.1016/j.jhydrol.2014.04.034, 2014.

Ahearn, D. S., Sheibley, R. W., Dahlgren, R. A., and Keller, K. E.: Temporal dynamics of stream water chemistry in the last free-flowing river draining the western Sierra Nevada, California, Journal of Hydrology, 295, 47-63, https://doi.org/10.1016/j.jhydrol.2004.02.016, 2004.

Ai, L., Shi, Z. H., Yin, W., and Huang, X.: Spatial and seasonal patterns in stream water contamination across mountainous watersheds: Linkage with landscape characteristics, Journal of Hydrology, 523, 398-408, https://doi.org/10.1016/j.jhydrol.2015.01.082, 2015.

Akaike, H.: A new look at the statistical model identification, IEEE transactions on automatic control, 19, 716-723, 1974.

Akritas, M. G., Murphy, S. A., and Lavalley, M. P.: The Theil-Sen estimator with doubly censored data and applications to astronomy, Journal of the American Statistical Association, 90, 170-177, 1995.

Ali, G., Wilson, H., Elliott, J., Penner, A., Haque, A., Ross, C., and Rabie, M.: Phosphorus export dynamics and hydrobiogeochemical controls across gradients of scale, topography and human impact, Hydrological Processes, 31, 3130-3145, 10.1002/hyp.11258, 2017.

Australian Water Technologies: Victorian water quality monitoring network and state biological monitoring programme manual of procedures, Australian Water Technologies, 68 Ricketts Rd, Mt Waverley VIC 3149, 1999.

Bailey, R. T., and Ahmadi, M.: Spatial and temporal variability of in-stream water quality parameter influence on dissolved oxygen and nitrate within a regional stream network, Ecological modelling, 277, 87-96, 2014.

Bende-Michl, U., and Hairsine, P. B.: A systematic approach to choosing an automated nutrient analyser for river monitoring, Journal of Environmental Monitoring, 12, 127-134, 2010.

~~Bengraïne, K., and Marhaba, T. F.: Using principal component analysis to monitor spatial and temporal changes in water quality, Journal of Hazardous Materials, 100, 179-195, https://doi.org/10.1016/S0304-3894(03)00104-3, 2003.~~

Bond, N. R., Lake, P. S., and Arthington, A. H.: The impacts of drought on freshwater ecosystems: an Australian perspective, Hydrobiologia, 600, 3-16, 10.1007/s10750-008-9326-z, 2008.

Borsuk, M. E., Higdon, D., Stow, C. A., and Reckhow, K. H.: A Bayesian hierarchical model to predict benthic oxygen demand from organic matter loading in estuaries and coastal zones, Ecological Modelling, 143, 165-181, https://doi.org/10.1016/S0304-3800(01)00328-3, 2001.

Bouza-Deaño, R., Ternero-Rodríguez, M., and Fernández-Espinosa, A. J.: Trend study and assessment of surface water quality in the Ebro River (Spain), Journal of Hydrology, 361, 227-239, https://doi.org/10.1016/j.jhydrol.2008.07.048, 2008.

Box, G. E., and Cox, D. R.: An analysis of transformations, Journal of the Royal Statistical Society: Series B (Methodological), 26, 211-243, 1964.

Bureau of Meteorology: Geofabric V2 2012.

~~Bureau of Meteorology: Water Data Online, 2019.~~

Bureau of Rural Sciences: 2005/06 Land use of Australia, version 4. . 2010.

Burt, T. P., Worrall, F., Howden, N. J. K., and Anderson, M. G.: Shifts in discharge-concentration relationships as a small catchment recover from severe drought, Hydrological Processes, 29, 498-507, 10.1002/hyp.10169, 2015.

Formatted: Footer, Line spacing: single, Tab stops: Not at 7.62 cm + 15.24 cm

Carey, R. O., and Migliaccio, K. W.: Contribution of wastewater treatment plant effluents to nutrient dynamics in aquatic systems: a review, Environ Manage, 44, 205-217, 10.1007/s00267-009-9309-5, 2009.
Caruso, B. S.: Temporal and spatial patterns of extreme low flows and effects on stream ecosystems in Otago, New Zealand, Journal of Hydrology, 257, 115-133, https://doi.org/10.1016/S0022-1694(01)00546-7, 2002.
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.
~~Chiew, F. H. S., Potter, N. J., Vaze, J., Petheram, C., Zhang, L., Teng, J., and Post, D. A.: Observed hydrologic non-stationarity in far south-eastern Australia: implications for modelling and prediction, Stochastic Environmental Research and Risk Assessment, 28, 3-15, 10.1007/s00477-013-0755-5, 2014.~~
Clark, J. S.: Why environmental scientists are becoming Bayesians, Ecology Letters, 8, 2-14, doi:10.1111/j.1461-0248.2004.00702.x, 2005.
Costelloe, J. F., Grayson, R. B., McMahon, T. A., and Argent, R. M.: Spatial and temporal variability of water salinity in an ephemeral, arid-zone river, central Australia, Hydrological Processes, 19, 3147-3166, 10.1002/hyp.5837, 2005.
DeFries, R., and Eshleman, K. N.: Land-use change and hydrologic processes: a major focus for the future, Hydrological Processes, 18, 2183-2186, doi:10.1002/hyp.5584, 2004.
Department of Environment Land Water and Planning Victoria: Victorian water measurement information system. . 2016.
Eidenshink, J. C.: The 1990 Conterminous U.S. AVHRR Data Set, Photogrammetric Engineering and Remote Sensing, 58, 1992.
Fraser, A. I., Harrod, T. R., and Haygarth, P. M.: The effect of rainfall intensity on soil erosion and particulate phosphorus transfer from arable soils, Water Science and Technology, 39, 41-45, https://doi.org/10.1016/S0273-1223(99)00316-9, 1999.
Frost, A. J., Ramchurn, A., and Smith, A.: The bureau's operational AWRA landscape (AWRA-L) Model, Bureau of Meteorology, 2016.
Fu, B., Merritt, W. S., Croke, B. F. W., Weber, T., and Jakeman, A. J.: A review of catchment-scale water quality and erosion models and a synthesis of future prospects, Environmental Modelling & Software, https://doi.org/10.1016/j.envsoft.2018.12.008, 2018.
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.
Geoscience Australia: Dams and water storages. . 2004.
Geoscience Australia: Environmental Attributes Dataset. 2011.
~~Glasgow, H. B., Burkholder, J. M., Reed, R. E., Lewitus, A. J., and Kleinman, J. E.: Real-time remote monitoring of water quality: a review of current applications, and advancements in sensor, telemetry, and computing technologies, Journal of Experimental Marine Biology and Ecology, 300, 409-448, https://doi.org/10.1016/j.jembe.2004.02.022, 2004.~~
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.
Granger, S. J., Bol, R., Anthony, S., Owens, P. N., White, S. M., and Haygarth, P. M.: Chapter 3 - Towards a Holistic Classification of Diffuse Agricultural Water Pollution from Intensively Managed Grasslands on Heavy Soils, in: Advances in Agronomy, Academic Press, 83-115, 2010.
Guo, D., Lintern, A., Webb, J. A., Ryu, D., Liu, S., Bende-Michl, U., Leahy, P., Wilson, P., and Western, A. W.: Key Factors Affecting Temporal Variability in Stream Water Quality, Water Resources Research, 55, 112-129, 10.1029/2018wr023370, 2019.
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.
Hirsch, R. M., Alexander, R. B., and Smith, R. A.: Selection of methods for the detection and estimation of trends in water quality, Water Resources Research, 27, 803-813, 10.1029/91wr00259, 1991.

Formatted: Footer, Line spacing: single, Tab stops: Not at 7.62 cm + 15.24 cm

Hrachowitz, M., Benettin, P., van Breukelen, B. M., Fovet, O., Howden, N. J. K., Ruiz, L., van der Velde, Y., and Wade, A. J.: Transit times—the link between hydrology and water quality at the catchment scale, Wiley Interdisciplinary Reviews: Water, 3, 629-657, doi:10.1002/wat2.1155, 2016.
Kaushal, S. S., Mayer, P. M., Vidon, P. G., Smith, R. M., Pennino, M. J., Newcomer, T. A., Duan, S., Welty, C., and Belt, K. T.: Land Use and Climate Variability Amplify Carbon, Nutrient, and Contaminant Pulses: A Review with Management Implications, JAWRA Journal of the American Water Resources Association, 50, 585-614, doi:10.1111/jawr.12204, 2014.
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.
Kingsford, R. T., Walker, K. F., Lester, R. E., Young, W. J., Fairweather, P. G., Sammut, J., and Geddes, M. C.: A Ramsar wetland in crisis – the Coorong, Lower Lakes and Murray Mouth, Australia, Marine and Freshwater Research, 62, 255-265, https://doi.org/10.1071/MF09315, 2011.
Kisi, O., and Parmar, K. S.: Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water pollution, Journal of Hydrology, 534, 104-112, https://doi.org/10.1016/j.jhydrol.2015.12.014, 2016.
Kurunç, A., Yürekli, K., and Çevik, O.: Performance of two stochastic approaches for forecasting water quality and streamflow data from Yeşilırmak River, Turkey, Environmental Modelling & Software, 20, 1195-1200, https://doi.org/10.1016/j.envsoft.2004.11.001, 2005.
Lannergård, E. E., Ledesma, J. L., Fölster, J., and Futter, M. N.: An evaluation of high frequency turbidity as a proxy for riverine total phosphorus concentrations, Science of the Total Environment, 651, 103-113, 2019.
Leblanc, M., Tweed, S., Van Dijk, A., and Timbal, B.: A review of historic and future hydrological changes in the Murray-Darling Basin, Global and Planetary Change, 80-81, 226-246, https://doi.org/10.1016/j.gloplacha.2011.10.012, 2012.
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, 2018a.
Lintern, A., Webb, J. A., Ryu, D., Liu, S., Waters, D., Leahy, P., Bende-Michl, U., and Western, A. W.: What Are the Key Catchment Characteristics Affecting Spatial Differences in Riverine Water Quality?, Water Resources Research, doi:10.1029/2017WR022172, 2018b.
Luca Scrucca: Package 'GA', The Comprehensive R Archive Network,, 2019.
May, R., Dandy, G., and Maier, H.: Review of input variable selection methods for artificial neural networks, in: Artificial neural networks-methodological advances and biomedical applications, InTech, 2011.
Mellander, P.-E., Jordan, P., Shore, M., Melland, A. R., and Shortle, G.: Flow paths and phosphorus transfer pathways in two agricultural streams with contrasting flow controls, Hydrological Processes, 29, 3504-3518, doi:10.1002/hyp.10415, 2015.
Meybeck, M., and Helmer, R.: The quality of rivers: From pristine stage to global pollution, Palaeogeography, Palaeoclimatology, Palaeoecology, 75, 283-309, https://doi.org/10.1016/0031-0182(89)90191-0, 1989.
Miller, C., Magdalina, A., Willows, R. I., Bowman, A. W., Scott, E. M., Lee, D., Burgess, C., Pope, L., Pannullo, F., and Haggarty, R.: Spatiotemporal statistical modelling of long-term change in river nutrient concentrations in England & Wales, Science of The Total Environment, 466-467, 914-923, https://doi.org/10.1016/j.scitotenv.2013.07.113, 2014.
Moatar, F., Abbott, B. W., Minaudo, C., Curie, F., and Pinay, G.: Elemental properties, hydrology, and biology interact to shape concentration-discharge curves for carbon, nutrients, sediment, and major ions, Water Resources Research, 53, 1270-1287, 10.1002/2016wr019635, 2017.
Mosley, L. M.: Drought impacts on the water quality of freshwater systems; review and integration, Earth-Science Reviews, 140, 203-214, https://doi.org/10.1016/j.earscirev.2014.11.010, 2015.
Murdoch, P. S., Baron, J. S., and Miller, T. L.: POTENTIAL EFFECTS OF CLIMATE CHANGE ON SURFACE-WATER QUALITY IN NORTH AMERICA1, JAWRA Journal of the American Water Resources Association, 36, 347-366, 10.1111/j.1752-1688.2000.tb04273.x, 2000.

Formatted: Footer, Line spacing: single, Tab stops: Not at 7.62 cm + 15.24 cm

Musolff, A., Schmidt, C., Selle, B., and Fleckenstein, J. H.: Catchment controls on solute export, Advances in Water Resources, 86, 133-146, https://doi.org/10.1016/j.advwatres.2015.09.026, 2015.
NASA LP DAAC: MOD13A3: MODIS/Terra Vegetation Indices Monthly L3 Global 1km V005. 2017.
Onderka, M., Wrede, S., Rodný, M., Pfister, L., Hoffmann, L., and Krein, A.: Hydrogeologic and landscape controls of dissolved inorganic nitrogen (DIN) and dissolved silica (DSi) fluxes in heterogeneous catchments, Journal of Hydrology, 450-451, 36-47, https://doi.org/10.1016/j.jhydrol.2012.05.035, 2012.
Outram, F. N., Lloyd, C. E. M., Jonczyk, J., Benskin, C. M. H., Grant, F., Perks, M. T., Deasy, C., Burke, S. P., Collins, A. L., Freer, J., Haygarth, P. M., Hiscock, K. M., Johnes, P. J., and Lovett, A. L.: High-frequency monitoring of nitrogen and phosphorus response in three rural catchments to the end of the 2011–2012 drought in England, Hydrol. Earth Syst. Sci., 18, 3429-3448, 10.5194/hess-18-3429-2014, 2014.
Ouyang, W., Hao, F., Skidmore, A. K., and Toxopeus, A. G.: Soil erosion and sediment yield and their relationships with vegetation cover in upper stream of the Yellow River, Science of The Total Environment, 409, 396-403, https://doi.org/10.1016/j.scitotenv.2010.10.020, 2010.
Parmar, K. S., and Bhardwaj, R.: Statistical, time series, and fractal analysis of full stretch of river Yamuna (India) for water quality management, Environmental Science and Pollution Research, 22, 397-414, 10.1007/s11356-014-3346-1, 2015.
Pellerin, B. A., Stauffer, B. A., Young, D. A., Sullivan, D. J., Bricker, S. B., Walbridge, M. R., Clyde Jr., G. A., and Shaw, D. M.: Emerging Tools for Continuous Nutrient Monitoring Networks: Sensors Advancing Science and Water Resources Protection, JAWRA Journal of the American Water Resources Association, 52, 993-1008, 10.1111/1752-1688.12386, 2016.
Poor, C. J., and McDonnell, J. J.: The effects of land use on stream nitrate dynamics, Journal of Hydrology, 332, 54-68, https://doi.org/10.1016/j.jhydrol.2006.06.022, 2007.
Poudel, D. D., Lee, T., Srinivasan, R., Abbaspour, K., and Jeong, C. Y.: Assessment of seasonal and spatial variation of surface water quality, identification of factors associated with water quality variability, and the modeling of critical nonpoint source pollution areas in an agricultural watershed, Journal of Soil and Water Conservation, 68, 155-171, 10.2489/jswc.68.3.155, 2013.
Poulsen, D. L., Simmons, C. T., Le Galle La Salle, C., and Cox, J. W.: Assessing catchment-scale spatial and temporal patterns of groundwater and stream salinity, Hydrogeology Journal, 14, 1339-1359, 10.1007/s10040-006-0065-9, 2006.
Qin, B., Zhu, G., Gao, G., Zhang, Y., Li, W., Paerl, H. W., and Carmichael, W. W.: A Drinking Water Crisis in Lake Taihu, China: Linkage to Climatic Variability and Lake Management, Environmental Management, 45, 105-112, 10.1007/s00267-009-9393-6, 2010.
Raupach, M., Briggs, P., Haverd, V., King, E., Paget, M., and Trudinger, C.: Australian water availability project (AWAP): CSIRO marine and atmospheric research component: final report for phase 3, 67, 2009.
Raupach, M., Briggs, P., Haverd, V., King, E., Paget, M., and Trudinger, C.: Australian Water Availability Project. CSIRO Marine and Atmospheric Research, Canberra, Australia., 2012.
~~Ritchie, J. C., Zimba, P. V., and Everitt, J. H.: Remote Sensing Techniques to Assess Water Quality, Photogrammetric Engineering & Remote Sensing, 69, 695-704, 10.14358/PERS.69.6.695, 2003.~~
~~Robertson, D. M., Hubbard, L. E., Lorenz, D. L., and Sullivan, D. J.: A surrogate regression approach for computing continuous loads for the tributary nutrient and sediment monitoring program on the Great Lakes, Journal of Great Lakes Research, 44, 26-42, https://doi.org/10.1016/j.jglr.2017.10.003, 2018.~~
Ren, W., Zhong, Y., Meligrana, J., Anderson, B., Watt, W. E., Chen, J., and Leung, H.-L.: Urbanization, land use, and water quality in Shanghai: 1947–1996, Environment International, 29, 649-659, https://doi.org/10.1016/S0160-4120(03)00051-5, 2003.
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.
Saft, M., Peel, M. C., Western, A. W., and Zhang, L.: Predicting shifts in rainfall-runoff partitioning during multiyear drought: Roles of dry period and catchment characteristics, Water Resources Research, 52, 9290-9305, doi:10.1002/2016WR019525, 2016.

Formatted: Footer, Line spacing: single, Tab stops: Not at 7.62 cm + 15.24 cm

Schilling, K. E., Kim, S. W., and Jones, C. S.: Use of water quality surrogates to estimate total phosphorus concentrations in Iowa rivers, Journal of Hydrology: Regional Studies, 12, 111-121, https://doi.org/10.1016/j.ejrh.2017.04.006, 2017.

Schwarz, G.: Estimating the dimension of a model, The annals of statistics, 6, 461-464, 1978.

Sharpley, A. N., Kleinman, P. J. A., McDowell, R. W., Gitau, M., and Bryant, R. B.: Modeling phosphorus transport in agricultural watersheds: Processes and possibilities, Journal of Soil and Water Conservation, 57, 425-439, 2002.

Smith, A. P., Western, A. W., and Hannah, M. C.: Linking water quality trends with land use intensification in dairy farming catchments, Journal of Hydrology, 476, 1-12, 2013.

Smith, R. A., Alexander, R. B., and Wolman, M. G.: Water-Quality Trends in the Nation's Rivers, Science, 235, 1607-1615, 1987.

Smyth, A. R., Thompson, S. P., Siporin, K. N., Gardner, W. S., McCarthy, M. J., and Piehler, M. F.: Assessing nitrogen dynamics throughout the estuarine landscape, Estuaries and coasts, 36, 44-55, 2013.

Stan Development Team: RStan: the R interface to Stan. R package version 2.18.1, 2018.

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

Sueker, J. K., Clow, D. W., Ryan, J. N., and Jarrett, R. D.: Effect of basin physical characteristics on solute fluxes in nine alpine/subalpine basins, Colorado, USA, Hydrological Processes, 15, 2749-2769, 10.1002/hyp.265, 2001.

Tang, Z., Engel, B. A., Pijanowski, B. C., and Lim, K. J.: Forecasting land use change and its environmental impact at a watershed scale, Journal of Environmental Management, 76, 35-45, https://doi.org/10.1016/j.jenvman.2005.01.006, 2005.

Terrestrial Ecosystem Research Network: Soil and landscape grid of Australia. 2016.

Tian, J. R., and Zhou, P. J.: Phosphorus fractions of floodplain sediments and phosphorus exchange on the sediment–water interface in the lower reaches of the Han River in China, Ecological Engineering, 30, 264-270, https://doi.org/10.1016/j.ecoleng.2007.01.006, 2007.

Tramblay, Y., Ouarda, T. B. M. J., St-Hilaire, A., and Poulin, J.: Regional estimation of extreme suspended sediment concentrations using watershed characteristics, Journal of Hydrology, 380, 305-317, https://doi.org/10.1016/j.jhydrol.2009.11.006, 2010.

Ukkola, A. M., Prentice, I. C., Keenan, T. F., van Dijk, A. I. J. M., Viney, N. R., Myneni, Ranga B., and Bi, J.: Reduced streamflow in water-stressed climates consistent with CO2 effects on vegetation, Nature Climate Change, 6, 75, 10.1038/nclimate2831

https://www.nature.com/articles/nclimate2831#supplementary-information, 2015.

van Dijk, A. I. J. M., Beck, H. E., Crosbie, R. S., de Jeu, R. A. M., Liu, Y. Y., Podger, G. M., Timbal, B., and Viney, N. R.: The Millennium Drought in southeast Australia (2001–2009): Natural and human causes and implications for water resources, ecosystems, economy, and society, Water Resources Research, 49, 1040-1057, 10.1002/wrcr.20123, 2013.

van Vliet, M. T. H., and Zwolsman, J. J. G.: Impact of summer droughts on the water quality of the Meuse river, Journal of Hydrology, 353, 1-17, https://doi.org/10.1016/j.jhydrol.2008.01.001, 2008.

Victorian water measurement information system: http://data.water.vic.gov.au/monitoring.htm, 2016.

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.

Vörösmarty, C. J., McIntyre, P. B., Gessner, M. O., Dudgeon, D., Prusevich, A., Green, P., Glidden, S., Bunn, S. E., Sullivan, C. A., Liermann, C. R., and Davies, P. M.: Global threats to human water security and river biodiversity, Nature, 467, 555, 10.1038/nature09440

https://www.nature.com/articles/nature09440#supplementary-information, 2010.

Wang, Q. J., and Robertson, D. E.: Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences, Water Resources Research, 47, 10.1029/2010wr009333, 2011.

Formatted: Footer, Line spacing: single, Tab stops: Not at 7.62 cm + 15.24 cm

Wang, Q. J., Shrestha, D. L., Robertson, D. E., and Pokhrel, P.: A log-sinh transformation for data normalization and variance stabilization, Water Resources Research, 48, doi:10.1029/2011WR010973, 2012.

~~Real-time water data: https://realtimedata.waternsw.com.au/, access: 12/07/2018, 2018.~~

Webb, J. A., and King, L. E.: A Bayesian hierarchical trend analysis finds strong evidence for large-scale temporal declines in stream ecological condition around Melbourne, Australia, Ecography, 32, 215-225, doi:10.1111/j.1600-0587.2008.05686.x, 2009.

Whitworth, K. L., Baldwin, D. S., and Kerr, J. L.: Drought, floods and water quality: Drivers of a severe hypoxic blackwater event in a major river system (the southern Murray–Darling Basin, Australia), Journal of Hydrology, 450-451, 190-198, https://doi.org/10.1016/j.jhydrol.2012.04.057, 2012.

Zhang, Q., and Ball, W. P.: Improving riverine constituent concentration and flux estimation by accounting for antecedent discharge conditions, Journal of Hydrology, 547, 387-402, https://doi.org/10.1016/j.jhydrol.2016.12.052, 2017.

Zhang, Q.: Synthesis of nutrient and sediment export patterns in the Chesapeake Bay watershed: Complex and non-stationary concentration-discharge relationships, Science of The Total Environment, 618, 1268-1283, https://doi.org/10.1016/j.scitotenv.2017.09.221, 2018.

Zhao, T., Schepen, A., and Wang, Q. J.: Ensemble forecasting of sub-seasonal to seasonal streamflow by a Bayesian joint probability modelling approach, Journal of Hydrology, 541, 839-849, https://doi.org/10.1016/j.jhydrol.2016.07.040, 2016.

Zhou, T., Wu, J., and Peng, S.: Assessing the effects of landscape pattern on river water quality at multiple scales: A case study of the Dongjiang River watershed, China, Ecological Indicators, 23, 166-175, https://doi.org/10.1016/j.ecolind.2012.03.013, 2012.

Formatted: Footer, Line spacing: single, Tab stops: Not at 7.62 cm + 15.24 cm

Tables

Table 1. ~~Comparison of model performance for all records~~Key factors affecting the spatial and ~~only the above-LOR records~~temporal variability for each ~~constituent~~of six constituents, as identified in Lintern et al. (2018) and Guo et al. (2019b), respectively.

Constituent	~~Above-LOR records only~~Key factors that affect spatial variability	~~All records~~Key factors that affect
TSS	Hottest month maximum temperature	Same-day streamflow
	Percentage area covered by grass	7-day antecedent streamflow
	Percentage area covered by shrub	Water temperature
	Percentage cropping area	Soil moisture root
	Maximum elevation	Soil moisture deep
	Dam storage	
	Percentage clay area	
TP	Erosivity	Same-day streamflow
	Percentage area covered by grass	30-day antecedent streamflow
	Percentage area covered by shrub	Water temperature
	Percentage area made up of roads	Soil moisture root
	Percentage cropping area	Soil moisture deep
	Average soil TP content	
FRP	Percentage area covered by shrub	Same-day streamflow
	Percentage cropping area	Water temperature
	Catchment area	Soil moisture deep
	Average soil TP content	
	Mean channel slope	
TKN	Percentage clay area	Same-day streamflow
	Warmest quarter mean temperature	30-day antecedent streamflow
	Coldest quarter rainfall	NDVI
	Percentage cropping area	Water temperature
	Percentage pasture area	Soil moisture root
	Average soil TP content	Soil moisture deep
NO_x	Annual radiation	Same-day streamflow
	Warm quarter rainfall	30-day antecedent streamflow
	Hottest month maximum temperature	NDVI
	Average soil TP content	Water temperature
	Mean channel slope	Soil moisture root
		Soil moisture deep
EC	Annual radiation	Same-day streamflow
	Annual rainfall	14-day antecedent streamflow
	Wettest quarter rain	Water temperature
	Hottest month maximum temperature	Soil moisture root
	Percentage agriculture area	Soil moisture deep
	Percentage cropping area	
	Percentage area covered by shrub	
	Average soil TN content	

Table 2. The key catchment landscape characteristics that are related to the varying relationships of water quality and same-day streamflow across space, which were selected as the two predictors for the streamflow effect in our model. Two characteristics were selected to summary the variability of streamflow effects across space for each constituent, see Section 2.3 for details of the selection method. The corresponding Spearman's correlation (R, at p<0.05) between the effect of streamflow and each catchment characteristic is presented.

Constituent	Key factors that affect spatial variability in temporal effects	Spearman's ρ ($p<0.05$)
TSS	Annual rainfall	0.722
	Hottest month maximum temperature	-0.575
TP	Annual rainfall	0.695
	Percentage area used for cropping	-0.556
FRP	Percentage agriculture area	0.392
	Percentage area underlain by mixed igneous bedrock	0.314
TKN	Annual rainfall	0.713
	Hottest month maximum temperature	-0.618
NO_x	Total storage capacity of dams in catchment	-0.493

Formatted: Footer, Line spacing: single, Tab stops: Not at 7.62 cm + 15.24 cm

	Mean soil TN content	0.458
EC	Percentage area covered by grassland	-0.347
	Percentage area covered by woodland	-0.317

Table 3. Comparison of model performance for all records and only the above-DL records for each constituent.

Constituent	Above-DL records only	All records
TSS	0.225	0.397
TP	0.433	0.445
FRP	-1.920	0.199
TKN	0.658	0.630
NO_x	0.216	0.382
EC	0.907	0.886

Table 4. Model ability to capture observed water quality trends across all monitoring sites for each constituent. The percentages of sites where observed positive and negative trends are captured by the model are presented separately. Values in brackets indicate numbers of sites where corresponding positive or negative trends are observed. For detailed estimation of these percentages please refer to Sect. 2.2.

Constituent	% positive trends captured	% negative trends captured
TSS	33.3 (12)	85.0 (20)
TP	82.1 (28)	16.7 (12)
FRP	47.1 (17)	55.6 (9)
TKN	81.1 (37)	40.0 (10)
NO_x	68.6 (35)	66.7 (27)
EC	82.6 (23)	77.3 (22)

Table 5. Comparison of model performances (as NSE) of the full model (Column 2) and the ~~five~~50 partial models (Columns 3 to ~~7~~5) with each calibrated to 80% randomly selected monitoring sites. ~~In~~ Columns 3 to ~~7~~5 summarize the mean, minimum and maximum NSE values across the 50 runs, where for each constituent, the top row showing calibration performance and the bottom row showing the validation performance (i.e. at the 20% sites that were not used for calibration).

Constituent	Full model	~~80% sites split 1~~50 CV mean	~~80% sites split 2~~50 CV min	~~80% sites split 3~~50 CV max	~~80% sites split 4~~	~~80% sites split 5~~
TSS	0.~~397~~225	0.~~406~~413	0.~~431~~376	0.~~390~~439	~~0.428~~	~~0.423~~
		0.~~348~~382	0.~~441~~292	0.~~443~~513	~~0.446~~	~~0.434~~
TP	0.~~445~~433	0.~~440~~461	0.~~422~~427	0.~~440~~501	~~0.472~~	~~0.456~~
		0.~~386~~411	0.~~444~~151	0.~~454~~575	~~0.449~~	~~0.444~~
FRP	~~0.199~~-1.92	0.~~141~~168	0.~~219~~067	0.~~244~~232	~~0.216~~	~~0.177~~
		0.~~041~~129	- 0.~~350~~078	0.~~337~~272	~~0.356~~	~~0.344~~
TKN ~~0.630~~ ~~0.664~~ ~~0.643~~ ~~0.630~~	0.658	0.~~669~~654	0.622	0.670		
		0.~~639~~622	0.~~589~~468	0.~~581~~691	~~0.584~~	~~0.587~~
NO_x	0.~~382~~216	0.~~410~~453	0.~~464~~414	0.~~438~~489	~~0.476~~	~~0.466~~
		0.~~419~~397	0.~~593~~258	0.~~603~~563	~~0.597~~	~~0.597~~
EC	0.~~886~~907	0.~~895~~893	0.~~894~~882	0.~~875~~903	~~0.900~~	~~0.892~~
		0.~~796~~875	0.~~828~~809	0.~~837~~924	~~0.828~~	~~0.826~~

Deleted Cells
Deleted Cells
Deleted Cells
Deleted Cells
Deleted Cells
Deleted Cells
Deleted Cells
Deleted Cells
Inserted Cells
Inserted Cells
Deleted Cells
Deleted Cells
Formatted: Footer, Line spacing: single, Tab stops: Not at 7.62 cm + 15.24 cm

Table 3. Comparison of model performances (as NSE) of the full model and the three models that were calibrated to the pre-drought (1994-1996), drought (1997-2009) and the post-drought (2010-2014) periods. For each of the models, the calibration performance is shown on the top row and the validation performance (i.e. over the periods that were not used for calibration) is shown on the bottom row.

Constituent	Full model	Pre-drought calibration	During drought calibration	Post-drought calibration
TSS	0.~~397~~225	0.495	0.399	0.499
		0.208	0.402	0.390
TP	0.~~445~~433	0.477	0.438	0.525
		0.421	0.474	0.411
FRP	~~0.199~~-1.92	-1.336	0.187	0.204
		-1.406	0.197	0.024
TKN	0.~~630~~658	0.649	0.650	0.711
		0.566	0.648	0.610
NOx	0.~~382~~216	0.443	0.426	0.509
		0.394	0.471	0.393
EC	0.~~886~~907	0.854	0.901	0.901
		0.887	0.873	0.884

Formatted: Footer, Line spacing: single, Tab stops: Not at 7.62 cm + 15.24 cm

Supplementary Materials

Table S1. Data sources of the potential spatial predictors for water quality (i.e. catchment characteristics). See Lintern et al. (~~20182~~2018b) for details.

	Catchment characteristic	Data Source
Climate	Average annual radiation (MJ~~/m2/day~~ $m^{-2}day^{-1}$)	(Geoscience Australia, 2011)
	Average temperature ~~(degrees Celsius~~(°C)	(Geoscience Australia, 2011)
	Average temperature of warmest quarter ~~(degrees Celsius~~(°C)	(Geoscience Australia, 2011)
	Average temperature of coldest quarter ~~(degrees Celsius~~(°C)	(Geoscience Australia, 2011)
	Maximum temperature of hottest month ~~(degrees Celsius~~(°C)	(Geoscience Australia, 2011)
	Minimum temperature of coldest month ~~(degrees Celsius~~(°C)	(Geoscience Australia, 2011)
	Annual average rainfall (mm)	(Geoscience Australia, 2011)
	Average rainfall of the wettest quarter (mm)	(Geoscience Australia, 2011)
	Average rainfall of the driest quarter (mm)	(Geoscience Australia, 2011)
	Average rainfall of the coldest quarter (mm)	(Geoscience Australia, 2011)
	Average rainfall of the warmest quarter (mm)	(Geoscience Australia, 2011)
	Annual average catchment rainfall erosivity (MJ mm~~/ha hr yr~~$^{-1}$$ha^{-1}hr^{-1}yr^{-1}$)	(Geoscience Australia, 2011)
Hydrology	Average annual runoff (mm)	(Geoscience Australia, 2011)
	Average of average daily flow (ML~~/~~ d^{-1})	Calculated using instantaneous flows from DELWP (2016)
	Standard deviation of average daily flow (ML~~/~~ d^{-1})	Calculated using instantaneous flows from DELWP (2016)
	Pereniality of runoff (%) (proportion of "contribution to mean annual discharge by the driest six months of the year" (Geoscience Australia, 2011))	(Geoscience Australia, 2011)
	Mean number of days where there is no flow annually (days~~/~~ $year^{-1}$)	Calculated using daily flows from DELWP (2016)
	Mean 7-day low flow (ML~~/~~ d^{-1})	Calculated using instantaneous flows from DELWP (2016)
	Mean Base Flow Index	Calculated using method outlined in Grayson et al. (1996)
	Maximum distance upstream to dam wall or reservoir (km)	(Geoscience Australia, 2011)
	Area of catchment comprised of farm dams (%)	(Department of Environment Land Water and Planning Victoria, 2016)
	Total storage capacity of dams in catchment ~~normalised~~normalized to average daily flow (ML~~/~~ ML~~/d~~$^{-1}d^{-1}$)	(Geoscience Australia, 2004)
Land use	Area of catchment ~~urbanised~~urbanized (%)	(Bureau of Rural Sciences, 2010)
	Area of catchment made up of roads (%)	(Bureau of Rural Sciences, 2010)
	Area of catchment used for horticulture (%)	(Bureau of Rural Sciences, 2010)
	Area of catchment used for agriculture (%)[1]	(Bureau of Rural Sciences, 2010)
	Area of catchment used for pastures (grazing) (%)	(Bureau of Rural Sciences, 2010)
	Area of catchment used for cropping (%)[2]	(Bureau of Rural Sciences, 2010)
Land cover	Mean width of vegetated riparian zone (m)	(Department of Environment Land Water and Planning, 2014)
	Average fragmentation of riparian zone (%)	(Department of Environment Land Water and Planning, 2014)

	Area of catchment covered with grass (%)[3]	(Geoscience Australia, 2011)
	Area of catchment covered with forest (%)[4]	(Geoscience Australia, 2011)
	Area of catchment covered with shrubs (%)[5]	(Geoscience Australia, 2011)
	Area of catchment covered with woodland (%)[6]	(Geoscience Australia, 2011)
	Area of catchment bare (%)	(Geoscience Australia, 2011)
Soil type and geology	Area of catchment underlain by unconsolidated bedrock (%)	(Geoscience Australia, 2011)
	Area of catchment underlain by igneous bedrock (%)	(Geoscience Australia, 2011)
	Area of catchment underlain by sedimentary bedrock (%)	(Geoscience Australia, 2011)
	Area of catchment underlain by mixed igneous and sedimentary bedrock (%)	(Geoscience Australia, 2011)
	Average soil TP content (mg/kg^{-1})	(Terrestrial Ecosystem Research Network, 2016)
	Average soil TN content (mg/kg^{-1})	(Terrestrial Ecosystem Research Network, 2016)
	Average soil clay content (%)	(Terrestrial Ecosystem Research Network, 2016)
	Area of catchment with saline aquifers (%)	(Department of Agriculture and Water Resources, 2013)
Topography	Catchment area ($km^{2)}$	(Geoscience Australia, 2011)
	Mean catchment elevation (m)	(Geoscience Australia, 2011)
	Maximum catchment elevation (m)	(Geoscience Australia, 2011)
	Area of catchment made up of valley bottoms (%)	(Geoscience Australia, 2011)
	Total catchment length (km)	(Geoscience Australia, 2011)
	Mean catchment slope (%)	(Geoscience Australia, 2011)
	Mean channel slope (%)	Calculated using BOM (2012)

1. Agricultural activities include all primary production activities including plantation forests, grazing pastures, cropping and horticulture. This includes both dryland and irrigation agricultural activities.

2. Cropping refers to the production of commodities such as cereals, beverage and spice crops, hay, oilseeds, sugar, cotton, alkaloid poppies and pulses.

3. Grass refers to grasslands with tussock, hummock, reeds/rushes.

4. Forest refers to rainforests, Eucalypt forests, mangroves and low closed forests (e.g., Acacia, Melaleuca or Banksia species). Areas with high density of vegetation (>30% cover) and tall trees (>10 m).

5. Shrubs refers to open and dry woodlands and shrublands with hummock or tussock grass, Melaleuca shrublands, lignum shrublands, saltbush and chenopods. Areas with vegetation <2 m tall.

6. Woodlands refer to areas with medium trees (<10 m) at medium density (<30% cover).

Table S2. Data sources of the potential temporal predictors for water quality. See Guo et al. (2019) for details.

Data		**Source**
Daily rainfall (mm)		Australian Water Availability Project (AWAP) (Raupach et al., 2009, 2012) Available from: http://www.csiro.au/awap; http://www.bom.gov.au/jsp/awap/index.jsp
Daily average temperature (°C)		
Daily actual ET (mm)		Australian Water Resources Assessment (Frost et al., 2016) Available from: http://www.bom.gov.au/water/landscape
Daily average root zone soil moisture		
Daily average deep soil moisture		
Monthly NDVI	January 1994 – December 1999	Advanced Very High Resolution Radiometer product (AVHRR) (Eidenshink, 1992) Available from: https://earthdata.nasa.gov/
	January 2000 – December 2013	Moderate Resolution Imaging Spectroradiometer (MODIS); MOD13A3 (NASA LP DAAC, 2017) Available from: https://earthdata.nasa.gov/

Table S3. Log-~~Sinh~~sinh transformation parameter (a and b) values for 50 potential spatial predictors for stream water quality (i.e. catchment characteristics).

Catchment characteristics	*a*	*b*
Annual radiation (MJ $m^{-2}day^{-1}$)	3.458	2.052
Annual temperature (°C)	2.425	3.133
Annual rainfall (mm)	0.008	0.001
Erosivity (MJ $mm^{-1}ha^{-1}hr^{-1}yr^{-1}$)	0.030	0.000
Driest quarter rain (mm)	0.099	0.003
Wettest quarter rain (mm)	0.002	0.003
Warmest quarter rainfall (mm)	0.039	0.005
Coldest quarter rainfall (mm)	0.001	0.001
Coldest month minimum temperature (°C)	4.999	0.000
Hottest month maximum temperature (°C)	0.000	0.002
Coldest quarter mean temperature (°C)	4.986	4.996
Warmest quarter mean temperature (°C)	3.805	2.193
Average of average daily flow (ML~~/~~ d^{-1})	0.002	0.001
Average of average daily flow (ML~~/~~ d^{-1})	0.034	0.002
Standard deviation of average daily flow (ML~~/~~ d^{-1})	0.012	0.430
Pereniality of runoff (%) (proportion of 'contribution to mean annual discharge by the driest six months of the year'	0.106	0.152
Mean number of days where there is no flow annually (days~~/~~ $year^{-1}$)	0.000	0.066
Mean 7-day low flow (ML~~/~~ d^{-1})	0.045	3.319
Mean Base Flow Index	4.896	0.000
Maximum distance upstream to dam wall or reservoir (km)	0.034	0.006
Area of catchment comprised of farm dams (%)	0.000	5.000
Total storage capacity of dams in catchment ~~normalised~~normalized to average daily flow (ML~~/~~ ML~~/d~~$^{-1}d^{-1}$)	0.003	0.002
Area of catchment ~~urbanised~~urbanized (%)	0.000	0.135
Area of catchment made up of roads (%)	0.055	0.729
Area of catchment used for agriculture (%)	4.998	4.995
Area of catchment used for pastures (grazing) (%)	0.174	0.114
Area of catchment used for cropping (%)	0.000	0.079
Area of catchment used for horticulture (%)	0.000	0.373
Mean width of vegetated riparian zone (m)	0.293	0.013
Average fragmentation of riparian zone (%)	0.174	0.132
Area of catchment covered with grass (%)	0.000	0.158
Area of catchment covered with forest (%)	0.238	0.020
Area of catchment covered with shrubs (%)	0.000	0.403
Area of catchment covered with woodland (%)	0.002	0.108
Area of catchment bare (%)	0.000	5.000
Area of catchment underlain by unconsolidated bedrock (%)	0.024	0.050
Area of catchment underlain by igneous bedrock (%)	0.034	0.068
Area of catchment underlain by sedimentary bedrock (%)	4.998	4.995
Area of catchment underlain by mixed igneous and sedimentary bedrock (%)	0.000	0.032
Average soil TP content (mg~~/~~ kg^{-1})	0.044	4.744
Average soil TN content (mg~~/~~ kg^{-1})	0.213	1.733

Average soil clay content (%)	0.000	0.021
Area of catchment with saline aquifers (%)	0.001	0.000
Catchment area (km^2)	0.177	0.001
Mean catchment elevation (m)	0.044	0.001
Area of catchment made up of valley bottoms (%)	0.002	0.074
Total catchment length (km)	0.003	0.001
Mean catchment slope (%)	0.078	0.068
Mean channel slope (%)	0.029	4.899
Average soil clay content (%)	0.103	0.040

Table S4. Box-Cox transformation parameter (~~*lambda*~~*λ*) values for the six water quality constituents and the nineteen potential temporal predictors. Values in bracket show the standard deviation of individual site-level *λ*.

Water Quality Constituent	***~~lambda~~λ***
TSS	-0.249 (0.287)
TP	-0.058 (0.181)
FRP	-0.836 (1.056)
TKN	0.141 (0.342)
NO_x	0.107 (0.305)
EC	-0.024 (0.921)
Temporal predictors	***~~lambda~~λ***
Rainfall (mm)	~~-~~0.~~243~~106 (0.041)
Rainfall on previous day (mm)	0.~~107~~108 (0.028)
Averaged rainfall over previous 3 days (mm)	0.~~108~~157 (0.022)
Averaged rainfall over previous 7 days (mm)	0.~~157~~220 (0.025)
Averaged rainfall over previous 14 days (mm)	0.~~220~~192 (0.046)
Averaged rainfall over previous 30 days (mm)	0.~~193~~116 (0.075)
Streamflow (mm d^{-1})	~~-~~0.~~115~~015 (0.225)
Streamflow on previous day (mm d^{-1})	-0.~~014~~027 (0.207)
Averaged Streamflow over previous 3 days (mm d^{-1})	-0.~~028~~032 (0.207)
Averaged Streamflow over previous 7 days (mm d^{-1})	-0.~~033~~030 (0.2)
Averaged Streamflow over previous 14 days (mm d^{-1})	-0.~~032~~021 (0.198)
Averaged Streamflow over previous 30 days (mm d^{-1})	-0.~~023~~004 (0.195)
Dry spell length in the past 14 days (days)	~~-~~0.~~005~~257 (0.089)
NDVI for the month	~~0.258~~3.715 (1.998)
Water temperature (°C)	~~3.712~~0.357 (0.269)
Air temperature (°C)	0.~~234~~231 (0.244)
Evaporation (mm)	0.~~021~~019 (0.13)
Root zone soil moisture (%)	0.~~094~~913 (0.648)
Deep soil moisture (%)	0.~~910~~357 (0.269)

Table S5. ~~Key~~The key temporal predictor for each water quality constituent, and the two key factors ~~affecting~~that are mostly closely related to the spatial ~~variability for each~~ variation of ~~six constituents~~each temporal predictor (see Section 2.3 in the main text~~, and~~ for detailed selection process). The corresponding Spearman's correlation coefficients (R) are also ~~Lintern et al. (2018)).~~shown in the last column.

Constituent	**Key factors that affect temporal variability**	**Key factors that affect spatial variability in temporal effects**	**Spearman' R**
TSS	Same-day streamflow	Annual rainfall	0.722
		Hottest month maximum temperature	-0.575

Inserted Cells

Inserted Cells

	7-day antecedent streamflow	Annual runoff	-0.536
		Mean elevation	-0.465
	Water temperature	Daily flow standard deviation	0.204
		Total catchment length	0.177
	Soil moisture root	Percentage area with saline aquifers	0.507
		Hottest month maximum temperature	0.495
	Soil moisture deep	Maximum distance upstream to dam wall or reservoir	-0.275
		~~Hottest month maximum temperature~~ Percentage area covered by ~~grass~~ ~~Percentage area covered by shrub~~ grassland~~Percentage cropping area~~ ~~Maximum elevation~~ ~~Dam storage~~ ~~Percentage clay area~~	-0.24
TP	Same-day streamflow	Annual rainfall	0.695
		Hottest month maximum temperature	-0.556
	30-day antecedent streamflow	Erosivity	-0.675
		Percentage cropping area	0.626
	Water temperature	Percentage agricultural area	0.382
		Percentage area used for roads	0.274
	Soil moisture root	Percentage pasture area	0.564
		Hottest month maximum temperature	0.557
	Soil moisture deep	Percentage area underlain by mixed igneous bedrock	-0.23
		Maximum distance upstream to dam wall or reservoir	-0.21
FRP	Same-day streamflow	Percentage agriculture area	0.392
		Percentage area underlain by mixed igneous bedrock	0.314
	Water temperature	Total catchment length	-0.28
		Coldest quarter mean temperature	0.232
	Soil moisture deep	Percentage area used for roads	-0.21
		Percentage aea covered by woodland	0.204
TKN	Same-day streamflow	Annual rainfall	0.713
		Hottest month maximum temperature	-0.618
	30-day antecedent streamflow	Erosivity	-0.823
		Percentage cropping area	0.694
	NDVI	Mean_7daylowflow	0.42
		Maximum distance upstream to dam wall or reservoir	-0.366
	Water temperature	Coldest quarter rainfall	-0.386
		Maximum distance upstream to dam wall or reservoir	0.374
	Soil moisture root	Warmest quarter mean temperature	0.6
		Percentage pasture area	0.588
	Soil moisture deep	Hottest month maximum temperature	-0.274
		Warmest quarter mean temperature	-0.269

Inserted Cells

NOx	Same-day streamflow	Total storage capacity of dams in catchment	-0.493
		Mean soil TN content	0.458
	30-day antecedent streamflow	Coldest quarter rainfall	-0.413
		Hottest month maximum temperature	0.396
	NDVI	~~Erosivity~~ Percentage area covered by ~~grass~~ ~~Percentage area covered by shrub~~ ~~Percentage area made up of roads~~ woodland~~Percentage cropping area~~ ~~Average soil TP content~~	-0.442
		Maximum elevation	-0.428
	Water temperature	Percentage area underlain by mixed igneous bedrock	0.266
		Percentage urbanized area	-0.2
	Soil moisture root	Annual temperature	0.44
		Warmest quarter average temperature	0.338
	Soil moisture deep	Percentage horticulture area	0.341
		Wettest quarter rainfall	-0.334
~~FRP~~EC	Same-day streamflow	Percentage area covered by ~~shrub~~ ~~Percentage cropping area~~ ~~Catchment area~~ ~~Average soil TP content~~ ~~Mean channel slope~~grassland	-0.347
		Percentage ~~clay~~ area covered by woodland~~Warmest quarter mean temperature~~ ~~Coldest quarter rainfall~~ ~~Percentage cropping area~~ ~~Percentage pasture area~~ ~~Average soil TP content~~	-0.317
	14-day antecedent streamflow	Percentage area covered by forest	0.324
		PerForest_Ext	0.276
	Water temperature	~~Annual radiation~~ ~~Warm quarter rainfall~~ ~~Hottest~~Coldest month ~~maximum~~minimum temperature ~~Average soil TP content~~ ~~Mean channel slope~~	-0.328
		Mean catchment slope	0.28
	Soil moisture root	Mean 7-day low flow	0.33
		Average soil TN content	0.303
	Soil moisture deep	Maximum elevation	0.366
		~~Annual radiation~~ ~~Annual rainfall~~ ~~Wettest quarter rain~~ ~~Hottest month maximum temperature~~ ~~Percentage agriculture area~~ ~~Percentage cropping area~~ Percentage area covered by ~~shrub~~	0.312

Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells

		~~Average soil TN content~~woodland	

~~Table S6. Key factors affecting~~

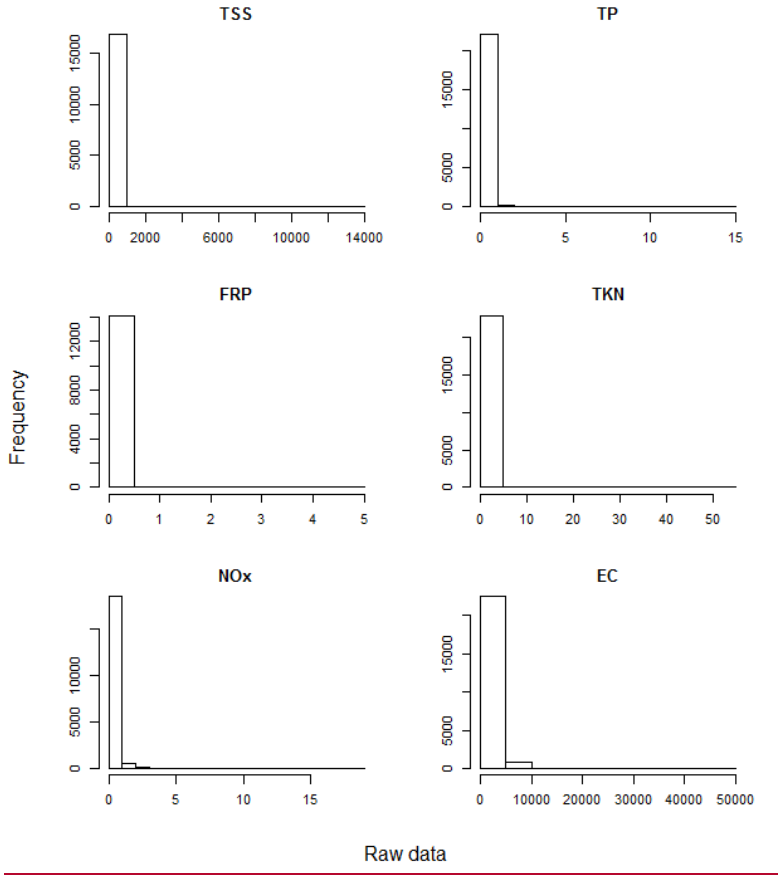


Figure S1. Distribution of the raw water quality data across all catchments. Each panel shows one constituent with only the above-DL data.

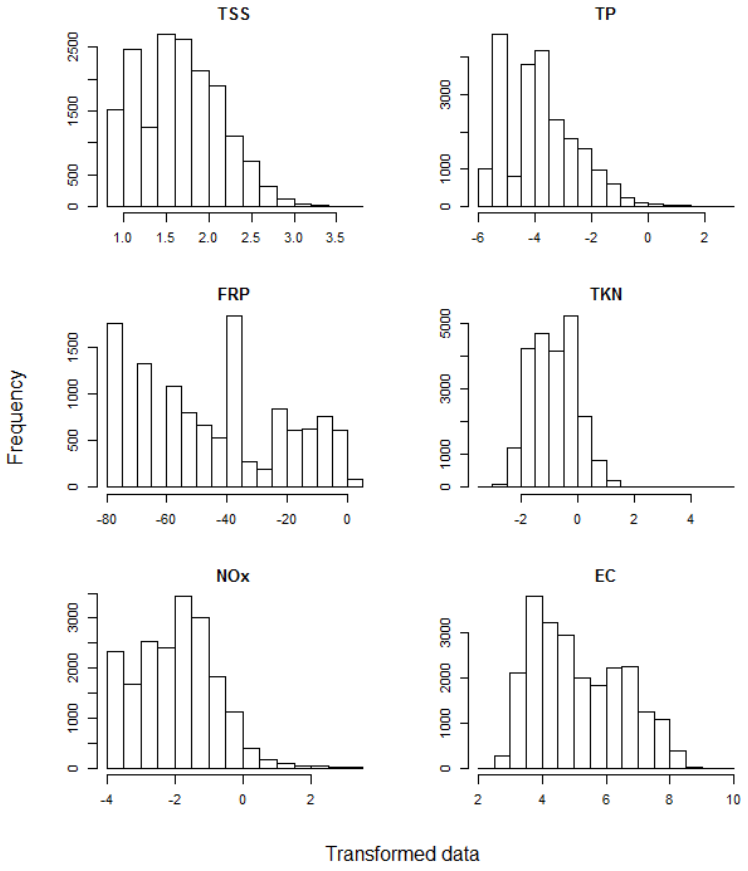


Figure S2. Distribution of the transformed water quality data across all catchments. Each panel shows one constituent with only the above-DL data.

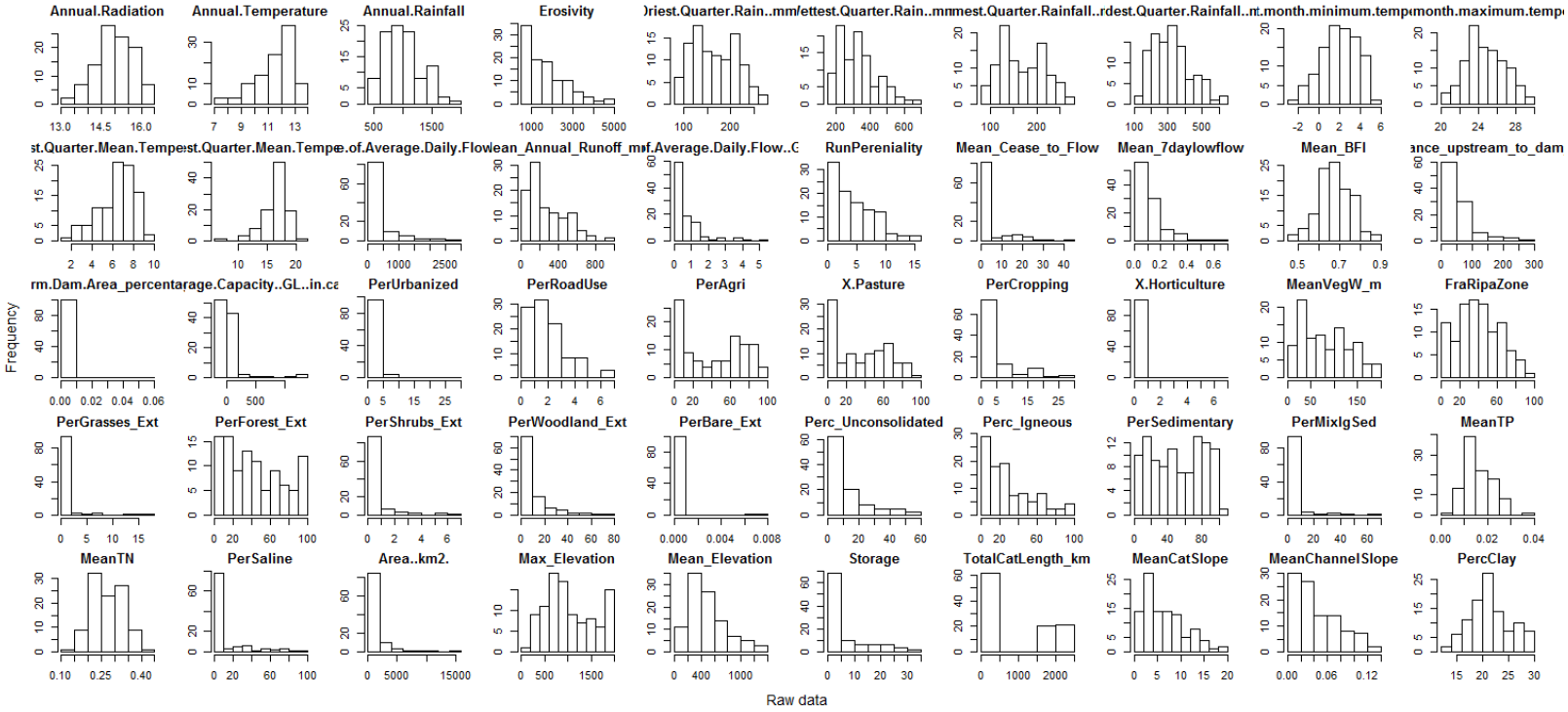


Figure S3. Distribution of the raw data for catchment characteristics included as potential spatial predictors in the model.

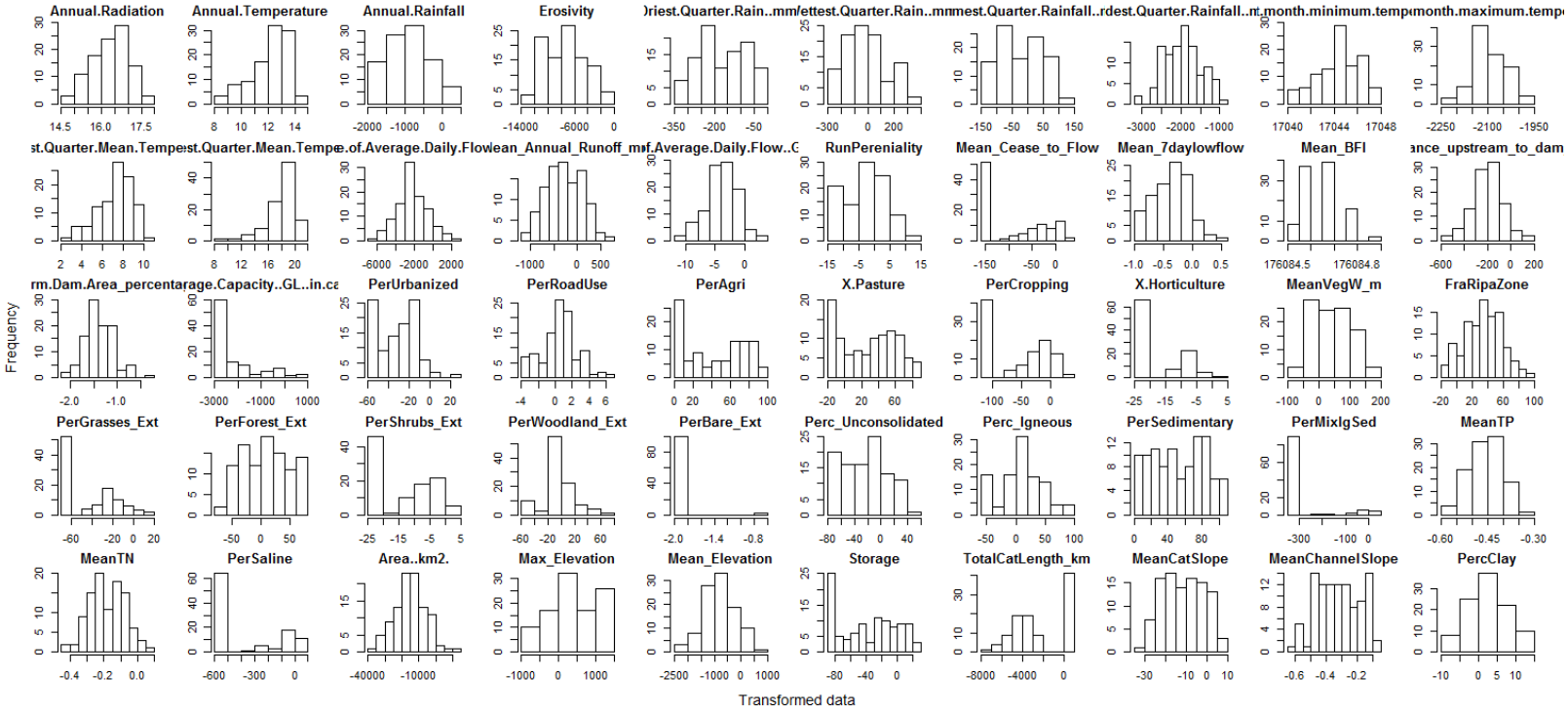


Figure S4. Distribution of the transformed data for catchment characteristics included as potential spatial predictors in the model.

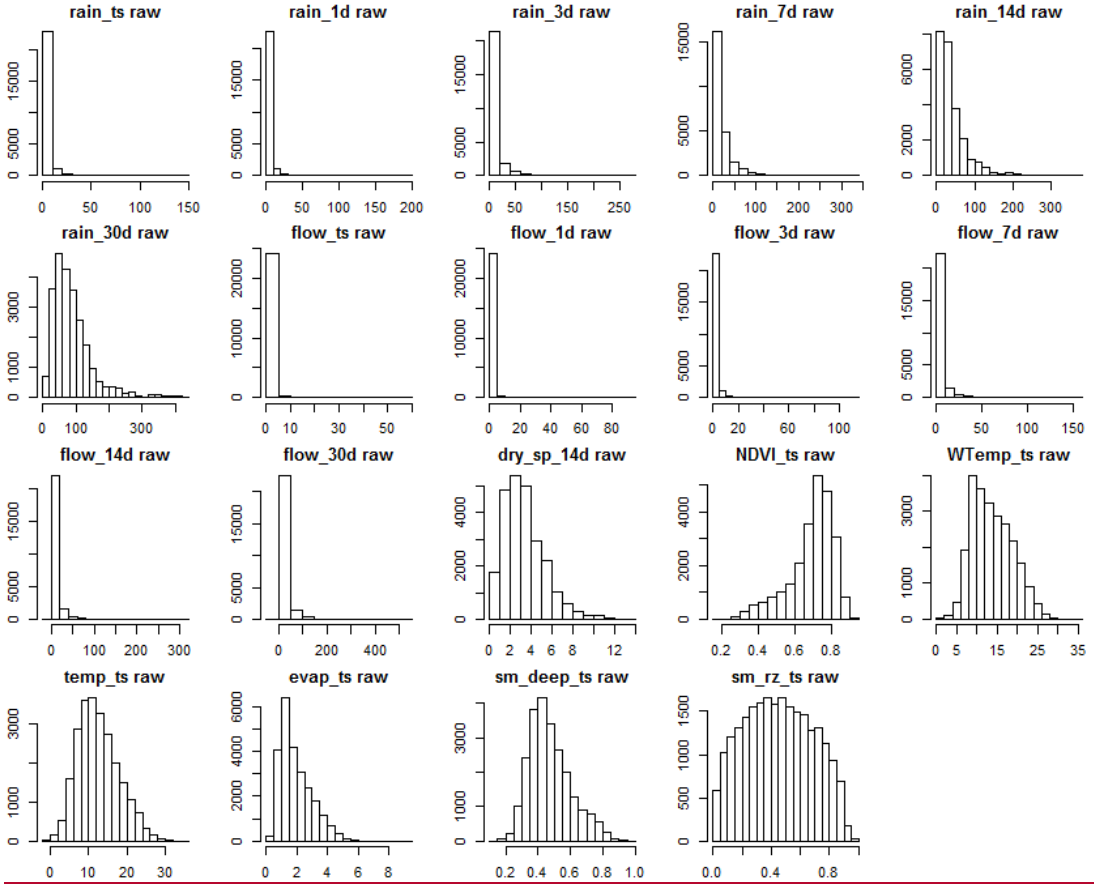


Figure S5. Distribution of the raw data for hydro-climatic and vegetation variables included as potential temporal ~~variability for~~ predictors in the model.

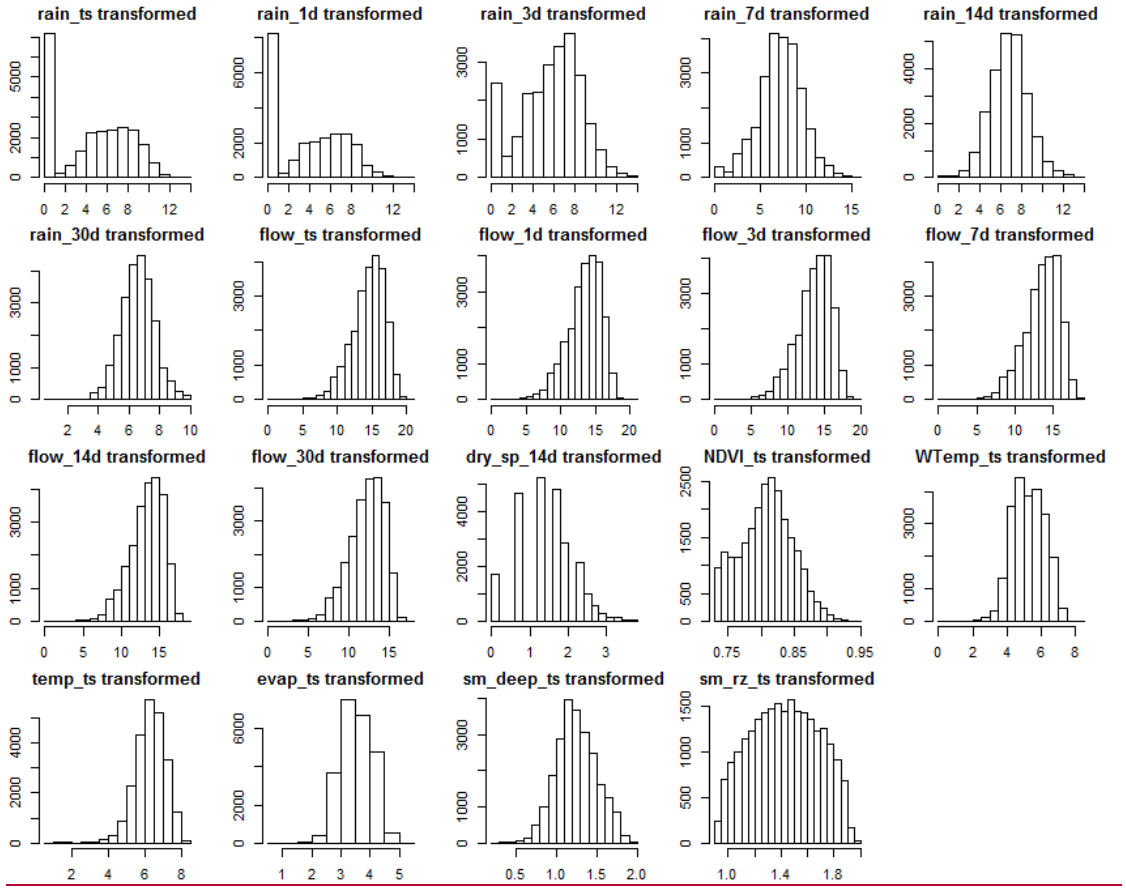


Figure S6. Distribution of the transformed data for transformed (Box-Cox) hydro-climatic and vegetation variables included as potential temporal predictors in the model.

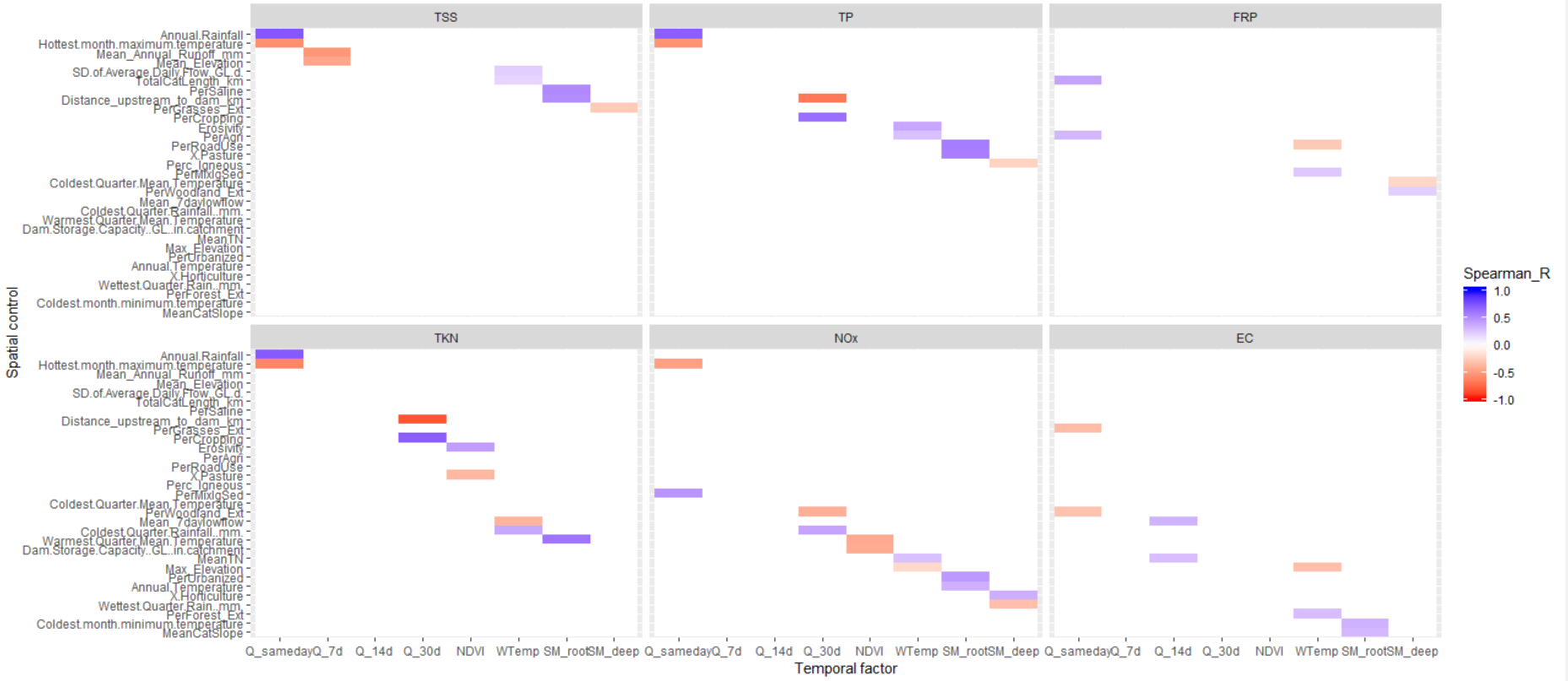


Figure S4. The two key factors that are mostly closely related to the spatial variation of each ~~of six~~temporal predictor of each water quality constituents, as highlighted in the coloured cells (see Section 2.3 in the main text~~, and also Guo et al. (2019)). The third column shows the two key catchment characteristics that affect the spatial variability in each temporal factor, which were selected by~~ for detailed selection of the two key factors). Colours indicate the corresponding Spearman's correlation ~~analyses between the coefficient values of the temporal~~ coefficients (R) from -1 (red) to 1 (blue).

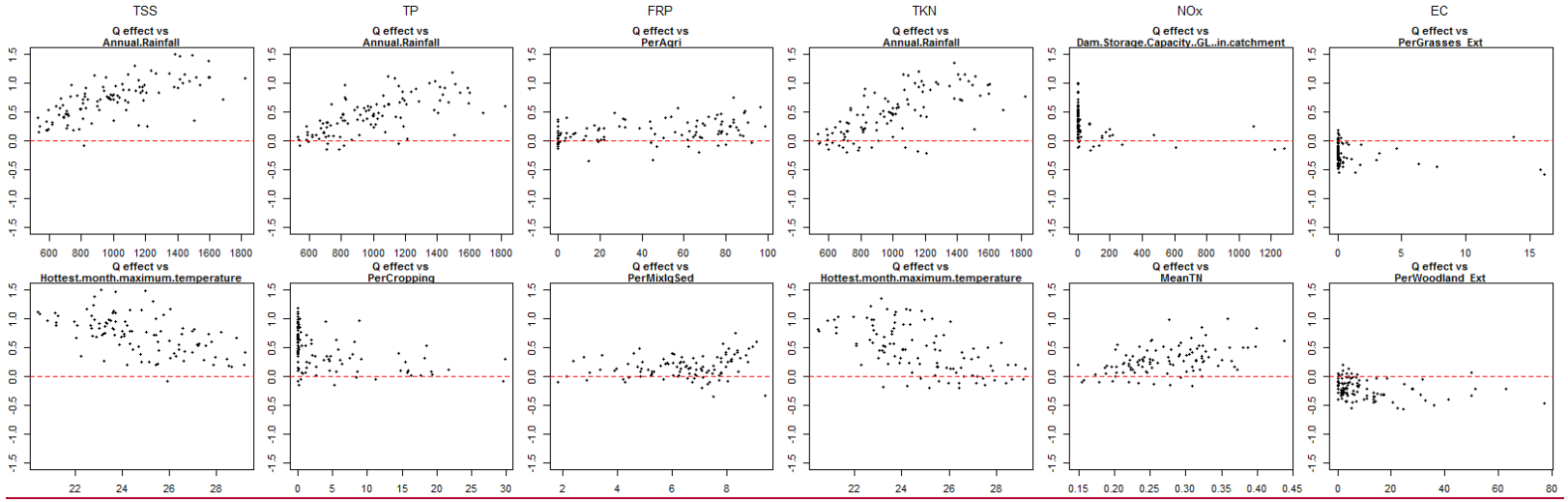


Figure S5. Effects of streamflow across catchments against the two most important catchment landscape characteristics, for each constituent (see Section 2.3 in the main text for detailed selection of the two key factors). Red dash lines indicate the zero levels, and thus differentiate positive and negative streamflow effects ~~and~~

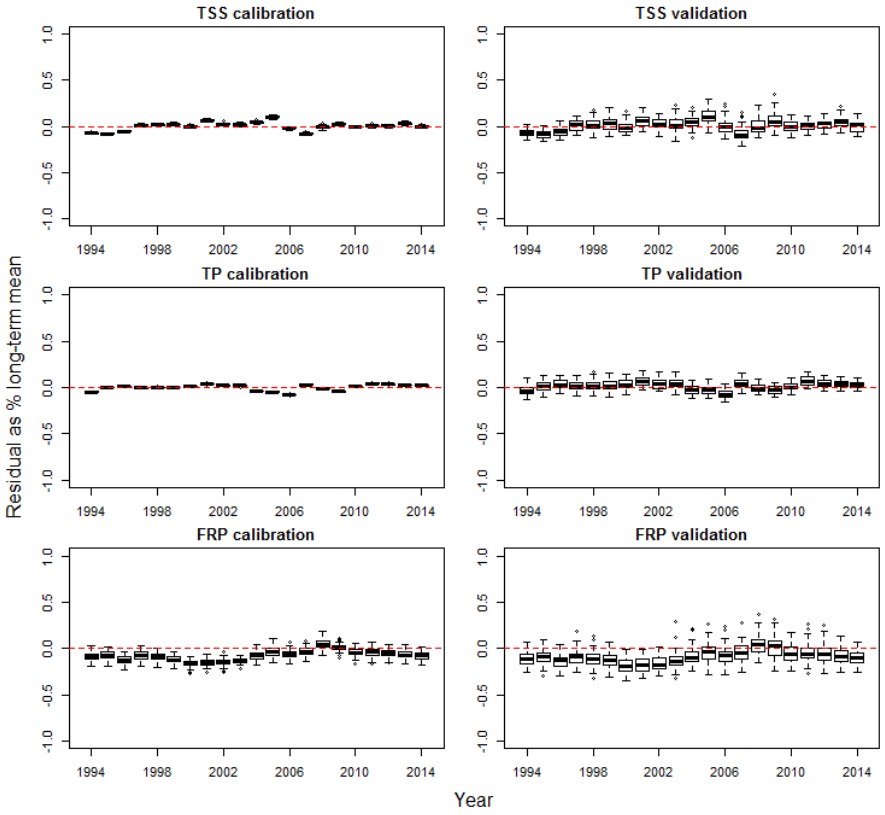
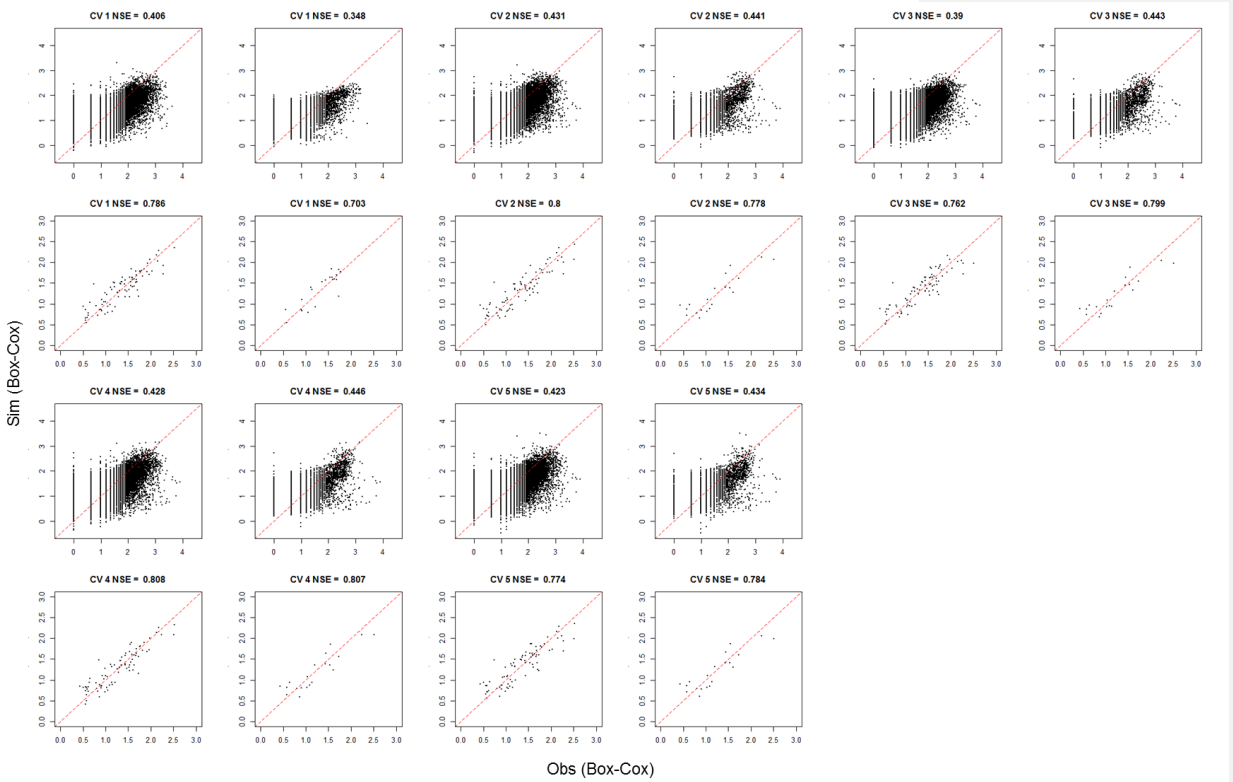


Figure S6. Annual average residuals of the ~~catchment characteristics.~~

~~Constituent~~	~~Key factors that affect temporal variability~~	~~Key factors that affect spatial variability in temporal effects~~
~~TSS~~	~~Same-day streamflow~~	~~Annual rainfall,~~ ~~Hottest month maximum temperature~~
	~~7-day antecedent streamflow~~	~~Annual runoff,~~ ~~Mean elevation~~
	~~Water temperature~~	~~Daily flow standard deviation,~~ ~~Total catchment length~~
	~~Soil moisture root~~	~~Percentage area with saline aquifers,~~ ~~Hottest month maximum temperature~~
	~~Soil moisture deep~~	~~Maximum distance upstream to dam wall or reservoir, Percentage area covered by grassland~~
~~TP~~	~~Same-day streamflow~~	~~Annual rainfall,~~ ~~Hottest month maximum temperature~~
	~~30-day antecedent streamflow~~	~~Erosivity~~ ~~Percentage cropping area~~
	~~NDVI~~	~~Mean 7-day low flow,~~ ~~Maximum distance upstream to dam wall or reservoir~~
	~~Water temperature~~	~~Coldest quarter rainfall,~~ ~~Maximum distance upstream to dam wall or reservoir~~
	~~Soil moisture root~~	~~Warmest quarter average temperature,~~ ~~Percentage pasture area~~

	Soil moisture deep	Hottest month maximum temperature, Warmest quarter average temperature
FRP	Same-day streamflow	Percentage agriculture area, Coldest quarter mean temperature
	Water temperature	Total catchment length, Coldest quarter mean temperature
	Soil moisture deep	Percentage area used for roads, Percentage aea covered by woodland
TKN	Same-day streamflow	Annual rainfall, Hottet month maximum temperature
	30-day antecedent streamflow	Erosivity, Percentage cropping area
	NDVI	Mean 7-day low flow, Maximum distance upstream to dam wall or reservoir
	Water temperature	Coldest quarter rainfall, Maximum distance upstream to dam wall or reservoir
	Soil moisture root	Warmest quarter mean temperature, Percentage pasture area
	Soil moisture deep	Hottest month maximum temperature, Warmest quarter mean temperature
NOx	Same-day streamflow	Total storage capacity of dams in catchment, Mean soil TN content
	30-day antecedent streamflow	Coldest quarter rainfall, Hottest month maximum temperature
	Water temperature	Percentage area covered by woodland, Maximum elevation
	NDVI	Percentage area underlain by mixed igneous bedrock, Percentage urbanized area
	Soil moisture root	Annual rainfall, Warmest quarter average temperature
	Soil moisture deep	Percentage horticulture area, Wettest quarter rainfall
EC	Same-day streamflow	Percentage area covered by grassland, Percentage area covered by woodland
	14-day antecedent streamflow	Mean 7-day low flow, Percentage area covered by forest
	Water temperature	Coldest month minimum temperature, Mean catchment slope
	Soil moisture root	Mean 7-day low flow, Average soil TN content
	Soil moisture deep	Maximum elevation, Percentage area covered by woodland



~~Figure S1. Fittings of the five partial~~ **models for TSS** ~~(see Section 2.4 in text for calibration/validation approaches), each within a 2x2 panel and showing the calibration and validation fittings in the left and right columns, respectively. Within each partial model, top row shows the fitting to all data whereas bottom row shows fitting to site-level mean concentrations.~~, TP and FRP, as % of long-term average. **All**

values are presented in a Box-Cox transformed ~~space and the dashed red lines indicate 1:1 (perfect fit).~~scale.

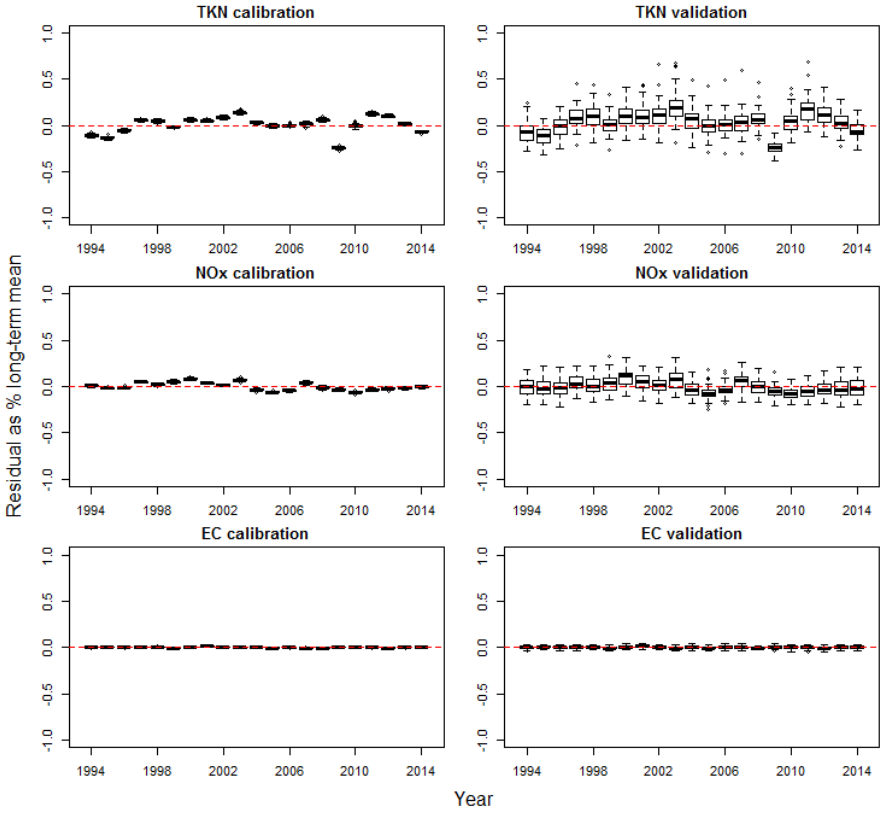
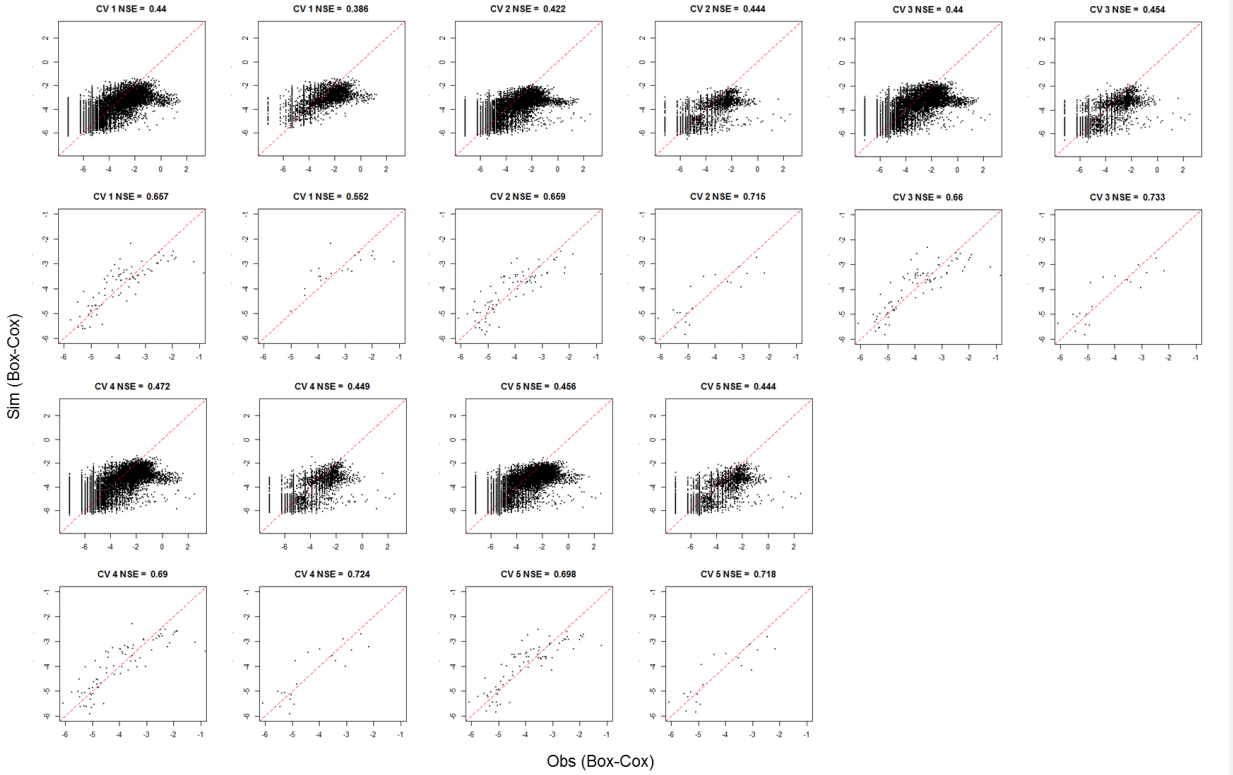


Figure S7.



Annual average residuals of the models for TKN, NOx and EC, as % of long-term average. All values are presented in a Box-Cox transformed scale.

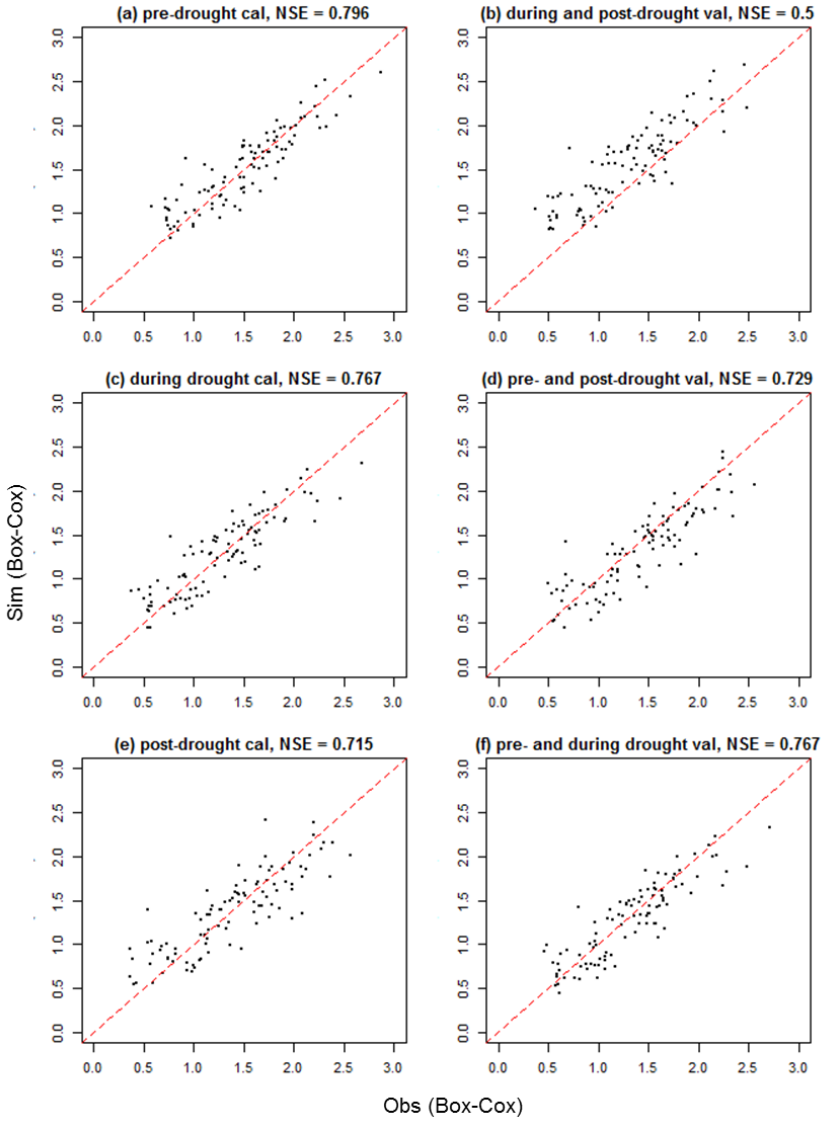
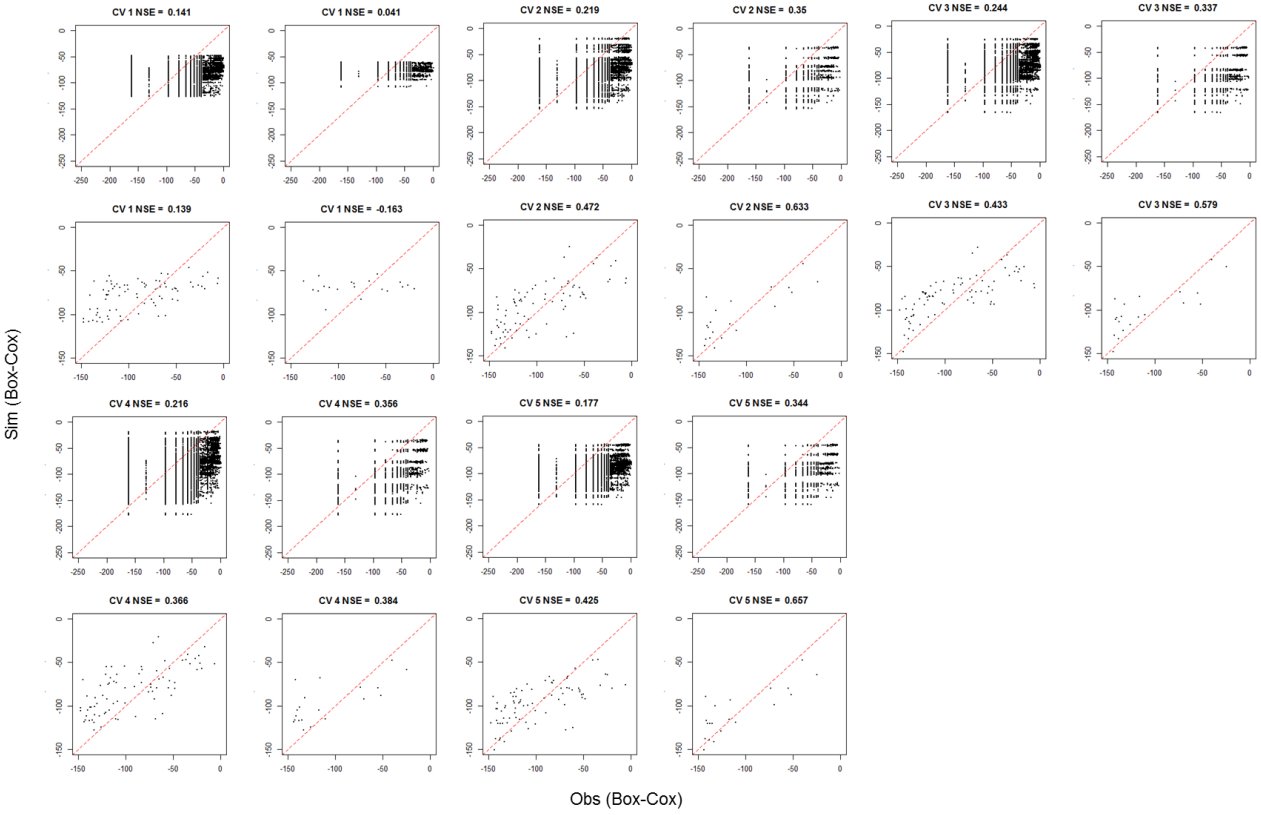
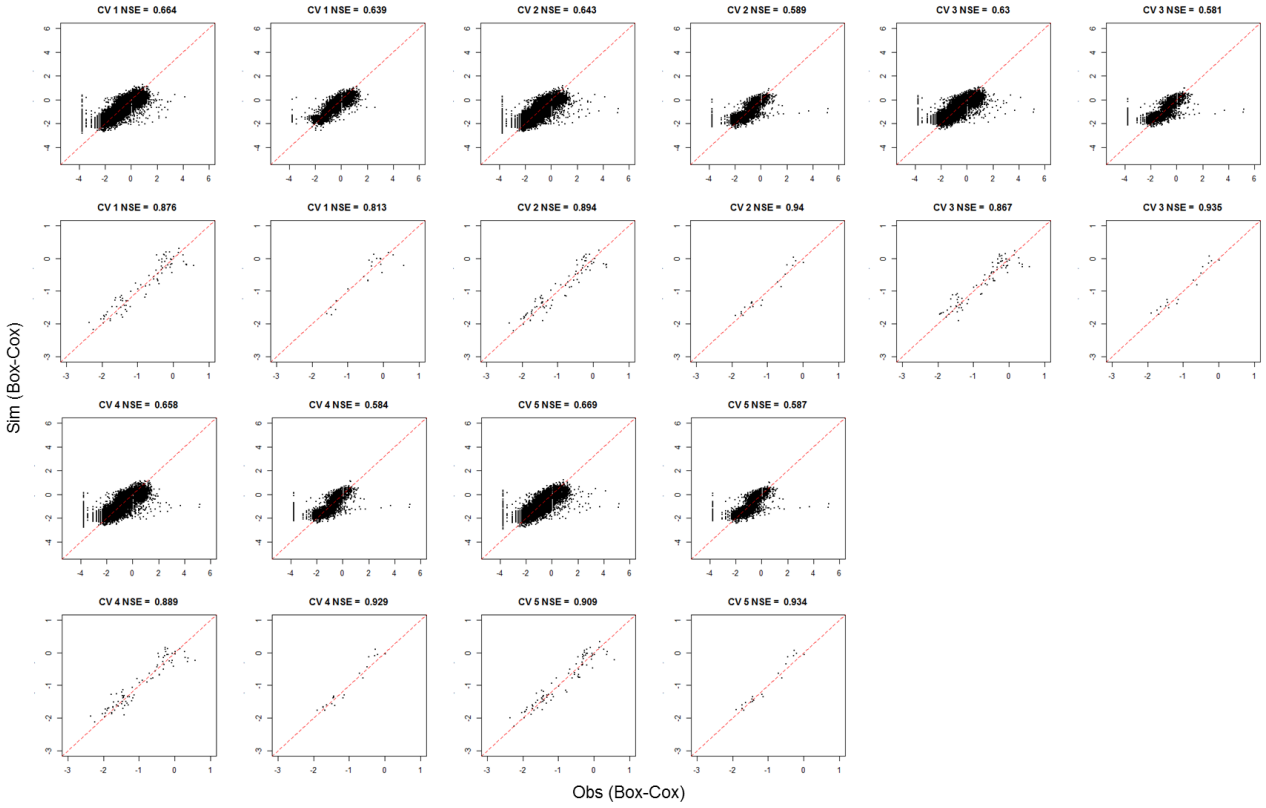

(a) pre-drought cal, NSE = 0.796
(b) during and post-drought val, NSE = 0.5
(c) during drought cal, NSE = 0.767
(d) pre- and post-drought val, NSE = 0.729
(e) post-drought cal, NSE = 0.715
(f) pre- and during drought val, NSE = 0.767
Sim (Box-Cox)
Obs (Box-Cox)
0.0
0.5
1.0
1.5
2.0
2.5
3.0

Figure S8. ~~**Figure S2. Fittings of the five partial models for TP (see Section 2.4 in text for calibration/validation approaches), each within a 2x2 panel and showing the calibration and validation fittings in the left and right columns, respectively. Within each partial model, top row shows the fitting to all data whereas bottom row shows fitting to site-level mean concentrations. All values are presented in Box-Cox transformed space and the dashed red lines indicate 1:1 (perfect fit).**~~



~~Figure S3.~~ ~~Fittings of the five partial models for FRP (see Section 2.4 in text for calibration/validation approaches), each within a 2x2 panel and showing the calibration and validation fittings in the left and right columns, respectively. Within each partial model, top row shows the fitting to all data whereas bottom row shows fitting to site-level mean concentrations. All values are presented in Box-Cox transformed space and the dashed red lines indicate 1:1 (perfect fit).~~



~~Figure S4.~~ ~~Fittings of the five partial models for TKN (see Section 2.4 in text for calibration/validation approaches), each within a 2x2 panel and showing the calibration and validation fittings in the left and right columns, respectively. Within each partial model, top row shows the fitting to all data whereas bottom row shows fitting to site-level mean concentrations. All values are presented in Box-Cox transformed space and the dashed red lines indicate 1:1 (perfect fit).~~

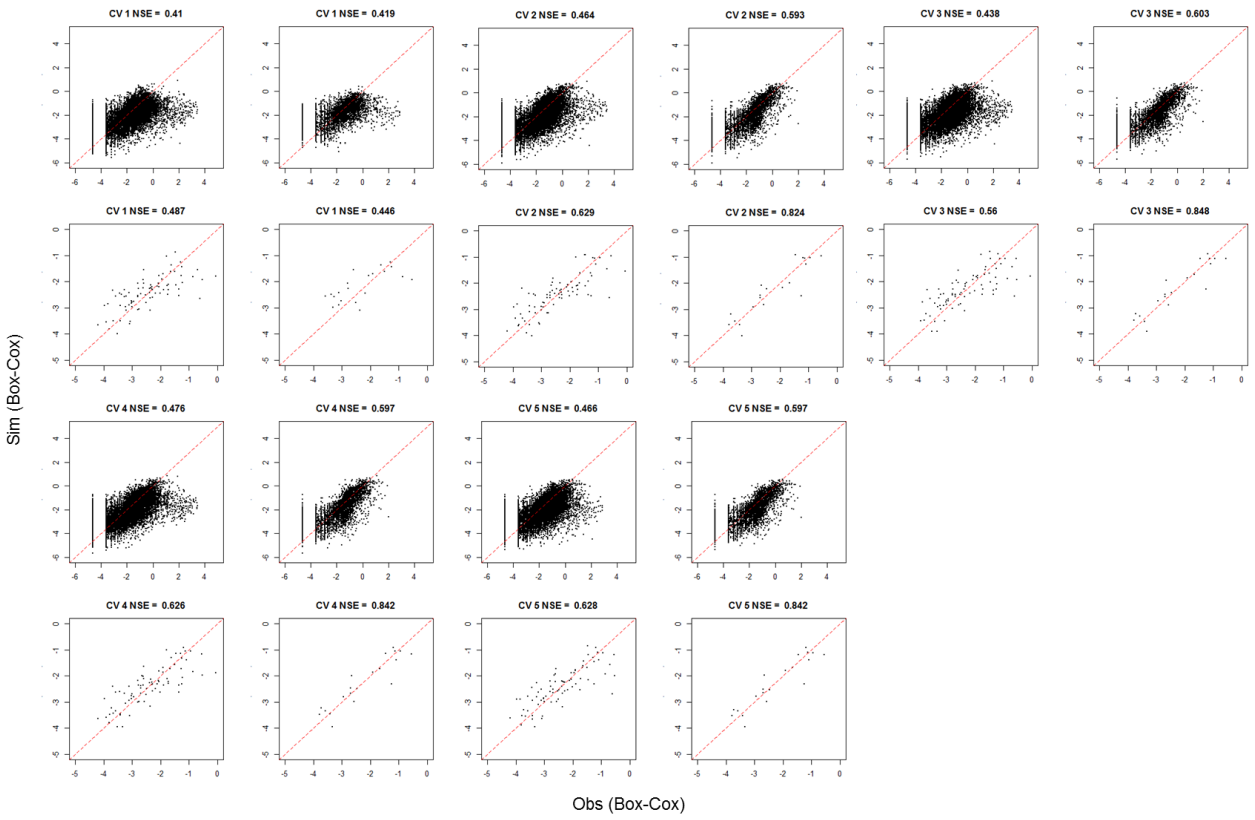
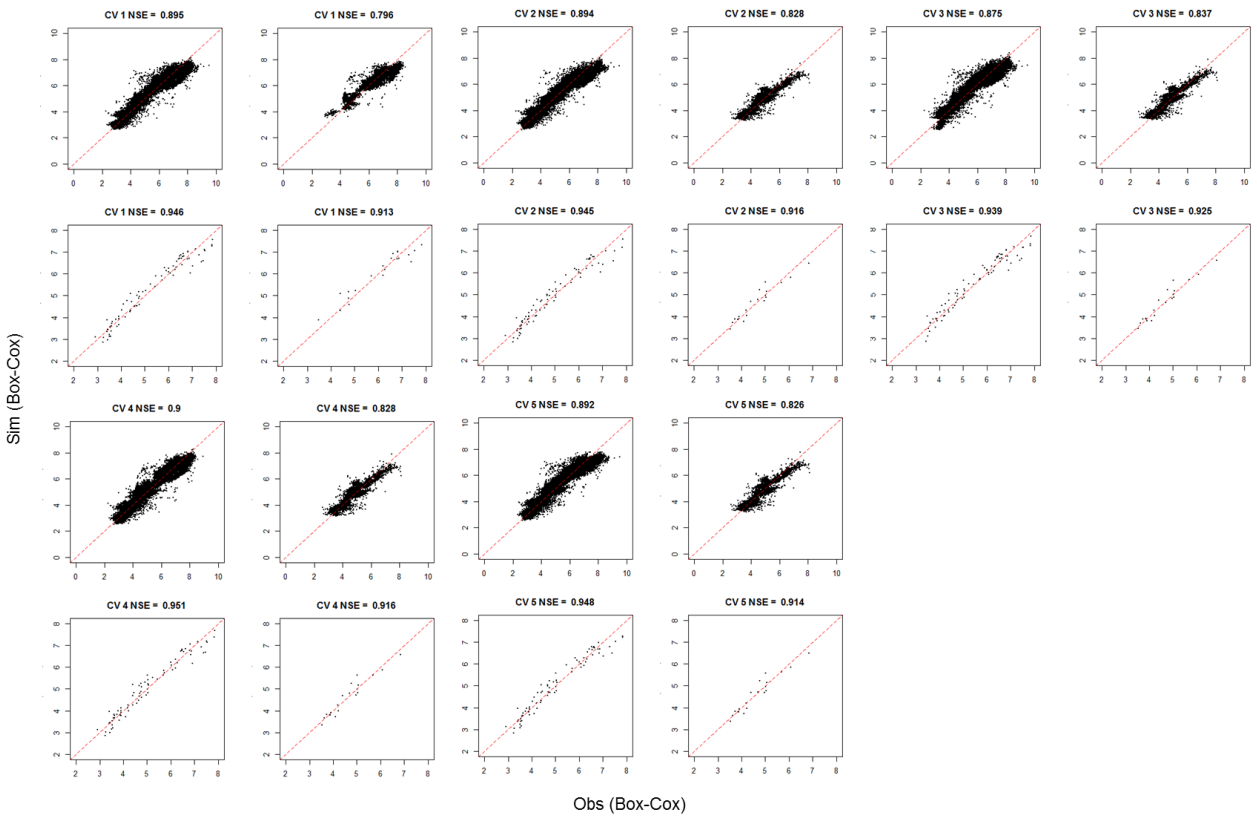
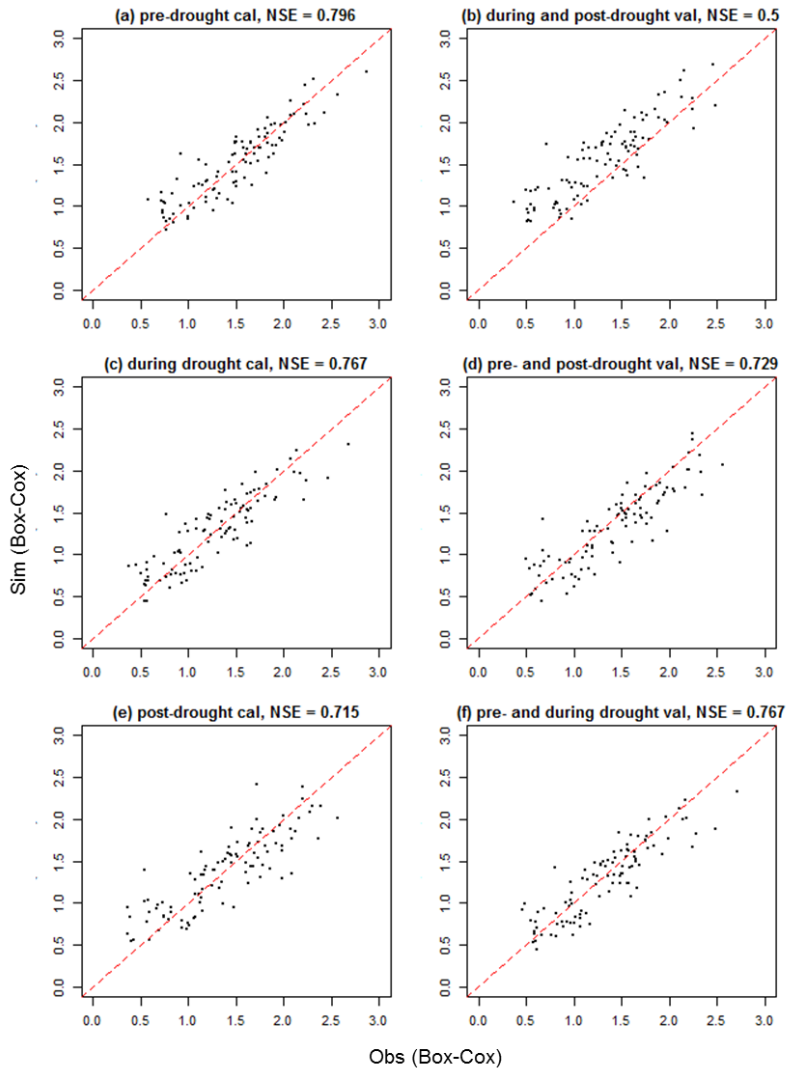


Figure S5. Fittings of the five partial models for NO_x (see Section 2.4 in text for calibration/validation approaches), each within a 2x2 panel and showing the calibration and validation fittings in the left and right columns, respectively. Within each partial model, top row shows the fitting to all data whereas bottom row shows fitting to site-level mean concentrations. All values are presented in Box-Cox transformed space and the dashed red lines indicate 1:1 (perfect fit).



~~Figure S6. Fittings of the five partial models for EC (see Section 2.4 in text for calibration/validation approaches), each within a 2x2 panel and showing the calibration and validation fittings in the left and right columns, respectively. Within each partial model, top row shows the fitting to all data whereas bottom row shows fitting to site-level mean concentrations. All values are presented in Box-Cox transformed space and the dashed red lines indicate 1:1 (perfect fit).~~



~~Figure S7.~~ Comparison of the TSS model performance, as the simulated against observed site-level mean concentrations across three different calibration/validation periods for calibrations on the pre-drought (1994-1996), drought (1997-2009) and the post-drought (2010-2014) periods, respectively, see Section 2.4 for details of the calibration and validation approach.

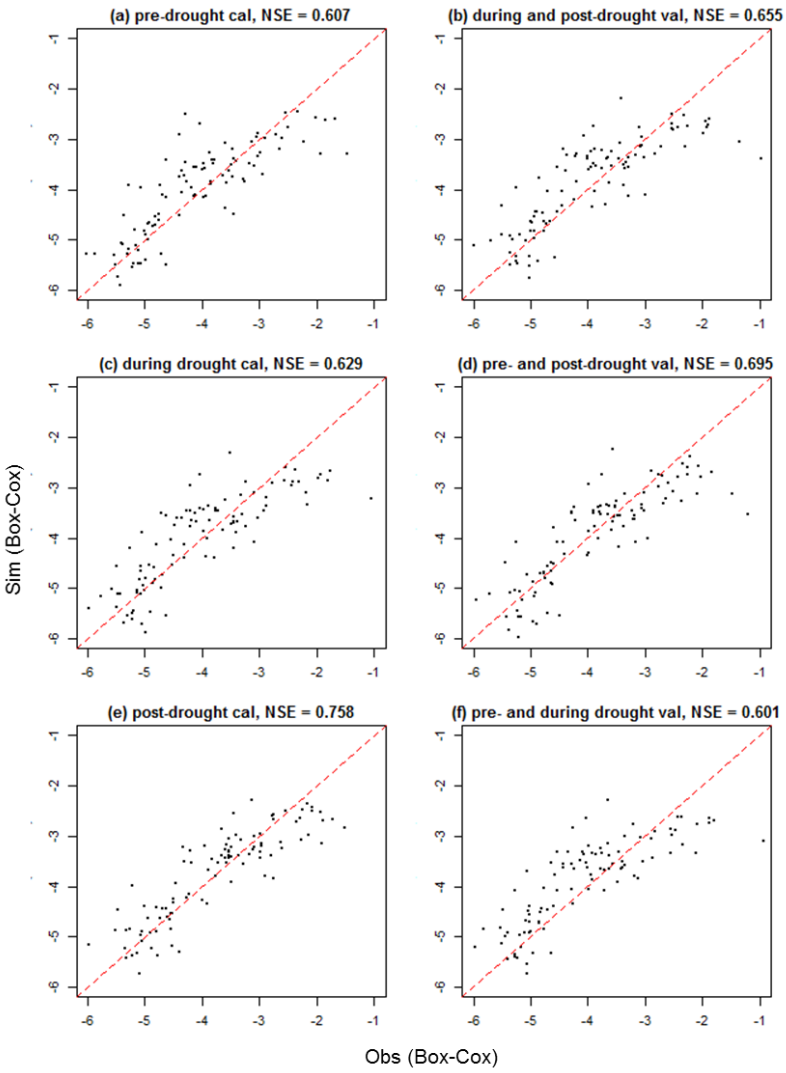


Figure ~~S8~~S9. Comparison of the TP model performance, as the simulated against observed site-level mean concentrations across three different calibration/validation periods for calibrations on the pre-drought (1994-1996), drought (1997-2009) and the post-drought (2010-2014) periods, respectively, see Section 2.4 for details of the calibration and validation approach.

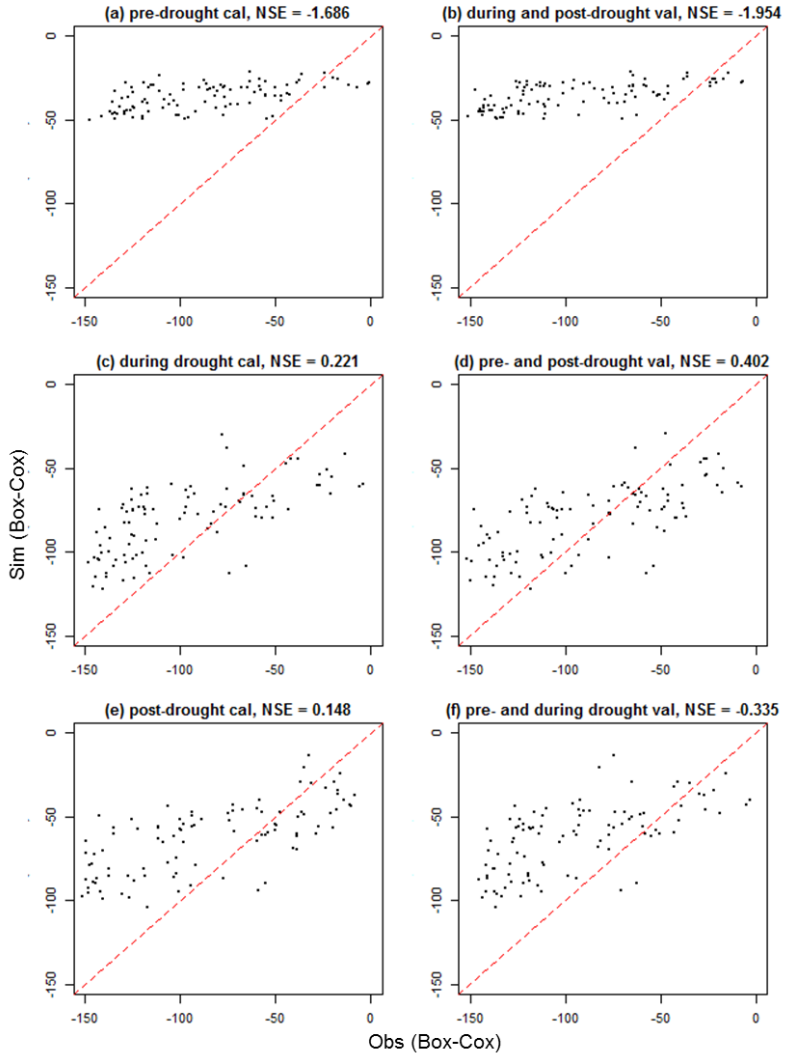


Figure ~~S9~~S10. Comparison of the FRP model performance, as the simulated against observed site-level mean concentrations across three different calibration/validation periods for calibrations on the pre-drought (1994-1996), drought (1997-2009) and the post-drought (2010-2014) periods, respectively, see Section 2.4 for details of the calibration and validation approach. Note that the unstable performance can be resulted by the poor performance for the full model, see Section 3.1.

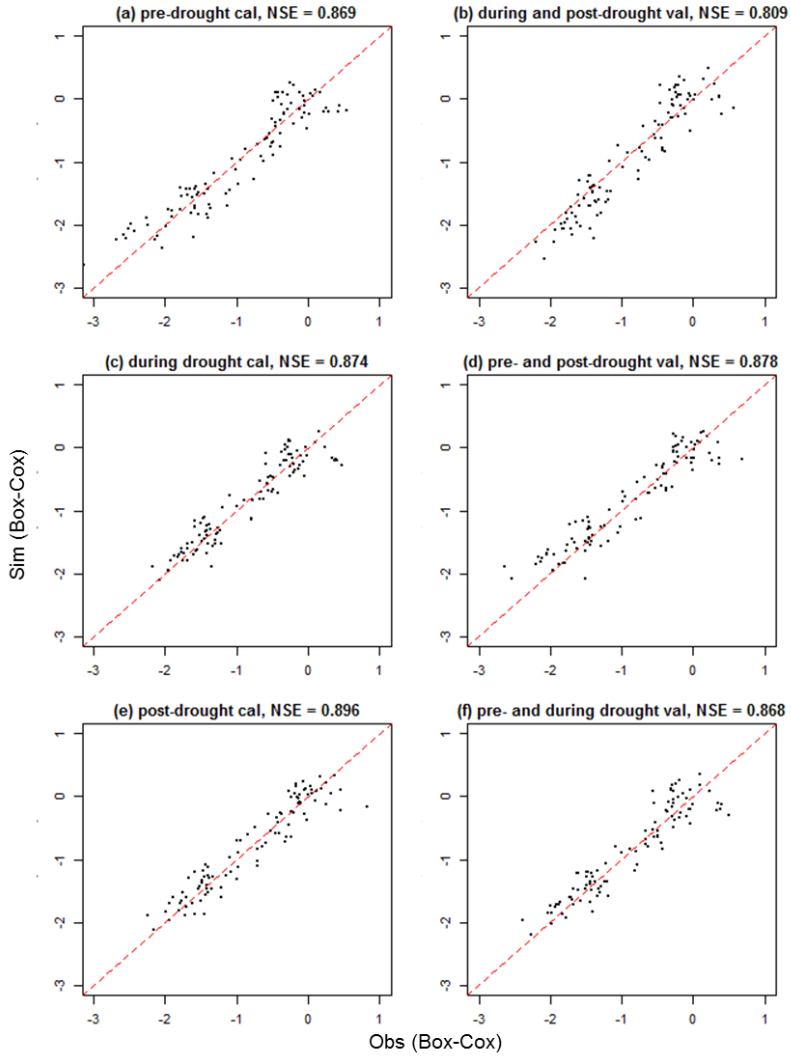


Figure ~~S10~~S11. Comparison of the TKN model performance, as the simulated against observed site-level mean concentrations across three different calibration/validation periods for calibrations on the pre-drought (1994-1996), drought (1997-2009) and the post-drought (2010-2014) periods, respectively, see Section 2.4 for details of the calibration and validation approach.

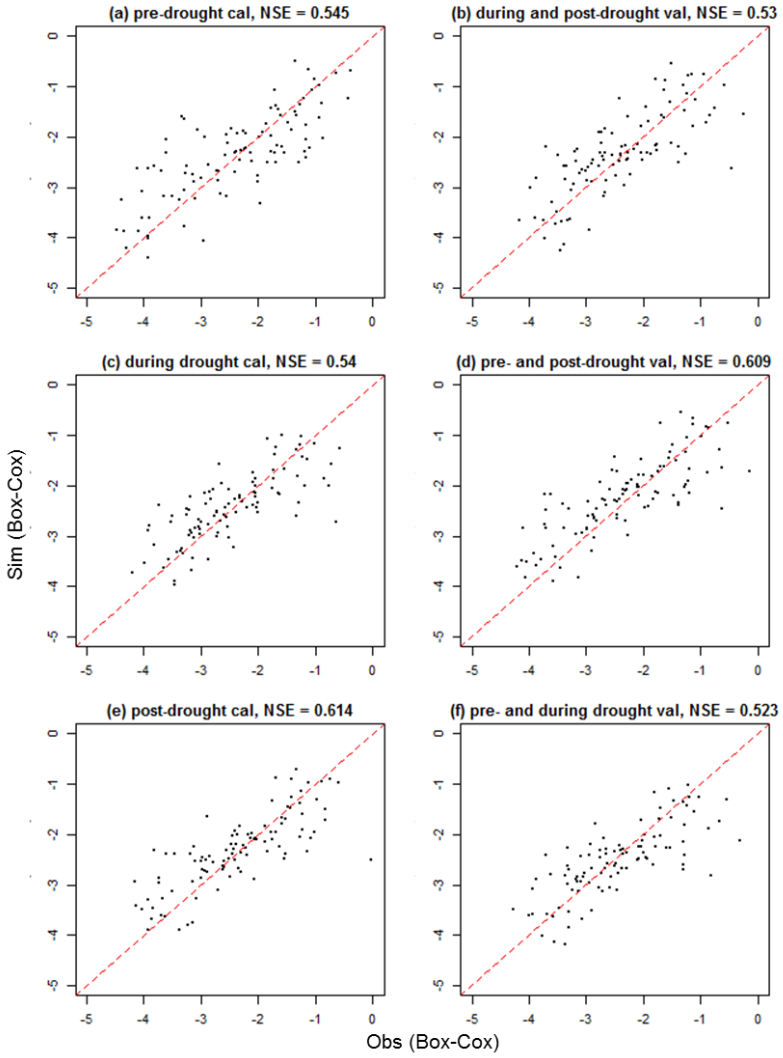


Figure ~~S11~~S12. Comparison of the NO_x model performance, as the simulated against observed site-level mean concentrations across three different calibration/validation periods for calibrations on the pre-drought (1994-1996), drought (1997-2009) and the post-drought (2010-2014) periods, respectively, see Section 2.4 for details of the calibration and validation approach.

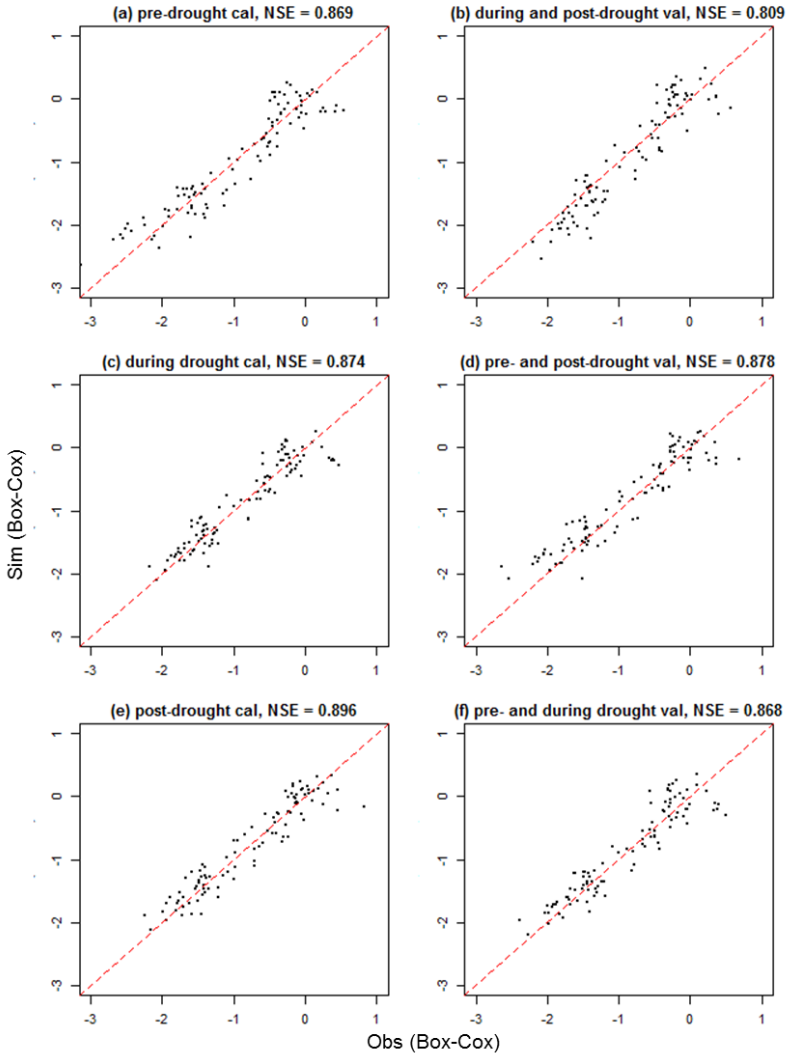
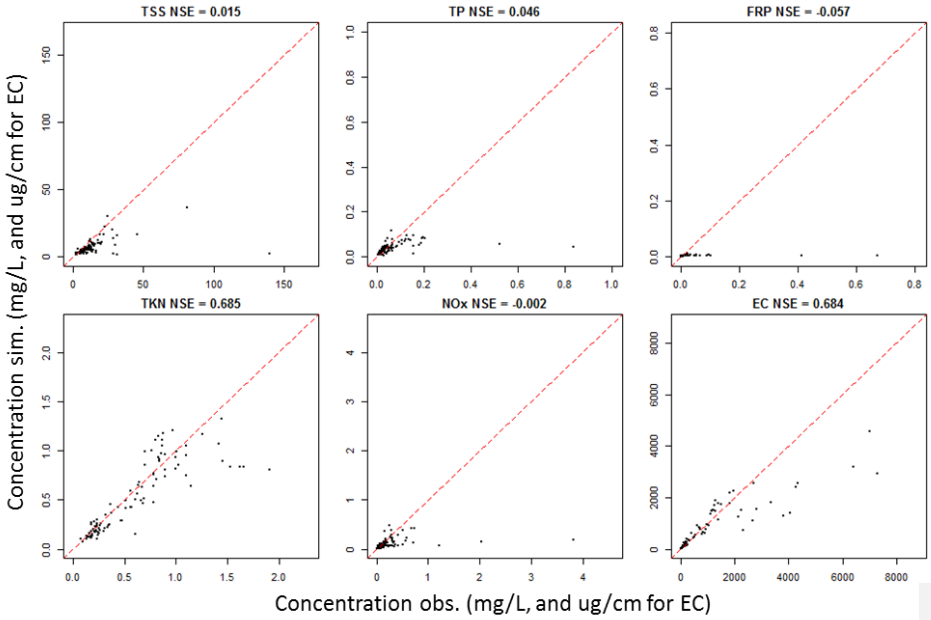


Figure ~~S12~~S13. Comparison of the EC model performance, as the simulated against observed site-level mean concentrations across three different calibration/validation periods for calibrations on the pre-drought (1994-1996), drought (1997-2009) and the post-drought (2010-2014) periods, respectively, see Section 2.4 for details of the calibration and validation approach.



~~**Figure S13. Back-transformation of the model simulations to the measurement scale emphasizes influences of unusually high concentrations and thus heavily affects model fitting, illustrated by simulated against observed site-level mean concentrations of each constituent in a back-transformed scale.**~~

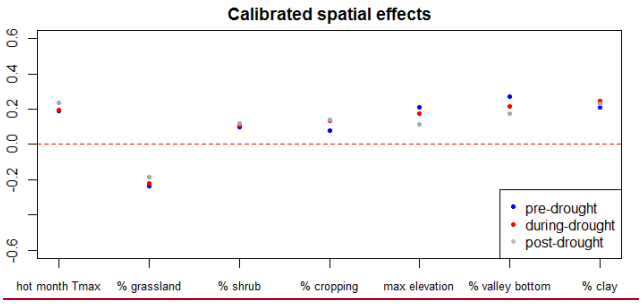


Figure S14. Effects of the seven key predictors for the spatial variability in TSS across 102 sites, summarized by the posterior mean of the calibrated parameter values for each predictor, to the pre-, during- and post-drought periods (differentiated by colour). The seven key predictors are, from left: hottest month maximum temperature, percentage catchment area as grassland, percentage catchment area as shrub, percentage catchment area as cropping land, maximum catchment elevation, percentage catchment area made up of valley bottoms, and average soil clay content.