Responses to Reviewer #2 (hess-2021-353)

Our responses are in blue and proposed revisions are underlined.

This manuscript presents a synthesis of baseflow effects on C-Q relationships in watersheds across Australia. The authors have leveraged a Bayesian Hierarchical Model in this research. Overall, I think the research is solid, the manuscript is well written, and it can become an important contribution to the literature on riverine C-Q relationships. I provide below some comments to the author, which I hope can help improve the manuscript.

We thank you for providing valuable feedback on the study for further improvement. We provide a point-to-point response to your comment below.

1. The use of Bayesian Hierarchical Model should be more fully justified. I am aware of the research the team has done in the past few years involving Bayesian approaches, but why is it used in this work on C-Q relationships. Please provide your reasoning in the Introduction, probably the last paragraph.

Thank you for raising this point. <u>We propose to improve justification of the Bayesian Hierarchical Models (BHM) as follows:</u>

1) <u>In the end of the Introduction, we will add a brief discussion on the uneven distribution of the available data across sites to highlight the value of applying BHM in this analysis.</u>
2) <u>In the introduction paragraph of Section 2.2 (which introduces the model), we will add a brief justification for applying BHM. Specifically, we will discuss the key advantages of BHM in effectively handling data-limited situations and spatio-temporal data with uneven coverage. We will also add references to our previous works (as below) which have illustrated these advantages.</u>

References:

Guo, D., Lintern, A., Webb, J. A., Ryu, D., Bende-Michl, U., Liu, S. & Western, A. W. (2020). A data-based predictive model for spatiotemporal variability in stream water quality. Hydrology and Earth System Sciences, 24(2), pp. 827-847.doi:10.5194/hess-24-827-2020

Guo, D., Lintern, A., Webb, J. A., Ryu, D., Liu, S., Bende-Michl, U., . . . Western, A. W. (2019). Key Factors Affecting Temporal Variability in Stream Water Quality. Water Resources Research, 55(1), 112-129. doi:10.1029/2018wr023370

Liu, S., Ryu, D., Webb, J. A., Lintern, A., Guo, D., Waters, D., & Western, A. W. (2021). A Bayesian approach to understanding the key factors influencing temporal variability in stream water quality – a case study in the Great Barrier Reef catchments. Hydrol. Earth Syst. Sci., 25(5), 2663-2683. doi:10.5194/hess-25-2663-2021

2. The authors reported that the Bayesian Hierarchical Model can explain over half of the observed variability in concentration of TSS, EC and P species across all catchments (93% for EC, 63% for TP, 63% for SRP, and 60% for TSS). I feel the intention has switched here from understanding C-Q relationship to predicting water-quality concentrations, which seems to be a distraction to me. Moreover, what's the benefit of adopting the Bayesian Hierarchical Model, given that many statistical models (e.g., WRTDS) have been developed and can probably provide more accurate estimates?

Thank you. We propose to move the focus away from predicting water quality concentration, which can help better focus on our objective to understand the impact of BFI on CQ slopes. To achieve this, we will remove the current results in Section 3.3.1 and instead include new results to compare the performance of models with i) climate-specific impact of BFI; and ii) a single impact of BFI across all climate zones. This update will help to remove the previous confusion on our communication of the model performance, while also illustrating the value of considering climate-specific impacts of BFI in our study.

In addition, we acknowledge that our Bayesian Hierarchical Model (BHM) predicts across different catchments and is thus different to the catchment-specific WRTDS model. BHM is fitted to all catchments simultaneously and thus has a huge advantage of 'borrowing power' – to inform parameter estimation for one group of data by information from other groups (with 'group' being climate zone in our model). The hierarchical model architecture is ideally suited to grouped datasets, which enables data within the same group to share common features. The authors' previous studies illustrated the effectiveness of this model for simulating water quality temporal variability across multiple catchments (Liu et al., 2021; Guo et al., 2020; Guo et al., 2019). Therefore, BHM is useful to conceptualize the nation-wide C-Q dataset across multiple catchments and climate zones. We will further strengthen these to justify the use of BHM in the Introduction and Method sections (as detailed in our response to your Comment #1).

3. Figure 3b: It is not a strictly positive relationship for the entire range of BFI_m. The variability continues to increase with BFI_m up to ~ 0.5 and then starts to decrease with BFI_m. The latter part of the curve seems largely ignored in the manuscript, including Discussion. The same observation holds true for the individual constituents (Figure S5).

We agree with your observation. Our further investigation suggests that the non-linear relationship between BFI_m and the range of BFI is that both BFI_l and BFI_h increase with higher BFI_m, but the difference between BFI_l and BFI_h decreases at higher BFI_m > 0.6, which only occurs at a few catchments (see Figure R3 below). This highlights the need to explicitly consider the impact of the range of BFI. We plan to explore this by including an alternative model structure in the manuscript, where the range of BFI (BFI_range) is used as the key predictor. We also propose to replace the current Figure 3b) in the manuscript with Figure R3, which better explains the nonlinear relationship between BFI_m and BFI_range.

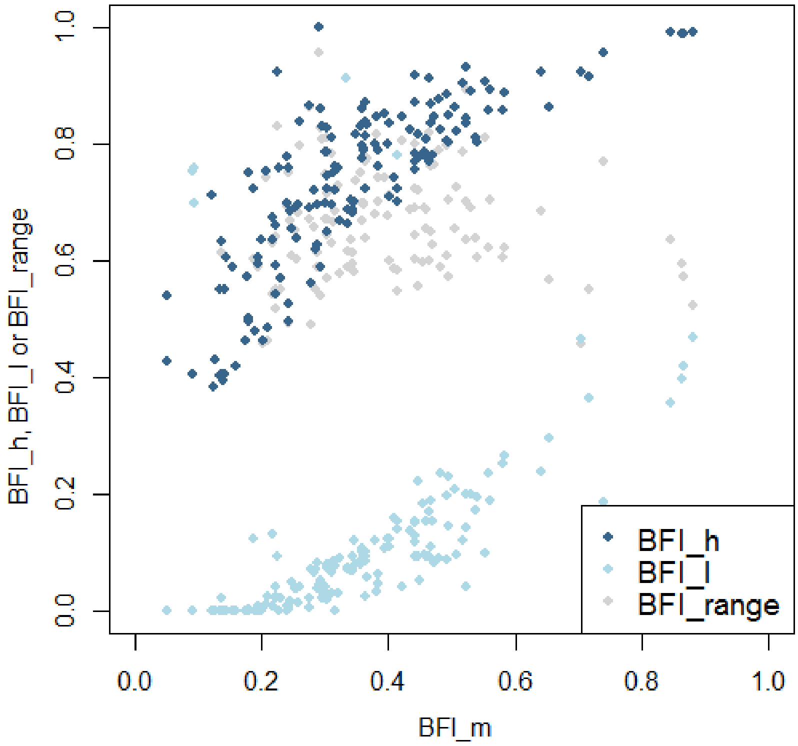

Figure R3. The low (10th percentile) and high (90th percentile) of instantaneous BFI (BFI_l and BFI_h), and the range of instantaneous BFI (BFI_h - BFI_l) versus BFI_m. The plot includes all 157 catchments across six water quality variables studied.

4. It would be interesting to investigate the effects of seasons and antecedent discharge conditions (wet vs. dry), both of which may change the response of C-Q slope to the BFI_m metric. There may be strong contrast among, for example, growing vs. non-growing seasons. Toward the end of manuscript, the authors have briefly pointed out the possibility of season effects. I think it is probably beyond your scope to look into these effects in this paper, but I encourage the authors to provide a brief discussion to point out that the response of C-Q slope to the BFI_m metric can vary among different seasons, among different antecedent discharge conditions, and even among different periods. In the latter regard, it is reported that anthropogenic disturbances and/or management actions occurred in the catchment can cause the C-Q relationship to change. For example, Zhang (2018) provides an investigation of C-Q relationship for different river flows and years: https://doi.org/10.1016/j.scitotenv.2017.09.221.

Thank you for the excellent suggestion. We will add further discussions on the seasonal changes of BFI with respect to existing studies (including the suggested reference), and integrate these with our results to highlight further study directions.

5. The term BFI_m (median BFI) is not self-evident in the Abstract. Given the importance of this metric, I encourage the authors to define it more clearly in the Abstract.

We will revise the abstract using 'catchment median BFI' throughout, instead of 'BFI_m'.

6.

a) The authors have used BFI_l and BFI_h to represent the variability of BFI, which makes sense to me. I may have used 2.5% and 97.5% instead but 10% and 90% are fine.
b) By the way, have you considered using standard deviation to capture the spread, which may help shorten the manuscript in terms of text and figures presented? I think an argument can

be added to the end of Section 2.1, which favors the use of BFI_l and BFI_h, that these are percentile based and hence are more robust to outliers.

a) Using 2.5% and 97.5% quantiles have potential risk to capture some outliers considering our catchment selection criteria (Section 2.1). Specifically, catchments with a minimum of 300 CQ-pairs can be included in our analysis, which means that the BFI_l and BFI_h are each if 2.5% and 97.5% quantiles are used to calculated BFI_l and BFI_h, each index can be calculated with <10 data points at the most data-scarce catchments. Therefore, we intend to keep the original decision of having BFI_l and BFI_h as the 10% and 90% quantiles of BFI for each catchment.

b) We thank you for suggesting an additional justification for the use of BFI_l and BFI_h and we will add this to the revised manuscript. A future relevant change we propose is to add a new model with the same structure as our original model based on BFI_m, which is based on BFI_range (the difference between BFI_h and BFI_l). The modelled effects of BFI_range of CQ slopes (see Figure R1 below) are highly consistent with the modelled effect of BFI_m (see Figure R2 below, with minor updates from the original Figure 6 after correcting some NOx data errors). These new results provide more concrete evidence on our previous 'speculation' on how the BFI range in a catchment can impact its CQ slope. We will add the new modelled effects to the Results section and update our discussion accordingly.

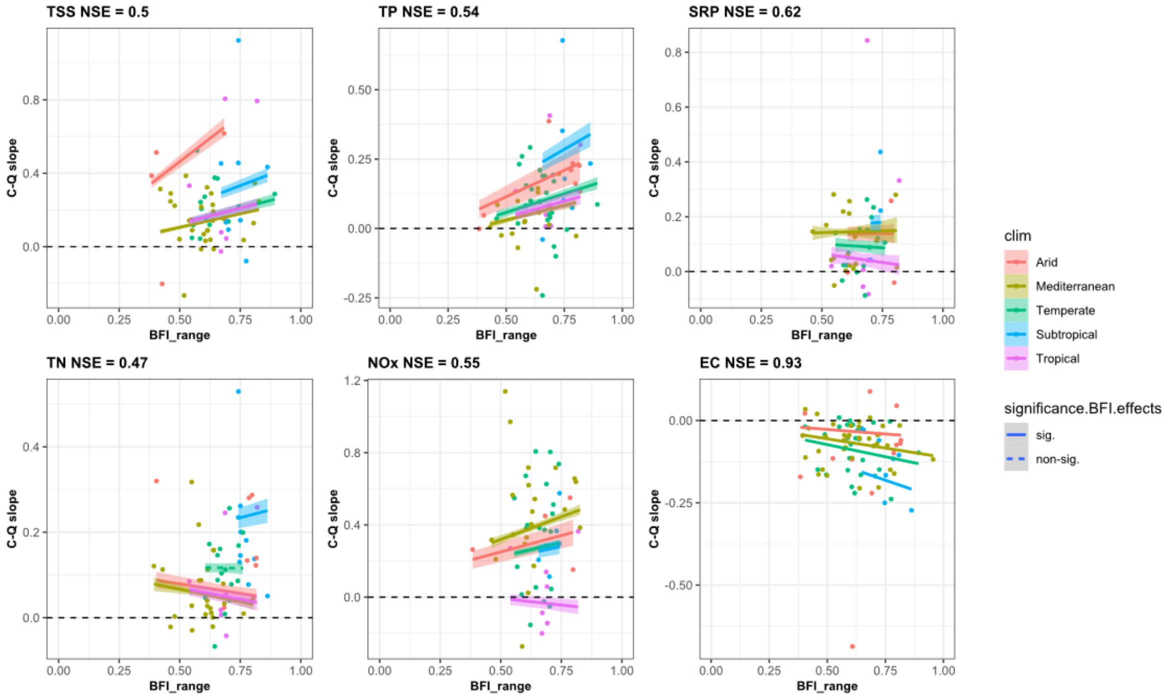

Figure R1. Catchment C-Q slope vs. catchment BFI range (BFI_range, as the difference between BFI_h and BFI_l), coloured by climate zones. The lines represent the modelled C-Q slope~BFI_range regression lines for individual climate zones,

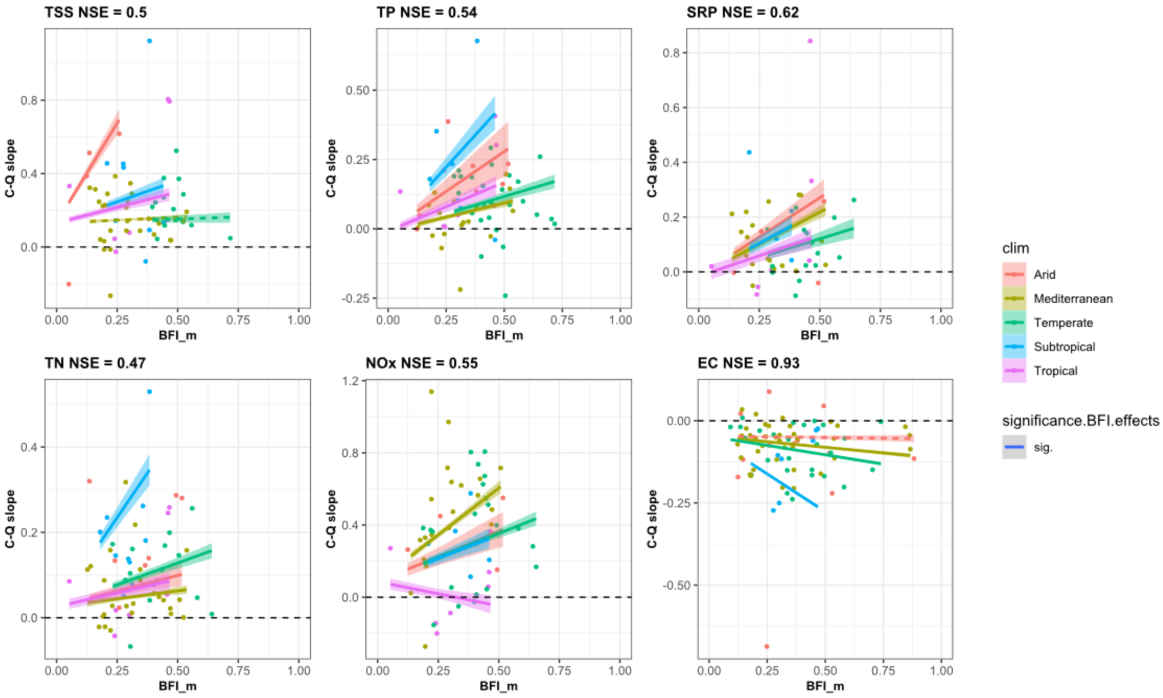

Figure R2. Catchment C-Q slope vs. catchment median BFI (BFI_m), coloured by climate zones. The lines represent the modelled C-Q slope~BFI_m regression lines for individual climate zones,

7. For days with multiple samples, is it necessary to pre-calculate the average concentration? Why not keeping all the samples in the analysis? In addition, it may be helpful to provide a table that quantifies the fraction of such days in the record.

Streamflow records are generally in daily timestep (instead of sub-daily) across the nation, which limits us to use all high-frequency water quality samples to perform CQ analysis. To clarify this, we propose to add a sentence to justify this averaging process in the current Line 90 as:

- *"A daily average is taken if more than one water quality sample was collected for any day at any site. This is because streamflow records in Australia are largely in daily timestep, which limits our ability to analyse all high-frequency water quality samples."*

We will add a table to the supplementary information to specify the fraction of days with multiple samples – this generally occurs rarely except for some EC sites which collect high-frequency samples.

8. BFI calculation: I am curious about the use of 0.98 for alpha in the baseflow filter. Did Ladson et al. (2013) recommend this value? What is the rationale?

Yes, this was a recommendation in Ladson et al. (2013). The study stated: "recent comparisons of modelled and measured baseflow values in the Murray Darling Basin suggest a value of 0.98 produces more reasonable results.". Murray Darling Basin is a region which a large proportion of our study catchments are located (Figure 1), and we already clarified this with L130:

- *"The daily BFIs were estimated using a Lynne-Hollick baseflow filter with Alpha = 0.98 … as recommended for Murray-Darling Basin in the south-eastern Australia (Ladson et al., 130 2013), within which a large number of the study catchments are located."*

9. Figure 2: Please add numbers and units (even if hypothetical) on the y-axis for panels a and b.

We will revise the figure as suggested.

10. Equations 2-3: Consider changing $\delta BFIclimate$ to δBFI_$climate$. (Move "BFI" to the subscript.) At first glance, I thought this is the product of two variables (δ and BFIclimate).

We will revise this term as suggested ($\delta_{\mathrm{BFI}\_climate}$) throughout the paper.

11. Section 3.3.1, including Table 1: I would like to refer back to my comment above. The NSE values do not seem to be comparable to more established approaches such as WRTDS. What is the value of showing these results? Should the baseline model or the BFI-based model be used for predicting concentrations? Why not those other established approaches?

As detailed in our response to your Comment #2, we propose to remove the current results in Section 3.3.1 and instead include new results to compare the performance of models with i) climate-specific impact of BFI; and ii) a single impact of BFI across all climate zones. We believe that this update will help to remove the previous confusion on our communication of the model performance.

Furthermore, since the model is fitted to all data across multiple catchments, this NSE value essentially describes how well the model explained the total spatio-temporal variability in water quality within the national dataset. Predicting spatio-temporal variability is a much more difficult task than predicting the temporal variability alone – as achieved by WRTDS for individual catchments (Zhang et al., 2021; Sprague et al., 2019). Thus we believe that the NSE values from BHM is not comparable with WRTDS. However, any related confusion should be removed with the proposed revision of Section 3.3.1.

References:

Zhang, Q., Webber, J. S., Moyer, D. L., & Chanat, J. G. (2021). An approach for decomposing river water-quality trends into different flow classes. Science of The Total Environment, 755, 143562. doi:https://doi.org/10.1016/j.scitotenv.2020.143562

Sprague, L. A., Mitchell, R. M., Pollard, A. I., & Falcone, J. A. (2019). Assessing water-quality changes in US rivers at multiple geographic scales using results from probabilistic and targeted monitoring. Environmental Monitoring and Assessment, 191(6), 348. doi:10.1007/s10661-019-7481-5

12. Section 3.3.2: According to published literature on many catchments around the world, SRP is a minor component of TP, whereas NOx is a major component of TN. It is quite interesting that in these Australian catchments, NOx/TN is quite small. This presents a strong contrast to many regions and may be discussed with a couple of sentences.

Thank you for highlighting this point, we will add some discussion on the contrasting NOx:TN pattern with other regions.