Response to comments from Reviewer 2

Summary

2.1 In this paper, the authors introduce a Multi-Temporal Hydrological Residual Error (MuTHRE) model that enables the production of seamless streamflow forecasts (e.g., daily, weekly, fortnightly, monthly) within the range of 1-30 days. The approach is described and compared against a non-seamless streamflow post-processing (QPP) model implemented by the Australian Bureau of Meteorology's Dynamic Forecasting System. The comparison is performed in 11 Australian catchments in terms of several forecast attributes, and the authors conclude that the MuTHRE model is not only capable of providing good performance for daily streamflow forecasts and cumulative volumes, but also similar performance to that obtained with the non-seamless QPP model for monthly flows.

Overall, this is an interesting manuscript that contributes with encouraging results on the use of seamless streamflow forecasting frameworks. The motivation is clearly stated and the results are nicely presented. There is, nevertheless, a lot of room for improving explanations of the model formulation, streamflow forecast generation, and verification, so that any reader could reproduce the results presented here. There are other minor comments and editorial suggestions that may also help the authors to improve the quality of their manuscript.

We thank Reviewer 2 for their encouraging feedback and detailed review of our paper. In particular, we appreciate their thoughtful suggestions for improving the description of streamflow post-processing models in Section 2, which will make this material easier to follow.

Main comment:

2.2 Model description (section 2): I found this section very hard to understand. I think the manuscript would greatly benefit from re-organizing the material and improving definitions and descriptions. For example:

It seems that the two approaches compared here follow the same general model structure (equations 1 and 2). Is that what you mean with "both QPP models"? Can you please be more explicit?

Yes the two approaches have the same general structure. We will rewrite this sentence to be more explicit about this.

We will also perform a comprehensive review of terminology across the paper, which will also help avoid confusion in this section.

2.3 Also, Qt is described as a "probability model for streamflow" (equation 1), and then as a "residual error model" (L103, equation 2) when it is, in reality, the sum of deterministic model output and a residual error term. I wonder if you actually need equation (1) in this description.

The sentence describing Q_t as a residual error model was indeed incorrect and confusing - thanks for highlighting this error. As the reviewer notes, Q_t is the sum of deterministic model output and a residual error term (which is modelled using a residual error model). We will rectify this in the revised paper.

Equation 1 is important in summarising the overall structure of the forecasting models, notably their probabilistic nature, their inputs and parameters.

2.4 I think it would be better to have the information presented in L191-214 (differences between MuTHRE and monthly QPP model) right after section 2.1.

We appreciate this suggestion.

A major benefit would be that Figure 1a, which shows structures of two post-processing models, is presented earlier on. This would make it easier to follow the descriptions of the post-processing models in Sections 2.2 and 2.3.

However, moving all of the information in Lines 191-214 (Section 2.4.1) to immediately after Section 2.1 would introduce a new set of problems, because the text in Section 2.4.1 relies on material presented in the earlier Sections 2.2 and 2.3.

Based on these considerations, we will

i. Introduce Figure 1a much earlier on, when we first describe the overall approach for the MuTHRE and monthly QPP models (Sections 2.2.1 and 2.3.1).
ii. Keep most of the summary of structural differences between the post-processing models in Section 2.4.1, in order to avoid forward referencing.

2.5 The authors should consider separating Figure 1 (which is very nice) into two figures: one for model structure (which could include model equations for more direct comparisons between model structures), and another figure for model calibration and forecasting.

There are pros and cons to this suggestion.

Separating out the model structure figure would allow us to present this schematic earlier, and make it easier to follow the description of the post-processing models (see comment 2.4 above). On the other hand the schematics can be more easily compared to each other when they are presented as multiple panels within the same figure.

In terms of including equations in the figure, it is not immediately clear whether it is feasible to include all relevant equations in this figure without creating clutter and potential confusion. Note that the key differences between the post-processing models – namely aggregating to monthly scale and taking the median – are better described schematically rather than through equations.

2.6 The meaning of z should be included after presenting equation (2) (perhaps in line 107).

Good suggestion. We will move the meaning of z earlier, so that it comes directly after equation (2).

2.7 Since Xt is also used to describe state variables in the hydrology literature (especially in data assimilation books/papers), I think ut would be more appropriate for meteorological forcings (e.g., Liu and Gupta 2007).

Thanks for this suggestion. Ideally our notation would be consistent with the broad hydrological literature. However, since the previous papers on the MuTHRE model (McInerney et al., 2020;

McInerney et al., 2021) use $\mathbf{x}_t$ for meteorological forcings we prefer to also use this notation for consistency.

2.8 Additionally, in L125 you describe st as a time-varying scaling factor, while the same variable is used to describe hydrological model states in L100.

Good catch. We will change the symbol for the time-varying scaling factor

2.9 L113: I presume that the raw streamflow forecasts do not account for uncertainty in hydrologic model parameters. Can you please clarify?

Correct. We will clarify this.

2.10 L135: Is the ensemble size still Nfoc after adding the residual term?

Yes it is. We will clarify this.

2.11 L141: What do you mean by "individual raw forecast"? Each ensemble member produced with the ensemble of rainfall forecasts?

Yes this is what we meant. We will ensure that consistent terminology is used for ensemble members throughout the paper.

2.12 L148: how is m* determined?

It is computed as the mean of the residual $\boldsymbol{\eta}$ after the seasonality and dynamic bias terms are removed. We will mention this in the text.

2.13 Since the paper should be self-contained, additional information on the calibration procedures referred to in L162 and L190 should be provided (what are the calibration period, objective functions, and optimization algorithms?). A couple of sentences should suffice.

We agree with the reviewer that additional information on the approaches used to calibrate the post-processing models should be included here. In the revised text we will describe how the parameters in the streamflow post-processing models are estimated (e.g. method of moments, maximum likelihood).

We note that Section 3.3 provides details of cross-validation procedure (including calibration periods), as well as the objective function and optimization algorithm for calibrating the hydrological model.

2.14 L167: "and then collapsed to a deterministic forecast by taking the median". Is this current operational practice?

This step is performed in the Bureau of Meteorology's dynamic forecasting system (see Section 2.3.2 in Woldemeskel et al., 2018).

2.15 L111, L112, L135, L168, L169, and elsewhere: is "replicate" the same as "ensemble member"?

Yes "replicate" is the same as "ensemble member". We will ensure "ensemble member" is used throughout paper

Additional minor comments

2.16 L33-35: It makes more sense to me to describe common practice before referring to the need for seamless forecasts. Also, it would be worth highlighting that non-seamless forecasting efforts have been (and are being) conducted in South America (e.g., Souza Filho and Lall 2003; Mendoza et al. 2014), Europe (e.g., Ionita et al. 2008; Hidalgo-Muñoz et al. 2015), Asia (e.g., Pal et al. 2013) and everywhere else around the world, with appropriate citations.

Thanks for this suggestion. We originally referred to "seamless forecasts" first, because it made it easier to define a "non-seamless forecast". However, it does make sense to start with common practice, and we plan to implement this change.

We also appreciate the references for other non-seamless forecasts around the world, and will include these in the introduction where appropriate.

2.17 L37: "This is the focus of our study". This reads out of place here. I recommend deleting this sentence or moving it toward the end of the introduction.

We will remove this sentence.

2.18 Figure 2: How many values are contained in each boxplot? One per basin?

There are 11 values in each boxplot, corresponding to the 11 catchments. We will clarify this in the caption.

2.19 Since you have only 11 catchments, I think it would be better to show one line per basin.

Thanks for the suggestion. We did try this, but preferred the boxplots because

- They are cleaner (fewer lines), which makes it easier to see variability between months and years
- All catchments show roughly the same patterns for both monthly and annual variability, so there is not much to be gained from showing each catchment

2.20 Further, it would be informative for readers to have a table with the name of the station, basin-averaged elevation, area, mean annual runoff, mean annual precipitation, mean annual temperature, annual runoff ratio, and aridity index.

Good suggestion. We intend to include a table will relevant catchment information in the revised manuscript.

2.21 L273: Are you working with calendar years or water years?

Calendar years. We will clarify this in the text.

2.22 Are daily forecasts produced each day in year j with MuTHRE, or only at the beginning of each month?

Forecasts for the MuTHRE model are produced at the start of the month only. We will clarify this in the text.

2.23 What is the final ensemble size of your forecasts?

The size of the post-processed streamflow ensembles is 100, which is the same as size of the forecast rainfall ensembles. We will clarify this in the text.

2.24 L274-275: The problem of hydrologic memory in Australian catchments and its implications for cross-validation has been previously documented (e.g., Robertson et al. 2013; Pokhrel et al. 2013). I recommend the authors read and cite these papers here. The following blog article is also relevant: https://hepex.inrae.fr/how-good-is-my-forecasting-method-some-thoughts-on-forecast-evaluation-using-cross-validation-based-on-australian-experiences/

Thank you for these references. We will add appropriate citations regarding the importance of dealing with hydrologic memory in the cross-validation procedure.

2.25 L292: Perhaps it would be better to replace the word "uncertainty" with "spread".

Good suggestion. We will make this change.

2.26 Also, it would be informative to state that sharpness is a forecast attribute only (i.e., it does not depend on the observations).

Good suggestion. We will add this.

2.27 L296: Since CRPS measures the difference between forecast and observation CDFs, it would be better to refer to “probability forecast errors” instead of “combined performance”.

We describe CRPS as a metric for "combined" performance because, as shown in Hersbach (2000), the CRPS can be decomposed into terms representing individual performance aspects, namely reliability, and uncertainty/resolution (related to sharpness). We agree this may not be obvious to a general reader and will clarify this in the text. We will also clarify that for a single observation, the CRPS represents the error between the forecast distribution and the observed data point.

Note that the CRPS serving as a combined performance metric is very relevant to our study because the other three metrics – namely reliability, sharpness and bias - focus on fundamentally more specific performance characteristics. If we do not highlight this, readers may form the erroneous impression that we have four independent performance metrics.

2.28 Figure 4: How are confidence limits generated?

These are 10th and 90th percentiles of metric values based on values for the 11 catchments. We will clarify this in the caption.

2.29 Do you compute the metric merging forecasts from all basins? Please clarify these points in the manuscript.

The metrics are computed separately for each catchment, and the distribution of these metric values is summarized using the median, 10th and 90th percentiles. We will clarify this in the caption.

2.30 L368-370: You mention that reliability results are similar, although the boxplots look different. I recommend applying a statistical test to determine whether the distributions of these metrics are significantly different.

We do use a statistical test to evaluate difference in distributions of metric values. Specifically, we use “practical significance testing” (described in Section 3.4.3) to determine whether differences between models are not just statistically significant, but are statistically significantly larger than some pre-defined practically relevant margin (chosen as 20% of metric value).

Although the distributions in Figure 6a appear different, these differences are not practically significant based on the criteria defined in Section 3.4.3.

2.31 L370 and elsewhere: “practically significant” or “significant”. Are the authors referring to a statistically significant result? If not, I suggest re-wording or deleting the word ‘significant’.

Following on from the above point (2.30), the term “practically significant” refers to cases where the difference in metric values are statistically significantly larger than a pre-defined margin.

We note that we did not explicitly define the term “practically significant differences” in Section 3.4.3, which may have led to confusion.

We will make the following changes in order to make it clearer what “practically significant” refers to

1. We will explicitly define “practically significant” in Section 3.4.3
2. We will refer to Section 3.4.3 when we first mention “practically significant” in the Results section.

2.32 Figure 6: I think you should say "overall monthly performance" in the caption, and perhaps remind readers here what "overall" means.

This is a good suggestion. We will do this.

2.33 Are you grouping the results of all basins? In the left panels, how many points are contained in each boxplot?

Yes we are showing the results from all catchments in the boxplots. Each boxplot represents the distribution of metric values from the 11 catchments. We will clarify this in the caption.

2.34 In the center and right panels, how are the confidence limits computed?

These are 10th and 90th percentiles based on values for the 11 catchments. We will clarify this in the caption.

2.35 L421: Shall we expect persistence in rainfall, given the chaotic nature of the atmosphere?

Good question.

In this sentence we are

- Referring to the day-to-day persistence in rainfall, which will be important for reliable monthly rainfall forecasts.
- Not referring to longer-term persistence, which will be much smaller due to the chaotic nature of the atmosphere.

We will clarify that we are referring to day-to-day persistence in the revised paper.

2.36 L426: I encourage the authors to replace the last sentence of this paragraph (which reads a lot like "propaganda") with a more quantitative statement regarding the performance of MuTHRE.

We will rewrite this sentence.

2.37 L467: “This was not feasible in this study”. If you cannot provide an explanation on why was not feasible, I suggest deleting this sentence.

We will remove this sentence.

2.38 L479: “High-quality forecasts”. Note that the quality depends a lot on the forecast attributes you are analyzing. I think it would be good to provide a brief discussion (maybe in section 5) about tradeoffs between the metrics included here (e.g., how your forecast system can improve reliability at the cost of losing sharpness), and what makes a forecast “good” or “high-quality”.

We agree that our classification of “high-quality forecasts” is dependent on the forecast attributes considered in this study. We will make this point by changing this phrase to “high-quality forecasts, based on the metrics considered in this study” (or something similar).

The results of our case study show that differences in reliability for the two models are not practically significant. As such, there is no (practically relevant) trade-off between the reliability and sharpness of forecasts to discuss.

Suggested edits

2.39 L51: ‘Hydro-electric’ -> ‘hydropower’.
L70: ‘drop in’ -> ‘loss of’.
L312: ‘which’ -> ‘who’.
L377: ‘in 1 month (September)’ -> ‘in September’.
L380: delete ‘similar/better performance in all months, with practically’.
L382: delete ‘similar/better performance in 19 out of 22 years, with practical’.
L451: ‘Simplifies’ -> ‘A simplified’.
L473: ‘to forecasts’ -> ‘compared to forecasts’.

Thank you for these suggested edits. We intend to implement these changes.

References

Hersbach, H. 2000. Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather and Forecasting,* 15**,** 559-570.

McInerney, D., Thyer, M., Kavetski, D., Laugesen, R., Tuteja, N. & Kuczera, G. 2020. Multi-temporal hydrological residual error modelling for seamless sub-seasonal streamflow forecasting. *Water Resources Research,* 56**,** e2019WR026979.

McInerney, D., Thyer, M., Kavetski, D., Laugesen, R., Woldemeskel, F., Tuteja, N. & Kuczera, G. 2021. Improving the Reliability of Sub-Seasonal Forecasts of High and Low Flows by Using a Flow-Dependent Nonparametric Model. *Water Resources Research,* 57**,** e2020WR029317.

Woldemeskel, F., McInerney, D., Lerat, J., Thyer, M., Kavetski, D., Shin, D., Tuteja, N. & Kuczera, G. 2018. Evaluating residual error approaches for post-processing monthly and seasonal streamflow forecasts. *Hydrol. Earth Syst. Sci. Discuss.,* 2018**,** 1-40.