Reviews of paper "Simple Box-Cox probabilistic models for hourly streamflow predictions"

We thank Reviewer 2 for dedicating his/her time and expertise to review our manuscript, and his/her constructive comments.

The remainder of this document provides detailed responses to each of R2's comment and proposes actions to address his/her comments. The reviewer comments are in black and our responses are in *red italics*.

Reviewer 2

The manuscript explores practical approaches for uncertainty quantification within a probabilistic framework for streamflow prediction. It focuses on two widely used strategies, that are (i) representing the heteroscedasticity of streamflow errors through the Box–Cox (BC) transformation, and (ii) capturing the temporal persistence of errors using autoregressive (AR and ARMA) models. The study examines seven catchments located in Spain, Switzerland, and the United States, covering a gradient from humid to semi-arid climates.

For the analysis, the authors employ a simple conceptual hydrological model that balances descriptive accuracy with model parsimony, and they implement it at an hourly time step. As the manuscript highlights, hourly predictions—and particularly their uncertainty quantification—have been relatively understudied, largely due to the scarcity of high-quality hourly data and the challenges involved in characterizing the associated prediction errors.

Against this background, the study investigates how different sources of uncertainty influence model predictions, considering various error model configurations that combine different BC transformations with AR structures of different orders. The authors also evaluate how these uncertainties propagate across temporal scales, from hourly to daily and monthly predictions, in a "seamless" prediction framework. Finally, by analyzing catchments from diverse climatic settings, the manuscript aims to identify results that can be generalized across contrasting hydrological regimes.

The topic is highly relevant for the scientific community. The innovative contribution of the manuscript is clear. The paper is generally well structured and clearly written, presenting its results and conclusions in a coherent and accessible manner. While the findings are not groundbreaking, they are—as the authors themselves point out—highly relevant for the practical application of probabilistic prediction frameworks. The manuscript deserves to be published after minor revisions.

Thank you for this positive feedback and constructive comments

R2.1. There is, indeed, one aspect that would benefit from a more detailed explanation: the seamless modeling framework. I encourage the authors to expand on how the transition from hourly to daily and monthly time scales is handled, as this additional information would greatly aid in interpreting the results.

We agree this concept deserves some elaboration.

The transition from hourly to daily and monthly time scales is handled by aggregation. For example, daily streamflow predictions are obtained by aggregating hourly streamflow for each hour of the (calendar) day. Similarly, monthly streamflows are obtained by aggregating together hourly streamflows for each hour of the calendar month.

This aggregation approach is distinct from the common approach of constructing a streamflow model directly at the daily scale.

We will review the text and elaborate where necessary to make this clear.

R2.2. Some additional effort could be devoted to making the paper fully readable on its own, by providing a few more details rather than directing readers to previous publications. This is, however, a very minor point, and I leave it to the authors to decide whether such additions are necessary.

This comment is similar to comment R1.2. We will review options for making the manuscript more self-consistent without distracting the reader from the key contributions of this work.

I also have a few minor corrections, listed below.

R2.3. Line 255. How the value of A was established?

The purpose of the offset parameter A is to avoid numerical problems when $q \approx 0$.

McInerney et al., 2018) recommend setting the value of A to 1% of the mean observed streamflow. In our study this yields A = 0.0013 mm/h for the Lasarte catchment. We will undertake sensitivity analysis to ascertain the impact of this choice across the case study catchments.

R2.4. Line 347. Table 2, not 1.

Thank you, we will correct – same issue was raised in R1.6

R2.5. Figure 1. Capital letters in the legends.

Thank you, we will correct this

R2.6. Figure 3. The Caption end with “Persistence model: PACF analysis of innovations”, which - I suppose - is the subtitle of the following paragraph (4.3).

Thank you, we will correct.

R2.7. Figures 4-6. I suggest to remove comments from the captions that should only describe the figures.

We appreciate this comment. We consider a short summary of messages to be useful for readers, however we will shorten where possible to keep the caption concise.

R2.8. Table 3. Are reliability and precision comparable across different data sets (time series at the hourly, daily and monthly scales) so that we can compare numerical values

Reviewer 1 also asked details of this procedure in comment R1.11. For a given performance metric, its values are calculated individually at each catchment.

We then report several types of metric averages

Averaging across catchments – these values are reported in the interior cells of Table 3. This is a straightforward operation, but we will provide equations to avoid confusion

Averaging across models – these values are reported in the last row of Table 3, labelled "Average AR1, AR2, AR3". This is also a straightforward operation, and will also provide equations

Averaging across timescales – these values are reported in the last columns of Table 3, labelled "Average (h,d,m)". We agree this averaging is less obvious, as we are averaging a metric calculated from hourly prediction with a metric calculated from daily predictions, and a metric calculated from monthly predictions.

Our current view is that this last averaging is still useful to compare the overall performance of the models. Notably, all metrics are dimensionless, which avoids dependence on scale-specific flow ranges, etc. An analogue for comparing deterministic models A and B would be to calculate, for each model, the average of Nash-Sutcliffe values for hourly, daily and monthly predictions.

We will review the current text during the revisions and will provide clarifications where required.