Replies to reviews

"Towards robust seasonal streamflow forecasts in mountainous catchments: impact of calibration metric selection in hydrological modeling"

Diego Araya, Pablo A. Mendoza, Eduardo Muñoz-Castro and James McPhee

We thank the three reviewers for their time in commenting on our paper. We provide responses to each individual point below. For clarity, comments are given in italics, and our responses are given in plain blue text.

Reviewer #1

The manuscript aims to evaluate the role of calibration metrics (objective function for calibration and performance evaluation metrics) on the seasonal streamflow forecasts in 22 mountainous river basins in Chile based on CAMELS-CL datasets. The quantum of work done by the authors needs appreciation, as well as the framing of the scientific questions. The manuscript has enough scientific content to be published in HESS after revision. The problems framed in the manuscript are tested using scientifically sound methodology.

However, I feel the manuscript is a complicated read due to the multiple parameters, metrics and lack of clarity, especially in the methods section. I believe the manuscript can benefit from reorganizing the content. The main result must be better highlighted, and others could be moved to the supplementary section to improve readability. The result sections do not highlight the overall conclusion or takeaway in each section. Therefore, I was pretty confused, even after multiple reads, about what the authors were trying to communicate.

We greatly appreciate the reviewer's comments, and provide detailed responses below.

A detailed flow chart can be used to convey the method. Parts of the methodology are distributed across different sections, including the introduction.

We have designed a new and more detailed flow chart. Additionally, we have added a diagram inspired by Figure 3 in Crochemore et al. (2020) to explain the Ensemble Streamflow Prediction (ESP) method, which is now detailed in section 3.2.

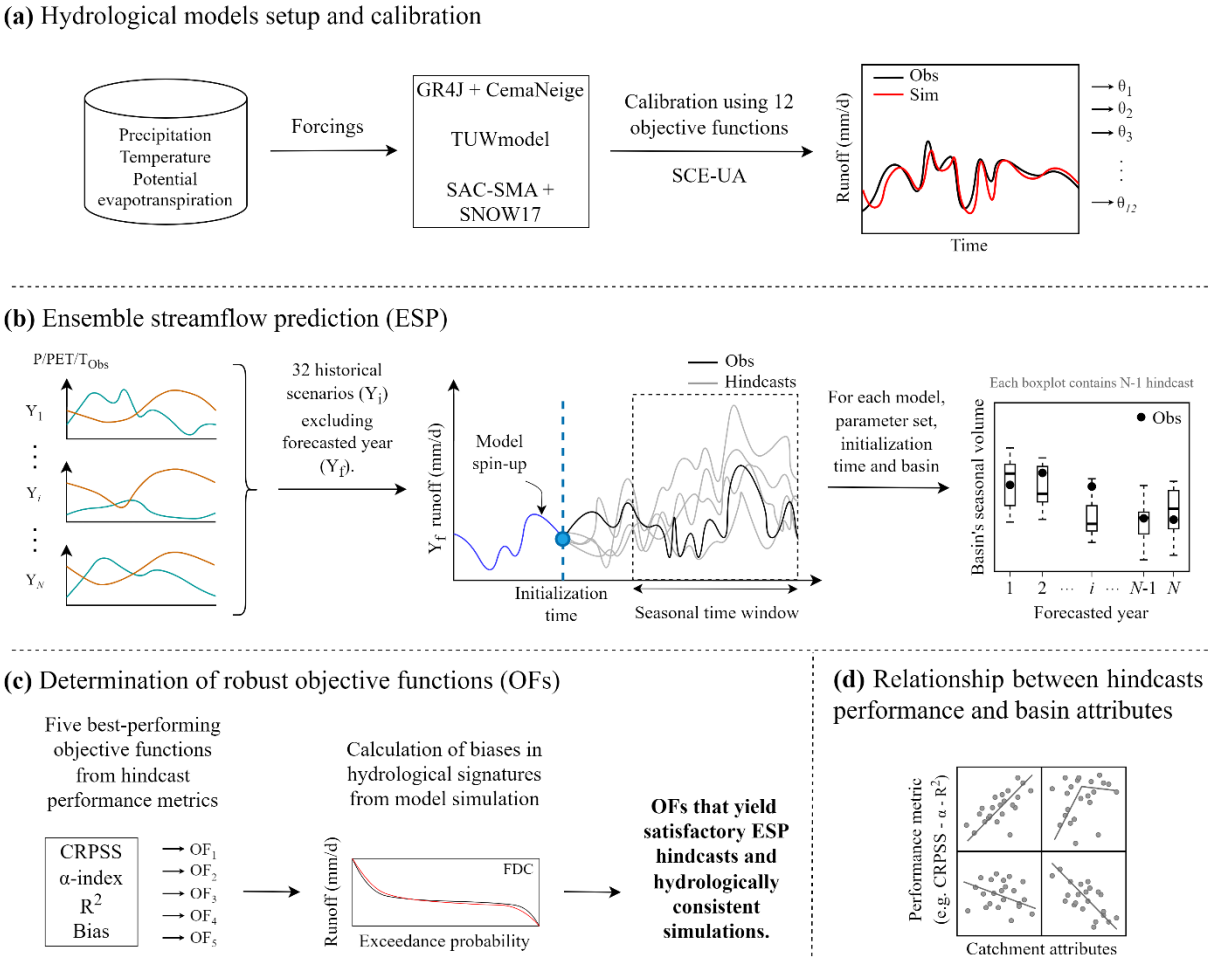

Figure 3. Flowchart describing the approach used in this study. See text for details.

The concept of Ensemble Streamflow Prediction used in the study is defined in the introduction section. I would appreciate elaborating on it in the methodology section instead. The introduction section should better focus on existing gaps in the literature and highlight the need for the present study.

We have moved the full description of the Ensemble Streamflow Prediction (ESP) method to section 3.2 ("Hindcast generation and verification"). We choose to maintain a brief description of the method in the introduction, only to highlight how our study contributes to the existing literature by exploring the impact of calibration metric selection on the quality of seasonal streamflow forecasts.

There is a lack of consistency in the terms used, which makes it more confusing. For instance, though hindcasts are performed in the paper, at certain places, forecasts are used.

In response to this observation and the comments from Reviewer #2:

- We have modified the term "forecasts" by "hindcasts" when referring to our methods and results, since this work presents an assessment of retrospective forecasts obtained from the application of different model calibration metrics.
- We use the term 'forecasts' when referring to past studies and operational applications.
- We use the term 'verification' when referring to the assessment of retrospective seasonal streamflow hindcasts.
- We use the term "evaluation" to the assessment of streamflow simulations outside the calibration period. The term "validation" is no longer used in this paper.

We will clarify this terminology at the beginning of section 3 in the revised manuscript.

Similarly, I did not understand what the authors meant by the first and second validation periods in the Figure 8 caption. Did you mean the calibration period, where the model is calibrated using different parameters, and the hindcast period, where the ensemble streamflow prediction method is employed?

As pointed above, we have replaced the term "validation" by "evaluation", which refers to the process of evaluating the quality of streamflow simulations outside the calibration period. Additionally, we have merged the originally proposed first and second evaluation periods into a single evaluation data set (which spans April/1987 – March/1994 and April/2013 – March/2020) to assess model simulations graphically and quantitatively. Such evaluation is illustrated for three objective functions in Figure 4 (former Figure 8):

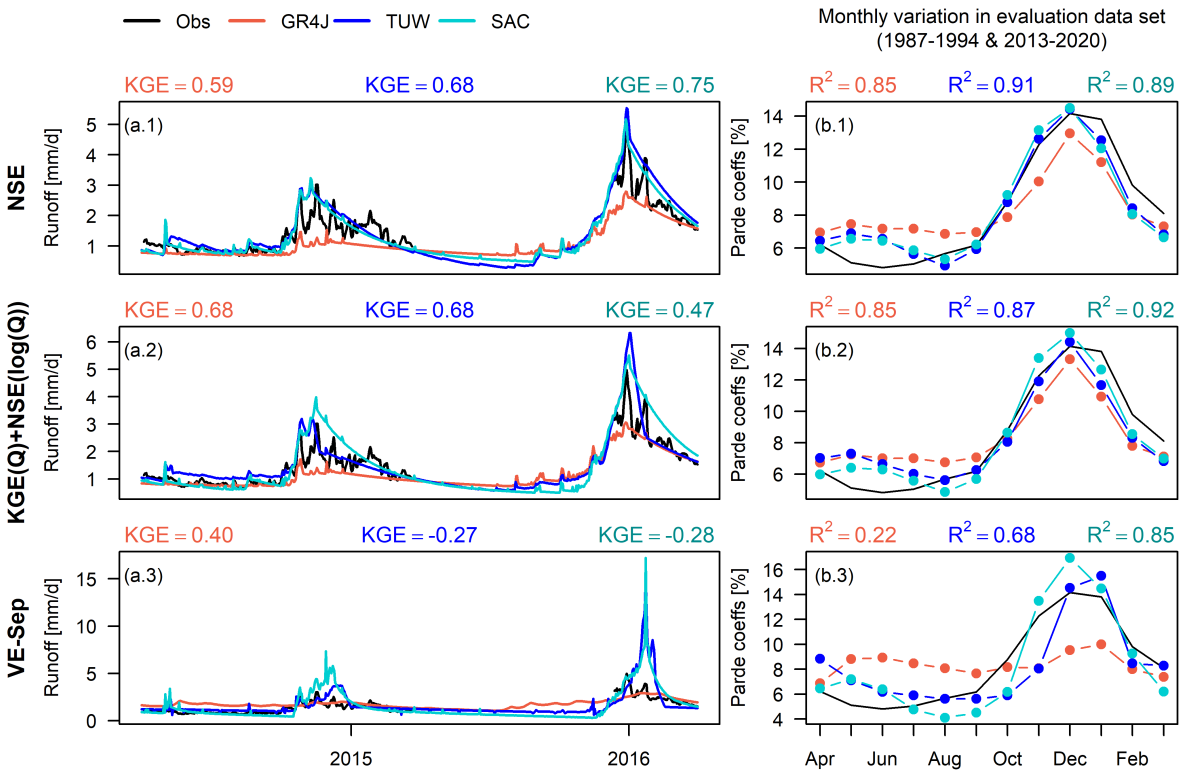

Figure 4. (Left) Daily hydrographs (April/2009 – March/2011) and (right) monthly variation curves for the evaluation dataset (April/1987 – March/1994 and April/2013 – March/2020) at the Maipo en el Manzano River basin, obtained with the three models and three objective functions: (1) NSE, (2) KGE(Q)+NSE(log(Q)) and (3) VE-Sep. The daily KGE obtained with each model is displayed in the left panels, while right panels include the coefficient of determination (R^2) between mean monthly simulated and observed runoff averages.

I think the manuscript will also benefit from redesigning the figures. The multiple boxplot figures create a lot of complexity in analyzing. Reducing the amount of noise while focusing on a particular science question could considerably improve the manuscript's readability and merit.

We appreciate this observation, and we agree that the figures included the original manuscript where unnecessarily busy. Hence, we have redesigned Figures 3 to 10, and we have moved Figure 6 in the original manuscript to Supplementary Material following the recommendation of Reviewer #2.

For instance, I would suggest focusing on the median result of all model combinations while showing the effect of initialization time and performance metrics for each calibration OF (Figure 7).

In response to the comments from Reviewers #1 and #2, we have redesigned Figure 7, showing three hindcast performance metrics and five initialization times from only one model structure (TUW),

with parameters calibrated with only one objective function (NSE). We display the median (solid line) and the 5th & 95th percentiles from the 22 case study basin basins as a light-blue shade for each metric. The results presented in this figure communicate the same findings obtained for the remaining representative objective functions and models: CRPSS and R^2 (α-index) values increase (degrades) as hindcast initializations approach Sep. 1. We decided to keep all five initialization times to clearly show the progression of seasonal (i.e., September-March) hindcast quality during the austral winter. The extended version of the new Figure 6 (which contains all five representative objective functions) is now included in the Supporting Information document.

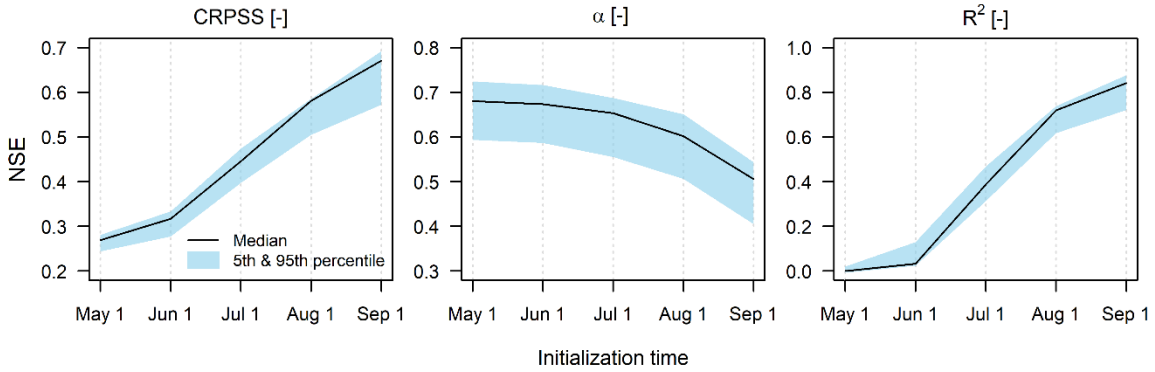

Figure 7. Impact of initialization time on hindcast verification metrics for NSE calibration objective function and the TUW model. The shades represent the 5th and 95th percentiles in each metric from the 22 case study basins, and the solid line represents the median of each metric.

Figure S2 is not cited, and supplementary figure S3 is wrongly numbered.

We thank the reviewer for the detailed revision of our manuscript. Figure S2 is now cited in the revised manuscript. Additionally, we have made sure that all the figures contained in the Supplementary material are correctly cited in the main manuscript.

In respect of results, '(not shown)' is used multiple times in the manuscript. I would suggest it will be better to include it in the supplementary section if the results are an important part of the argument.

We now include in the supplementary material most of the results that were referred to as 'not shown', in order to better support the arguments exposed here.

I reiterate the scientific questions in the manuscript intend to improve the seasonal ensemble streamflow prediction by assessing its sensitivity to calibration metrics is an important question. However, improving the organization and presentability of results are required to understand the manuscript outcomes better.

We agree with the Reviewer and thank him/her for the positive feedback and for his/her constructive review.

References

Crochemore, L., Ramos, M. H. and Pechlivanidis, I. G.: Can Continental Models Convey Useful Seasonal Hydrologic Information at the Catchment Scale?, Water Resour. Res., 56(2), 1–21, doi:10.1029/2019WR025700, 2020.