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 #2

This study evaluates the effect of calibration, initial condition and choice of model structure on forecasting seasonal volumes in mountainous areas. To that end, the parameters of three conceptual models were calibrated for 22 river basins in Chile using 12 objective functions. The calibrated models were then used to produce ESP forecasts considering five initial conditions. The authors evaluated the quality of the spring-summer forecasts using 33 years of data given the different combinations of models, parameters and initial conditions. An evaluation of the links between forecast quality, model performance in simulating different streamflow signatures, and catchment characteristics was conducted. The authors found that the choice of objective function has an impact on forecast quality and that a high performance in simulating hydrological signatures does not ensure good forecast quality. The authors also found that these results depend to some extent on the hydrological model.

As mentioned by the first reviewer, this study represents a significant amount of work. The subject is very relevant and has not been covered much in the literature. The differences in forecast quality and hydrological consistency for the different forecast combinations demonstrate that this study covers an important topic that needs to be considered for operational use.

Given the amount of model/calibration/HIC combinations and the choice of presentation by the authors, the article is not very easy to understand. I also have several questions and comments on some of the methodological choices that were made. Because the questions, comments and corrections I am making below might require a lot of work, I think they should be considered in light of the choices that will be made to reorganize the paper.

We express our gratitude to the referee for his/her meticulous review of our manuscript and for the constructive suggestions. We provide our responses below.

Major comments:

Twenty-two catchments with seasonal snowmelt contributions to total runoff were selected from the CAMELS-CL dataset for this study. It is stated (line 105) that "the selected basins are included in the CAMELS-CL dataset … and meet the following criteria…". Were all the catchments meeting these criteria selected (resulting in a total of 22 catchments) or were there other mountainous catchments of CAMELS-CL meeting the same criteria? Although these catchments encompass a large variety of hydroclimatic conditions, a larger dataset would enable more general conclusions to be drawn from this study. If other catchments of CAMELS-CL were to meet the same criteria, I would suggest including them in this study. Since they come from the same open-source database, it would mean running the same calculations but for more catchments.

Although we agree with the reviewer in that a larger sample of catchments would be desirable, there were no other Andean catchments meeting the six conditions. The most restrictive criteria are: (v) at least 75% of days with streamflow observations during the period April/1987 – March/2020, and (vi) at least 20 water years with seasonal (Sep-Mar) streamflow observations for hindcast verification purposes. We consider that these two requirements are essential for proper hydrologic model calibration and evaluation (since seasonal objective functions rely solely on Sep-Mar data availability) and a robust verification of seasonal streamflow hindcasts. We have added the following lines to section 2 to clarify this:

"The most restrictive conditions are (v) and (vi), which hinder the possibility to include additional mountainous catchments from CAMELS-CL; nevertheless, we consider that both requirements are essential for proper hydrologic model calibration and evaluation (since seasonal objective functions rely solely on Sep-Mar data availability) and a robust verification of seasonal streamflow hindcasts."

If the choice of selecting 22 catchments was driven by computation time restrictions, please consider mentioning it in the manuscript.

The choice of catchments was not driven by computation time restrictions. Please see our previous response.

As pointed out by the first reviewer, the manuscript can benefit from reorganizing the content. One way to reorganize the content would be to keep only one figure presenting the results for all models/objective functions/HICs (e.g. fig. 5, extended in height to enhance visualisation) and change the other figures so that they better highlight the conclusions of the paper.

We agree with the reviewers that the original set of figures has a high degree of complexity. Hence, we have redesigned Figures 3 to 10 to make them easier to understand and to better highlight the key results and conclusions of this work. Following the reviewer's recommendation, we now have only one figure (i.e., Figure 6, which contains CRPSS values) presenting the results for all models, all (12) calibration objective functions and hindcast initializations. Additionally, we have improved the methodology flow chart, and offer more detailed explanations in the text about the approach and results.

A few related suggestions:

Fig. 4 and section 4.1: I would have put this section at the end of the results section to illustrate the results. This figure could be reduced in terms of results by only keeping one model and two OFs (a popular OF and KGE(Q)+NSE(log(Q)))

We appreciate the reviewer's suggestion. However, we believe that it is important to illustrate first the individual basin calibration (Figure 3a) and the hindcast generation/verification (Figure 3b) steps in one case study basin. The extension of steps 3a and 3b (Figure 3) to the 22 case study catchments is the basis for subsequent analyses aiming to explore connections between seasonal hindcast performance, the hydrological consistency of streamflow simulations obtained with the various calibration metrics, and catchment characteristics. We now clarify this at the beginning of section 3.

Additionally, we have simplified the original Figure 8 (as also suggested by Reviewer #1) and placed it before the original Figure 4 (now Figure 5). These results are included in Section 4.1, which has been renamed to "Example: hydrologic model calibration and ESP at the Upper Maipo River basin". Nevertheless, we fully agree that the information content of the original Figure 4 was excessive, and have hence simplified it from a 6 x 3 panel figure, to a 3 x 2 panel figure. We now show results from one model structure (TUW model) and three calibration metrics: NSE (a popular OF as the reviewer suggests), KGE(Q)+NSE(log(Q)), and VE-Sep, due to their relevance in the subsequent results and findings.

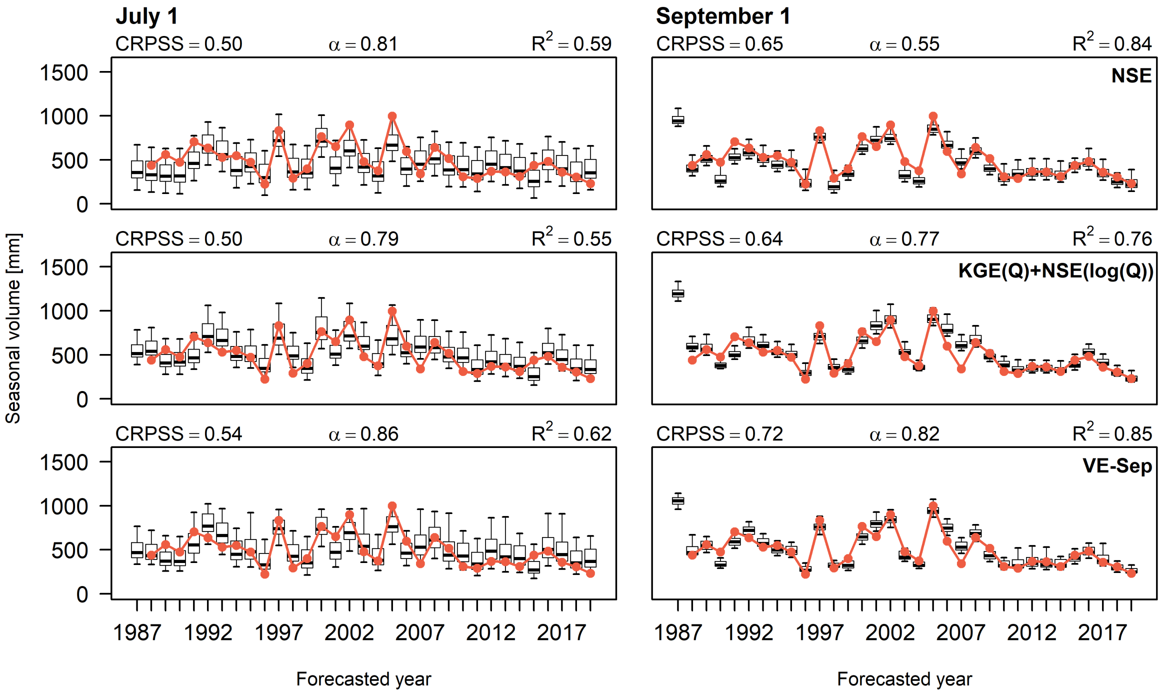

Figure 5. Time series with ESP seasonal hindcasts (i.e., September-March runoff) initialized on July 1 (left panels), and September 1 (right panels) for the Maipo at El Manzano basin. The boxes correspond to the interquartile range (IQR, i.e., 25th and 75th percentiles); the horizontal line in each box is the median, whiskers extend to the ±1.5·IQR of the ensemble, and the red dots represent the observations. The results were produced with the TUW model, using parameters obtained from calibrations conducted with NSE, KGE(Q)+NSE(log(Q)) and VE-Sep (see details in Section 3.1). Each panel displays the CRPSS, the reliability index α, and the coefficient of determination R2 (computed using the ensemble forecast median).

Fig. 5: to extend in height to enhance visualisation.

We have modified this figure (now Figure 6) following the reviewer's suggestion. We have also changed the colours to enhance visualization.

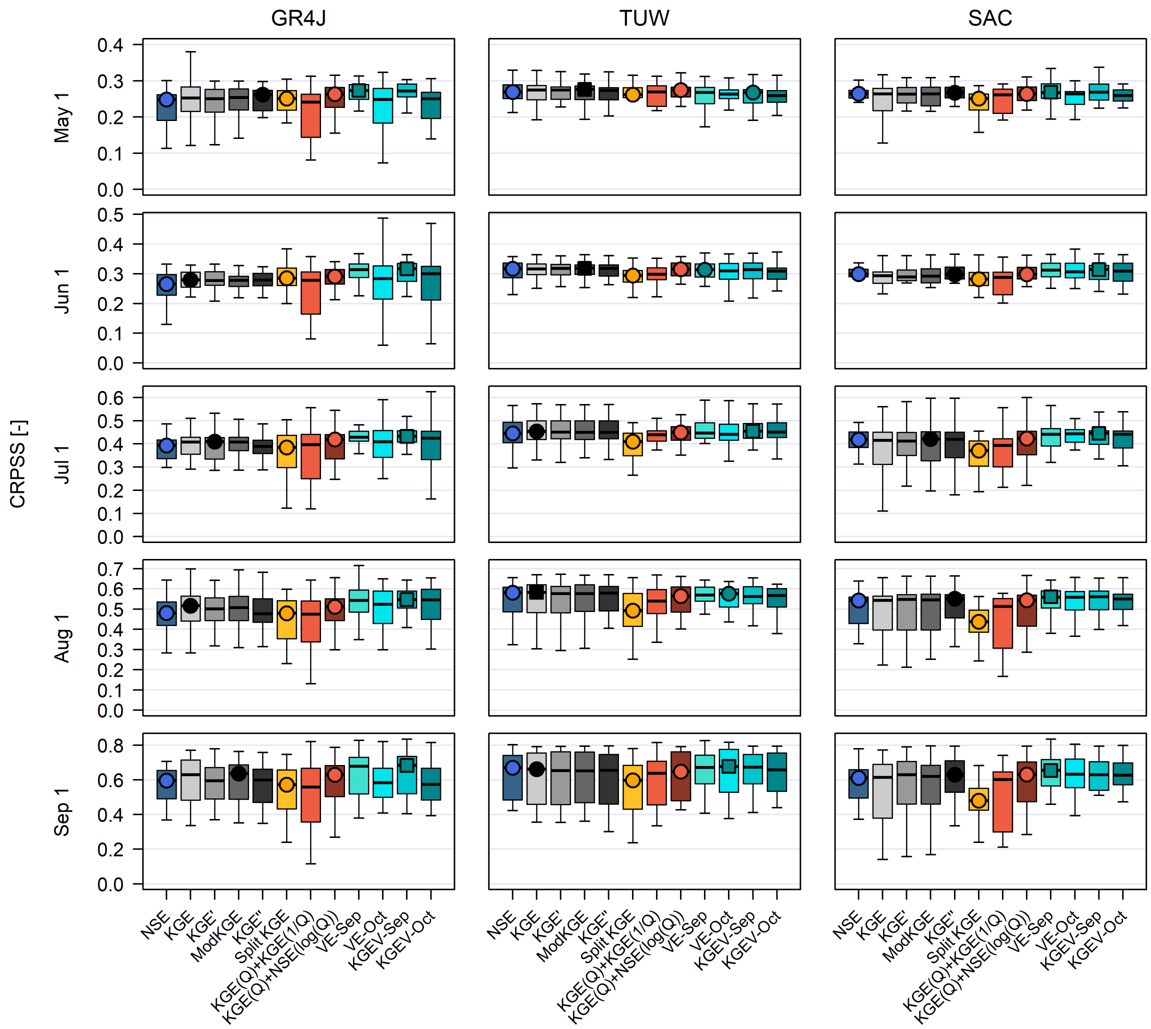

Figure 6. Comparison of CRPSS obtained with different calibration objective functions. Each panel contains results for a specific combination of initialization time (rows) and hydrological model (columns), and each boxplot comprises results from the 22 case study basins. The boxes correspond to the interquartile range (IQR, i.e., 25th and 75th percentiles), the horizontal line in each box is the median, and whiskers extend to the ±1.5·IQR of the ensemble. The circle indicates the objective function providing the highest median within each family of calibration metric (identified with different colors), and the square indicates the objective function that delivers the best set of metric values using a specific combination of initialization time and hydrological model.

Fig. 6: to move to the supplementary materials.

We have moved this figure to the Supporting Information section, following the reviewer's recommendation.

Fig. 7: show only three OFs (pick the most relevant to highlight your conclusions), May 1 and Sep 1 for initialization times and two forecast criteria.

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. We believe that the results from one model and one calibration metric are enough to communicate the same findings obtained for the remaining representative OFs 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) will be included in the Supporting Information.

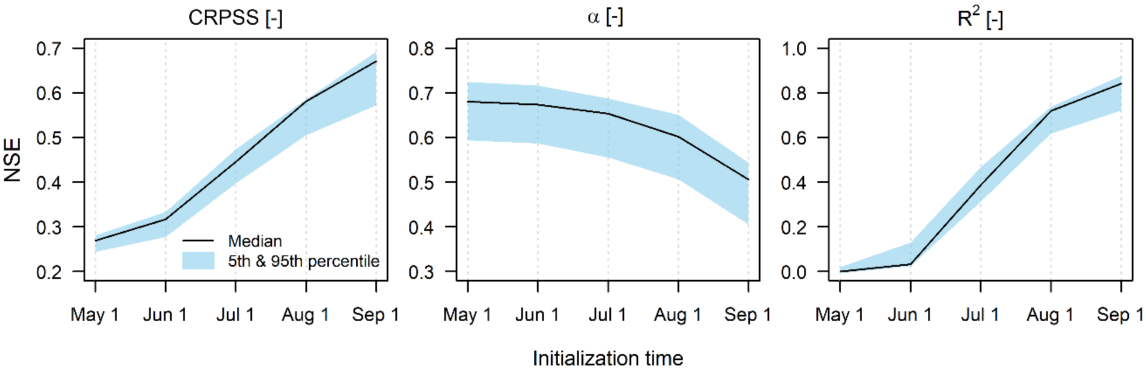

Figure 1. Impact of initialization time on hindcast verification metrics for NSE calibration objective function and the TUW model. The shades represent the 5th and 95th percentiles in each metric from the 22 case study basins, and the solid line represents the median of each metric.

Fig. 8: keep only two OFs and present the daily hydrographs for two years in one of the validation periods.

In response to the recommendations provided by reviewers #1 and #2, 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). Hence, We show the daily hydrographs for two water years from the evaluation data set, and the runoff seasonality curves are displayed jointly for these two periods. We decide to show results for three representative objective functions (the same as in Figure 5), and place this figure before.

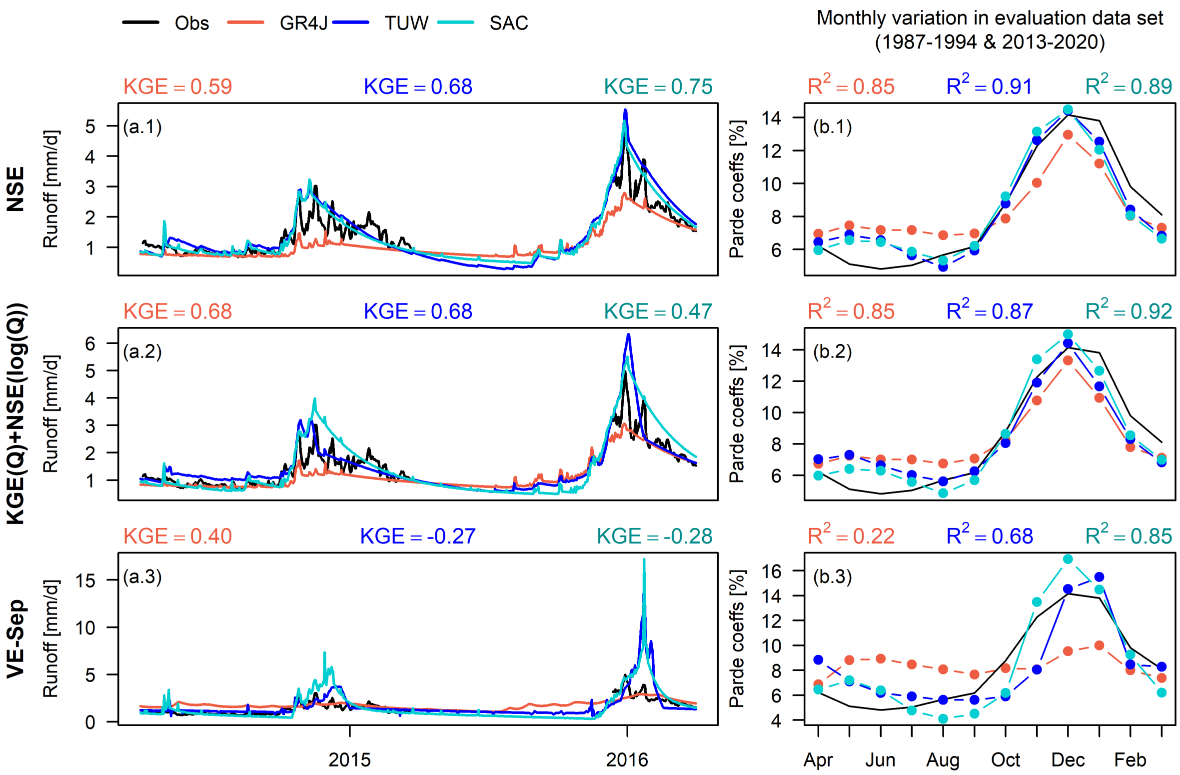

Figure 4. (Left) Daily hydrographs (April/2009 – March/2011) and (right) monthly variation curves for evaluation data set (April/1987 – March/1994 and April/2013 – March/2020) at the Maipo en el Manzano River basin, obtained with the three models and three objective functions: (1) NSE, (2) KGE(Q)+NSE(log(Q)) 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.

Fig. 9: keep only two or three OFs and extend the figure in height.

We have extended the figure (now numbered as Figure 8) in height and have simplified it by removing one of the signatures (F9M, since it contained very similar information of high flows than FHV), and showing the results from only one model (TUW). Nevertheless, we decided to keep the five representative objective functions for consistency with Figure 6, which highlights them as the best performing OFs per family. Additionally, we now show percent biases in hydrological signatures for the calibration period and for the combined evaluation data set, which combines the periods April/1987 – March/1994 and April/2013 – March/2020.

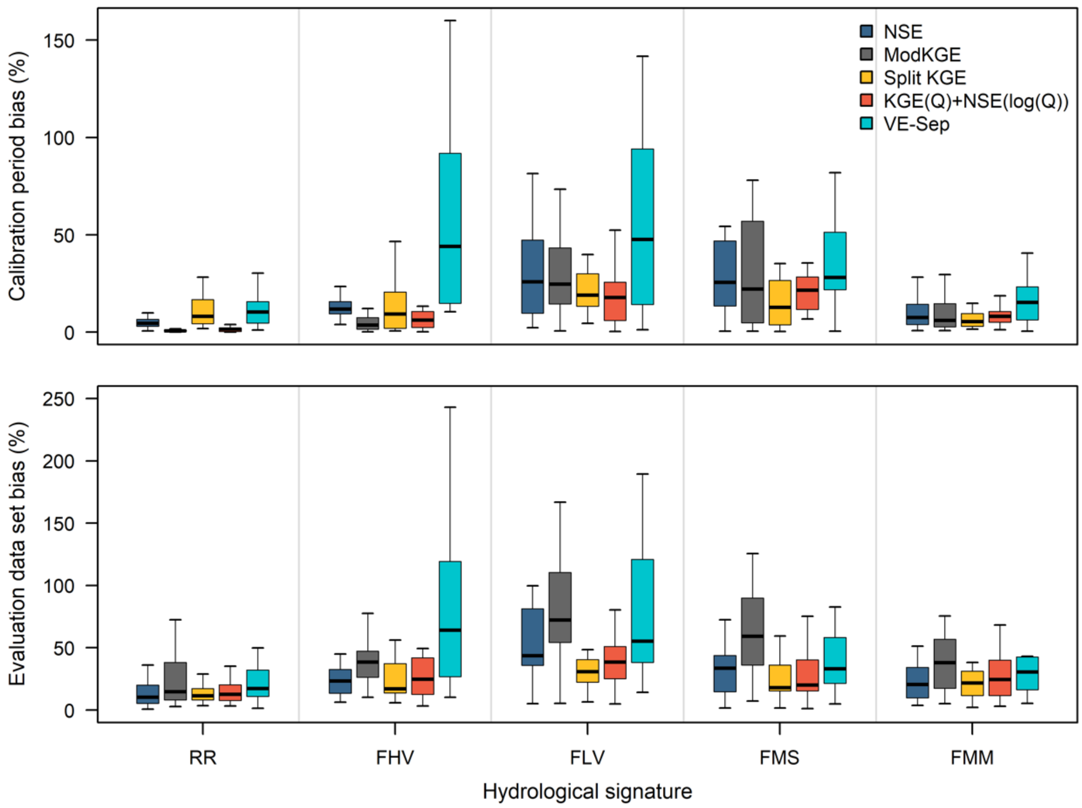

Figure 8. Percent biases in hydrologic signatures (x-axis) obtained with the five representative objective functions and the three hydrological models (shown in different columns) for the (upper row) calibration (April/1994 – March/2013) and (bottom row) evaluation data set (April/1987 – March/1994 and April/2013 – March/2020). Each boxplot comprises results for our 22 case study basins. The boxes correspond to the interquartile range (IQR, i.e., 25th and 75th percentiles), the horizontal line in each box is the median, and whiskers extend to the ±1.5·IQR of the ensemble.

Fig. 10: remove the alpha and bias lines. Show only KGE or NSE for popular OFs.

We have modified this figure (now numbered as Figure 9) following the reviewer's suggestion. We have also removed the results from the KGE(Q)+KGE(1/Q) objective function because it did not add any relevant information relative to what is illustrated with the remaining objective functions.

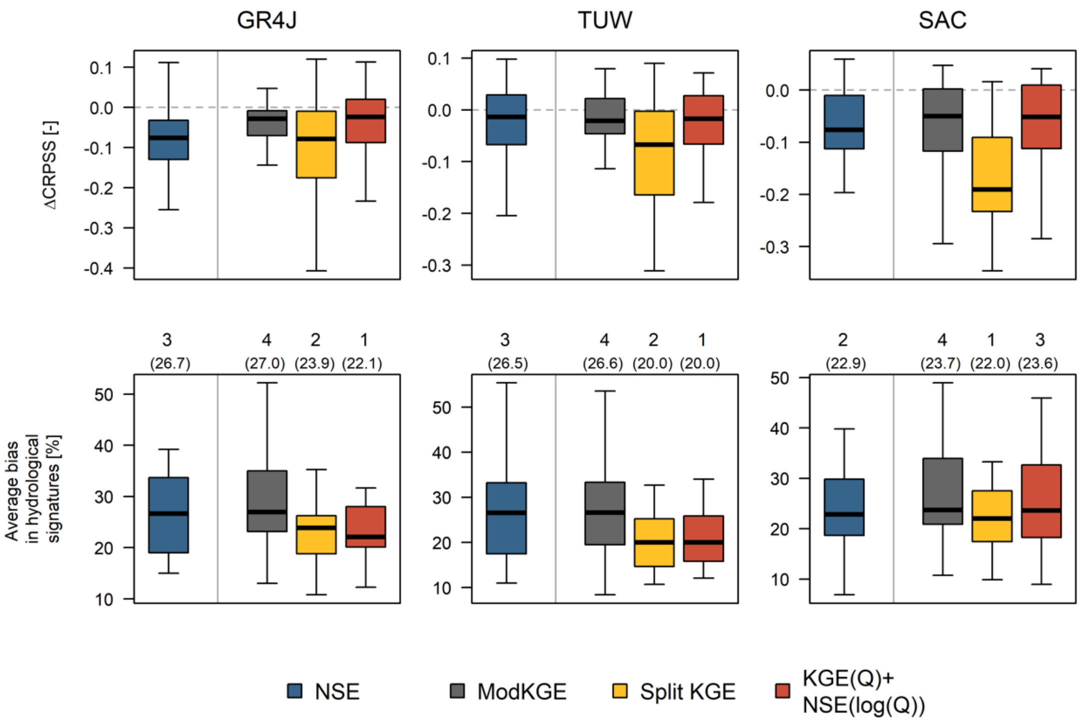

Figure 9. Variations in September 1 CRPSS (top panels) due to the choice of popular and alternative objective functions (shown in different boxplots), relative to the best performing OF in terms of forecast quality (VE-Sep). The dashed line indicates no difference (i.e., no loss) in forecast performance. The bottom panels display the average bias in hydrological signatures (computed over the calibration period and evaluation data set) with the associated ranking (being 1 the best in terms of hydrological consistency), and median average bias obtained from the sample of basins (in parentheses).

I did not understand why the ESP forecasts were evaluated on both calibration and evaluation periods without separation in the result analyses. Even if there are 32 meteorological inputs for each forecasted year, and that they are different from the meteorological input of the forecasted year, the streamflow data used to calculate seasonal performance has already been "seen" by the model during calibration. Since one of the goals of this study is to evaluate the impacts of calibration on seasonal forecasts, I think the authors should consider the evaluation of forecast quality only on the evaluation periods (or better explain why it was done that way). 33-1 years of meteorological data can still be used for the ensemble forecasts of each forecasted year. To improve the analysis of the hydrological consistency of model simulations, the authors could perform a split-sample test.

We agree with the reviewer in that, if the hindcasted year was also used for model calibration, the model has already "seen" meteorological inputs, even if the climate time series observed during that year is excluded from the generation of ESP hindcasts. However, the limited sample size has been widely recognized as a limitation in the seasonal forecasting literature (e.g., Shi et al., 2015; Trambauer et al., 2015; Mendoza et al., 2017; Lucatero et al., 2018; Wood et al., 2018). so we decided to take advantage of the entire 33-year period for hindcast verification. Further, we consider a sample size of 14 observations (i.e., the number of WYs used for model evaluation) to be insufficient for a robust assessment of seasonal hindcast quality. To demonstrate this point and to address the reviewer's concerns, we characterized the impact of sample size on the spread of CRPSS results by performing a bootstrap analysis with 1000 realizations for the Maipo River basin, using hindcasts produced with the TUW model and KGE(Q)+NSE(log(Q)) as the calibration metric (Figure 10). The analysis was conducted for the following verification samples: (a) full period (i.e., 33 WYs) using the same parameter set, obtained by calibrating the model with data from the period April/1994 – March/2013 (blue); (b) full period, using parameter sets re-calibrated with all data except the hindcasted year (i.e., 33 parameter sets to produce 33 seasonal hindcasts, gray); (c) 19 WYs (calibration periods), using a single parameter set obtained with data from the same period (red); (d) 14 WYs (i.e., evaluation data set April/1987 – March/1994 and April/2013 – March/2020), using the same parameter set as in case (c) (orange); and (e) 14 WYs (April/2006 – March/2020), using the same parameter set as in case (c) (cyan).

The results in Figure 10 show a considerable spread in CRPSS arising from sampling uncertainty when using 14-year verification periods (orange and cyan boxes). Additionally, the median CRPSS results are lower than those obtained with 19 and 33 WYs in July 1, August 1 and September 1 (i.e., when predictability from land conditions increases). An interesting result is the similarity of CRPSS values obtained with scenarios (a) and (b), suggesting that the hindcasting generation and verification approach adopted here (i.e., using a single parameter set) is a good proxy to characterize the hindcast quality that would be obtained with an operational setup that considers parameter re-calibration. We thank the reviewer for this observation, and will include these results in the discussion section of the revised manuscript.

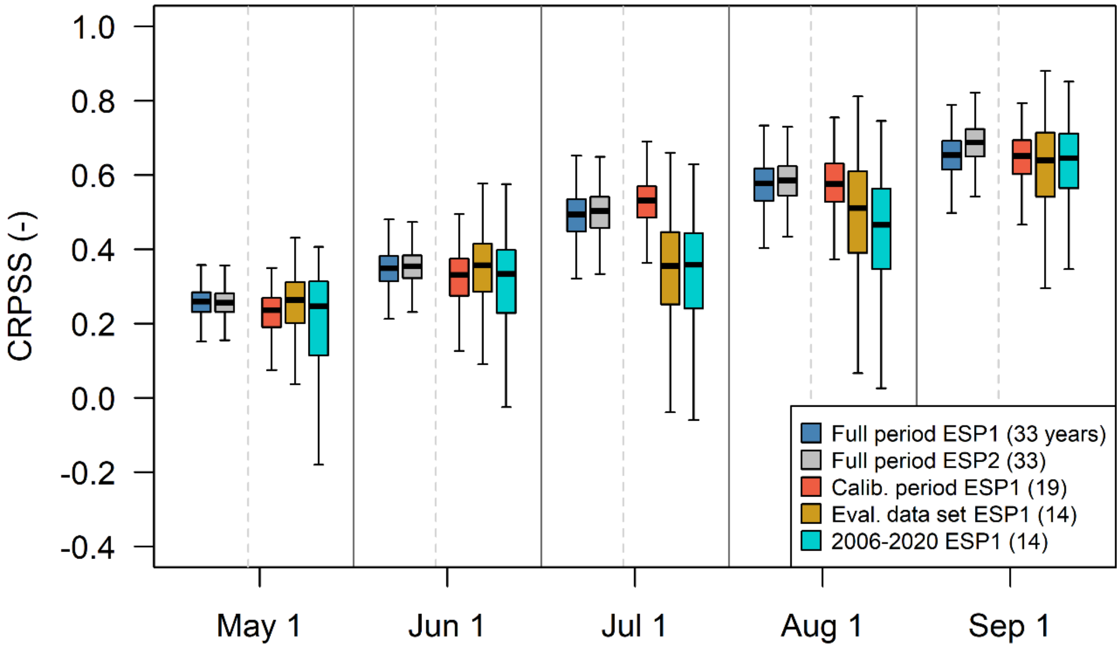

Figure 10: Comparison of CRPSS values for seasonal (i.e., September-March) streamflow hindcasts produced at the Maipo River basin with the TUW model and KGE(Q)+NSE(log(Q)) as calibration metric. Each box comprises results from 1000 bootstraps with replacement applied to different verification sample sizes (i.e., number of hindcast-observation pairs): (a) full period (i.e., 33 WYs) using the same parameter set, obtained by calibrating the model with data from the period April/1994 – March/2013 (blue); (b) full period, using parameter sets re-calibrated with all data except the hindcasted year (i.e., 33 parameter sets to produce 33 seasonal hindcasts, gray); (c) 19 WYs (calibration periods), using a single parameter set obtained with data from the same period (red); (d) 14 WYs (i.e., evaluation data set April/1987 – March/1994 and April/2013 – March/2020), using the same parameter set as in case (c) (orange); and (e) 14 WYs (April/2006 – March/2020), using the same parameter set as in case (c) (cyan). The boxes correspond to the interquartile range (IQR, i.e., 25th and 75th percentiles); the horizontal line in each box is the median, and the whiskers extend to the $\pm 1.5 \cdot IQR$ of the ensemble.

Lines 21 and 435, the term "hydrologically consistent parameter set" is used. No analyses of parameter sets were made in this study, only the ability of the models to reproduce streamflow signatures was evaluated. A high performance for specific streamflow signatures may imply more consistency in simulating streamflow than using a "popular" metric. However, I would argue that it does not necessarily imply that the parameter sets of the models are more hydrologically consistent, *as a model can be wrongly parameterized and the hypotheses behind not fit for the studied catchments. Even when a specific parameter set leads to better signature performance, equifinality of parameters can remain high, especially if the parameter sets yielding good performance vary between periods. As the manuscript already includes many results, I suggest considering a small analysis of the parameter sets of one of the models (e.g. TUWmodel that seems to be giving the highest forecast quality). This analysis would only be conducted for three objective functions (the one associated with the lowest hydrological consistency, the one associated with the highest hydrological consistency and KGE(Q)+NSE(log(Q)) which is the best compromise between forecast quality and hydrological consistency). In TUW model, not all the parameters would need to be assessed but, for instance, only the ones related to baseflow (in the TUWmodel package, it would be "param k2") or/and snow, to relate the results to catchment attributes that have a strong correlation with forecast quality.*

We agree with the reviewer's appreciation and, therefore, we have removed any references to "hydrologically consistent parameter sets" in the revised manuscript, replacing by "hydrologically consistent simulations", as the reviewer suggests below. Despite we recognize that parameter equifinality can be substantial, characterizing such effects is out of the scope of this study. Recently, Muñoz-Castro et al. (2023) examined the effects of calibration metric selection and parameter equifinality on the level of (dis)agreement in parameter values across 95 catchments in Chile, finding that (i) the choice of objective function has smaller effects on parameter values in catchments with low aridity index and high mean annual runoff ratio, in contrast to dryer climates, and (ii) catchments with better parameter agreement also provide better performance across model structures and simulation periods. Future work could explore whether such performance in streamflow simulations translates well into seasonal forecast quality attributes. We will make these points in section 5.5 ("Limitations and future work") of the revised manuscript.

The variations of parameters between periods could then be evaluated (if you were to follow the previous comment about periods of calibration and evaluation).

We appreciate the reviewer's sentiment. Nevertheless, the assessment of temporal stability in hydrological model parameters is itself a topic worth of detailed investigation (e.g., Merz et al., 2011; Coron et al., 2012; Gharari et al., 2013; Fowler et al., 2018; Duethmann et al., 2020) that we prefer to leave for future work. We will make this point in section 5.5 ("Limitations and future work") of the revised manuscript.

Of course, this would need to be considered after a reorganisation of the manuscript (to better highlight the results already presented). It should not come in the same format as the other results. The content of the manuscript (in terms of results) needs to be reduced first. That being said, the publications needs some extra work. The main points that need attention are argumentation for hydrological model aggregation, the structure of text and figures, additional reflection on the meaning of study results, and the archiving of code and data.

We have re-organized and re-designed most of the figures, reducing the amount of results included in the original submission, and we have revised the text accordingly. The data and codes used to produce the results presented here are correctly archived in a Zenodo repository (see our response below).

Minor comments:

The link to the Zenodo repository does not seem to be working anymore (last checked: 30/06).

The Zenodo repository is working again (last checked: August 15, 2023). We attach the message from Zenodo Support.

Dear Diego

In Zenodo we have automated mechanisms to block spam, however this system can sometimes make mistakes that lead to an issue with user blocking. We apologize for the inconvenience it has caused. Your account has been fully restored, and we have taken the appropriate measures to ensure it won't happen again.

Best regards,
Lars

Zenodo Support
https://zenodo.org

Lines 27 and 86: the term "for the right reasons" is a bit strong, as no additional data to streamflow was used for model evaluation. I suggest using "more hydrologically consistent simulations" as you did in the remaining of the paper.

We have deleted any reference to "the right reasons", and now refer to "hydrologically consistent simulations", following the reviewer's recommendation.

Line 108: how were the seasonal snowmelt contributions calculated?

We have deleted any reference to snowmelt contributions, since we did not estimate these. What we actually meant for criteria (iv) is that the selected basis have the requirement of snowmelt influence on runoff seasonality (i.e., they must have snow-driven, nivo-pluvial or pluvio-nival regimes, as described by Baez-Villanueva et al., 2021). We have modified the text in Section 2 to clarify this.

Line 155: the CemaNeige model also partitions total precipitation into liquid and solid precipitation. Liquid precipitation and snowmelt are fed to the soil moisture store.

In response to the reviewer's observation, we have re-worded the text in section 3.1.1 as follows:

"(…) The CemaNeige module first partitions total precipitation into liquid and solid, and then simulates snow accumulation and melt over five or more (user-defined; here we use 10) elevation bands, using a two-parameter degree-day based scheme (Valéry et al., 2014) that adds snowmelt and liquid precipitation to the soil moisture accounting reservoir. (…)".

Line 158: the GR4J model also includes a non-conservative function for water exchanges between topographical catchments.

In response to the reviewer's observation, we have added the following text in section 3.1.1:

"A groundwater exchange term acts on both flow components to represent water exchanges between topographical catchments."

Line 160: what do you mean by response area?

We meant response routine. We have modified the text to clarify.

Sect 3.1.1: were different elevation bands considered in TUWmodel and SNOW17?

No. These models were implemented in a lumped fashion. We clarify this at the end of section 3.1.1: "While the CemaNeige is configured with 10 elevation bands, the snow routines of TUW and SAC-SMA (i.e., SNOW-17) are implemented in a lumped fashion."

Sect 3.1.1: the three models used in this study were implemented within the R environment. These models and their exact implementation were described in (Astagneau et al., 2021; https://doi.org/10.5194/hess-25-3937-2021). In addition, the structures of the models are compared using a unified representation of the different storages and fluxes (Fig. 1). If, and only if, you used *this paper to choose, implement or understand these models, you should cite it. Otherwise, please ignore this comment.*

We did not select the model structures based on the paper mentioned by this reviewer. Instead, the models used here were selected because they are widely used by the hydrology community (Addor and Melsen, 2019), with a myriad applications to streamflow forecasting. For example, SAC-SMA has been applied for testing alternative approaches (e.g., Mendoza et al., 2017), and is used to produce operational streamflow forecasts in the US (Micheletty et al., 2021). GR4J has been applied to assess streamflow forecasting frameworks in large samples of catchments (e.g., Harrigan et al., 2018; Woldemeskel et al., 2018). HBV-like conceptual models have been used to assess short (e.g., Pauwels and De Lannoy, 2009; Verkade et al., 2013) to long (e.g., Peñuela et al., 2020) range streamflow forecasts, especially in European countries. This explanation is included in section 3.1.1.

Fig. 3 and sect. 3.1.2: I am not sure there is any validation of the models made in this study. For me, validation means that you are choosing one model over the other (or rejecting one) and evaluation means that you are evaluating the models outside the calibration period. Please consider replacing "validation" by "evaluation".

In response to the reviewer's observation, we have replaced the term "validation" by "evaluation" in the text and the figures.

Sect 3.2: it could be useful to add a more detailed presentation of the ESP method (for instance by adding a reference to Fig. 3 of Crochemore et al., 2020; https://doi.org/10.1029/2019WR025700) and extend Fig. 3 that really helps to understand your framework.

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

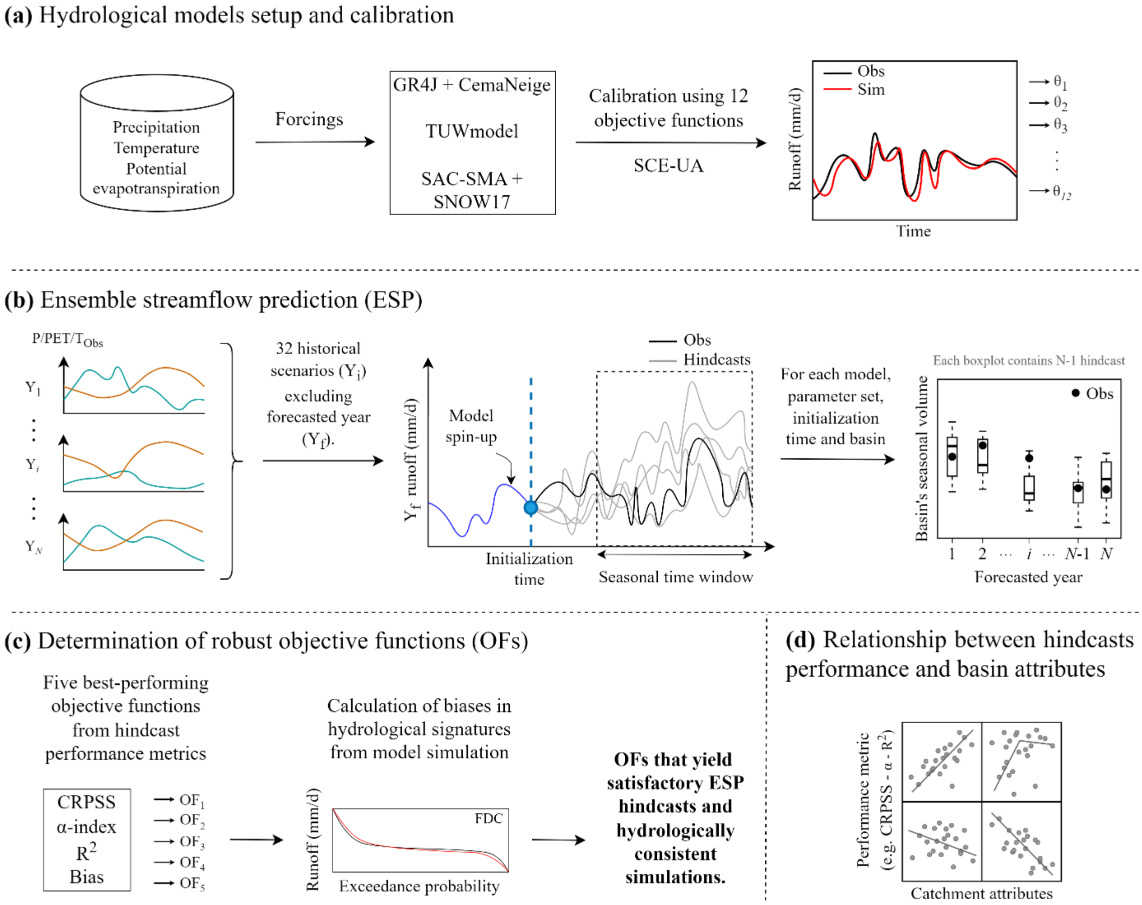

Figure 3. Flowchart describing the approach used in this study. See text for details.

References

Addor, N. and Melsen, L. A.: Legacy, Rather Than Adequacy, Drives the Selection of Hydrological Models, Water Resour. Res., 55(1), 378–390, doi:10.1029/2018WR022958, 2019.

Baez-Villanueva, O. M., Zambrano-Bigiarini, M., Mendoza, P. A., McNamara, I., Beck, H. E., Thurner, J., Nauditt, A., Ribbe, L. and Thinh, N. X.: On the selection of precipitation products for the regionalisation of hydrological model parameters, Hydrol. Earth Syst. Sci., 25(11), 5805–5837, doi:10.5194/hess-25-5805-2021, 2021.

Coron, L., Andréassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M. and Hendrickx, F.: Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments, Water Resour. Res., 48, W05552, doi:10.1029/2011WR011721, 2012.

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.

Duethmann, D., Blöschl, G. and Parajka, J.: Why does a conceptual hydrological model fail to correctly predict discharge changes in response to climate change?, Hydrol. Earth Syst. Sci., 24(7), 3493–3511, doi:10.5194/hess-24-3493-2020, 2020.

Fowler, K., Coxon, G., Freer, J., Peel, M., Wagener, T., Western, A., Woods, R. and Zhang, L.: Simulating Runoff Under Changing Climatic Conditions: A Framework for Model Improvement, Water Resour. Res., 54(12), 9812–9832, doi:10.1029/2018WR023989, 2018.

Gharari, S., Hrachowitz, M., Fenicia, F. and Savenije, H. H. G.: An approach to identify time consistent model parameters: Sub-period calibration, Hydrol. Earth Syst. Sci., 17(1), 149–161, doi:10.5194/hess-17-149-2013, 2013.

Harrigan, S., Prudhomme, C., Parry, S., Smith, K. and Tanguy, M.: Benchmarking ensemble streamflow prediction skill in the UK, Hydrol. Earth Syst. Sci., 22(3), 2023–2039, doi:10.5194/hess-22-2023-2018, 2018.

Lucatero, D., Madsen, H., Refsgaard, J. C., Kidmose, J. and Jensen, K. H.: Seasonal streamflow forecasts in the Ahlergaarde catchment, Denmark: The effect of preprocessing and post-processing on skill and statistical consistency, Hydrol. Earth Syst. Sci., 22(7), 3601–3617, doi:10.5194/hess-22-3601-2018, 2018.

Mendoza, P. A., Wood, A. W., Clark, E., Rothwell, E., Clark, M. P., Nijssen, B., Brekke, L. D. and Arnold, J. R.: An intercomparison of approaches for improving operational seasonal streamflow forecasts, Hydrol. Earth Syst. Sci., 21(7), 3915–3935, doi:10.5194/hess-21-3915-2017, 2017.

Merz, R., Parajka, J. and Blöschl, G.: Time stability of catchment model parameters: Implications for climate impact analyses, Water Resour. Res., 47(2), W02531, doi:10.1029/2010WR009505, 2011.

Micheletty, P., Perrot, D., Day, G. and Rittger, K.: Assimilation of Ground and Satellite Snow Observations in a Distributed Hydrologic Model for Water Supply Forecasting, J. Am. Water Resour. Assoc., doi:10.1111/1752-1688.12975, 2021.

Muñoz-Castro, E., Mendoza, P. A., Vásquez, N. and Vargas, X.: Exploring parameter (dis)agreement due to calibration metric selection in conceptual rainfall-runoff models, Hydrol. Sci. J., doi:10.1080/02626667.2023.2231434, 2023.

Pauwels, V. R. N. and De Lannoy, G. J. M.: Ensemble-based assimilation of discharge into rainfall-runoff models: A comparison of approaches to mapping observational information to state space, Water Resour. Res., 45(8), W08428, doi:10.1029/2008WR007590, 2009.

Peñuela, A., Hutton, C. and Pianosi, F.: Assessing the value of seasonal hydrological forecasts for improving water resource management: insights from a pilot application in the UK, Hydrol. Earth Syst. Sci., 24(12), 6059–6073, doi:10.5194/hess-24-6059-2020, 2020.

Shi, W., Schaller, N., MacLeod, D., Palmer, T. N. N. and Weisheimer, A.: Impact of hindcast length on estimates of seasonal climate predictability, Geophys. Res. Lett., 42(5), 1554–1559, doi:10.1002/2014GL062829, 2015.

Trambauer, P., Werner, M., Winsemius, H. C., Maskey, S., Dutra, E. and Uhlenbrook, S.: Hydrological drought forecasting and skill assessment for the Limpopo River basin, southern Africa, Hydrol. Earth Syst. Sci., 19(4), 1695–1711, doi:10.5194/hess-19-1695-2015, 2015.

Valéry, A., Andréassian, V. and Perrin, C.: 'As simple as possible but not simpler': What is useful in a temperature-based snow-accounting routine? Part 2 – Sensitivity analysis of the Cemaneige snow accounting routine on 380 catchments, J. Hydrol., 517(Supplement C), 1176–1187, doi:https://doi.org/10.1016/j.jhydrol.2014.04.058, 2014.

Verkade, J. S., Brown, J. D., Reggiani, P. and Weerts, A. H.: Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales, J. Hydrol., 501, 73–91, doi:10.1016/j.jhydrol.2013.07.039, 2013.

Woldemeskel, F., McInerney, D., Lerat, J., Thyer, M., Kavetski, D., Shin, D., Tuteja, N. and Kuczera, G.: Evaluating post-processing approaches for monthly and seasonal streamflow forecasts, Hydrol. Earth Syst. Sci., 22(12), 6257–6278, doi:10.5194/hess-22-6257-2018, 2018.

Wood, A. W., Sankarasubramanian, A. and Mendoza, P.: Seasonal Ensemble Forecast Post-processing, in Handbook of Hydrometeorological Ensemble Forecasting, pp. 1–27, Springer Berlin Heidelberg, Berlin, Heidelberg, Heidelberg., 2018.