FROSTBYTE: A reproducible data-driven workflow for probabilistic seasonal streamflow forecasting in snow-fed river basins across North America

Response to RC1

In this paper, the authors have presented a data-driven workflow for ensemble seasonal streamflow forecasting using snow water equivalent as predictors. The findings offer valuable insights relevant to various stakeholders, such as forecasters and decision-makers, effectively merging scientific precision with practical workflow development insights. The subject matter is of current interest and contributes insights to hydrological forecasting, both the workflow, and the knowledge of the predictability of streamflow from late-season snowpack. For a deeper comprehension of the study, I propose additional discussions with the authors, detailed below.

We thank the reviewer for their positive and constructive comments on our manuscript. Their comments are copy-pasted below verbatim in black, and our responses are underneath each comment in blue.

Line 169 and 206, could you clarify whether an independent regression model is employed for each target period at every initialization date, or if a single model is capable of generating multiple outputs for all target periods at the same initiation date?

Hindcasts are indeed generated using an independent regression model for each river basin, initialization date, target period, and year left out. This is mentioned in Section 2.2.4, on L224-225 of the initial manuscript: "We conduct a PCA and fit a new model for each predictor-predictand combination".

We have clarified this point in the text: "An independent regression model is used to produce an ensemble hindcast for each river basin, initialization date, target period, and year left out".

Line 221, could you provide the information of the average explained variance of the PCs?

We agree that this would be useful information for the readers and will add this information in the revised manuscript.

Line 293, would be interested to know what are general reliability of other methods, to better understand how much improvement the proposed method obtains.

Comparing our reliability results to the reliability of other methods is not straightforward given the variety of approaches used to evaluate reliability in the literature. However, our results could be compared to those from Mendoza et al. (2017), who compared the reliability index of statistical, process-based and hybrid methods for seasonal streamflow forecasting at five case study sites across the USA Pacific Northwest region. They found that, overall, the statistical methods yielded the most reliable hindcasts.

We have added some more detail about the reliability index values obtained with the various methods in the revised manuscript: "For five case study sites across the USA Pacific Northwest, their regression-based methods achieved reliability index values ranging between 0.6 and nearly 1, while the reliability of the process-based ensemble streamflow prediction (ESP) hindcasts declined when approaching the April 1st initialization date, with an overall reliability index ranging between 0.4 and 0.9. Our approach yielded reliability index values comparable to those obtained from the statistical methods developed by Mendoza et al. (2017)".

Line 300, please considering to rescale the y axis to make the difference more visible.

We have changed the figure's y-axis limits so that results are clearer - thank you.

Line 310, here the authors mentioned the limitation of comparing between systems with different ensemble members. Would it be more comparable to use fairCRPSS instead?

We thank the reviewer for this idea and will calculate the fair-CRPSS as well to compare results against the standard CRPSS, prior to adding these results to the manuscript.

Line 321, two "and" here.

Thank you for catching this - we have corrected this in the revised manuscript.

Line 323, please specify which basins are the ones that display low to no skill throughout all initialization dates. And does this refer back to Fig.6 since there is no initialization date information in Fig.7.

As described in the text, basins with low to no skill are situated in the northwest and in the east. We will however add a few basin names to add more in-depth information to the manuscript.

We have additionally removed the sentence "On the other hand, some basins display low to no skill throughout all initialization dates" on L323-324, as this was a repetition of the first sentence of this paragraph on L319-320. Further, as pointed out by the reviewer, initialization dates are not shown on Fig. 7 (at least not explicitly).

We now also provide some clarifying information on the lead months displayed on Fig. 7 in the revised manuscript, in both the text and the caption respectively: "Note that the lead months are different from the initialization dates of the hindcast, where lead month refers to the number of months between the hindcast initialization date and the target period start", and "Note that initialization dates and periods of interest are not shown explicitly here. For instance, the first map, showing the CRPSS for hindcasts with zero months lead time, will include results from hindcasts of January 1st to September 30th initialized on January 1st, as well as from hindcasts of February 1st to September 30th initialized on February 1st, etc. On the other hand, the last map, showing the CRPSS for hindcasts with six months lead time, will include results from hindcasts of July 1st to September 30th or later, initialized on January 1st or later".

Line 332, an additional interesting pattern from this figure is that the peak skill for each target period appears when the initialization month is at the start of the target period, e.g. the target period of April to September, the peak occurs when initialized at April 1st.

This is indeed a noteworthy finding which we mention, albeit differently, on L332-334: "For all target periods/periods of interest, the ROC AUC increases with later initialization dates, reaching a peak with hindcasts generated on May 1st/June 1st. This suggests that the potential usefulness of these hindcasts increases as lead time decreases. It is however not the case for hindcasts for July 1st to September 30th, where the ROC AUC decreases after June 1st."

We have rephrased so it is clearer and the revised text now reads: "For most target periods/periods of interest, peak ROC AUC is obtained for hindcasts with zero months lead time. For example, the ROC AUC of hindcasts for May 1st to September 30th is the highest when the hindcasts are initialized on May 1st. The ROC AUC, and therefore the potential usefulness, of most hindcasts decreases with increasing lead time. This is however not the case for hindcasts for July 1st to September 30th, where the ROC AUC is highest when hindcasts are initialized on average on May 1st (with two months lead time)".