Dear Reviewers,

Thank you for your valuable feedback. Please find below your referee comments (in black) and our responses (in blue).

With regards, the authors.

RC: Reviewer #1

Review of "The Value of Hydroclimatic Teleconnections for Snow-based Seasonal Streamflow Forecasting in Central Asia" (revised manuscript) by Umirbekov et al.:

This applied research article synthesizes and integrates a variety of methods and data types to create a convincing set of seasonal water supply forecast models for several rivers in Central Asia. The work is innovative, well-thought-out, technically sound, and socially relevant given its potential to support water management in an under-served region that would benefit from this kind of operational forecast system. Free public availability of the author's code and all the required input data means the forecast model could be quickly and easily implemented in production systems. The paper additionally presents a number of original research outcomes that should inform other studies around interactions between snowpack data, atmosphere-ocean circulation patterns, water supply forecasting, and practical machine learning systems.

In my opinion, this paper will be a strong contribution to HESS, and I recommend publication as-is (or with some very minor additional revisions). The revised article fully addresses the review comments I provided on the original submission, and I have only a few small additional suggestions to make here:

We appreciate your overall feedback on the revised manuscript.

• First sentence of the abstract: the phrasing is awkward – I suggest improving this to read something more like, "Due to the long memory of snow processes, statistically based seasonal streamflow prediction models in snow-dominated catchments can successfully leverage, but also typically rely on, snowpack estimates."

We have amended the sentence accordingly.

• Lines 36-44: This is much-improved over the original manuscript but would still benefit from a little additional work. A few points the authors might bear in mind: both process-simulation and data-driven water supply forecast models can ingest seasonal to subseasonal climate forecasts as input (see for example Lehner et al., Geophysical Research Letters, 2017, https://doi.org/10.1002/2017GL076043); process-based models have the advantage of explicitly capturing process physics, enhancing credibility and interpretability; data-driven models have the advantage of not requiring assumptions about the relevant physics and how to represent it in a pragmatic computational model. Rewriting the passage just a little more to acknowledge these points would lend greater credibility to the article overall.

We have amended the paragraph with additional excerpt along those lines (lines 44-48 in the track-changes version of the manuscript)

• Lines 176-186: it might be worth explicitly mentioning here that while several models were used in the ensemble, each of them individually is quite simple and parsimonious, with a single target variable (seasonal discharge volume) and three or fewer input variables (as subsequently noted on lines 215-216). That means there's only a small number of parameters to estimate from the limited sample size. In other words, the methodology used here is suitable for application to short datasets. There is some precedent in water supply forecast modeling for deliberately fine-tuning statistical and machine learning architectures to maximize parsimony and minimize the number of parameters to be estimated, enabling application to short

datasets with good out-of-sample performance, as well as improved regularization and geophysical explainability (some examples are Fleming et al., 2021, 2024, which are already cited in the manuscript).

Thank you for this suggestion. We have amended the Method section with additional note (lines 261-263 in the track-changes version of the manuscript).

• I really like Figure 5, but I think there's a typographical error. The caption reads, "For example, the April 1st forecast models for the Amudarya use as predictors the SWE estimate as of the beginning of March, the state of the PDO index in November, and the SCAN index in February." I think the figure states, though, that these models use the SCAN index in January, not February, for that forecast date and river.

Thank you for pointing at this mistake. We have corrected the wording in the caption.

Reviewer #2

I want to thank the authors for incorporating most of the suggestions provided in the first round of revisions. Despite the manuscript has been improved considerably, I have a suite of comments that I would like the authors to address before this paper is accepted for publication.

We appreciate your feedback on our previous changes to the manuscript.

Minor comments

1. L41: I think it would be more appropriate re-writing as "…some process-based models have higher computational demand (e.g., Oleson et al., 2010; Niu et al., 2011; Clark et al., 2021)…".

Thank you for this suggestion. We believe that the current version of the sentence is justified, as the context compares the computational demand of process-based and data-driven models. The following references may serve as supporting evidence (Clark et al., 2017; Tsai et al., 2021; Chen et al., 2022; Rahman et al., 2022; Bennett et al., 2024).

2. L88, L312 and everywhere else: please define the winter season (DJF?), since most readers will not be familiar with your study domain. The same comment applies to the remaining seasons.

This sentence is part of a general introductory paragraph on common hydrological cycle pattern in Central Asia. We define all relevant seasons we focus on as a range of respective months in the following paragraph (lines 99-101), such as "*cold season*" and "*growing season*", and use these terms throughout the text. In our opinion, specifying months for the terms "*winter*", "*spring*", and "*summer*" in the preceding sentence may introduce confusion for readers, as they do not fully align with the "*cold*" and "*growing*" seasons that are the focus of our study.

3. L131, L230 and everywhere else: I advise replacing "prediction" with "hindcast", since you are actually presenting results from retrospective forecasting (i.e., hindcasting) experiments. Please be precise and consistent with the terminology.

Thank you for this suggestion. The sentence in question provides a general explanation of suggested approach for seasonal streamflow forecasting in the region. In our opinion, using "*hindcast*" in this context would be confusing. For the same reason, we continue to use the term "*prediction*" in the Methods section, which details the seasonal streamflow model based on ensemble stacking. However, we acknowledge that we incorrectly referred to results as "*forecasts*" and have ensured that in the revised version of the manuscript this has been corrected to "*hindcasts*" where appropriate.

4. L233-234: if I understood well, you assess – for each model – cross-validated (deterministic?) hindcasts to get 16 evaluation metrics, and then select the k < 16 that fulfill the requirement R2>0.2, right? Please clarify.

Yes, that is correct. For each of the 16 base models, we compute a leave-one-out cross-validated (LOOCV) R^2 value based on deterministic hindcasts. Only base models that achieve an LOOCV $R^2 > 0.2$ are retained for further analysis. We have amended these lines (240-243) in the revised text for a better clarity.

5. Section 4.1: Did you try correlating seasonal averages (i.e., temporally averaged 2-month, 3-month, etc.) of your climate indices against seasonal precipitation for predictor screening?

Thank you for guiding question. No, we did not apply temporal averaging for predictor screening. However, we agree that averaging may be more appropriate for oscillations with relatively slower dynamics, such as SOI and PDO. Implementing this approach would require modifications to our code and could alter predictor lags, potentially affecting model structure and comparability. To maintain consistency and avoid introducing additional complexity at this stage, we prefer to keep the current methodology unchanged.

7. I strongly advise the authors to move Figure S1 to the main manuscript and move Figure 4 to the supplement, since your paper is about seasonal streamflow (and NOT precipitation) forecasting. Also, the current Figures S1 and 4 are nearly identical.

Thank you for this suggestion. However, we think that Figure 4 should remain in the main manuscript, as it is directly tied to our conceptual framework. Our approach uses climate oscillations as approximators for precipitation variability in the upcoming season. We believe that emphasizing the climate oscillation–precipitation–streamflow link provides a clearer justification for incorporating climate oscillations. Figure S1, on the other hand, may be viewed as supporting evidence that climate oscillations influence streamflow by modulating precipitation. We acknowledge that the initial version of the manuscript contained a paragraph explicitly explaining this rationale, which was removed during a previous revision. To address this, we have now revised the Introduction in lines 130-134 to better clarify our conceptual framework.

8. L339-342: the statements concerning the numbers of models are very hard to visualize. I recommend adding those numbers in Figure 6 for each forecast initialization.

We appreciate this suggestion. Since Figure 6 is already stacked, adding additional numbers directly to the figure may reduce readability. Instead, we propose moving information on the number and types of models to the Supplementary Material, ensuring that it is appropriately referenced in the text for clarity (lines 364-365).

9. L351: I presume you are referring to R2 results here, right? I do not think you should refer to "uncertainty", since you are not quantifying forecast spread or providing confidence intervals. Please be more precise and refer to what you are actually showing.

Yes, this comment refers to the evaluation of models in terms of R^2 and nMAE. We have revised the sentence to use the term "*lower accuracy*" for better clarity. Thank you for this correction.

10. Figure 7: Please clarify whether you are showing the results from the meta-learner model.

Yes, these results are from the meta-learner models. We have revised the caption accordingly.

11. Figure 8: Are you displaying the ensemble hindcast mean along the y-axis? Please clarify. Also, these results should not be in sub-section 5.4, since you are not illustrating any predictive uncertainty.

Figure 8 displays hindcasts produced by the meta-learner model. These results are now moved to sub-section 5.3

12. L398-399: This description should be in the methods section. How many times did you resample the data? Note that this step would be redundant if the SVM meta learner produced ensemble hindcasts.

We have moved the relevant description to the Methods section as a new sub-section titled "Uncertainty Estimation" (lines 265-268). In the earlier approach, we bootstrapped only the base model predictions, which primarily captured uncertainty stemming from the ensemble structure. We have now implemented a different method: after resampling the input data, we retrain both the base models and the meta-learner for each of the 500 bootstrap iterations.

13. L401: I strongly advise the authors to be more quantitative when judging the "width of uncertainty bounds". To this end, they can compute the alpha reliability index (Renard et al., 2010), as in previous seasonal hindcasting studies (e.g., Mendoza et al., 2017; Araya et al., 2023).

Thank you for this suggestion. We have now amended the wording in that and other lines. We have also amended this section with the following sentence: "*It should be noted that the uncertainty intervals are estimated by bootstrapping a relatively short streamflow time series and do not account for uncertainty caused by the potential limited representativeness of the actual natural variability by the observations."*

14. L409: please clarify what you mean with "more consistent".

We have now replaced "*more consistent*" with "*relatively less uncertain*".

15. L139, L391, L406 and everywhere else: please avoid using "significant" or "significance", unless you refer to statistically significant result.

We have now replaced "*significant*" in those lines with other words.

16. L458-459: "The resulting forecast models generate credible simulations…". Your models are producing hindcasts and NOT simulations (please revise Beven and Young, 2013). Also, the sentence reads as overselling, since your results are not good for all lead times. I suggest deleting.

We now replaced "*simulations*" with "*hindcasts*". Please note that the sentence notes "*albeit with performance variations depending on lead time.*". We do not consider the message as overselling, given the results we present and multiple restrictions we encounter within this study.

17. L464-465: "In most catchments, the SOI, PDO, or both were utilised, indicating the dominant influence of ENSO". The Pacific Decadal Oscillation (PDO) can modulate ENSO, but PDO and ENSO are different modes of variability. I suggest re-wording to avoid confusing readers.

The sentence is amended by adding: "*..and other climate variability patterns in the Pacific Ocean*."

18. L490: I think what you should write here is "more accurate seasonal streamflow hindcasts". Note that the term "reliable" has a very specific connotation in probabilistic forecasting, and is related with the degree to which forecast probabilities match relative observed frequencies (see Wilks, 2019).

We have replaced "*reliable*" with "*accurate*".

Suggested edits

19. L28: delete "other".
20. L33: Add "Additionally," before "accurate".
21. L42: revise "data-drivenapproaches".
22. L44 and L47: I suggest deleting "primarily".
23. L63: "multi ensemble" -> "multi-model ensemble".
24. L75: higher prediction accuracy and better quantify -> "the quantification of".
25. L94-95: "One approach" -> "The first approach".
26. L111-112: "for forecasting" -> "to forecast".
27. L331: I suggest rewriting as "…with varying R2…", since this is what you are actually showing.

Thank you for these suggestions. We have edited the text accordingly in most of the suggested instances, except the comment #20.

References:

Bennett, A. *et al.* (2024) 'Spatio-Temporal Machine Learning for Regional to Continental Scale Terrestrial Hydrology', *Journal of Advances in Modeling Earth Systems*, 16(6), p. e2023MS004095. doi: https://doi.org/10.1029/2023MS004095.

Chen, X. *et al.* (2022) 'Comparison of deep learning models and a typical process-based model in glacio-hydrology simulation', *Journal of Hydrology*, 615, p. 128562. doi: https://doi.org/10.1016/j.jhydrol.2022.128562.

Clark, M. P. *et al.* (2017) 'The evolution of process-based hydrologic models: historical challenges and the collective quest for physical realism', *Hydrology and Earth System Sciences*, 21(7), pp. 3427–3440. doi: 10.5194/hess-21-3427-2017.

Rahman, K. U. *et al.* (2022) 'Comparison of machine learning and process-based SWAT model in simulating streamflow in the Upper Indus Basin', *Applied Water Science*, 12(8), p. 178. doi: 10.1007/s13201-022-01692-6.

Tsai, W.-P. *et al.* (2021) 'From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling', *Nature Communications*, 12(1), p. 5988. doi: 10.1038/s41467-021-26107-z.