Dear Reviewer,

Thank you for your detailed and valuable feedback. In response to your comments, we will refine the text to better reflect the specific contributions of our study. We acknowledge that teleconnections and the use of SWE in streamflow forecasting are well-established, and we will clarify our findings by focusing on the operational insights specific to Central Asia. The literature review will be expanded to better contextualize our contribution, and we will more comprehensively cite other relevant work, including studies from North America and other regions.

Regarding forecast uncertainty (comment 2), we will incorporate analysis bootstrapped prediction intervals and Q-Q plots to better quantify forecast uncertainty. In addition, to address concerns about the limited sample size (comment 3), we will reduce the number of predictors to a maximum of three and revise the validation approach by removing the hold-out sample. Instead of the hold-out validation sample, we will perform full-sample LOOCV, ensuring more robust evaluation. We believe that the suggested changes, along with updates to figures and more consistent terminology, will improve the clarity and rigor of the manuscript. Please find below our responses (in blue) to your referee comments (in black).

With regards,

The authors.

> In this manuscript, the authors explore the relative contribution of large-scale climate oscillation predictors and snow water equivalent on the quality of April-September seasonal streamflow forecasts in eight catchments located across the Pamir and Tian-Shan mountains (central Asia). To this end, the authors first examine the correlation between climate modes of variability and (i) catchment-scale precipitation over the peak precipitation season (February-July), and (ii) April-September seasonal streamflow. Then, the authors adjust 16 models resulting from the combination of four statistical models and four SWE products, using SWE (at four forecast initialization times) as one of the predictors, and large scale climate indices as additional predictors. The total sample size (i.e., 18 points obtained from 18 years with data) is split into a sample of 15 points for cross-validation, and the remaining points are used for additional testing. The authors conclude that their technique is “a novel way to reduce uncertainties in seasonal discharge predictions in data-scarce snowmelt-dominated catchments”.
>
> This is basically a seasonal hindcasting study, generally well written and concisely presented. Nevertheless, my main critiques with this work are (1) the overselling, especially in the title, abstract and conclusions, (2) the lack of forecast uncertainty characterization (which is highlighted by the authors as a key contribution), and (3) the

limited sample size, and the way the authors address this problem in their analyses. Therefore, I think that the manuscript needs major revisions before being considered for publication in HESS.

Major comments

1. Title, abstract and conclusions: it is well known that the value of hydroclimatic teleconnections on seasonal streamflow forecasts is huge in snowmelt-driven catchments, especially during the preceding Fall season, when initial hydrologic conditions have not been fully developed (e.g., Mendoza *et al.*, 2017) – as the authors write in L22-24, and conclude in L403-404. There is a long history on the use of large-scale climate information for seasonal streamflow forecasting (e.g., Piechota *et al.*, 1998), and what the authors state in L20-21 and other parts of the manuscript was neatly shown nearly two decades ago using custom-based climate indices in two western US catchments (see Figure 8 in Grantz *et al.*, 2005; and also Regonda *et al.*, 2006; Opitz-Stapleton *et al.*, 2007; Bracken *et al.*, 2010; Mendoza *et al.*, 2014, etc.). Additionally, the use of simulated catchment-averaged SWE as a predictor to feed statistical models (L105-106) is not new either (e.g., Rosenberg *et al.*, 2011; Mendoza *et al.*, 2017). In other words, the findings reported by the authors are not novel and, based on this, I think that they should refine the title, abstract and conclusions to make them more specific to their actual contribution to the existing literature.

Thank you for your detailed feedback and for highlighting that teleconnections have been explored in seasonal streamflow forecasting, including for snowmelt-driven catchments. As you correctly noted, we mention this in the manuscript, although those excerpts were succinct and warrant to be expanded. Our manuscript offers additional contributions that expand on the valuable insights from Mendoza et al. (2017) and other researchers. As noted in the abstract and discussion, our study identifies specific instances when teleconnections may become more influential: at extended lead times, during strong in-season climate variability, or when catchment snow estimates are less reliable. We believe that this context, which has received little attention in previous studies (except for the first instance, which aligns with findings of Mendoza et al. (2017), helps refine the use of teleconnections for operational water supply forecasting.

Furthermore, while acknowledging that the use of large-scale climate indices and snow data for such forecasting has been previously explored, most of the cited references focus on North America. We believe our study offers novelty by demonstrating how teleconnections are pertinent to Central Asia and how their inclusion can aid seasonal water supply forecasting.

We will expand literature review on use of climate teleconnections in seasonal water supply forecasting, and clarify the abovementioned distinctions in the manuscript’s title, Abstract, and Discussion to better reflect our study’s contributions.

2. L25: the authors declare that their approach "provides a novel way to reduce uncertainties in seasonal discharge prediction". Do they refer to the spread of seasonal forecasts? Although they describe an ensemble stacking framework to produce a final forecast, only deterministic evaluation metrics (coefficient of determination and normalized mean absolute error) are reported, and no characterizations of hydrological prediction uncertainties are presented. A popular to do so is through ensembles (Georgakakos *et al.*, 2004; also, see publications produced by the HEPEX community on this topic), analyzing, for example, the statistical consistency of seasonal forecasts with graphical devices like rank histograms (Hamill, 2001) or Q-Q plot (Renard *et al.*, 2010), complementing with ensemble verification metrics (e.g., De Lannoy *et al.*, 2006). Therefore, I recommend the authors to take advantage of the multiple models developed to characterize forecast uncertainty or, alternatively, delete any references to "forecast uncertainty" from their manuscript (which I think would diminish the quality of their research).

Thank you for your valuable feedback. To address this well-grounded point, we propose implementing an uncertainty assessment using a bootstrapping approach. Our ensemble stacking approach involves different numbers of models per basin and issue date, which could complicate direct comparisons of ensemble spread. We suggest using bootstrapping to both resample the data and retrain the SVM meta-learner for each bootstrap sample to fully capture forecast uncertainty. In this framework, LOOCV will be used for training and assessing the generalizability of both the base models and the meta-learner. Afterwards the bootstrapping will be applied by training SVM meta-learner on each bootstrapped sample to generate 90% prediction intervals based on the variability in bootstrapped predictions. To complement this uncertainty characterization, we will also implement Q-Q plots to visually assess the consistency between predicted and observed discharge values.

3. Sample size (L126-127): this is a major issue in seasonal streamflow forecasting, since only one training/verification point is available per year. Therefore:In my opinion, the sample size is not large enough to support – being extremely generous – more than three predictor variables in their models (the authors report up to five predictors in Figure 5 for the Chu River basin), given the high risk of overfitting (see Wilks, 2011 or any other book on Statistics). Hence, I think that the authors should revisit their statistical models, removing combinations of predictors that may introduce multicollinearity.

Thank you for highlighting the issue of small sample size. Most of the data we used is derived from a previous study by Apel et al. (2018) . Unfortunately, we do not have recent updates extending beyond that period until today. Moreover, available historical data spans from 1970 to the 1990s for most rivers, with more complete observations for the largest rivers (Amudarya and Naryn) up to around 2018. However, using this extended dataset is problematic because the two datasets used to derive SWE estimates, FLDAS and GPM, are only available starting from 2000. Extending the observations further back would limit our ensemble to only ERA5 and MSWX, reducing the diversity of the ensemble, as MSWX is generated by bias-correcting ERA5 and may exhibit similar predictions for some catchments.

We acknowledge the limitations imposed by a small sample size, which could impact the generalizability of our results. To address this, we integrate several strategies: we employ an ensemble approach which is particularly effective for addressing the challenges associated with small datasets by combining the strengths of multiple models (Dietterich, 2000; Zounemat-Kermani et al., 2021). Furthermore, we adhere to parsimony in model selection and parametrization, and therefore we employ relatively simple machine learning models with parameters fixed at conservative level to minimize overfitting. To further enhance the robustness of the framework given the limited length of observations, we incorporate multiple independent data sources into the ensemble model.

We realize that descriptions of these approaches have been succinct (e.g. lines 94-101); we will explicitly highlight these strategies in a revised version of the manuscript, linking them directly to the limitations posed by the small data sample. To address your concern and ensure a balance between model complexity and interpretability, we will also reduce the number of predictors to a maximum of three in a revised version of the manuscript.

While we acknowledge that multicollinearity can distort the interpretation of individual predictor effects, evidence suggests it is less problematic for predictive performance (Kiers and Smilde, 2007).The selected model types, especially Support Vector Machines (SVM) and Random Forests (RF), are inherently more robust to multicollinearity and can accommodate more predictor variables than observations without a loss in predictive power. A preliminary check of collinearity among the predictors revealed that Pearson's correlation coefficients are generally below 0.1, except for PDO and SOI at their selected months (used in two basins), where the coefficient reaches 0.55. While we are unsure if this constitutes strong multicollinearity, to be cautious with the interpretation of results, we propose showing variable importance (Figure 5) aggregated into two classes: SWE and climate indices.

> I do not think it is appropriate to split their sample of points (n = 18) into a smaller sample for leave-one-out cross validation (with n =15), and another sample for verification that contains three (L314) or even two points. I recommend the authors using the entire sample to perform cross-validation and compute verification metrics. Further, they should characterize the impact of sampling uncertainty, which could be done by adding confidence intervals created through bootstrapping with replacement (see section 5.5 in Araya et al., 2023). This is a critical point that the authors should address, given the very small sample size.

Thank you for your concern regarding the adequacy of the training sample and the subsequent suggestions. Our two-tiered validation approach, combining LOOCV on the training sample with hold-out validation on the testing data, was intended as additional element for checking forecast reliability. However, we acknowledge that the hold-out validation sample, consisting of only 2 to 3 observations, may appear unrepresentative. In line with your suggestions, we will extend the training sample by removing the hold-out validation and incorporating bootstrap-based prediction intervals for the predictions.

Specific comments

4. L13: The authors use the term "predictions", which is an excessively ample word for what they really do. In this line, I recommend the authors using the word "forecasts", and consider using the words "hindcasts" and "hindcasting" in the remainder of the manuscript, especially when describing their methods and results (please see section 3 in Beven and Young, 2013).

Thank you for your suggestion. We will replace "predictions" with "forecasts" and "hindcasts" as appropriate.

5. L30: This population estimate is for almost ten years old. I suggest updating the number and the reference.

We could not find updates to this estimate in the given context. Immerzeel et al. (2020) providea similar estimate of ~1.9 billion people, though they focus on populations dependent on mountains. If you are aware of newer estimates, we would appreciate it if you could share the relevant reference with us.

6. L35: Sometimes you use "dynamic", and sometimes "dynamical". Please pick one term and be consistent.

Thank you for highlighting this inconsistency. We will revise the manuscript to use 'process-based' instead of "dynamic/dynamical" and "data-driven" instead of "statistical".

7. L36-37: This sentence is incorrect. Climate forecasts are not used until the IHCs have been produced by running a model with a historical meteorological dataset up to the forecast initialization time.

We appreciate this comment and apologize for the confusion. We intended to convey the same point, but used incorrect wording. In the revised version of the manuscript, the sentence will read: "*Process-based forecasts use a hydrological or land-surface model to estimate current hydrologic conditions, typically with assimilation of observational data, followed by the use of climate forecasts to project future conditions.*"

> 8. L39: I disagree with the authors' statement, since computational demand depends on model complexity and, therefore, a model simulation might take from seconds (e.g., GR4J, SAC-SMA) to several minutes (e.g., VIC, SUMMA) in a home PC.

Thank you for highlighting this. We agree that computational demand depends on model complexity. However, depending on the type of model and the level of spatial resolution, a simulation can take significantly longer than just a few minutes. Since the paragraph compares process-based and data-driven modelling approaches for hydrological forecasting, we suggest splitting the sentence into two, with the new sentence reading as: "*Process-based models typically exhibit higher computational demands.*"

> 9. L40: Note that meteorological variables obtained from numerical climate models ARE prone to uncertainties.

Thank you for bringing this to our attention. We will revise the wording accordingly.

> 10. L45-46: I think that the authors should cite more papers when referring to the relevance of SWE as a predictor in mountainous catchments (e.g., Garen, 1992; Rosenberg et al., 2011; Mendoza et al., 2014). In general, I recommend the authors strengthening the literature review in this paragraph.

Thank you for this suggestion. We admit that the initial submission lacked overview of existing and /or similar practices for seasonal hydrological forecasting based on accumulated snowpack, especially using data-driven methods. We will expand the literature review to better place our study in a global context, particularly by referencing relevant work from North America, as you suggest, and possibly from regions which share similar hydroclimatic and data challenges as Central Asia.

> 11. L46: "statistical forecasts of seasonal streamflow often rely solely on accumulated snowpack". I disagree with this statement. The current operational systems managed by the NRCS for the western US and the DGA for Chile use, besides SWE, in situ measurements of precipitation, air temperature and streamflow measured in the preceding months.

Thank you for the correction; we apologize for the confusion. What we intended to convey is that accumulated terrestrial water storage is the main determinant of seasonal water supply, with snowpack being its key component. We will revise this section to avoid confusion and include a mention of other commonly used predictors. In our study, we limited the predictors to two groups—SWE and teleconnections—because we aimed to assess the added value of teleconnections compared to SWE-based predictions. Additionally, we sought to keep the model parsimonious given the limited number of observations.

12. L74: Are the authors referring to hydrological droughts? I think that any paper by Anne Van Loon (e.g., Van Loon, 2015) may be useful to clarify this point.

Thank you for the suggested references. In this sentence we are referring to seasonal precipitation levels lower than the historical norm, which may represent droughts. However as this paragraph aims to overview climate teleconnections relevant to the Central Asian region, rather than droughts, we find it challenging to refer to Van Loon (2015) in this specific context.

13. L88-89: This approach was proposed and tested more than two decades ago (e.g., Piechota et al., 1998).

14. L97: It would be good clarifying here that SWE can be directly obtained from reanalysis, or estimated by combining satellite remotely sensed snow depth and a snow density model.

15. L99-101: Please note that ensemble techniques have been used for decades in seasonal streamflow forecasting (e.g., Twedt et al., 1977; Day, 1985; Regonda et al., 2006; Wang et al., 2011; Arnal et al., 2018; Emerton et al., 2018; Lucatero et al., 2018; Girons Lopez et al., 2021; Araya et al., 2023).

Thank you for these suggestions. We will expand the literature review accordingly, with appropriate referencing to earlier studies.

16. Table 1: I suggest adding the period used to compute the variables and more hydroclimatic descriptors, like mean annual runoff (mm/yr), mean annual runoff ratio and aridity index. Please change the units of seasonal discharge to mm/yr,

Thank you for these suggestions, we will amend descriptive statistics accordingly.

17. L173: what link function did you use in your GLM?

We applied a Gaussian family link function for the GLM model.

18. L189: Looks like the SVR works as a post-processor, right?

Yes, the SVR functions as a post-processor in our ensemble stacking approach.

19. L190-191: given the small sample size, I recommend deleting this step from your workflow (see comment #3).

Thank you for the suggestion. We agree to revise our validation strategy in line with your recommendation in comment #3. All relevant sections of the manuscript will be amended accordingly.

> 20. L205, L206, L297, L351 and L353 and everywhere else: the authors use the term "assimilate" when referring to the use of modeled SWE as a predictor in their statistical model. Nevertheless, such term is typically used when referring to a family of techniques that combine imperfect models with uncertain observations to improve dynamical model estimates (e.g., Liu and Gupta, 2007; Reichle, 2008; Kumar et al., 2016; Smyth et al., 2022). Since the authors do not refer to the former concept anywhere in this manuscript, I suggest deleting the words "assimilate" or "assimilation".

Thank you for highlighting this inconsistency in used terms. We will amend the terms used accordingly throughout the text.

> 21. L221: what do you mean with the word "underperforming"?

Thank you for your comment. By "underperforming" we refer to the fact that certain SWE products exhibit lower predictive accuracy in specific catchments compared to other products. We will clarify this in the revised manuscript to ensure the sentence is clear.

> 22. L221-222: I think that this sentence contradicts the previous one. Also, if ERA5-L and MSWX are better, why don't you just pick one of these products for subsequent analyses? Some of your subsequent figures are unnecessarily complicated.

Thank you for this comment. Our intention was to convey that, in general, ERA5-L and MSWX-based estimates show higher correlations with seasonal streamflow. However, different SWE products perform better or worse depending on the catchment, which is why we have not selected a single product for subsequent analyses. Figure 3 is intended not only to illustrate the association between snowpack and seasonal streamflow and how this relationship changes across forecast issue dates, but also to highlight the differences between the snow estimates. For these reasons, we would like to retain both the figure and its explanations in the text.

> 23. Section 5.2 and Figure 4: since your target variable is seasonal streamflow, you could show correlation results between this variable and climate indices here, and move the correlation results with precipitation to supplementary material.

We appreciate this suggestion. We propose retaining the correlation graph between peak-season precipitation and climate indices in the main text, while moving the streamflow correlation graph to the supplementary material. We believe this graph provides valuable context, which we will elaborate on, regarding the associations between climate oscillations and interannual precipitation variability, which in turn influences interannual fluctuations in streamflow levels.

> 24. Figure 5: I do not think you can support more than three predictors with a sample size n = 18 (see comment #3).

The new version of Figure 5 will display only three predictors, as noted in our response to the comment #3.

> 25. L295: Do you mean winner among statistical models? Can you please be more specific?

Thank you for your comment. To clarify, by 'best-performing,' we meant the model that most accurately predicts streamflow across all catchments and forecast lead times. We will revise the manuscript to make this clear.

> 26. L301-302: I do not think that the authors are quantifying uncertainty (see comment #2).

> 27. Figure 6 is quite difficult to read. Since the focus of the paper is on the relevance of climate information in seasonal streamflow forecasting, why don't you just show the best-performing statistical model, with the best SWE product? Further, you should include the assessment period in each figure caption.

Thank you for your suggestion. We would like to retain Figure 6, as it not only displays the accuracy of both the base model forecasts and the final ensemble forecast, but it also illustrates the varying performance of the base models across different issue dates. It also conveys the message that the ensemble forecast outperforms single model forecasts. We believe this broader comparison is useful for demonstrating the added value of the ensemble approach. Regarding the assessment periods, since they will be indicated in a previous figure/table (see our response to the comment #16), we do not see the need to repeat them again in this figure. However, we will ensure the figure captions are clear and include all necessary information.

> 28. L313: This is not true for all catchments. See, for example, the red bars for the Kashkadarya and Chu basins.

Thank you for highlighting this. Since the revised version will now include all observations in the LOOCV with no hold-out validation sample, this and related text exerts will likely be removed from the text.

29. L322: Do you mean larger errors? Are you comparing against the results obtained with SWE and climate information? In that case, I really think you should define a Skill Score for a comparative assessment.

Thank you for the comment. Yes, we are comparing two configurations of the same models—one using only SWE and the other using both SWE and climate indices. To ensure a more consistent comparison, and in light of previous suggestions (removing hold-out validation), we propose including a single graph that displays the MAEs of both configurations for each basin and issue date. We will also consider, as an alternative, displaying only the incremental differences between the two configurations as the percentage reduction in MAE for each basin and issue date.

30. Figure 8: I recommend presenting these results using scatter plots (eight panels), along with the 1:1 line, percent bias, MAE and R^2.

We appreciate this suggestion. We will display these results as Q-Q plots with embedded percent bias, MAE and R^2.

31. L348: What do you mean with 'effectively'? That near real-time SWE estimates are actually useful for seasonal streamflow forecasting?

By 'effectively,' we intended to convey that SWE estimates derived from or modelled using global sources, despite their biases and spatio-temporal inconsistencies, can still provide added value for seasonal streamflow forecasting. While this point may seem trivial, we believe it is relevant in the context of forecasting without in-situ data on predictors. We will consider revising this sentence to ensure clarity.

32. L350: I do not think the authors have presented any uncertainty or error propagation analysis (please see comment #2)

This sentence will be revised and updated in accordance with uncertainty analysis we proposed above (our response to the comment #2).

33. L352: Did you actually assess the accuracy of SWE products using in-situ observations?

No, as we note in the Introduction (L94-101) systematic in-situ SWE measurements are absent in the region

34. L420: In my opinion, models adjusted with such a small sample cannot be regarded as “reliable”.

Suggested edits

35. L28: “where it sustains” -> “sustaining”.

36. L32: “Accurate water availability forecasts” -> “accurate water supply forecasts”.

37. L36: “current hydrologic conditions” -> “initial hydrologic conditions”.

38. L42: “multiple variables” -> “multiple predictor variables”.

39. L43: delete “the context of”.

40. L61 and L63: replace “from now on” by “hereafter”.

41. L67: delete “from satellite”.

42. L73: “ENSO in its cold phase” -> “the cold phase of ENSO”.

43. L74: delete “ENSO’s”.

44. L95-96: “used to conduct” -> “conducted”.

45. L124: I think that the right word is “predictand”.

46. L130: delete “in near real-time”.

47. L132: “we simulated” -> “we obtained”.

48. L155-156: “precipitation levels” -> “precipitation amounts”.

49. L174: add “SVR” after “support vector regression”.

50. L214-215: I suggest deleting this sentence.

Thank you for the proposed edits. We will update the wording in line with your suggestions.

References:

Apel, H., Abdykerimova, Z., Agalhanova, M., Baimaganbetov, A., Gavrilenko, N., Gerlitz, L., Kalashnikova, O., Unger-Shayesteh, K., Vorogushyn, S., and Gafurov, A.: Statistical forecast of seasonal discharge in Central Asia using observational records: development of a generic linear modelling tool for operational water resource management, Hydrol. Earth Syst. Sci., 22, 2225–2254, https://doi.org/10.5194/hess-22-2225-2018, 2018.

Dietterich, T. G.: Ensemble Methods in Machine Learning, in: Multiple Classifier Systems, 1–15, 2000.

Immerzeel, W. W., Lutz, A. F., Andrade, M., Bahl, A., Biemans, H., Bolch, T., Hyde, S., Brumby, S., Davies, B. J., Elmore, A. C., Emmer, A., Feng, M., Fernández, A., Haritashya, U., Kargel, J. S., Koppes, M., Kraaijenbrink, P. D. A., Kulkarni, A. V., Mayewski, P. A., Nepal, S., Pacheco, P., Painter, T. H., Pellicciotti, F., Rajaram, H., Rupper, S., Sinisalo, A., Shrestha, A. B., Viviroli, D., Wada, Y., Xiao, C., Yao, T., and Baillie, J. E. M.: Importance and vulnerability of the world's water towers, Nature, 577, 364–369, https://doi.org/10.1038/s41586-019-1822-y, 2020.

Kiers, H. A. L. and Smilde, A. K.: A comparison of various methods for multivariate regression with highly collinear variables, Stat. Methods Appl., 16, 193–228, https://doi.org/10.1007/s10260-006-0025-5, 2007.

Van Loon, A. F.: Hydrological drought explained, Wiley Interdiscip. Rev. Water, 2, 359–392, https://doi.org/10.1002/wat2.1085, 2015.

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, 3915–3935, https://doi.org/10.5194/hess-21-3915-2017, 2017.

Zounemat-Kermani, M., Batelaan, O., Fadaee, M., and Hinkelmann, R.: Ensemble machine learning paradigms in hydrology: A review, J. Hydrol., 598, 126266, https://doi.org/https://doi.org/10.1016/j.jhydrol.2021.126266, 2021.