Dear Reviewer,

Thank you for your valuable and constructive comments on our manuscript. We have carefully considered all the suggestions. We will expand the literature review to provide a broader context by referencing relevant work from North America and other regions, especially focusing on similar hydroclimatic and water supply forecasting studies. We will also refine the terminology to clarify distinctions between process-based and data-driven approaches and ensure the manuscript uses consistent, field-appropriate language. Additionally, we will revise the methodological descriptions to better explain the rationale behind the chosen models and improve the clarity and precision of our wording throughout the paper. We believe that the revised version will be more comprehensive and better aligned with current research in the field. Please find below your referee comments (in black) and our responses (in blue).

With regards,

The authors.

Review of 'The value of hydroclimatic teleconnections for snow-based seasonal streamflow forecasting,' Umirbekov et al., HESS Discussions

This submission summarizes the motivation for, and development, implementation, and performance of, a data-driven model for seasonal river runoff volume forecasting in Central Asia. The method uses as predictors a combination of SWE data from existing large-scale operational remote sensing and land surface modeling products, and indices of various atmosphere-ocean circulation patterns, as predictors; it employs a multi-model ensemble model structure, which has some machine learning elements; and is intended to serve as an actual operational forecasting tool for directly supporting water management decision-making.

My overall recommendation is for publication pending minor revisions.

The article will be an excellent contribution to HESS. The paper is succinct and well-organized. The study location considered has been historically understudied, though admittedly not so severely as some other regions of the developing world. The candidate predictor data combinations are geophysically sensible, but also to some degree original in the particular way they are used here. Though it is possible to quibble with certain aspects of the predictor selection process used, it is a reasonable and defensible approach for the task at hand. Its multi-model ensemble philosophy is fully consistent with a large body of evidence demonstrating its value, yet it is only one of a tiny handful of hydrologic modeling examples where several separate data-driven/statistical/machine learning modeling systems are used and their results

pooled to form a best estimate, and it is also novel as implemented here. Unlike some research papers claiming to present a forecast model, this article makes a point of clearly confirming that the predictor datasets considered are available going forward on a near-real-time basis, a necessity if a modeling system is actually going to be useful for real-world operational forecasting. The submission also clearly identifies (e.g., lines 400 to 410) where, when, and why each of the candidate predictor datasets are or are not useful for water supply forecasting, which is crucially important for physical credibility of the forecasts and the systems and data generating them, and which is sometimes overlooked in research around data-driven models.

Thank you for your positive and detailed feedback. We appreciate your recognition of our study's contributions and the practicality of our model for real-world forecasting. We value your recommendation for publication pending minor revisions and look forward to addressing your comments below in the revised manuscript.

That said, I think a few improvements will be needed before the paper is ready for publication. I hope the following comments will be helpful to the authors if they move forward with submitting a revised manuscript:

1. Though in general the article is well-crafted, some passages are written poorly enough that their meaning is unclear. For example, the wrong word is used, or words are used incorrectly, or elaborate vocabulary or phrasing is used when simpler wording would do.

We will carefully revise sections where wording is unclear or overly complex to improve readability and ensure the meaning is precise.

2. The literature review is not quite adequate. While there seems to be sufficient reference to prior work in the study area, this article is not just a case study, and HESS is an international journal. More broadly, the methods described here need to be placed in the wider and deeper context of previous work, not just locally but globally, in order for readers to understand its contributions and wider implications – the methods used here may be applicable in entirely different regions of the world. The wider literature does not need to be discussed in detail, nor do the methods used in this submission need to be compared against them, but the paper does need to leave some clues for readers about relevant prior publications. What stood out for me is that prior research (and practice) around seasonal water supply forecasting in western Canada and the western US, directly relevant to this study, has not been adequately acknowledged. Some points of particular note are the following (full citations are provided at the end of this review):

2.a. It would probably be helpful to note in the article that the type of seasonal river discharge volume forecast modeling considered here is widely referred to as "water supply forecasting" (WSF) in the western North American operational hydrology and water management communities. They don't need to use the term throughout the article, but just pointing out at the start of the paper that they're working on what is commonly called WSF will help readers connect the study to a large existing body of prior research and practice.

Thank you for this suggestion. We admit that the initial submission lacked overview of existing and /or similar practices for seasonal hydrological forecasting, 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 from regions which share similar hydroclimatic and data challenges as Central Asia. We will also ensure to incorporate the cited research on water supply forecasting and teleconnection indices, ensuring our work is connected to this body of literature.

2.b. Contrary to what seems to be implied in this article, given the way certain passages are phrased and the sparseness of literature citations, combining teleconnection indices with snow data as inputs to statistical seasonal water supply forecasting models is neither new nor rare. It appears to have first been implemented decades ago (Garen, 1998) in the large-scale (hundreds of forecast locations) operational forecasting systems of the US Department of Agriculture's Natural Resources Conservation Service (NRCS) (Perkins et al., 2009), predictions from which are a staple for water managers across the American West. These principal component regression models (Garen, 1992) have used a combination of SWE, accumulated precipitation, and in some cases antecedent streamflow and El Niño-Southern Oscillation indices (Garen, 1998) as predictors of seasonal river flow volumes; these methods have since be adopted widely across western North America by other operational forecast agencies. Furthermore, continued applied R&D on combined use of snowpack observations and teleconnection indices as input variables to statistical seasonal discharge forecast models has been continued by many, such as Gobena et al (2013) in western Canada, and Moradkhani and Meier (2010) and Regonda et al. (2006a, 2006b) in the western US, to give just a few examples. It has also been extended to discovering new climate prediction skill using nonlinear methods or new indices in areas where conventional linear teleconnections are weak, such as southern Oregon and northern California (Kennedy et al., 2009; see also Fleming and Dahlke 2014).

We appreciate your highlighting the relevant studies in 2.c. After viewing through the papers, we believe they will be valuable additions to our literature review. We will also aim to expand it with additional relevant studies.

2.c. Though the computational model presented here appears to be novel, major elements of its philosophy and structure are strongly reminiscent of other recent advances in data-driven predictive modeling of seasonal river discharge volumes. Of note here is the multi model machine-learning metasystem (M4), which was developed for and is currently being operationally implemented by the US Department of Agriculture NRCS as its new western US-wide seasonal river discharge volume forecast model. This system has been run using in-situ SWE, precipitation, and antecedent streamflow data, as well as combinations of in-situ and remotely sensed snow data, as predictors (Fleming et al., 2021, 2024). It uses a multi-model ensemble approach in which six data-driven (statistical and machine learning) forecast systems are run independently and the results are pooled to form a best estimate, closely analogous to the modeling philosophy used in this HESS contribution. There are also significant differences between M4 and the method used in this submission, but citing M4 will better-place this HESS article's contributions in the larger research and applications literature, and provide literature support to the methods the submitted paper uses. By the same token, some related exploratory work on methods for combining outputs from multiple data-driven seasonal river discharg forecast models by Najafi and Moradkhani (2016) should be cited in this regard as well.

Thank you for these suggestions. Upon reviewing the noted models/studies, we found many concept-wise similarities and believe these will be valuable addendums to the literature review, particularly regarding methods and methodological advancements in the field.

2.d. The foregoing are just some examples I happen to be familiar with. I'd suggest that the authors scour the literature for other prior work, including work in other regions globally, that ought to be at least briefly cited in their revised paper.

3. The following are a few additional suggestions for improvements:

3.a. Line 35 and elsewhere: this paper distinguishes between what it calls "dynamic" vs. "statistical" approaches. This jargon tends to be used more in other (broadly related) disciplines like regional climate modeling, with "process-based" vs. "data-driven" being more common in the operational hydrology literature. Also, it's usually "dynamical" not "dynamic", and "data-driven" also tends to be preferable to "statistical" today because of the increasing popularity of machine learning techniques (including this submission). The authors can use "dynamical" and "statistical" if they like, but to better orient readers, including the operational water resource forecasting community, to which this article seems to be in part addressed, please provide some synonyms where the terms are first introduced (line 35). It could read something like "generated using either dynamical (process-based, physics-oriented) or statistical (data-driven including machine learning and conventional statistical) modeling approaches" or something similar.

Thank you for highlighting this issue. We agree with your suggestions and will revise the manuscript to use 'process-based' instead of 'dynamic' and 'data-driven' instead of 'statistical.' We will also include synonyms for these terms where they are first introduced to better orient readers, particularly those from the operational water resource forecasting community.

> 3.b. Line 46, "statistical forecasts of seasonal streamflow often rely solely on accumulated snowpack." Yes and no. Yes, data on winter-spring seasonal snowpack provides the primary source of predictive skill in data-driven forecast models of spring-summer river runoff volume in snowmelt-dominated rivers. But these models, in both the research literature and (in particular) in operational practice, at least in western North America, also almost always use additional predictor data types. Examples include wintertime accumulated precipitation, early-season precipitation, and at some locations, antecedent streamflow and/or El Niño indices. See point 2.b above.

Thank you for the correction, and we apologize for the confusion. What we intended to convey is that accumulated terrestrial water storage is the main determinant of seasonal water supply in snowmelt-driven basins, 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.

> 3.c. Line 50: excellent point!

> 3.d. Lines 52-53 and elsewhere: if the authors want to call the Apr-Sep target period the "vegetation season," that's fine I suppose, but it's not standard nomenclature. Typically this would be called either the "growing season," looking at it from an agricultural water supply or broader ecological perspective, or the "runoff season", looking at it from a hydrological perspective. And given that they call Nov-Mar the "cold season" rather than the "snowpack accumulation season", it might also be more consistent to simply call Apr-Sept the "warm season." Overall, "growing season" seems like it might be the best fit here?

Thank you for your feedback. To ensure consistency and clarity, we will revise the manuscript to use 'growing season' for the Apr-Sep period, which is resembling the term used by local hydrometeorological agencies in Central Asia. We will also ensure that the terminology for the Nov-Mar period remains as 'cold season' for consistency with our current nomenclature.

3.e. Figure 1: this figure is good, but for a wide international readership, please provide an additional map showing the location of the study area within the larger geographic context of Eurasia.

Thank you for your suggestion. We will include an additional map to provide geographical context for the study region.

3.f. Line 124: predictand, not predicant

Our apologies for the mistake, we will correct this and other mistakes in the text.

3.g. Line 125: An 18 year data record – in other words, 18 samples - is pretty short; it's enough to defensibly create one of these models, but just barely. Commensurate limitations to the authors' ability to train model parameters and validate model predictions could be viewed as a source of uncertainty in this study; the counterargument, of course, is that with rapid climate change in mountain regions such as this study area, the statistical nonstationarity in a longer data record would have reduced its value anyway. This might be worth a sentence or two here. A brief explanation of why the record doesn't go back further or continue to the present could be helpful to readers as well. My understanding of the political history of this region isn't great, but I think this was part of the Soviet Union, which (its grave misdeeds notwithstanding) wasn't too bad at keeping streamflow records, so one might have been forgiven for guessing that there might be some usable historical data here?

Thank you for highlighting this issue. 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 such an 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 and reduce the diversity of the ensemble, as MSWX is generated by bias-correcting ERA5 and may exhibit similar predictions for some catchments.

While we agree that overly lengthy data records can introduce non-stationarity concerns, we believe that extending the study period by including observations up to the present would have been beneficial and likely safe from such issues in our case. We acknowledge the short length of records in the current version of the discussion section and will highlight this limitation more explicitly in the introduction or data section.

3.h. Lines 129-130: excellent point re: near-real time input data availability – this is a prerequisite for an operational forecasting model, and it's sometimes overlooked in research articles.

Thank you for your feedback.

3.i. Lines 173-175: a little more information about the constituent models ("base models") is needed here. What link function was used in the GLM? And why were linear kernels used in the GP and SVR models? Does this imply that most of the base models are essentially variants of standard, multiple linear regression? If so, what are the pros and cons? Note that work in the western US has shown that the relationships between winter-spring hydroclimatic forcing and spring-summer runoff response in data-driven WSF models range from nearly linear to moderately nonlinear, with clear physical explanations for these inferred functional forms (see Fleming et al., 2021).

Thank you for these guiding questions. We applied a Gaussian family link function for the GLM model. Indeed, the selected base models are linear, except for the RF model, though they differ in their approach to estimation and optimization. We experimented with several other model approaches, including using the same models with non-linear kernels. In most cases, the presented combination of models yielded similar accuracy in terms of RMSE and R-squared coefficients during LOOCV, but outperformed non-linear alternatives during testing on the hold-out sample. In some instances, depending on the basin or issue date, certain non-linear models produced slightly better predictions. However, when generalizing across all basins and issue dates, the existing structure still showed superior performance. We assume this may be due to two major and non-exclusive factors: (1) a relatively smaller number of observations and predictors, which makes non-linear machine learning models less efficient and prone to overfitting, and (2) the selection of predictors based on a primarily linear metric (Pearson's correlation). We will reflect on these findings and their implications in the revised version of the manuscript.

3.j. Lines 181-185: in defense of their methodological choice, which has no literature citations attached to it in the submission, the authors might wish to note that LOOCV is standard practice in western US WSF modeling; see references in point 2.b above.

Thank you for this suggestion and the references. We will appropriately refer to LOOCV as a standard practice in western US WSF modelling (Mallick et al., 2022; Granata and Di Nunno, 2024; Xu et al., 2024; Li et al., 2019).

3.k. Lines 187-190: to improve accessibility to a broad readership which may not be uniformly well-versed in machine learning, it might be helpful to add just a sentence or two, with an additional reference or two, explaining the concept of a meta-learner. It might also be helpful, in terms of connecting this concept to prior work in

data-driven WSF, to refer to the work of Najafi and Moradkhani (2016) on exploring different methods for creating multi-model ensembles from the predictions of several data-driven models.

Thank you for the suggestion. To improve accessibility, we will expand the description of the ensemble stacking concept in the introduction. This will be preceded by insights from the comprehensive overview by Zounemat-Kermani *et al.* (2021), which details the evolution and application of ensemble methods in hydrological prediction. We will also incorporate references to the work of Najafi and Moradkhani (2016), who explored various strategies for creating multi-model ensembles in the context of WSF.

In addition, we will cite more recent studies that demonstrate the growing application of ensemble stacking techniques in hydrological forecasting, such as: *For example, Li et al. (2019) investigated stacked ensemble models for long-term streamflow forecasting using advanced pre-processing techniques to enhance model stability. Mallick, Talukdar and Ahmed (2022) applied stacking models for real-time flood forecasting, showing significant improvement in prediction accuracy compared to individual models. Similarly, Granata and Di Nunno (2024) demonstrated the benefits of meta-learners in complex streamflow prediction tasks, while Xu et al. (2024) employed stacking methods with hybrid feature selection strategies for improved water resource management in Central Italy.*

3.l. Line 205: in the context of operational hydrologic prediction models, data "assimilation" has a very specific connotation: formal methods for using new observational data, such as observed snowpack, to update the internal states, such as predicted snowpack, of a process-based (dynamical, physics-oriented) streamflow simulation model, often using fairly complex methods like ensemble Kalman filtering. It is not normally used to refer to the use of some particular data type, such as snow data, as an input predictor variable in a data-driven (statistical or machine-learning) streamflow model.

Thank you for highlighting this inconsistency in the use of terminology. We will amend the terminology following your suggestions throughout the text.

3.m. Lines 287-288: excellent point. The authors might wish to cite literature that backs up this result, such as the excellent overview article of Hagedorn et al. (2005) and the multi-model ensemble WSF modeling article of Fleming et al. (2021).

We appreciate your positive feedback. We will place these findings in the context of the suggested references, including those previously noted on ensemble techniques in hydrology.

3.n. Figure 6: this a great illustration! I do have one question though: are all the base models used for the Vaksh and Kashkadarya rivers? It's hard to tell from the figure panels.

We appreciate your positive feedback on Figure 6. For the meta-learning model, we use base model predictions that meet a 0.2 R-squared coefficient threshold (lines 181-185). As a result, *the resulting ensembles typically consist of fewer than 16 base models (4 different models x 4 different snow inputs). We observed two trends: 1) the later the issue date, the larger the number of base models in the ensemble, and 2) larger catchments tend to include more base models, possibly due to the coarse resolution of snow products. For some rivers, such as the Vaksh and Kashkadarya, this leads to significantly fewer base models being used for ensemble stacking.* We will highlight these peculiarities in the revised version of the manuscript.

3.o. Line 348, might suggest rephrasing this in a more specific way, such as "suggest that useful near-real time SWE estimates, suitable for operational seasonal river discharge volume forecasting, can be effectively"

3.p. Line 350: "and enlarge during the snow ablation phase" – confusing wording

3.q. Lines 365-370: the entire paragraph (except for the excellent final sentence) is muddled. Please rewrite more simply and clearly.

3.r. Line 373: "confirms this assumption" – what assumption?

3.s. Lines 398, "is assumingly reasoned by their compensation" – this is meaningless, please rewrite.

3.t. Lines 400-410: excellent points.

Thank you for these suggestions. We will revise the text throughout to enhance clarity and incorporate the recommended rephrasing where applicable, ensuring that the manuscript is easily understandable.

References:

Granata, F. and Di Nunno, F.: Forecasting short- and medium-term streamflow using stacked ensemble models and different meta-learners, Stoch. Environ. Res. Risk Assess., 38, 3481–3499, https://doi.org/10.1007/s00477-024-02760-w, 2024.

Li, Y., Liang, Z., Hu, Y., Li, B., Xu, B., and Wang, D.: A multi-model integration method for monthly streamflow prediction: modified stacking ensemble strategy, J. Hydroinformatics, 22, 310–326, https://doi.org/10.2166/hydro.2019.066, 2019.

Mallick, J., Talukdar, S., and Ahmed, M.: Combining high resolution input and stacking ensemble machine learning algorithms for developing robust groundwater potentiality models in Bisha watershed, Saudi Arabia, Appl. Water Sci., 12, 77, https://doi.org/10.1007/s13201-022-01599-2, 2022.

Najafi, R. M. and Moradkhani, H.: Ensemble Combination of Seasonal Streamflow Forecasts, J. Hydrol. Eng., 21, 4015043, https://doi.org/10.1061/(ASCE)HE.1943-5584.0001250, 2016.

Xu, X., Chen, F., Wang, B., Harrison, M. T., Chen, Y., Liu, K., Zhang, C., Zhang, M., Zhang, X., Feng, P., and Hu, K.: Unleashing the power of machine learning and remote sensing for robust seasonal drought monitoring: A stacking ensemble approach, J. Hydrol., 634, 131102, https://doi.org/https://doi.org/10.1016/j.jhydrol.2024.131102, 2024.

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.