Author's replies

Understanding the relationship between streamflow forecast skill and value across the western US

[Manuscript ID #2024-4046]

Our responses to the reviewer's comments appear in *blue italics,* and the proposed revisions (at this stage) in *red italics*.

Referee 1

The manuscript by Modi et al. presents a study on the link between forecast skill and value, in the case of a sample of unmanaged, snow-dominated stations in the United States. The authors focus on the prediction of low AMJJ volumes issued based on the ESP method with a distributed and an LSTM models, or taken directly from the NRCS operational forecasts. Synthetic forecasts based on streamflow climatology and introducing deviations in mean and in standard deviation serve as a reference to assess errors in true forecasts and derive skill and value for controlled forecast errors. Results reveal a symmetry in forecast skill, but an asymmetry in forecast value, and discuss the inadequacy of the initially chosen skill metric to explain value.

The paper is of very high quality, well written and very well illustrated. It tackles a lot of different scientific objectives, which include comparing the chosen LSTM and distributed models, and studying the relationship between skill in value in controlled and real forecast systems. I was unsure to which extent the first objective serves the second, or not, because the paper becomes lengthy with information that is secondary to the skill-value relationship. Nevertheless, I recommend this paper for publication provided that the points below are addressed or commented on.

Response: Thank you for your careful review of our manuscript and the constructive comments that helped us improve our work.

General comments

1) Both WRFH and the LSTM generate daily streamflow volumes summed up to generate AMJJ volumes. In the case of the LSTM, it is not clear why the model was not trained on AMJJ volumes directly.

Response: Thank you for this question. This study builds upon our earlier work, extending the LSTM-ESP framework and further evaluating its applicability. We used ESP for both models to maintain a consistent methodological approach in our study. While NRCS statistical forecasts are trained directly on AMJJ volumes, our approach focuses on generating daily streamflow volumes and summing them to seasonal totals, which is consistent with the WRFH approach. Training the LSTM directly on AMJJ volumes remains a promising avenue for research currently, as it could be computationally more efficient. We will add a sentence in section 2.4 for further clarification on LSTM approach. More details on our previously published work on this topic can be accessed at: https://doi.org/10.1029/2024MS004582.

Original: "Both forecast systems generate daily streamflow volumes from April to July, which are summed up to generate AMJJ volumes. However, it is important to note that WRFH is run on an hourly timescale, and its outputs are aggregated to daily values."

Proposed: "It is important to note that WRFH is run on an hourly timescale, and its outputs are aggregated into AMJJ volumes. Similarly, the LSTM follows the WRFH approach but runs on a daily timescale, with its outputs aggregated into AMJJ volumes."

2) Section 2.1.3: The same cost is used for hits and false alarms. One could argue that a false alarm does more damage than just the preventive cost since it may deteriorate trust in or reputation of the decision-making institutions. This is not something accounting for here, but that would be worth discussing.

Response: Thank you for raising this important point regarding the asymmetric impact of false alarms versus hits. In the current PEV methodology, we use the same cost for hits and false alarms to maintain a simplified and standardized assessment approach. However, we acknowledge that in real-life decision-making, false alarms can have broader consequences beyond the immediate preventive costs (C), such as spoiling trust in forecasting systems and decision-making authorities. While this aspect is not explicitly accounted for in our analysis, it is a valuable consideration for future work. Exploring ways to incorporate reputation-related costs or adjusting the cost-loss ratio to reflect symmetric impact could enhance the framework's applicability. We will add the following paragraph to the Discussion section to address this limitation.

Proposed: "We also recognize that while the PEV framework assumes equal costs for hits and false alarms, real-life decision-making may be more sensitive to false alarms due to their potential to damage trust in forecasting systems and decision-making authorities." (L. 694-696)

Future work could explore ways to incorporate asymmetric cost structures or impacts of reputation to better reflect these considerations in operational settings." (L. 713-714)

3) Throughout the results section, and related to Figure 8 and L237, it was not clear to me which tau value was chosen, or if a range of tau's were used in the assessment of the APEVmax. I think this point requires clarification in Section 2.1.3, and potentially reminders in the interpretation of results.

Response: Thank you for your feedback. To clarify, the PEV framework is applied across a range of critical probability thresholds ($0<\tau<1$), generating a set of possible PEV values for each cost-to-loss ratio ($0<\alpha<1$), as mentioned in L202-203. This approach allows us to identify PEV_{max} from the range of curves and compute the area under the curve ($APEV_{max}$). We will revise Section 2.1.3 to explicitly state this process and include reminders in the Results section to ensure clarity in the interpretation of Figure 8 and related discussions.

Original: "To represent the forecast value in this study, we calculated the area under the PEV_{max} curve (now "$APEV_{max}$") using the trapezoidal rule (Amlung et al., 2015). This method approximates the area by dividing the curve into trapezoids and integrating their areas. The resulting metric can be used as "forecast value of a given forecast system" when the maximum economic benefits across all α are obtained at their respective τ (shown by the red shading in Fig. 2)."

Proposed: "To represent the forecast value in this study, we calculated the area under the PEV_{max} curve (now "$APEV_{max}$") using the trapezoidal rule (Amlung et al., 2015). This method approximates the area by dividing the curve into trapezoids and integrating their areas. While negative PEV_{max} values are possible, they are excluded from the area calculation. Note that the PEV framework is applied iteratively across a range of critical probability thresholds ($0<\tau<1$) to identify PEV_{max} and to compute $APEV_{max}$ by integrating over the corresponding curve. The resulting metric can be used as the "forecast value of a given forecast system" when the maximum economic benefits across all α (shown by the red shading in Fig. 2)."

4) Section 2.1.1 could benefit from a few clarifications. In particular the phrases “percentile of dryness” or “driest 2% conditions” were a bit unclear. The variables are listed, but the time step or period to be considered are unclear, and would be interesting to have for reference. It would also be interesting to add a sentence to state why this choice of methodology here, and why deviate from the methodology proposed by the USDM. The length of the historical period would also be interesting to have at this stage.

Response: Thank you for your comment. The phrase “percentile of dryness” refers to the percentile ranking of hydrologic variables (such as precipitation, streamflow, etc.) within their historical distribution, indicating how dry the current conditions are relative to past events. The analysis is calculated over a standard scale of 1-3 months based on the historical period of record. To avoid confusion here, we will clearly mention in Section 2.3 the historical period used for this analysis since that is associated with the availability of operational forecasts. Furthermore, we will clarify why we focused on AMJJ streamflow volumes and deviated from the USDM’s multi-variable approach, as this choice was made to directly capture hydrologic drought conditions and maintain consistency with the objectives of this study.

Original: “The U.S. Drought Monitor (USDM) classifies drought into five categories based on threshold percentiles in key hydroclimate quantities, (e.g., precipitation, soil moisture, streamflow) – D0 (Abnormally dry), D1 (Moderate drought), D2 (Severe drought), D3 (Extreme drought), and D4 (Exceptional drought), with D0 being the least intense and D4 the most intense (Svoboda et al., 2002). Each category corresponds to specific percentile ranges of historical drought severity, with D0 indicating conditions in the 21st to 30th percentile of dryness, D1 in the 11th to 20th percentile, D2 in the 6th to 10th percentile, D3 in the 3rd to 5th percentile, and D4 representing the driest 2% of conditions. This study uses a categorical definition of hydrologic drought, occurring when the AMJJ streamflow volume that falls below the 25th percentile (P_{25}) of the historical record. To assess the skill-value relationship across different drought severities, we also consider two additional hydrological thresholds where the AMJJ volume falls below the 35th percentile and a severe drought where it falls below the 15th percentile.”

Proposed: “The U.S. Drought Monitor (USDM) classifies drought into five categories based on threshold percentiles in key hydroclimate quantities, e.g., precipitation, soil moisture, streamflow, over a standard 1-3 month period, based on a historical period of record – D0 (Abnormally dry), D1 (Moderate drought), D2 (Severe drought), D3 (Extreme drought), and D4 (Exceptional drought), with D0 being the least intense and D4 the most intense (Svoboda et al., 2002). Each category corresponds to specific percentile ranges of historical drought severity, with D0 indicating conditions in the 21st to 30th percentile of dryness, D1 in the 11th to 20th percentile, D2 in the 6th to 10th percentile, D3 in the 3rd to 5th percentile, and D4 representing the driest 2% of conditions based on the historical distribution of hydrologic variables. For clarity, the term “percentile of dryness” refers to the relative position of the observed value within this historical distribution. This study uses a categorical definition of hydrologic drought, occurring when AMJJ streamflow volume falls below the 25th percentile (P_{25}) of the historical record. To assess the skill-value relationship across different drought severities, we also consider two additional hydrological thresholds: one where AMJJ volume falls below the 35th percentile and another where it falls below the 15th percentile, indicating severe drought conditions. This approach deviates from the USDM methodology, which typically uses a range of hydroclimatic variables for its classification. We chose to focus specifically on AMJJ streamflow volumes to capture hydrologic drought conditions more directly and to maintain consistency with the study’s objectives.”

5) Section 2.1.2: A sum may be missing (in the equation or in the text) to compute losses for several forecasts. Related to this, the term n is not defined. Regarding notations, z is rather a probability of exceedance/non-exceedance associated with quantile y_z. Related to the final discussion on the inadequacy of this skill metric to reflect value, the equal weighting of the 3 quantiles is probably not suitable, nor resembling actual decision-making contexts (reflecting unequal importance on high/low volumes or asymmetrical decision thresholds). Could the author discuss this? Could another weighting or picking of quantile values be enough to match value patterns?

Response: Thank you for pointing that out. There is a sum missing in equation 1, and n corresponds to the number of observations (n~5 per basin since we are considering only drought years). We will revise the equation in the next version.

Further, we thank the reviewer for bringing up the issue of the equal weighting of the three quantiles in the context of decision-making. While the use of quantile loss as a skill metric is employed for assessing the accuracy of probabilistic forecasts, we acknowledge that this approach does not fully capture the unequal importance of quantiles, especially when the decision-making context involves asymmetrical thresholds or varying cost-loss functions associated with different forecast outcomes. We acknowledge that certain quantiles may indeed be more critical than others, particularly in scenarios where higher or lower values carry disproportionately more weight in terms of socioeconomic impact. This is especially true during drought events. We agree that exploring alternative weighting schemes for quantile loss or selectively focusing on specific quantiles could better match the decision-making contexts. Future work could investigate different functions for weighted quantile loss, assigning higher weights to quantiles that reflect their context in decision-making.

Proposed (in the Discussion): "To better reflect decision-making contexts, it may be beneficial to explore weighted quantile loss metrics, where different quantiles receive different weights depending on their relative importance in decision-making. Such a weighting scheme would better align with situations where high or low values have disproportionate consequences, as is often the case in hydrologic forecasting."

6) Section 2.3: This section may benefit from some discussion points about the choice of a normally distributed ensemble, which later appears to be a limit, about the fact that forecasts often overestimate in dry conditions and underestimate in wet conditions which is not mimicked here, about the fact that deviations applied reflect error in mean (bias) or characteristics in terms of spread (sharpness), but the likely important feature here is rather discrimination, which is not experimented on. Related to Figure 4, a comment on the year-to-year variation in the forecast would be helpful. Are they solely due to the exclusion of the forecast year?

Response: We thank the reviewer for their insightful comment. Using a normally distributed ensemble is indeed a simplifying assumption, and we recognize that this choice does not fully capture the complexities of real-world forecasts. In practice, forecasts often exhibit systematic biases, such as overestimating in dry conditions and underestimating in wet conditions (Modi et al., 2021). This is a potential limitation of the current approach, as it does not explicitly incorporate such biases. Moreover, while we applied errors to the mean (bias) and changes in standard deviation (variability) to simulate forecast errors, we agree that the ability of the forecast to discriminate between different scenarios—i.e., its ability to identify and differentiate between drought and non-drought conditions—is a crucial feature that was not explicitly explored in this study. Discrimination, or the forecast's ability to correctly classify extreme events, is indeed an important aspect of forecast skill that warrants further attention but was outside the scope of this manuscript, given the already lengthy treatment of each topic. That said, we believe this to be a

viable area of future inquiry to investigate discrimination-based metrics to complement the mean and spread-based error metrics. We have added a brief paragraph in the Discussion section

Regarding Figure 4, the minor year-to-year variations in the forecast are solely due to the exclusion of the forecast year. This method ensures that the forecast is independent of the year being evaluated, but it introduces minor variations between years.

Proposed: "The use of a normally distributed ensemble to develop synthetic forecasts is a simplification that allows us to model forecast uncertainty in a controlled manner. While real-world forecasts often exhibit more complex, irregular distributions and biases. For example, these may be overestimated in dry conditions and underestimated in wet conditions (Modi et al., 2021). A normal distribution was chosen to solely isolate the impact of mean and standard deviation. We recognize that this assumption does not fully capture the nuances of real-world forecast errors, such as skewness or non-normality in extreme conditions, which would require detailed treatment outside the scope of this analysis."

Reference: Modi, P. A., Small, E. E., Kasprzyk, J., & Livneh, B. (2021). Investigating the role of snow water equivalent on streamflow predictability during drought. Journal of Hydrometeorology, 23(10), 1607–1625. https://doi.org/10.1175/JHM-D-21-0229.1

7) Section 2.5: Please clarify the RMAD criterion, in particular how errors between true and synthetic forecasts are calculated given that they are ensemble forecasts.

Response: Thank you for your comment. The Relative Median Absolute Deviation (RMAD) criterion compares the variability between synthetic and true forecasts by measuring the median of the relative absolute errors. Given that both the true and synthetic forecasts are ensemble forecasts, the errors between the two are calculated by first comparing the corresponding ensemble members. Specifically, for each forecast, the absolute difference between the corresponding ensemble members of the true and synthetic forecasts is computed, and these errors are then normalized by the true forecast values. The median of these normalized errors is then taken to quantify the RMAD, with lower values indicating better alignment between the two ensembles.

Original: "We use the Relative Median Absolute Deviation (RMAD) to compare the variability between synthetic and true forecasts. RMAD measures the median of the relative absolute errors between the true and synthetic forecasts, with values closer to 0 indicating smaller deviations and better alignment between the forecasts."

Proposed: "We use the Relative Median Absolute Deviation (RMAD) to compare the variability between synthetic and true forecasts. RMAD measures the median of the relative absolute errors between the true and synthetic forecasts. Since both the true and synthetic forecasts are ensemble forecasts, the errors are calculated by first determining the absolute difference between corresponding ensemble members. These absolute errors are then normalized by the true forecast values to compute relative errors. The median of these relative errors across the ensemble members is then used to quantify RMAD, with values closer to 0 indicating smaller deviations and better alignment between the true and synthetic forecasts."

8) Throughout the manuscript and more specifically L521 "higher forecast skill and value were associated with negative errors in standard deviation": "negative errors in standard deviation" can be misleading. Changes in sharpness, in themselves, are not errors if they are not associated with an absence of bias (seen in the skill matrices). Sharpness is not a performance metric. Here negative errors in standard deviation rather mean that the ensemble is close to a deterministic forecast, which is only associated with high forecast skill if, and only if, the forecasts are not biased. I recommend changing the phrase error in

standard deviation throughout the paper, and recommend the following paper: Gneiting, T., Balabdaoui, F., Raftery, A.E., 2007. Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69, 243–268.

Response: Thank you for the insightful comment. We will revise the phrase "error in standard deviation" to "change in variability" throughout the manuscript. We agree that the reduced variability is linked to high forecast skill only if the ensemble is unbiased, and we will reword the relevant sentences to better reflect that distinction. Additionally, we will incorporate a reference you suggested in the paper to further support our discussion.

9) Figure 12: I don't understand why the three models appear in Figure 12a if this figure shows the synthetic forecasts. If my understanding is correct, only observations are used to generate the synthetic forecasts. Could you please clarify? Also L611 "the true LSTM and the corresponding synthetic forecast" has me confused.

Response: Thank you for your comment. To clarify, the background heatmaps in Figures 10 and 11 represent the median skill and value from synthetic forecasts, while the scatter points represent the true forecast systems. Each dot in those figures represents a basin. We use this overlap between synthetic and true forecasts from Figures 10 and 11 and plot the corresponding synthetic forecasts in Figure 12a. We describe this in L567 in section 3.3.1. We will revise the figure legend and caption to better explain this distinction.

Original: "Figure 12: Scatter plots depicting the relationship between skill (NMQloss) and value (APEVmax) for synthetic and true forecast systems. The points in (a) and (b) represent the synthetic forecast (the grid of the heatmap) that overlap with true forecast systems (scatter point) in Figures 4.10 and 4.11. Each point represents a basin, with the fitted exponential curves (dashed lines) indicating general trends and values in round brackets correlation."

Proposed: "Figure 12: Scatter plots depicting the relationship between skill (NMQloss) and value (APEVmax) for synthetic and true forecast systems. The points in (a) and (b) represent the synthetic forecast (the grid of the heatmap) that overlap with true forecast systems (scatter point) in Figures 4.10 and 4.11. Each point represents a basin, with the fitted exponential curves (dashed lines) indicating general trends and values in round brackets correlation. It should be noted that we use the overlap from Figures 10 and 11 to plot synthetic forecasts (corresponding to true forecasts) in Fig. 12a."

10) The use of term "skill" in the paper is not always consistent: L630 "better captured in categorical measures than skill": In the way the word "skill" is used in this study, it is not a metric (sometimes the case when it is the comparison of the performance of a forecast system with the performance of a benchmark) but rather a term used to qualify the performance of the forecast. Based on the use of the word "skill" in this study, categorical measures could very well be metrics used to define forecast skill. I suggest rephrasing. Also L654 "forecast skill generally reflects the accuracy of forecasts" can be unclear as accuracy can be perceived as one feature of forecast skill.

Response: Thank you for pointing that out. In this study, we define forecast skill as the model's ability to accurately predict observed values rather than as an improvement over a reference forecasting system. We recognize that this definition of "skill" differs from the standard approach of comparing to a reference forecast, but we elect to use "forecast skill" here for consistency.

11) Key references are well used to corroborate or discuss results or limits of this work. Some references about asymmetry in decision-making, about synthetic forecasts, or about the need for adequacy between

skill and value metrics can be found in the following works. I let the authors consider their relevance for their work:

Peñuela, A., Hutton, C., Pianosi, F., 2020. Assessing the value of seasonal hydrological forecasts for improving water resource management: insights from a pilot application in the UK. Hydrology and Earth System Sciences 24, 6059–6073. https://doi.org/10.5194/hess-24-6059-2020

Rouge, C., Peñuela, A., Pianosi, F., 2023. Forecast Families: A New Method to Systematically Evaluate the Benefits of Improving the Skill of an Existing Forecast. Journal of Water Resources Planning and Management 149, 04023015. https://doi.org/10.1061/JWRMD5.WRENG-5934

Crochemore, L., Materia, S., Delpiazzo, E. et al. 2024. A framework for joint verification and evaluation of seasonal climate services across socio-economic sectors. Bulletin of the American Meteorological Society. https://doi.org/10.1175/BAMS-D-23-0026.1

Response: Thank you for suggesting additional references. We will review and integrate the suggested works to further support and strengthen our discussion.

Proposed: "The relationship between forecast skill and value in seasonal streamflow forecasting is not only influenced by the operational characteristics of the water management system but also by the intrinsic qualities of the true forecasts themselves, particularly during extreme events like drought (Giuliani et al., 2020; Peñuela et al., 2020)."

Proposed: "Synthetic forecasts are used to more clearly understand the role of forecast errors on economic value (Rougé et al., 2023)."

Proposed: "However, in an increasingly complex economy with a growing and diverse user base, the relationship between forecast skill – the accuracy of the forecast and the forecast value – the forecast's impact on decision-making and economic outcomes is far from straightforward (Crochemore et al., 2024)."

Detailed comments

L58-59: Please review the definition of probabilistic seasonal streamflow forecasts, as it is not necessarily volumes, the concept of season can be unclear, and the cited methods are not always combined.

Response: Thank you for pointing that out. We will revise the sentence to improve the definition of probabilistic seasonal streamflow forecasts.

Original: "Probabilistic seasonal streamflow forecasts predict the range of potential water flow volumes in rivers or streams over a season using a combination of process-based and data-driven models, historical data, and climate forecasts."

Proposed: "Probabilistic seasonal streamflow forecasts estimate the likelihood of different streamflow signatures over a given period, using various approaches such as process-based models, data-driven models, historical data, or climate forecasts, or a combination of these approaches."

L65: A definition and references for ESP would be necessary. Consider the following work: Day, G., 1985. Extended Streamflow Forecasting Using NWSRFS. J. Water Resour. Plann. Manage. 111, 157–170.

Response: Thank you for pointing that out. We will add a short definition of ESP and the reference.

Original: "Over time, forecasting frameworks like the Ensemble Streamflow Predictions (ESP) have significantly improved in predicting water volumes through advances in hydrological modeling, the use of more accurate meteorological inputs, and the adoption of more sophisticated forecasting methods (Clark et al., 2016; Li et al., 2017)."

Proposed: "Ensemble Streamflow Predictions (ESP) is a hydrological forecasting method that generates multiple streamflow simulations using historical meteorological data as inputs to a hydrologic model (Day, 1985). Over time, ESPs have significantly evolved in predicting water volumes through advances in hydrological modeling, the incorporation of outputs from dynamical meteorological and climate models, and the adoption of more sophisticated forecasting methods (Clark et al., 2016; Li et al., 2017)."

L66 "more accurate": I am unsure whether this is about accuracy since ESP relies on climatology. It is rather about using outputs from dynamical meteorological or climate model instead.

Response: Response: Thank you for pointing that out. We modify the language to 'evolved', and we add a short definition of ESP and the reference from previous comment.

Original: "Over time, forecasting frameworks like the Ensemble Streamflow Predictions (ESP) have significantly improved in predicting water volumes through advances in hydrological modeling, the use of more accurate meteorological inputs, and the adoption of more sophisticated forecasting methods (Clark et al., 2016; Li et al., 2017)."

Proposed: "Over time, ESPs have significantly evolved in predicting water volumes through advances in hydrological modeling, the incorporation of outputs from dynamical meteorological and climate models, and the adoption of more sophisticated forecasting methods (Clark et al., 2016; Li et al., 2017)."

L115-116 "particularly during extreme events like droughts": references would be needed to support this. Consider the following work: Giuliani, M., Crochemore, L., Pechlivanidis, I., Castelletti, A., 2020. From skill to value: isolating the influence of end user behavior on seasonal forecast assessment. Hydrology and Earth System Sciences 24, 5891–5902. https://doi.org/10.5194/hess-24-5891-2020

Response: Thank you for providing this excellent suggestion. This paper explicitly discusses the skill-value relationship during drought years and reinforces the argument made. We will add the reference in the revised manuscript.

Original: "The relationship between forecast skill and value in seasonal streamflow forecasting is not only influenced by the operational characteristics of the water management system but also by the intrinsic qualities of the true forecasts themselves, particularly during extreme events like drought."

Proposed: "The relationship between forecast skill and value in seasonal streamflow forecasting is not only influenced by the operational characteristics of the water management system but also by the intrinsic qualities of the true forecasts themselves, particularly during extreme events like drought (Giuliani et al., 2020; Peñuela et al., 2020)."

L124 "forecasts respond to fundamental statistical measures" Consider reformulating.

Response: Thank you for pointing that out. We will reformulate the sentence in the revised manuscript.

Original: "We then evaluate how both synthetic and true forecasts respond to fundamental statistical measures such as error in mean and standard deviation."

Proposed: "We then assess how the performance of both synthetic and true forecasts are affected by modifying forecast properties such as mean and variability."

L128 and elsewhere: the word "evaluate" may be ambiguous in a paper about forecast value if it is used for both skill and value. "assess" could be a more neutral option.

Response: Thank you for the suggestion. We will revise this in the next version.

L128: Section 2.1.2 is rather about defining drought

Response: Thank you for pointing out the inconsistency. We will restructure the statement in the revised manuscript.

Original: "We begin by outlining the process for evaluating forecast skill using a quantile loss metric (Sect. 2.1.1) and defining drought, which serves as the basis for the categorical decision used to calculate the forecast value (Sect. 2.1.2). Section 2.1.3 describes the PEV framework for evaluating forecast value."

Proposed: "We begin by defining drought, which serves as the basis for the categorical criterion used to calculate the forecast value (Sect. 2.1.1). Section 2.1.2 outlines the process for assessing forecast skill using a quantile loss metric, while Section 2.1.3 describes the PEV framework for assessing forecast value."

L135 "fundamental performance metrics" as above, the choice of the adjective "fundamental" is not clear to me. I would suggest reformulating or clarifying.

Response: Thank you for pointing that out. We will fix this in the revised version and elsewhere.

Original: "Section 2.5 provides an overview of fundamental performance metrics."

Proposed: "Section 2.5 provides an overview of key performance metrics."

L196-197: "rate of occurrence" I suggest introducing s here.

Response: Thank you for pointing that out. We will fix this in the revised version.

Original: "Where $-\infty<PEV<1$ and each expense term is the summation of the contingency table elements, each weighted by the rate of occurrence."

Proposed: "Where $-\infty<PEV<1$ and each expense term is the summation of the contingency table elements, each weighted by the rate of occurrences."

L224-225: This sentence is key for understanding these 3 parameters. I would suggest placing it earlier in the section.

Response: Thank you for pointing that out. We will reorder these suggested sentences earlier in the introductory paragraphs.

L234-239: How are negative values accounted for when calculating the area? Is it possible to have negative PEVmax values?

Response: Thank you for that question. It is possible to have negative PEV_{max} values, but they are excluded from the area calculation. We will add a sentence to indicate that in the revised version.

Original: "To represent the forecast value in this study, we calculated the area under the PEVmax curve (now "APEVmax") using the trapezoidal rule (Amlung et al., 2015). This method approximates the area by dividing the curve into trapezoids and integrating their areas."

Proposed: "To represent the forecast value in this study, we calculated the area under the PEVmax curve (now "APEVmax") using the trapezoidal rule (Amlung et al., 2015). This method approximates the area by dividing the curve into trapezoids and integrating their areas. While negative PEV_{max} values are possible, they are excluded from the area calculation."

L239: While 0 is the theoretical minimum, 0.9 seems to be the observed maximum. If that is correct, I suggest clarifying this by stating the theoretical maximum (infinity?) before this observed maximum.

Response: Thank you for the question. Since we exclude the negative PEV_{max} values for the area, the theoretical minimum and maximum are 0 and 0.9, respectively.

Original: "This value ranges from 0, showing the lowest overall economic value, to 0.9, being the highest overall economic value in this study."

Proposed: "This value ranges from 0, representing the theoretical minimum economic value, to 0.9, representing the highest overall economic value in this study, as negative PEV_{max} values are excluded from the area calculation."

L247 "snow-dominated basins (i.e., unmanaged headwater systems)": the correspondence between the two basin types is not direct. Some snow-dominated basins in areas with altitude gradients can be heavily managed/influenced by hydropower dams. Please clarify.

Response: Thank you for pointing that out. We are only interested in the intersection between unmanaged and snow-dominated basins, which are often the headwater catchments that supply flows to downstream managed systems. We will fix this statement in the revised version.

Original: "Water availability in snow-dominated basins (i.e., unmanaged headwater systems) depends heavily on snowmelt timing and volume, making accurate forecasts essential for managing water resources and mitigating drought risks."

Proposed: "Water availability in basins that are both unmanaged and snow-dominated are of interest here. These are often headwater catchments, with flows heavily driven by snowmelt timing and volume, making accurate forecasts essential for managing water resources and mitigating drought risks."

L269: Give the full name for SWE as this is the first occurrence.

Response: Thank you for pointing that out. We will add the full name in the revised version.

Original: "These basins are colored by the ratio of April 1 SWE to water-year to-date cumulative precipitation that is derived from gridded snow and meteorological forcings (as described in Table A1)."

Proposed: "These basins are colored by the ratio of April 1 Snow Water Equivalent (SWE) to water-year to-date cumulative precipitation, which refers to the accumulated precipitation from the beginning of the current water year, Oct 1, to April 1, derived from gridded snow and meteorological forcings (as described in Table A2)."

L269: "water-year-to-date" may be worth explaining in its first occurrence as well.

Response: Thank you for pointing that out. We will add an explanation in the revised version.

Original: "These basins are colored by the ratio of April 1 SWE to water-year to-date cumulative precipitation that is derived from gridded snow and meteorological forcings (as described in Table A1)."

Proposed: "These basins are colored by the ratio of April 1 Snow Water Equivalent (SWE) to water-year to-date cumulative precipitation, which refers to the accumulated precipitation from the beginning of the current water year, Oct 1, to April 1, derived from gridded snow and meteorological forcings (as described in Table A2)."

L270: The reference to Table A1 is not clear to me.

Response: Thank you for pointing that out. We have revised the statement to reference Table A2 instead of Table A1. We will also fix some inconsistencies concerning Table A2 (per later comments L315 and L377).

Original: "These basins are colored by the ratio of April 1 SWE to water-year-to-date cumulative precipitation that is derived from gridded snow and meteorological forcings (as described in Table A1)."

Proposed: "These basins are colored by the ratio of April 1 Snow Water Equivalent (SWE) to water-year to-date cumulative precipitation, which refers to the accumulated precipitation from the beginning of the current water year, Oct 1, to April 1, derived from gridded snow and meteorological forcings (as described in Table A2)."

L283 "WY2006-2022": this notation used throughout the manuscript should be explained here.

Response: Thank you for pointing that out. We will explain the notation in the revised version.

Original: "It should be noted that the standard deviation here is assumed to be equal to interannual variability seen in the observations based on the retrospective years available in the forecast system (for example, in Fig. 4, WY2006-2022)."

Proposed: "It should be noted that the change in standard deviation here is assumed to be with respect to interannual variability seen in the observations based on the retrospective years available in the forecast system. We generate the forecasts for the years WY2006-2022, where "WY" represents the water year, Oct. 1 – Sep. 30 (Fig. 4)."

L290: I suggest citing the number of ensemble members used in practice here

Response: Thank you for pointing that out. We will add the number of ensemble members in the revised version.

Original: "An ensemble of forecast members is then generated, normally distributed around the modified mean and standard deviation."

Proposed: "An ensemble of 39 forecast members (explained further in section 2.4) is then generated, normally distributed around the modified mean and standard deviation."

L315 "snowpack information in the form of snow water equivalent": Based on Figure 5, it seems only to be the case for the LSTM. Is this information also used in the case of WRFH?

Response: Thank you for pointing that out. We will revise the statement on the next version.

Original: "In these systems, the primary input data consists of historical meteorology, geospatial basin attributes, snowpack information in the form of snow water equivalent (SWE), and streamflow observations—also used for training and validation (Table A1)."

Proposed: "In these systems, the primary input data consists of historical meteorology, geospatial basin attributes, snowpack information in the form of SWE (only for LSTM model), and streamflow observations—also used for training and validation (Table A2)."

L315: The reference to Table A1 does not seem correct.

Response: Thank you for pointing that out. The reviewer is correct; instead, the reference should be to Table A2. We will revise the references.

Original: "In these systems, the primary input data consists of historical meteorology, geospatial basin attributes, snowpack information in the form of snow water equivalent (SWE), and streamflow observations—also used for training and validation (Table A1)."

Proposed: "In these systems, the primary input data consists of historical meteorology, geospatial basin attributes, snowpack information in the form of SWE (only for LSTM model), and streamflow observations—also used for training and validation (Table A2)."

L329: the probabilities extracted from the forecast ensembles can only be comparable if the ensembles have the same number of members. Here it seems to be the case (Section 2.4.1), but I suggest mentioning this here in Section 2.4 already for clarification.

Response: Thank you for pointing that out. We acknowledge that the exceedance probabilities are comparable when the forecast ensembles have the same number of members. We will make a clarification in the statement.

Original: "All true forecasts include five forecasted exceedance probabilities at 90, 70, 50, 30, and 10%."

Proposed: "All true forecasts have the same number of ensemble members, and five forecasted exceedance probabilities computed at 90, 70, 50, 30, and 10% are extracted."

L350 "~20-30 years": In Figure 6, 23 years are mentioned, and L387 and 419, you mention the period 1983-2022 (40 years minus the forecast year). Please clarify.

Response: Thank you for pointing that out. Figure 6 is simply for illustration purposes, showing an ESP forecast of 23 ensemble members. All the true forecasts and synthetic forecasts are based on members from WY1983-2022 (40 years minus the forecast year). This has been clarified as per comment L329. We will also revise the following statement to clarify the confusion.

Original: "The result is a daily probabilistic hydrologic forecast ranging from 30 days up to 180 days from the forecast date that uses the spread in historical data from the past ~20 to 30 years as an analogue for the uncertainty in meteorological conditions after the forecast date."

Proposed: "The result is a daily probabilistic hydrologic forecast ranging from 30 days up to 180 days from the forecast date that uses the spread in historical data from the past ~20 to 30 years (shown in Fig. 6 – for illustration purposes, we only show 23 years here) as an analogue for the uncertainty in meteorological conditions after the forecast date."

L377 Here as well, the reference to Table A1 does not seem to match its content.

Response: Thank you for pointing that out. As per the earlier comment, we will revise the references.

Original: "Meteorological forcings used to run the WRF-Hydro (WRFH) include precipitation, average wind speed, 2 m average air temperature, incoming longwave and shortwave radiation, near-surface air pressure, and vapor pressure obtained from Analysis of Records for Calibration (AORC, Fall et al., (2023) – as detailed in Table A1)."

Proposed: "Meteorological forcings used to run the WRF-Hydro (WRFH) include precipitation, average wind speed, 2 m average air temperature, incoming longwave and shortwave radiation, near-surface air pressure, and vapor pressure obtained from Analysis of Records for Calibration (AORC, Fall et al., (2023) – as detailed in Table A2)."

L389-390: I suggest the term "initial states" instead of "memory states"

Response: Thank you for pointing that out. We will revise the statement in the next version.

Original: "For ESP forecasts on April 1, the WRFH simulation begins at the start of the water year, i.e., October 1, using true meteorological forcings to obtain WRFH's memory states (e.g., snowpack, soil moisture) on the forecast date."

Proposed: "For ESP forecasts on April 1, the WRFH simulation begins at the start of the water year, i.e., October 1, using true meteorological forcings to obtain WRFH's initial states (e.g., snowpack, soil moisture) on the forecast date."

Sections 2.4.2 and 2.4.3: it would be helpful to state in these sections the years used for model training/calibration (now in Appendices), in addition to the years of historical meteorology inputs and for which the forecasts are generated (already clear).

Response: Thank you for that suggestion. We agree that clearly stating the years used for model training/calibration could enhance clarity. However, to maintain a streamlined presentation in the main text, we believe it is best to keep all calibration details in one place, as currently presented in the appendices.

Section 2.5: It is generally advised to use different metrics for calibration/training and verification/validation. It is not clear here which metrics are used for which purpose.

Response: Thank you for pointing that out. The metrics used for calibration/training for each of the models are mentioned in the Appendices. Section 2.5 only mentions the metrics used to compare the historical performance (Section 3.1). We will revise the first sentence and add a statement towards the end of the paragraph clarifying this confusion in section 2.5.

Original: "We employed four fundamental performance metrics to calibrate/train the models and evaluate forecast accuracy, drawing from those widely adopted to quantify streamflow accuracy."

Proposed: "We employed four key performance metrics to compare the historical performance of our designed true forecast systems, drawing from those widely adopted to quantify streamflow accuracy."

Proposed (towards the end): "The metrics used to calibrate/train the true forecast systems are described in Appendices A1 and A2."

L473: If my understanding is correct, there is a single AMJJ value per year for the period 2006-2022 (17 values). How many years remain once only dry years are selected? Can it really ensure robust results for the rest of the study?

Response: Thank you for raising this critical question. We recognize that estimating forecast skill and value for a smaller sample size (as you correctly mentioned, n~5) is not ideal, affecting the statistical power of the analysis. This limitation arises due to the limited availability of operational forecasts and the need for sufficient ensemble members for ESP. We extend our formal acknowledgement of this limitation in the Discussion section. We propose adding the following to the Discussion section.

Proposed: "We acknowledge that estimating forecast skill and value for drought years necessitates a smaller sample size (here n~5), which is not ideal, affecting the statistical power of the analysis. This limitation arises due to the limited availability of operational forecasts and the need for sufficient ensemble members for ESP. Therefore, it would be important to assess whether a broader selection criterion or longer span of forecast availability would help ensure robust results."

L481 "As error in mean or standard deviation increase beyond these ranges, forecast skill worsens": This is arguable. Here the standard deviation reflects the sharpness of the probabilistic forecast. However, sharpness is a forecast characteristic rather than a performance metric.

Response: Thank you for pointing that out. As you correctly pointed out, we will rewrite the sentence for better clarification.

Original: "As error in mean or standard deviation increase beyond these ranges, forecast skill worsens."

Proposed: "As error in mean increase beyond these ranges, forecast skill worsens. However, an increase in standard deviation reflects the variability of the probabilistic forecast, which is a characteristic of the forecast rather than a direct performance metric."

L483: In the text, values seem to reach 0.9, but in Figure 8, the color scale ends at 0.5.

Response: Thank you for pointing that out. The $APEV_{max}$ technically reaches 0.9, but the color scale is capped at 0.5 only for illustration purposes (i.e., to see the asymmetry better). We will update the caption to reflect that in the revised version.

Proposed: "Figure 8: Sensitivity of quantile loss (forecast skill) and APEVmax (forecast value) to error in mean and change in variability for synthetic forecasts. The background heatmaps (a and b) represent synthetic forecasts, with lower values showing better forecast skill (closer to yellow) and higher values better forecast value (closer to yellow). We also illustrate four synthetic forecasts (shown in blue) corresponding to different error in mean and change in variability (c-f). The black line and ribbon represent a synthetic forecast, with the mean equal to the observation and the standard deviation representing the interannual variability of the observations. The red dots indicate drought events, defined as AMJJ volumes below P_{25}, whereas the histograms represent the hit (H), False Alarm (F), and Miss (M) rates. Note that the color scale for forecast value is capped at 0.5, although the actual values reach up to 0.9."

L496-498: This asymmetry is indeed interesting, and would benefit from some further discussion as to which parameters in the methodology cause this effect (AMJJ variable bounded by 0, tau value, alpha-s relationship and frequency of occurrence below 0.5 for droughts, …).

Response: Thank you for pointing that out. The observed asymmetry in forecast value is primarily due to the interplay between categorical measures, such as hit and false alarm rates, and mainly our focus on events below the P_{25} drought threshold. These factors cause forecast value to respond differently than forecast skill.

Original: "This highlights the different sensitivities of skill and value to error in mean and standard deviation, likely due to the interplay of categorical measures, where forecast value responds differently than forecast skill."

Proposed: "This asymmetry is largely due to the interplay of categorical measures, such as hit and false alarm rates, as well as our focus on events below the P_{25} drought threshold. These factors lead to different sensitivities of skill and value to error in mean and change in variability."

L532 "Each dot in Fig. 10 represents a basin with colors showing the median skill and value": is it just skill?

Response: Thank you for pointing that out. We will revise the statement to reflect that the colors only show the skill and not the value.

Original: "Each dot in Fig. 10 represents a basin with colors showing the median skill and value only during drought years."

Proposed: "Each dot in Fig. 10 represents a basin with colors showing the median skill during drought years."

L580: Given that only a type of forecast skill is investigated here, I suggest the following "This skill-value comparison between synthetic and true forecast systems indicates that factors beyond forecast skill, as defined in this study…"

Response: Thank you for the suggestion. We will revise the sentence in the next version.

Original: "This skill-value comparison between synthetic and true forecast systems indicates that factors beyond forecast skill influence the value of true forecast systems which we analyze in the following sections to some extent."

Proposed: "This skill-value comparison between synthetic and true forecast systems indicates that factors beyond forecast skill, as defined in this study, influence the value of true forecast systems, which we analyze in the following sections to some extent."

L653: References would be helpful at the end of this sentence.

Response: Thank you for the suggestion. We will cite a few references in the next version.

Original: "This study was motivated by recent literature showing that the relationship between forecast skill and value in hydrology is multifaceted and context dependent"

Proposed: "This study was motivated by recent literature showing that the relationship between forecast skill and value in hydrology is multifaceted and context dependent (Maurer and Lettenmaier, 2004; Rheinheimer et al., 2016; Hamlet et al., 2002; Portele et al., 2021, Giuliani et al., 2020)."

Figures

Figure 1: Equation numbers preceded by minus signs can be confusing. In the caption, "forecast probabilities are calculated from probabilistic forecasts" is redundant. Do you mean "threshold exceedance probabilities are calculated from probabilistic forecasts"?

Response: Thank you for pointing that out. We will remove the minus signs and use square brackets for equation numbers to avoid confusion in Figure 1. The PEV workflow starts with Step 1, where we use the forecast ensembles as input. We first calculate the forecast probability based on a defined drought event

(values 0.4 and 0.6 in red are the forecast probabilities). Based on this forecast probability and critical probability threshold (prescribed range of $0<\tau<1$ – also explained in detail in general comment #3), we decide to act in Step 2. We will revise the text as follows:

Original: "Flowchart showing the workflow to quantify the PEV using the probabilistic forecasts. For the calculation of PEV, forecast probabilities are calculated from probabilistic the forecasts (Step 1), a critical probability threshold (τ) is applied (Step 2), a contingency table is created (Step 3), and lastly, PEV is calculated across the prescribed range of α (Step 4). The PEV relies on contingency table parameters (H and F), climatological frequency (s), and cost-loss ratio (α). The equations were adapted from Richardson (2000) and Jolliffe and Stephenson (2003)."

Proposed: "Flowchart showing the workflow to quantify the PEV using the probabilistic forecasts. For the calculation of PEV, forecast probabilities (for a given event) are calculated from the forecasts (Step 1), a critical probability threshold (τ) is applied (Step 2), a contingency table is created (Step 3), and lastly, PEV is calculated across the prescribed range of α (Step 4). The PEV relies on contingency table parameters (H and F), climatological frequency (s), and cost-loss ratio (α). The equations were adapted from Richardson (2000) and Jolliffe and Stephenson (2003)."

Figure 2: The arrows used to indicate the cases when C<0 and C>L point to ranges where C>0 and C

Response: Thank you for pointing that out. We will revise the figure in the next version.

Figure 4: This figure is helpful. It may be worth mentioning in sub-figure (a) that the AMJJ volume for the forecast year is excluded, if that is the case.

Response: Thank you for pointing that out. Figure 4a indicates the workflow of synthetic forecasts. The synthetic forecasts are created by introducing errors in the mean and adjusting the variability of observations, with the forecast year's AMJJ observed volume included in the process.

Figure 5: "Historical meteorology" and "Basin attributes" are rather vague. I suggest specifying these to better highlight the differences between the three types of modelling/forecasting chains.

Response: Thank you for pointing that out. While specifying these terms in more detail would indeed help clarify the differences between the workflows, doing so would also complicate the representation of workflow in Figure 5. Instead, we refer readers to Table A2, which provides detailed descriptions of each of these inputs, including their sources.

Figure 7: The caption should explain the difference between shaded areas (distributions over the 76 basins) and the vertical lines.

Response: Thank you for that suggestion. We will revise the caption in the next version.

Original: "Figure 7: Historical model performance of true forecast systems. (a) Daily NSE, (b) NRMSE of the total April-July streamflow volumes, (c) daily correlation, and (d) Ratio of the standard deviation against observations for WRFH (default and calibrated) and LSTM (initial and final) models. Comparison shown for the 76 basins during the testing period, WY2001-2010."

Proposed: "Figure 7: Historical model performance of true forecast systems. (a) daily NSE, (b) NRMSE of the total April-July streamflow volumes, (c) daily correlation, and (d) the ratio of the standard deviation against observations for calibrated WRFH and fully trained LSTM models. The shaded areas represent the distributions of model performance metrics over the 76 basins, while the vertical lines indicate the performance of individual basins during the testing period, WY2001-2010."

Figure 8e: It is not clear why there is a miss based on the time series plot.

Response: Thank you for pointing that out. The histograms are calculated based on all the ensemble members for every given year. In Figures 8e and 8f, some ensemble members do cause misses (M=0.13) and false alarms (M=0.01). We will add a sentence for clarification in the revised version.

Original: "Fig. 8e, featuring a negative error in mean, hits all events (H=0.87) but suffers from high false alarms (F=0.70), resulting in a value of 0.03, while Fig. 8f, with a positive error in mean, has almost no false alarms (F=0.01) but a lower hit rate (H=0.28) resulting in a value of 0.20."

Proposed: "Fig. 8e, featuring a negative error in mean, hits all events (H=0.87) but suffers from high false alarms (F=0.70), resulting in a value of 0.03, while Fig. 8f, with a positive error in mean, has almost no false alarms (F=0.01) but a lower hit rate (H=0.28) resulting in a value of 0.20. It should be noted that some ensemble members cause misses (M=0.13) and false alarm rates (F=0.01) in Figures 8e and 8f, respectively."

Figure 9: "Synthetic errors": if they are calculated from the true forecasts, these are no longer synthetic errors.

Response: Thank you for pointing that out. We will fix the type in the revised version.

Original: "Figure 9: Synthetic errors in (a) mean and (b) standard deviation of three true forecast systems (NRCS, WRFH, and LSTM). Each point represents a basin, and the errors are reported for drought years (below the P_{25}) between WY 2006 and 2022. 76 basins are divided across six ranges, with the square bracket representing the number of basins within each range."

Proposed: "Figure 9: (a) Error in mean and (b) change in variability (with respect to interannual variability during WY2006-2022) of three true forecast systems (NRCS, WRFH, and LSTM). Each point represents a basin, and the errors/changes are reported for drought years (below the P_{25}) between WY 2006 and 2022. 76 basins are divided across six ranges, with the square bracket representing the number of basins within each range."

Typos

L63: Give the full name for NRCS

Response: Thank you for pointing that out. We will fix this as well as other instances in the revised manuscript.

Original: "For example, the NRCS forecasts have been widely used for water management and agricultural planning (Fleming et al., 2021)."

Proposed: "For example, the Natural Resources Conservation Services (hereafter "NRCS") forecasts have been widely used for water management and agricultural planning (Fleming et al., 2021)."

L122 "performance of true forecasts against observations generated in this study" can be unclear as to what is generated in this study.

Response: Thank you for pointing that out. We will improve the sentence for better clarity.

Original: "We begin by assessing the historical model performance of true forecasts against observations generated in this study."

Proposed: "We start by assessing the historical performance of true forecasts generated in this study by comparing them to observations."

L123: "models"

Response: Thank you for pointing that out. We will fix this in the revised manuscript.

Original: "This involves comparing the calibrated WRF-Hydro and fully trained LSTM model to assess their effectiveness in simulating streamflow volumes."

Proposed: "This involves comparing the calibrated WRF-Hydro and fully trained LSTM models to assess their effectiveness in simulating streamflow volumes."

L144: "when the AMJJ streamflow volume falls below"

Response: Thank you for pointing that out. We will fix this in the revised manuscript.

Original: "This study uses a categorical definition of hydrologic drought, occurring when the AMJJ streamflow volume that falls below the 25th percentile (P_{25}) of the historical record."

Proposed: "This study uses a categorical definition of hydrologic drought, occurring when the AMJJ streamflow volume falls below the 25th percentile (P_{25}) of the historical record."

L205 "where the value of"

Response: Thank you for pointing that out. We will fix this in the revised manuscript.

Original: "Using this set of PEV estimates, we construct a PEVmax curve by taking the maximum value from this set for each α, where value of α is equal to the critical probability threshold (τ)."

Proposed: "Using this set of PEV estimates, we construct a PEV_{max} curve by taking the maximum value from this set for each α, where the value of α is equal to the critical probability threshold (τ)."

L232: "REV" instead of PEV.

Response: Thank you for pointing that out. We will fix this in the revised manuscript.

Original: "Negative REV values (grey boxes in Fig. 2) indicate decisions that were worse than using the climatology (Laugesen et al., 2023; Richardson, 2000; Wilks, 2001)."

Proposed: "Negative PEV values (grey boxes in Fig. 2) indicate decisions that were worse than using the climatology (Laugesen et al., 2023; Richardson, 2000; Wilks, 2001)."

L260 "with one or fewer" : not sure about what this means, maybe this is correct.

Response: Yes, this wording is correct, and we wish to keep it as written.

L320 "statistical forecasts (…) operational forecasts"

Response: Thank you for pointing that out. We will fix this in the revised manuscript.

Original: "In addition, we also use Natural Resources Conservation Services statistical forecasts (now "NRCS") operational forecasts over the study watersheds to benchmark true forecasts."

Proposed: "In addition, we also use NRCS operational forecasts over the study watersheds to benchmark true forecasts."

L452-454: “during WY2001-2010 » appears twice in this sentence.

Response: Thank you for pointing that out. We will fix this in the revised manuscript.

Original: “We first compared the performance of the calibrated WRFH and fully trained LSTM models against observations for 76 basins during the testing period, WY2001-2010, using four fundamental metrics: daily NSE, normalized root mean square error (NRMSE) of total AMJJ volume, daily correlation, and the ratio of the standard deviation with the observations during WY2001-2010 (Fig. 7).”

Proposed: “We first compared the performance of the calibrated WRFH and fully trained LSTM models against observations for 76 basins during the testing period, WY2001-2010, using four key metrics: daily NSE, normalized root mean square error (NRMSE) of total AMJJ volume, daily correlation, and the ratio of the standard deviation (Fig. 7).”

L456: “for both models”

Response: Thank you for pointing that out. We will fix this in the revised manuscript.

Original: “The median correlation was greater than 0.7 for all models, with LSTM showing the highest correlation of 0.85, demonstrating a capability to capture temporal dynamics in daily streamflow prediction.”

Proposed: “The median correlation was greater than 0.7 for both models, with LSTM showing the highest correlation of 0.85, demonstrating a capability to capture temporal dynamics in daily streamflow prediction.”

L458 “These results suggest that the LSTM models, particularly LSTM”

Response: Thank you for pointing that out. We will fix this in the revised manuscript.

Original: “These results suggest that the LSTM models, particularly LSTM, perform much better in simulating streamflow than the WRFH models.”

Proposed: “These results suggest that the LSTM model performs much better in simulating streamflow than the WRFH model.”

Figure A3 is titled Figure A2 and referenced as Figure A3 in the text.

Response: Thank you for pointing that out. We will fix this in the revised manuscript.

L473 “only for the drought years only”

Response: Thank you for pointing that out. We will fix this in the revised manuscript.

Original: “We estimate skill and value only for the drought years only (i.e., years below the 25th percentile based on observed AMJJ volumes between WY2006-2022).”

Proposed: “We estimate skill and value only for the drought years (i.e., years below the 25th percentile based on observed AMJJ volumes between WY2006-2022).”

L493: “the higher number of false alarms reduces”

Response: Thank you for pointing that out. We will fix this in the revised manuscript.

Original: "In Fig. 4.8c, with a -50% error in standard deviation, we observe the highest skill (0.05) and value (0.62), as most events are correctly forecasted (H=0.73), though a few ensemble members cause false alarms (F=0.06). In Fig. 8d, with a +50% error in standard deviation, all events are still hit (H=0.63), but the higher false alarms (F=0.20) reduce the forecast value from 0.62 to 0.42."

Proposed: "In Fig. 8c, with a -50% change in variability, we observe the highest skill (0.05) and value (0.62), as most events are correctly forecasted (H=0.73), though a few ensemble members cause false alarms (F=0.06). In Fig. 8d, with a +50% change in variability, all events are still hit (H=0.63), but the higher number of false alarms (F=0.20) reduces the forecast value from 0.62 to 0.42."

L529 and L571 "the three true forecasts"

Response: Thank you for pointing that out. We will fix this in the revised manuscript.

Original: "Figure 10 illustrates the normalized mean quantile loss (NMQloss) of three true forecast systems over the heatmaps developed for synthetic forecasts based on Fig. 8a."

Proposed: "Figure 10 illustrates the normalized mean quantile loss (NMQloss) of the three true forecast systems over the heatmaps developed for synthetic forecasts based on Fig. 8a."

Original: "Both scatter plots show the relationship between NMQloss (forecast skill) and APEVmax (forecast value) for three true forecast systems (WRFH, LSTM, and NRCS), with each point corresponding to a different basin."

Proposed: "Both scatter plots show the relationship between NMQloss (forecast skill) and APEVmax (forecast value) for the three true forecast systems (WRFH, LSTM, and NRCS), with each point corresponding to a different basin."

Author's replies

Understanding the relationship between streamflow forecast skill and value across the western US

[Manuscript ID #2024-4046]

Our responses to the reviewer's comments appear in *blue italics,* and the proposed revisions (at this stage) in *red italics*.

Referee 2

Thank you for the interesting paper, which, in my view, reports an extensive well-documented hydrological seasonal forecasting research, around the observation and concern that many forecast verification publications do not report performance in terms of potential added value for decision making (e.g. through PEV, HR, FR). I support your call in the final sentence of your paper to '..adopt more sophisticated forecast evaluation approaches that prioritize forecast value..'

I do have the following **general comments and questions**:

I miss in the Introduction and the Discussion and Conclusion the clear recognition that the various performance scores and skill scores have been designed for, and serve, their own purpose. Continuous scores assessing accuracy (e.g. mean error, NSE), reliability (BS), overall performance (CRPS), etc., have been designed and are primarily used to intercompare forecasting systems and measure progress. This is, I believe, an understandable reason why in most scientific literature introducing a new or updated forecasting system, focus has been on such metrics. The potential economic value metric, and others based on contingency tables, and on multi-class decision problems, have been designed to asses and analyse forecast performance for operation and decision making, e.g. for specific applications informing the forecast and user community whether the performance is potentially good enough to use the forecasts and provide guidance on how to use them.

Response: Thank you for that insightful comment. We acknowledge that the different performance scores indeed serve distinct purposes. In our study, we intentionally focused on establishing the link between forecast skill (as measured by continuous metrics) and forecast value (often a multi-class decision problem), recognizing that while accuracy metrics are critical for assessing system performance, their implication in real-world decision-making is limited. Our goal was to demonstrate how skill – often used to assess a model's performance in terms of accuracy – can be linked to value-oriented metrics, like PEV, which are more directly related to decision-making processes. In this sense, the intention was not to undermine accuracy metrics but rather to highlight that forecast value goes beyond mere accuracy. In the Discussion section (1st paragraph), we will clarify this distinction, emphasizing that accuracy metrics remain fundamental for comparing forecasting systems but that metrics like PEV provide a more complementary, more practical perspective on forecast utility in real-world decision-making.

Original: "This study was motivated by recent literature showing that the relationship between forecast skill and value in hydrology is multifaceted and context dependent. While forecast skill generally reflects the accuracy of forecasts relative to the observations, the value represents the economic benefits derived from utilizing those forecasts in the decision-making process."

Proposed: "This study was motivated by recent literature showing that the relationship between forecast skill and value in hydrology is multifaceted and context dependent (Maurer and Lettenmaier, 2004;

Rheinheimer et al., 2016; Hamlet et al., 2002; Portele et al., 2021, Giuliani et al., 2020). While forecast skill generally reflects the accuracy of forecasts relative to observations, forecast value represents the economic benefits derived from utilizing those forecasts in decision-making. In this context, we emphasize that while traditional accuracy metrics are fundamental for evaluating forecasting systems, they have limited ability to capture the full utility of forecasts. By linking skill to value, we demonstrate how these metrics offer a more complementary perspective on forecast utility."

I would kindly request the authors to reflect on which findings were as to be expected, and which were the surprising findings and why. E.g., with hit rate as positive and false alarm rate as negative term in the definition of potential economic value, they indeed explain the variability in PEV. And with PEV assessed for low flow warnings, it is perhaps as expected that error in mean and standard deviation do not work through to PEV in a consistent way?

Response: Thank you for this insightful comment and for pointing out the need to better reflect our findings. Some aspects of our results aligned with expectations, while others were more surprising. We anticipated that the asymmetry in $APEV_{max}$ (compared to skill) would arise resulting from the interplay between hit and false alarm rate, which indeed is discussed and shown in Fig. 8. Additionally, forecast skill followed anticipated patterns for both synthetic and true forecasts, primarily as a function of error in mean and change in variability (Fig. 10).

However, we found three aspects particularly surprising. First, the skill-value relationship was remarkably consistent for synthetic forecasts despite only controlling the mean and variability across the observations. This suggested that regular error structures allowed for a more predictable translation of skill into value. Second, and in contrast, the skill-value relationship was completely inconsistent for true forecasts, particularly in the context of drought examined in this study (Fig. 12). This indicates that in real-world conditions, forecast value is influenced by additional complexities beyond forecast skill. Third, even though some forecasts demonstrated high skill (like NRCS or LSTM), the weaker skill-value relationship for true forecasts meant that good forecast skill did not always translate into high forecast value.

In response to the reviewer's suggestion, we will emphasize these expected and surprising aspects more explicitly in the Discussion section, particularly the contrasting behavior of the skill-value relationship between synthetic and true forecasts.

Original: "We begin by assessing the historical model performance of true forecasts against observations generated in this study. This involves comparing the performance of the WRFH and LSTM models across 76 basins using fundamental metrics to assess their ability to satisfactorily capture streamflow dynamics. The LSTM models consistently outperformed the WRFH models, likely due to the advanced capabilities of deep learning to better capture input-output dynamics (Fig. 7). We then analyzed the sensitivity of forecast skill and value to errors during drought years with respect to fundamental statistical measures – error in mean and standard deviation. For synthetic forecasts, a key observation was that the forecast skill was symmetric around mean errors, whereas the value was asymmetric due to the influence of categorical measures (Fig. 8). For the true forecast systems, we examined actual error in mean and standard deviation against observations, observing a consistent pattern of overprediction in mean and standard deviation lower than interannual variability from historical records (Fig. 9).

We overlapped true forecasts over synthetic forecasts to systematically analyze the impact of error in mean and standard deviation on the skill and value of true forecasts. The correspondence in forecast skill between the synthetic and true systems, particularly for LSTM and NRCS, was small, indicating forecast

skill in either forecast was primarily a function of error in mean and standard deviation (Fig. 10). However, the correspondence in forecast value between synthetic and true forecasts was large (Fig. 11). Unlike forecast skill, which was primarily a function of error in mean and standard deviation, the forecast value appears to be influenced by more complex interactions within the true forecast system. These differences caused synthetic forecasts, with their regular error structures, to exhibit a strong and consistent skill-value relationship, whereas true forecasts showed a weaker and more variable relationship (Fig. 12)."

Proposed: "We begin by assessing the historical model performance of true forecasts against observations generated in this study, comparing the WRFH and LSTM models across 76 basins using key performance metrics. As expected, the LSTM model consistently outperformed the WRFH model, likely due to the advanced capabilities of deep learning to better capture input-output dynamics (Fig. 7). We then analyzed the sensitivity of forecast skill and value to errors during drought years, specifically focusing on error in mean and change in variability. For synthetic forecasts, we anticipated that forecast skill would be symmetric around mean errors, while value would exhibit asymmetry due to the influence of categorical measures such as hit and false alarm rates (Fig. 8). The use of a normally distributed ensemble to develop synthetic forecasts is a simplification that allows us to model forecast uncertainty in a controlled manner. While real-world forecasts often exhibit more complex, irregular distributions and biases. For example, these may be overestimated in dry conditions and underestimated in wet conditions (Modi et al., 2021). A normal distribution was chosen to solely isolate the impact of mean and standard deviation. We recognize that this assumption does not fully capture the nuances of real-world forecast errors, such as skewness or non-normality in extreme conditions, which would require detailed treatment outside the scope of this analysis. For the true forecast systems, we examined actual error in mean and variability against observations, observing a consistent pattern of overprediction in mean and variability lower than interannual variability from historical records (Fig. 9), as also reported in Modi et al., 2021. Additionally, we expected forecast skill for both synthetic and true forecasts to primarily follow patterns driven by error in mean and change in variability, and indeed, the correspondence of forecast skill for both synthetic and true forecasts showed small differences, indicating that forecast skill was largely a function of error in mean and variability (Fig. 10). We acknowledge that estimating forecast skill and value for drought years necessitates a smaller sample size (here n~5), which is not ideal, affecting the statistical power of the analysis. This limitation arises due to the limited availability of operational forecasts and the need for sufficient ensemble members for ESP. Therefore, it would be important to assess whether a broader selection criterion or longer span of forecast availability would help ensure robust results.

However, we found three aspects particularly surprising. First, the skill-value relationship was remarkably consistent for synthetic forecasts despite only controlling for mean and variability across the observations. This suggested that regular error structures allowed for a more predictable translation of skill into value (Fig. 12). Second, in contrast, the skill-value relationship was completely inconsistent for true forecasts, particularly in the context of droughts. This was unexpected, as we had anticipated some level of variability, but the degree of inconsistency indicated that in real-world conditions, forecast value is influenced by additional complexities beyond forecast skill (Fig. 12). Third, even though some true forecast systems, such as NRCS and LSTM, demonstrated high skill, the weaker skill-value relationship for true forecasts meant that good forecast skill did not always translate into high forecast value (Figs. 11 and 12)."

The results are presented for three different forecasting systems, such that perhaps the following question can also be addressed in the paper: Would the differences in error, quantile loss, and PEV lead to different

conclusions on which forecasting system best to use for low flow forecasting? And then for more general discussion/conclusion: Do your results indicate, or not, if there is a risk in potential users referring to papers intended for measuring performance progress (publishing only overall performance metrics) when selecting a forecasting system to use?

Response: Thank you for unfolding those important questions. Regarding the first question, while our results show that differences in error, quantile loss, and PEV could influence the evaluation of different forecasting systems for low-flow forecasting, the scope of this paper is not to determine which system best suits low-flow forecasting. Instead, our goal is to promote a more comprehensive evaluation of forecasting systems beyond traditional skill metrics. PEV provides a useful framework for linking forecasts to decision-making, but it remains a simplified representation that omits many important considerations. Therefore, while PEV can offer insights into forecast value, it should not be used in isolation to choose the optimal forecasting system. We will clarify this point in the Discussion section.

For the second question, our findings do suggest a risk in stakeholders relying solely on overall performance metrics when selecting a forecasting system. As demonstrated in our results, even when forecasts exhibit high skill, their economic value can vary significantly due to complexities and system interactions. This ties back to the first point – assessments based purely on skill metrics may overlook how forecast performance translates into decision-making benefits. Our study reinforces the need for a multi-faceted evaluation approach that considers both forecast skill and value while also acknowledging the limitations of the used PEV framework. In response to the reviewer's suggestion, we will emphasize these nuances more explicitly in our Discussion and Conclusions sections.

Original (Discussion): "These findings emphasize the need for more sophisticated approaches to forecast evaluation, focusing on value across varying conditions rather than solely improving forecast skill metrics like mean and standard deviation."

Proposed (Discussion): "Our findings highlight the risk of stakeholders relying solely on traditional performance metrics when selecting a forecasting system. While high forecast skill may indicate good performance, the economic value can vary significantly due to system complexities and interactions. This underscores the need for more sophisticated assessment approaches that consider forecast value, particularly in decision-making contexts, rather than focusing solely on skill metrics. Our study advocates for a multi-faceted assessment framework that integrates both skill and value while also recognizing the limitations of the PEV framework."

Original (Conclusions): "This suggests that value is influenced by factors beyond forecast accuracy, such as the specific types of error structures and user-specific decision-making. The findings emphasize the importance of adopting more sophisticated forecast evaluation approaches that prioritize forecast value under varying conditions rather than focusing exclusively on skill metrics."

Proposed (Conclusions): "Our findings emphasize that forecast value is influenced by factors beyond forecast accuracy, such as the error structures and user-specific decision-making. This suggests that simply relying on performance metrics can overlook important variations in economic value. To address this, a more sophisticated evaluation approach is needed—one that prioritizes forecast value under varying conditions rather than focusing exclusively on accuracy metrics. A comprehensive evaluation framework that integrates both skill and value is essential for more informed, impactful decision-making."

Detailed comments

Methodology: The paper seems to be quite long. The introduction and explanation of the development of LSTM and WRFH models and forecasts is extensive, while comparing them or whether one performance metric relates differently to the next for each of the models is not the focus of the paper. Consider further shortening the model and forecast descriptions and moving more information to the Annex.

Response: We appreciate the reviewer's feedback and acknowledge that the paper is quite long. We will move Table 2 (LSTM hyperparameters used in this study) to Appendix A2 as the new Table A3. However, we believe that further shortening or moving the model and forecast descriptions could disrupt the flow and omit essential details needed to fully understand the intricacies of our designed true forecast systems. That said, we will carefully review these sections to identify any content that can be more concisely presented or moved to the Appendix while ensuring clarity and completeness.

Consider leaving out Figure 1 as the procedure is explained in the text as well and references to PEV papers have been given.

Response: We appreciate the reviewer's suggestion. However, based on feedback from various audiences we have presented our work to (the coauthor list, seminar goers, as well as the other reviewer), we have found that including a figure greatly enhances the understanding of the multiple steps and assumptions involved in estimating PEV. We believe that visual representation of the PEV workflow really does help to clarify key steps and improve accessibility for readers. Therefore, we are electing to keep Figure 1 intact.

Consider leaving out Figure 5, as well as reducing the level of detail in explaining the steps of generating forecasts as this is I believe well-known by most people experienced in or interested in hydrometeorological forecasting.

Response: We appreciate the reviewer's suggestion. However, we believe this article will cater not only to those in hydrometeorological forecasting but also to readers from economics and decision-making communities. Including Figure 5 helps provide a clear, visual representation of how different components interact, making it easier for more visually oriented readers to follow the forecast generation process. Additionally, we have kept the figure as simple as possible to ensure clarity, as these details can easily get buried in the text.

Results: Heading 3.3.3 Consider just 'Hit and False Alarm Rate and forecast value', because HR and FR are per definition better estimators of forecast value (because expressed as PEV in this paper).

Response: Thank you for the suggestion. We will fix the heading in the revised version.

Original: "3.3.3 Hit and False Alarm Rate are better estimators of forecast value"

Proposed: "3.3.3 Hit and False Alarm Rate and Forecast Value"

Consider merging section 3.3.4 (value (defined as PEV) is per definition largely explainable by hit and false alarm rates) with 3.3.3

Response: Thank you for the suggestion. We will merge sections 3.3.3 and 3.3.4 under the new section 3.3.3 (Hit and False Alarm Rate and Forecast Value) in the revised version. We have proposed this new section in the previous comment.

Discussion:

p32 l653 - p33 l688 reads mostly as a summary of the paper, which fits better in a shortened version in Conclusions. Consider leaving out here.

Response: We appreciate the reviewer's feedback, and we also understand the concern about the paper's length. However, we believe that a brief summary at the beginning of the Discussion section helps orient the reader to the transition from the results to their broader implications. This overview reinforces key findings and frames the discussion. We will add a sentence at the start of the section to outline the Discussion section.

Proposed: "We begin with a brief summary of our results, followed by a transition into their broader implications."

Conclusions:

p35 l730 - 735. These sentences discuss LSTM outperforming the conceptual model, while such analysis is I believe not the main objective of this paper. Consider leaving out or reducing here.

Response: Thank you for pointing that out. While the comparison of historical performance is not the primary objective of the paper, it provides useful insights into how these models function and their capability in simulating streamflow. We will, however, make the paragraph more concise to better align with the paper's main objectives.

Original: "The comparison between the WRFH and LSTM models showed that the LSTM model significantly outperformed the WRFH model in simulating streamflow. Training had a much larger impact on the LSTM models, improving median daily NSE from 0.58 to 0.77, while the WRFH models saw minimal improvements across most metrics except the variability (ratio of standard deviation) post-calibration. The LSTM models also exhibited more stable structures, with lower NRMSE and better correlation, while the WRFH models had larger and more irregular error structures despite some improvement in variability after calibration."

Proposed: "The WRFH and LSTM models showed distinct responses to training and calibration in simulating streamflow. The LSTM model was more sensitive to training, with more stable structures, lower NRMSE, and better correlation. In contrast, the WRFH model showed minimal improvements post-calibration, with larger and more irregular error structures despite some improvement in variability."

l736: reformulate because now reads as if defining skill as error, while forecast skill is defined as improvement over a reference forecasting system (in this paper always climatology).

Response: Thank you for pointing that out. In this study, we define forecast skill as the model's ability to accurately predict observed values rather than as an improvement over a reference forecasting system. We recognize that this definition of "skill" differs from the standard approach of comparing to a reference forecast, but we elect to use "forecast skill" here for consistency.

l743-744: "This disconnect is further compounded.." I do not understand what is intended here. Please kindly clarify.

Response: Thank you for pointing that out. The phrase "disconnect" was intended to refer to the gap between forecast skill and value, which may not align, especially under real-world conditions. To clarify, we will revise the sentence to refer to this as a "gap" instead of "disconnect."

Original: "This suggests that overall model performance – how well a model handles variability and uncertainty – can significantly influence the disconnect between forecast skill and value. This disconnect is further compounded, not to mention the complexities introduced by operational structures."

Proposed: "This suggests that overall model performance – how well a model handles variability and uncertainty – can significantly influence the gap between forecast skill and value. This gap is further complicated by the complexities introduced by operational structures."

Editorial comments

p9 l232: REV should probably be PEV

Response: Thank you for pointing that out. We will fix this in the revised manuscript.

Original: "Negative REV values (grey boxes in Fig. 2) indicate decisions that were worse than using the climatology (Laugesen et al., 2023; Richardson, 2000; Wilks, 2001)."

Proposed: "Negative PEV values (grey boxes in Fig. 2) indicate decisions that were worse than using the climatology (Laugesen et al., 2023; Richardson, 2000; Wilks, 2001)."

p22: check caption. Default and calibrated, and initial and final models are mentioned in the caption, but not shown in the legend or figure.

Response: Thank you for pointing that out. We will fix this in the revised manuscript.

Original: "Figure 7: Historical model performance of true forecast systems. (a) Daily NSE, (b) NRMSE of the total April-July streamflow volumes, (c) daily correlation, and (d) Ratio of the standard deviation against observations for WRFH (default and calibrated) and LSTM (initial and final) models. Comparison shown for the 76 basins during the testing period, WY2001-2010."

Proposed: "Figure 7: Historical model performance of true forecast systems. (a) daily NSE, (b) NRMSE of the total April-July streamflow volumes, (c) daily correlation, and (d) the ratio of the standard deviation against observations for calibrated WRFH and fully trained LSTM models. The shaded areas represent the distributions of model performance metrics over the 76 basins, while the vertical lines indicate the performance of individual basins during the testing period, WY2001-2010."

p30 l605: consider "..a low false alarm rate limits unnecessary.."

Response: Thank you for pointing that out. We will fix this in the revised manuscript.

Original: "In decision-making, a high hit rate ensures timely actions for critical events like drought, while a low false alarm rate prevents unnecessary responses and maintains trust in the forecast system."

Proposed: "In decision-making, a high hit rate ensures timely actions for critical events like drought, while a low false alarm rate limits unnecessary responses and maintains trust in the forecast system."

p30 l611-612: "..LSTM forecasts.." "..synthetic forecasts.."

Response: Thank you for pointing that out. We will fix this in the revised manuscript.

Original: "For this analysis, we compare the true LSTM (shown in green) and the corresponding synthetic forecast (shown in black) based on the overlap shown in Figs. 10 and 11."

Proposed: "For this analysis, we compare the LSTM forecasts (shown in green) and the corresponding synthetic forecasts (shown in black) based on the overlap shown in Figs. 10 and 11."

p30 l616: "(Fig. 14a - right)"

Response: Thank you for pointing that out. We will fix this in the revised manuscript.

Original: "In terms of the False Alarm Rate, the synthetic forecast initially shows a lower rate compared to true forecast (LSTM), indicating fewer false alarms at higher thresholds (Fig. 14a - left)."

Proposed: "In terms of the False Alarm Rate, the synthetic forecast initially shows a lower rate compared to true forecast (LSTM), indicating fewer false alarms at higher thresholds (Fig. 14a - right)."

p31: Caption Figure 14, l636: consider "..to each forecast system."

Response: Thank you for pointing that out. We will fix this in the revised manuscript.

Original: "The values indicate the $APEV_{max}$ corresponding to each forecast."

Proposed: "The values indicate the $APEV_{max}$ corresponding to each forecast system."

p35 l737: "..- exhibit complex.."

Response: Thank you for pointing that out. We will fix this in the revised manuscript.

Original: "Our results showed that forecast skill — indicating how accurately forecasts match observations — and forecast value — representing the economic benefits derived from those forecasts in decision-making — exhibits a complex relationship for true forecasts due to their irregular error structures."

Proposed: "Our results showed that forecast skill — indicating how accurately forecasts match observations — and forecast value — representing the economic benefits derived from those forecasts in decision-making — exhibit complex relationships for true forecasts due to their irregular error structures."

p35 l738-739: kindly clarify, "skill was more sensitive to error and SD", compared to what?

Response: Thank you for pointing that out. We will fix this in the revised manuscript.

Original: "Our comparisons between synthetic and true forecasts revealed that forecast skill across the basins was more sensitive to error in mean and standard deviation."

Proposed: "Our comparisons between synthetic and true forecasts revealed that forecast skill across the basins was more sensitive to error in mean and change in variability than the forecast value."

I prefer Annex to be after Reference list, but this is probably governed by HESS.

Response: Thank you for pointing that out. We will contact the editor to confirm the manuscript's structure and restructure it accordingly.