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 quality 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 quality 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 quality 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 errors 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 errors in mean and changes 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 – errors 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 errors 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 errors 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 errors 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 errors 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 quality and value to errors during drought years, specifically focusing on errors in mean and changes 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). For the true forecast systems, we examined actual errors 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 quality for both synthetic and true forecasts to primarily follow patterns driven by errors in mean and changes in variability, and indeed, the correspondence of forecast quality for both synthetic and true forecasts showed small differences, indicating that forecast quality was largely a function of errors in mean and variability (Fig. 10).

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 quality 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 quality metrics. Our study advocates for a multi-faceted assessment framework that integrates both quality 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 quality 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 the Appendix. 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. To better align with standard definitions, we will revise the manuscript text as well as figures and tables to use the term "forecast quality" throughout. As per your earlier comment, below is one of the instances where we have made these changes in the Discussion section.

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 quality 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 quality 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 quality to value, we demonstrate how these metrics offer a more complementary perspective on forecast utility."

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. Comparison shown for the 76 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 errors in mean and standard deviation."

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