https://doi.org/10.5194/egusphere-2024-4046
Preprint. Discussion started: 14 January 2025

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

Parthkumar Modi[1,2], Jared Carbone[3], Keith Jennings[4], Hannah Kamen[3], Joseph Kasprzyk[1], Bill Szafranski[5], Cam Wobus[6] and Ben Livneh[1,2,7]

[1]Department of Civil, Environmental, and Architectural Engineering, University of Colorado Boulder, Boulder, 80309, USA
[2]Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado Boulder, Boulder, 80309, USA
[3]Economics and Business, Colorado School of Mines, Golden, 80401, USA
[4]Water Resources Institute, University of Vermont
[5]Lynker
[6]CK Blueshift LLC, University of Colorado Boulder, Boulder, 80309, USA
[7]Western Water Assessment, University of Colorado Boulder, Boulder, 80309, USA

Correspondence to: Parthkumar Modi (parthkumar.modi@colorado.edu)

Abstract.

Accurate seasonal streamflow forecasts are essential for effective decision-making in water management. In a decision-making context, it is important to understand the relationship between forecast skill— the accuracy of forecasts against observations – and forecast value, which is the forecast's economic impact assessed by weighing potential mitigation costs against potential future losses. This study explores how errors in these probabilistic forecasts can reduce their economic "value", especially during droughts when decision-making is most critical. This value varies by region and is contextually dependent, which often limits retrospective insights to specific operational water management systems. Additionally, the value is shaped by the intrinsic qualities of the forecasts themselves. To assess this gap, this study examines how forecast skill transforms into value for true forecasts (using real-world models) in unmanaged snow-dominated basins that supply flows to downstream managed systems. We measure forecast skill using quantile loss and quantify forecast value through the Potential Economic Value framework. The framework is well-suited for categorical decisions and uses a cost-loss model, where the economic implications of both correct and incorrect decisions are considered for a set of hypothetical decision-makers. True forecasts are included, made with commonly used models within an Ensemble Streamflow Prediction (ESP) framework using a process-based hydrologic modeling system, WRF-Hydro; a deep learning model, Long Short-term Memory Networks; as well as operational forecasts from the NRCS. To better interpret the relationship between skill and value, we compare true forecasts with synthetic forecasts that are created by imposing regular error structures on observed streamflow volumes. We evaluate the sensitivity of skill and value from both synthetic and true forecasts to fundamental statistical measures - errors in mean and standard deviation. Our findings indicate that errors in mean and standard deviation consistently explain variations in forecast skill for true forecasts. However, these errors do not fully explain the variations in forecast value across the basins, primarily due to irregular error structures, which impact categorical measures such as hit and false alarm rates, causing high forecast skill to not necessarily result in high forecast value. We identify two key insights: first, hit and false alarm rates effectively capture variability in forecast value rather than errors in mean and standard deviation; second, the relationship between forecast skill and value shifts monotonically with drought severity. These findings emphasize the need for a deeper understanding of how forecast performance metrics relate to both skill and value, highlighting the complexities in assessing the effectiveness of forecasting systems.

1 Introduction

Probabilistic seasonal streamflow forecasts are essential for informed decision-making in water resource management, including flood risk mitigation, agriculture, energy production, and in-stream ecosystem services. These forecasts enable stakeholders to plan for optimal water allocation, optimize reservoir operations, and prepare for extreme hydrological events like droughts or floods (Wood et al., 2015). 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. Forecast value is influenced by factors such as the cost of taking preventive action (e.g., investing in crop insurance), the potential losses from incorrect decisions (e.g., economic losses due to over or under-allocation of water resources), and the context of decision-making (e.g., hiring labor for an agricultural entity). This relationship is complex and varies by region, often restricting the retrospective insights gained to specific operational systems. As a result, there is limited understanding of the link between skill and value - especially concerning the quality of forecasting systems. The complexity of forecast value can be framed within simple economic models like the cost-loss ratio framework. In this model, decision-makers face a potential loss if an adverse event (e.g., a drought) occurs but can take preventive action at a cost to mitigate this loss. Understanding how forecast skill translates into forecast value is critical, as it highlights the importance of not only improving the accuracy of forecasts but also understanding how skill impacts decision-making outcomes. This study addresses the research question: How do errors in different forecasting systems affect forecast skill and decision-making value in unmanaged basins, and how can these insights guide improvements in forecast systems?

1.1 Forecast skill of probabilistic seasonal streamflow forecasts has evolved

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. Probabilistic seasonal streamflow forecasts have become a crucial tool in water resources management (Crochemore et al., 2016; Ficchì et al., 2016; Kaune et al., 2020; Turner et al., 2017; Watts et al., 2012), as they provide a range of possible outcomes rather than a single deterministic prediction (Demargne et al., 2014). This probabilistic approach helps decision-makers quantify forecast uncertainty, enabling more informed and flexible water management strategies (Pagano et al., 2014). For example, the NRCS forecasts have been widely used for water management and agricultural planning (Fleming et al., 2021).

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). Key developments include better representation of watershed processes in hydrologic models and the use of data assimilation techniques (Wood & Lettenmaier, 2006). Furthermore, the application of machine learning algorithms, such as the popular Long Short-term Memory (LSTM), has become instrumental in detecting complex patterns in data, leading to even greater refinement in forecast accuracy when combined with improved meteorological inputs (Modi et al., 2024; Mosavi et al., 2018). Among the various methods, the National Water Model (NWM) stands out as a state-of-the-art process-based forecasting framework, which provides high-resolution operation streamflow forecasts across the CONUS by incorporating improved hydrological representation and real-time meteorological data to enhance the forecast skill (Cosgrove et al., 2024). However, the model has limitations in certain regions, such as parts of the Intermountain West, where forecast skill remains a challenge. This study will test some of these methods, evaluating their effectiveness and applicability across various scenarios to provide comprehensive insights into their skill and value.

1.2 Seasonal streamflow forecasts provide economic benefit

Seasonal streamflow forecasts provide crucial information about water availability, enabling stakeholders such as water managers, energy producers, and farmers to make informed decisions about water allocation, crop planning, and reservoir operations. These forecasts play a substantial role in regions prone to hydrological variability, where early forecasts allow for better preparedness and can help mitigate the risk of extreme events like droughts or floods. In this context, the study focuses on streamflow volume during the April-July period (AMJJ), a predominant time window for water supply decisions across the snow-dominated basins in the western US (Livneh and Badger, 2020; Modi et al., 2021). Studies have shown that using streamflow forecasts can lead to tangible economic gains, though the percentage increase vary widely depending on the context. While some studies report modest gains of 1-2% (Maurer and Lettenmaier, 2004; Rheinheimer et al., 2016), others demonstrate much higher benefits. For example, Hamlet et al. (2002) showed a significant increase in hydropower revenue of 40% or $153 million per year in the Columbia River basin. Moreover, Portele et al. (2021) showed that seasonal streamflow forecasts can yield up to 70% of the potential economic gains in semi-arid regions from taking early and optimal actions during droughts. Across the US, improved water supply forecasts have been associated with annual economic benefits ranging from $1 billion to $3 billion, particularly in sectors like agriculture, energy, and flood prevention (EASPE, 2002). Given that economic benefits from these vary by context, it remains uncertain whether these benefits are primarily driven by the intrinsic quality of the forecast itself or by specific operational factors (e.g., reservoir storage buffer).

1.3 Forecast value

Traditionally, streamflow forecast skill has been assessed based on its accuracy and reliability in predicting water flow volumes. However, an additional layer of assessment can be introduced by incorporating economic evaluations. This contrast highlights not only the technical skill of forecasts but also their practical value in optimizing economic outcomes for decision-making. Hydrologists continue to show strong interest in assessing the value of forecasts to support decision-making using Potential Economic Value (Abaza et al., 2013; Portele et al., 2021; Thiboult et al., 2017; Verkade et al., 2017). Potential Economic Value quantifies the economic benefit of using a particular forecast system compared to solely relying on climatology or no forecast. It is a standard metric for assessing the economic utility of forecasts, particularly in categorical decision-making scenarios, typically modeled through a cost-loss framework (Richardson, 2000; Wilks, 2001). In a cost-loss framework, decision-makers face a choice between taking preventive action at a cost (C) based on the forecast or bearing the potential loss (L) if an adverse event, such as a drought, occurs. A major assumption is that the cost (C) is smaller than the loss (L). PEV is a non-dimensionalized measure that facilitates comparison across different decision-making contexts, making it a practical tool for evaluating forecast effectiveness (Wilks, 2001). Its straightforward application, ease of comparison across different forecasting systems, and ability to estimate the upper bound of forecast value make it a useful tool in evaluating seasonal streamflow forecasts. It remains particularly valuable in contexts where binary decisions are prevalent, and the economic impact of forecasts is a key concern. We apply this simple framework—the cost-loss model—to examine how forecast skill translates into economic value as a function of inherent quality of the different forecasting systems. This will help evaluate the economic implications of both correct and incorrect decisions for a set of hypothetical decision-makers in unmanaged basins.

1.4 Study Summary

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. Motivated by the nuanced and often inconsistent link between forecast skill and value, as well as a limited understanding of how this relationship behaves across different forecast systems, this study offers an assessment of how skill transforms into value, using PEV as a tool in unmanaged basins. To better interpret the relationship between skill and value, we compare true forecasts with synthetic forecasts that are generated by imposing regular error patterns on observed streamflow volumes. This approach helps to address the impact of irregular error structures present in true forecasts, which are often non-normally distributed and exhibit varying variances. We begin by assessing the historical model performance of true forecasts against observations generated in this study. This involves comparing the calibrated WRF-Hydro and fully trained LSTM model to assess their effectiveness in simulating streamflow volumes. We then evaluate how both synthetic and true forecasts respond to fundamental statistical measures such as errors in mean and standard deviation.

Lastly, we investigate the relationship between skill and value across different drought severities, considering the interplay of error structures from both synthetic and true forecasts and the factors influencing the PEV framework.

2 Methods

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. Section 2.2 describes the study domain and basin screening procedure. Section 2.3 outlines the "synthetic" forecast approach that imposes errors on April-July (now "AMJJ") streamflow volumes. Section 2.4 outlines the generation of true forecasts that use a process-based model, WRF-Hydro (now "WRFH"); and a deep learning model, LSTM; and describes the operational NRCS forecasts. This section also describes the model inputs, architecture, training/calibration, and their implementation in an ESP framework. Section 2.5 provides an overview of fundamental performance metrics.

2.1 Drought event, forecast skill and value

2.1.1 Defining a drought event using hydrological threshold categories

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.

2.1.2 Forecast Skill Metric: Normalized Mean Quantile Loss

Quantile Loss, also called pinball loss, evaluates the performance of a probabilistic forecast by measuring the difference between predicted quantiles (percentiles) and observed values (Eq. 1). In other words, it rewards situations in which the observed value is within quantiles of the ensemble forecast members. It is adopted widely operationally and recently used in the Bureau of Reclamation's water supply forecast challenge (Water Supply Forecast Rodeo: Forecast Stage, 2024). It provides an asymmetric error metric, i.e., it adjusts penalties based on whether the forecast overestimates or underestimates the observed values.

$$Qloss_z = \frac{2}{n} * \begin{cases} z * (y_{obs} - \hat{y}_z) & if\ y_{obs} \geq \hat{y}_z \\ (1-z) * (y_{obs} - \hat{y}_z) & if\ y_{obs} < \hat{y}_z \end{cases} \tag{1}$$

Where y_{obs} is the observed AMJJ streamflow volume, $\hat{y}$ is the predicted AMJJ streamflow volume, and z is the quantile. We use a scaled version of quantile loss, multiplied by a factor of 2, so that the loss at the 0.5 quantile (median) aligns with the mean absolute error (MAE), ensuring consistency in error interpretation across quantiles (Water Supply Forecast Rodeo: Forecast Stage, 2024). To represent forecast skill in this study, we calculate normalized mean quantile loss (NMQloss), an average of quantile loss calculated for each quantile $z \in \{0.1, 0.5, 0.9\}$ normalized by the mean of the observations (Eq. 2). These quantiles are based on the multiple ensemble members in the probabilistic forecasts. This approach allows us to assess error across different quantiles, comprehensively evaluating forecast skill. A lower mean quantile loss, closer to zero, indicates better forecast skill.

$$NMQloss = \frac{Qloss_{0.1} + Qloss_{0.5} + Qloss_{0.9}}{3 * \overline{y_{obs}}} \tag{2}$$

2.1.3 Forecast Value Metric: Area under PEV$_{max}$ curve

The PEV metric is based on the cost-loss ratio (α=C/L), where C represents the cost of taking preventive action (e.g., buying crop insurance) and L is the potential loss incurred if no action is taken and an adverse event occurs. The ratio helps decision-makers assess whether the benefit of preventing a loss outweighs the cost of taking preventive action. For instance, when α is low, the cost of action is small relative to the potential loss, making it more likely that preventive action will be taken. Conversely, a high α suggests that the cost of action outweighs the potential benefit, making action less justifiable. In practical terms, α reflects an aspect of the decision-maker's risk tolerance and serves as a threshold for action.

We use probabilistic forecasts of AMJJ volume as an input to PEV, which are based on ensemble predictions from multiple forecasting systems. These forecasts, discussed in detail in Sect. 2.3 and 2.4, provide a range of possible outcomes for the AMJJ volume, helping to capture uncertainty and variability. Figure 1 shows the PEV workflow where we first calculate the forecast probability of these forecasts for a future event, i.e., in our case, a P_{25} drought event when the AMJJ streamflow volume falls below the 25th percentile of the historical record (Step 1). For demonstration purposes, this calculation is shown by assuming five ensemble members representing AMJJ volume, and the future event is assumed to have volumes less than 2.5. These forecast probabilities are transformed into categorical forecasts by applying a critical probability threshold (τ). This threshold represents another aspect of the user's risk tolerance, i.e., the minimum probability at which a future event is considered likely enough to warrant action for a user. It should be noted that both α and τ represent different aspects of a user's risk tolerance, quantifying their willingness to act under uncertainty. As shown in step 2 of Fig. 1, a more conservative threshold of 0.5 would trigger an action in 2007 (only one of the years shown), while a looser threshold of 0.7 would not trigger action in 2007. In contrast, both thresholds would trigger no action in 2006, despite some of the ensemble members predicting flows below 2.5 for both years. This categorical forecast is used to create a 2x2 contingency table (Step 3; Fig. 1), which calculates the hit rate (H – the proportion of correctly predicted events), false alarm rate (F – the proportion of non-events incorrectly classified as events), miss rate (M – the proportion of events incorrectly classified as non-events), and correct rejection rate (Q – the proportion of correctly predicted non-events) based on the years available retrospectively in the forecast system we are assessing. Finally, the PEV metric is calculated by comparing the relative difference in the total long-run net expenses (i.e., for taking preventive action over the set of retrospective years in the forecast system) made using an actual forecast ($E_{forecast}$ – uses real-world data and models to generate forecasts – Eq. I), climatology ($E_{climate}$ – historical average of volumes in the record – Eq. II) and a perfect forecast ($E_{perfect}$ – complete knowledge of future volumes – Eq. III) over a prescribed range of cost-to-loss ratios (0<α<1) using equation IV (Step 4; Fig. 1).

$$E_{forecast} = F(1-s)C - Hs(L-C) + sL \qquad (I; Fig. 1)$$

$$E_{climate} = \min{(C, sL)} \qquad (II; Fig. 1)$$

$$E_{perfect} = sC \qquad (III; Fig. 1)$$

$$PEV = \frac{E_{climate} - E_{forecast}}{E_{climate} - E_{perfect}} \qquad (IV; Fig. 1)$$

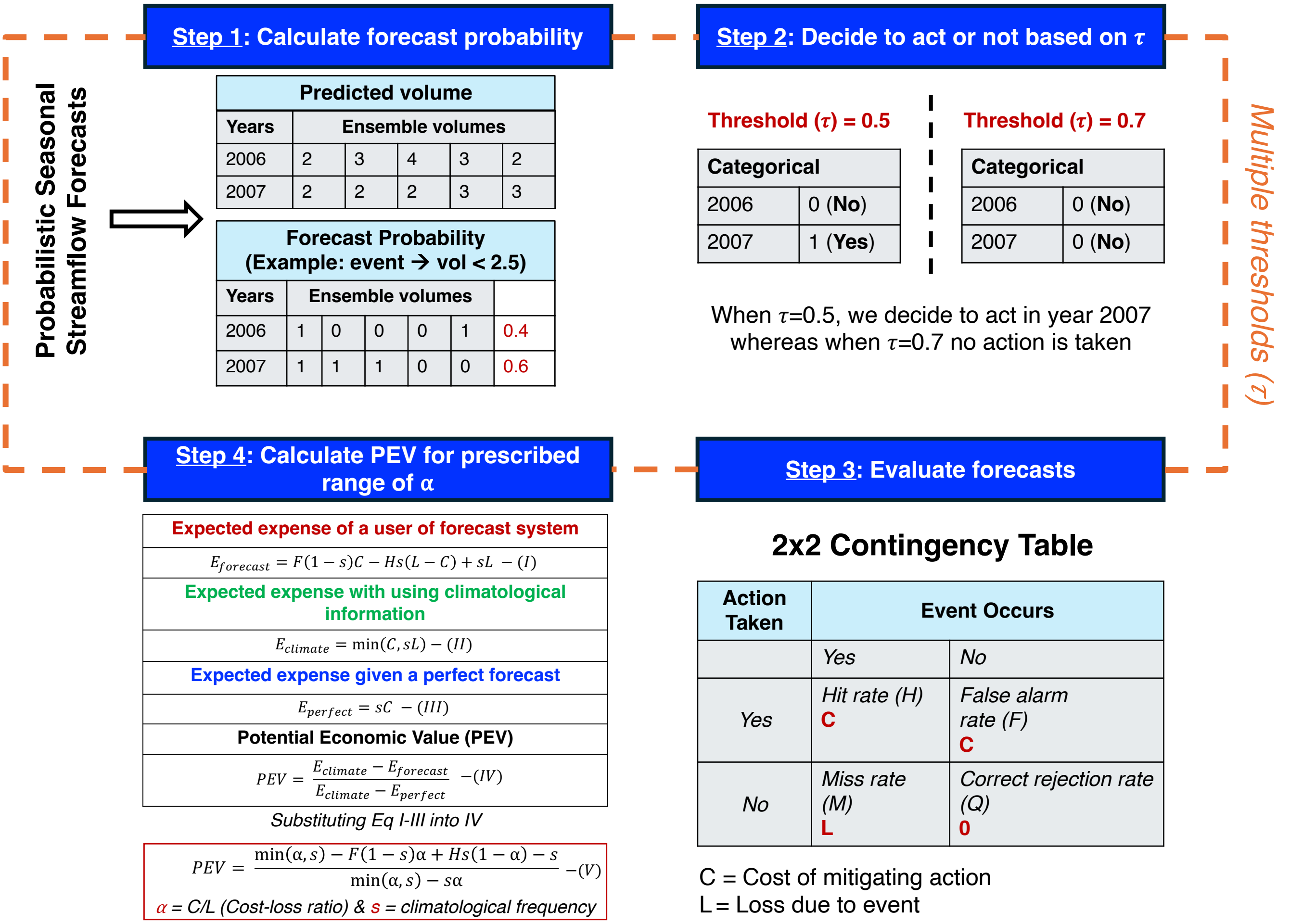
Where -∞<PEV<1 and each expense term is the summation of the contingency table elements, each weighted by the rate of occurrence. Equation V is used to calculate PEV based on Jolliffe and Stephenson (2003).

$$PEV = \frac{min(\alpha, s) - F(1-s)\alpha + Hs(1-\alpha) - s}{min(\alpha, s) - s\alpha} \qquad (V; Fig. 1)$$

Where α=C/L is the cost-loss ratio, s is the climatological frequency, i.e., the observed base rate of an event, and H and F are the hit and false alarm rates. A PEV of 1 indicates that the forecast system is perfect, providing maximum economic value, whereas a PEV of <0 indicates that the forecast offers no advantage over climatology (Murphy, 1993).

Steps 1, 2, 3, and 4 are repeated for multiple critical probability thresholds (τ) over the prescribed range of 0<τ<1 to generate a set of possible PEV values for each cost-to-loss ratio α (0<α<1). Multiple thresholds are adopted to account for varying risk tolerances among users and provide a more realistic evaluation of value. Using this set of PEV estimates, we construct a PEV_{max} curve by taking the maximum value from this set for each α, where value of α is equal to the critical probability threshold (τ). This approach assumes the user will adjust on their own, based on their specific α value (Laugesen et al., 2023; Richardson, 2000). The equations in the calculation workflow are adapted from Richardson (2000) and Jolliffe and Stephenson (2003).

https://doi.org/10.5194/egusphere-2024-4046
Preprint. Discussion started: 14 January 2025



Step 1: Calculate forecast probability

Probabilistic Seasonal Streamflow Forecasts

Predicted volume					
Years	Ensemble volumes				
2006	2	3	4	3	2
2007	2	2	2	3	3

Forecast Probability (Example: event → vol < 2.5)						
Years	Ensemble volumes					
2006	1	0	0	0	1	0.4
2007	1	1	1	0	0	0.6

Step 2: Decide to act or not based on τ

Threshold (τ) = 0.5

Categorical	
2006	0 (**No**)
2007	1 (**Yes**)

Threshold (τ) = 0.7

Categorical	
2006	0 (**No**)
2007	0 (**No**)

When τ=0.5, we decide to act in year 2007 whereas when τ=0.7 no action is taken

Multiple thresholds (τ)

Step 3: Evaluate forecasts

2x2 Contingency Table

Action Taken	Event Occurs	
	Yes	*No*
Yes	*Hit rate (H)* C	*False alarm rate (F)* C
No	*Miss rate (M)* L	*Correct rejection rate (Q)* 0

C = Cost of mitigating action
L = Loss due to event

Step 4: Calculate PEV for prescribed range of α

Expected expense of a user of forecast system

$$E_{forecast} = F(1-s)C - Hs(L-C) + sL \quad -(I)$$

Expected expense with using climatological information

$$E_{climate} = \min(C, sL) - (II)$$

Expected expense given a perfect forecast

$$E_{perfect} = sC \quad -(III)$$

Potential Economic Value (PEV)

$$PEV = \frac{E_{climate} - E_{forecast}}{E_{climate} - E_{perfect}} \quad -(IV)$$

Substituting Eq I-III into IV

$$PEV = \frac{\min(\alpha, s) - F(1-s)\alpha + Hs(1-\alpha) - s}{\min(\alpha, s) - s\alpha} \quad -(V)$$

α = C/L (Cost-loss ratio) & s = climatological frequency

Figure 1: Flowchart showing the workflow to quantify the PEV using the probabilistic forecasts. For the calculation of PEV, forecast probabilities are calculated from probabilistic 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).

Fig. 2 illustrates an economic value diagram that depicts a PEV$_{max}$ curve. This diagram visually represents the cost-loss ratio (α), on the X-axis, whereas PEV is on the Y-axis. At low values of α where the cost of preventive action is small relative to the potential loss, forecast systems tend to show higher economic value, as decision-makers can take advantage of accurate predictions to reduce potential losses with minimal expenditure. However, as α increases and the cost of preventive action becomes comparable to or exceeds the potential loss, the economic value of the forecast may decrease. In such cases, acting on the forecast becomes less advantageous because the cost of the preventive measure outweighs the potential benefit. The optimal economic value occurs when the α is balanced in a way that maximizes the benefit of acting on the forecast while minimizing unnecessary costs. This usually happens when the α is equal to the observed probability of the event (climatological frequency – s; Jolliffe and Stephenson (2003)). Unlike alpha and tau, which represent different aspects of the user's risk tolerance, s is a quantitative measure of the long-term probability of an event based on historical data. A value diagram, as shown in Fig. 2, will help decision-makers visualize and select appropriate actions based on their specific α (X-axis) and the performance of the forecast system compared to using climatology as PEV (Y-axis). In Fig. 2 on X-axis, α=0 indicates the cost of mitigation (C) is zero i.e., always beneficial, whereas α=1 indicates the cost of mitigation (C) equals the potential loss (e.g., a farmer paying $10,000 as insurance money to prevent a loss of $10,000 due to a future event). PEV=1 means forecast-based decisions perform as well as those using perfect information, while PEV=0 indicates the forecast offers no advantage over the baseline. A value of PEV=0.7 at a given α suggests a 70% improvement in decision-making compared to using the climatology. 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).

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) A larger APEV_{max} curve indicates that the forecast system delivers higher economic value over a broad range of decision-making scenarios, regardless of α. This value ranges from 0, showing the lowest overall economic value, to 0.9, being the highest overall economic value in this study.

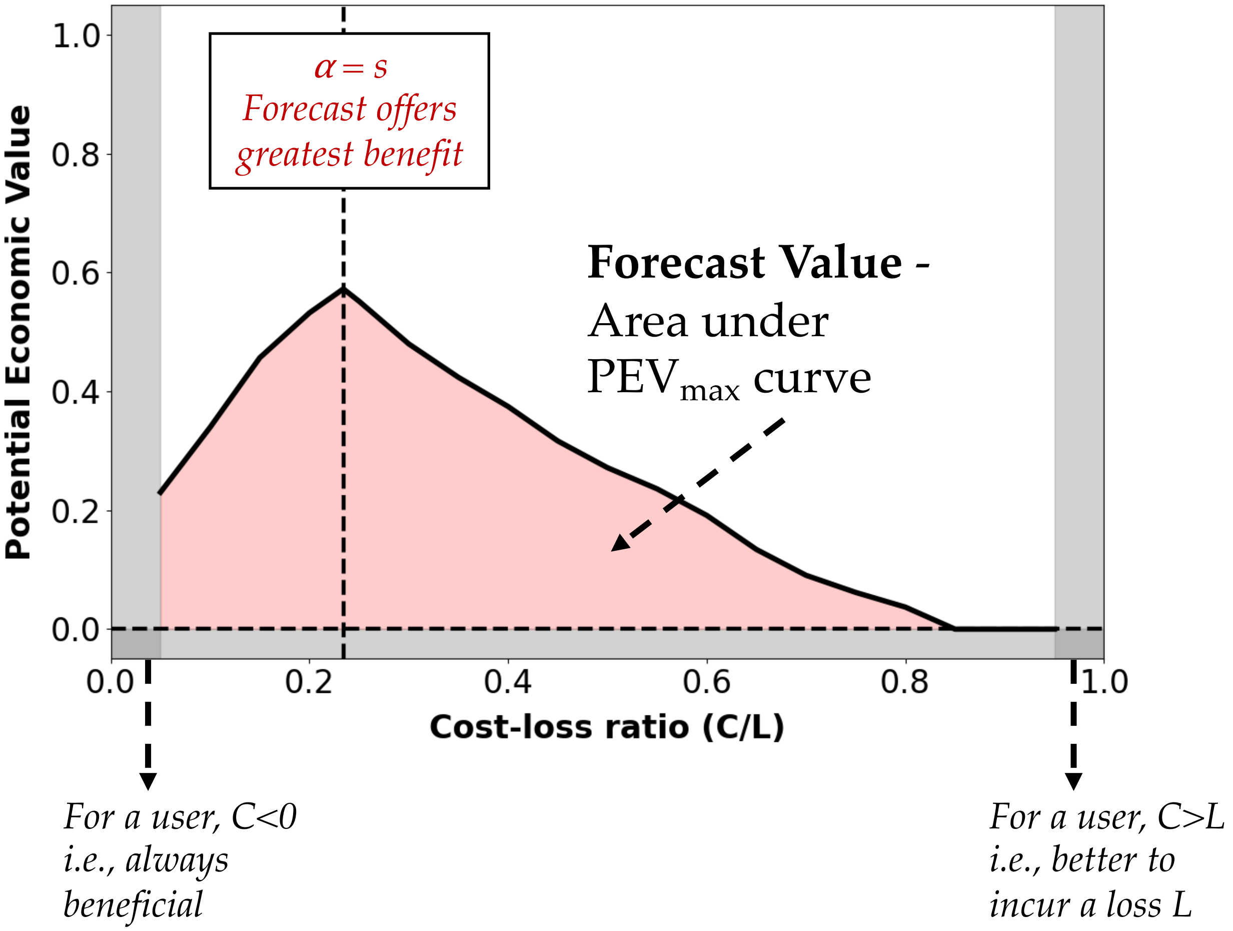


Figure 2: Economic value diagram showing cost-loss ratio (α) on the X-axis and the potential economic value on the Y-axis. The red shading shows the area under PEV$_{max}$ (APEV$_{max}$). It highlights the positive PEV values across α, indicating that the forecast is preferred over climatology, whereas the grey regions highlight negative PEV values, indicating that climatology should be preferred. The left vertical grey boxes indicate that the user is always beneficial when the preventive cost (C) is less than zero. In contrast, on the right, when the preventive cost (C) exceeds the potential loss (L), the user will always incur the loss L.

2.2 Study domain and basin screening procedure

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. Assessing forecast value in such basins is crucial since they often supply flows to downstream managed systems. We selected a diverse sample of drainage basins across the western US, representing a broad spectrum of hydroclimatic conditions. These basins were identified using geospatial attributes from three key sources: the USGS Geospatial Attributes of Gages for Evaluating Streamflow (GAGES-II) dataset, the Hydro-Climatic Data Network (HCDN; Slack and Landwehr, (1992)), and the Catchment

Attributes and Meteorology for Large-sample Studies (CAMELS; Addor et al., 2017; Newman et al., 2014). The basin screening procedure employed here was based on a similar approach to the CAMELS methodology (Addor et al., 2017; Newman et al., 2014) but with a slightly broader inclusion of basins from the GAGES-II dataset. Both the CAMELS basins and additional basins included in our analysis are subsets of the GAGES-II dataset. As a result, most of the basins are unmanaged basins with drainage areas smaller than 2500 km^2 with minimal anthropogenic influence and at least 30 years of streamflow observations to ensure records for model training/calibration and validation.

Additional screening criteria were applied to the additional basins sourced from GAGES-II. These include limiting basins to those with one or fewer major dams (defined as storage > 5000 acre-feet), ensuring the ratio of reservoir storage to average streamflow (1971-2000) was below 10%, and selecting basins with a GAGES-II hydro-disturbance index of less than 10 (Falcone et al., 2010). To further verify the accuracy of basin boundaries and drainage areas, we enforced additional criteria based on GAGES-II boundary attributes. This included a boundary confidence score (on a scale of 2-10, with 10 indicating high confidence) of at least 8, a percent area difference of no more than 10% compared to NWIS values, and a qualitative check ensuring the HUC10 boundaries were deemed at least “reasonable” or “good” (further described in Falcone et al., 2010; GAGES-II: geospatial attributes of gages for evaluating streamflow., 2021)). It should be noted that only 76 basins (out of 664 basins used for model training as described in Sect. 2.3.3) had NRCS forecasts available for the purpose of comparison. A majority of these basins lie within the US Environmental Protection Agency’s snow level III ecoregions labeled in Fig. 3. 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).

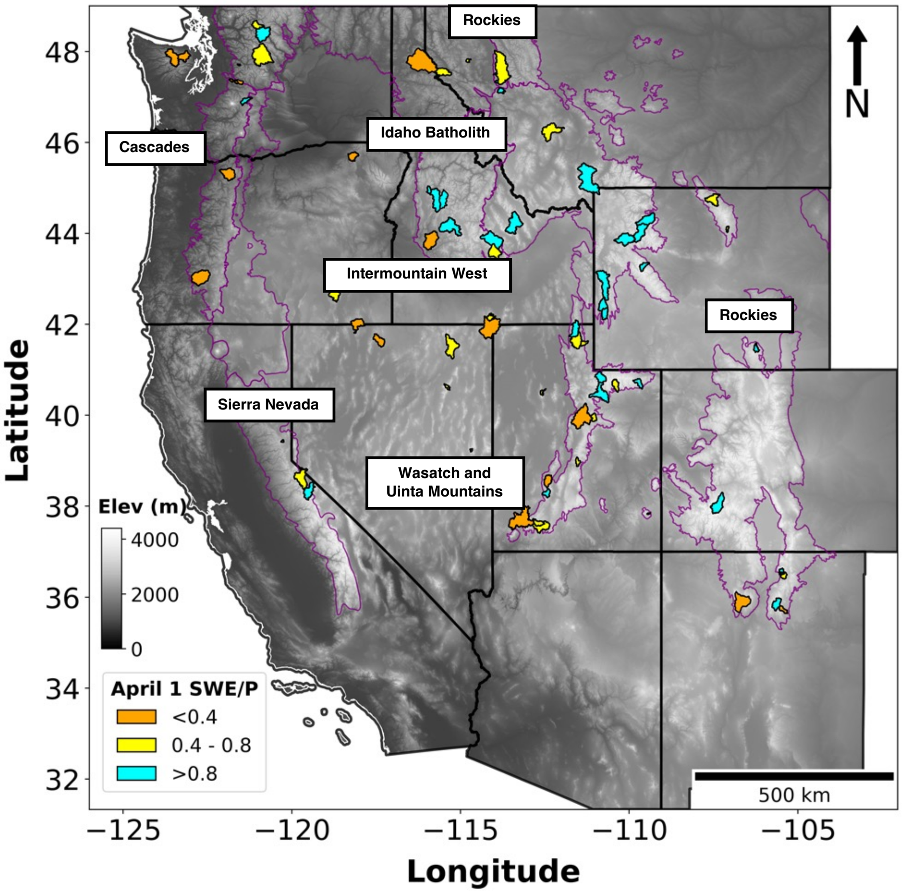


Figure 3: A map of the study domain, comprising 76 USGS drainage basins across the western US colored by the ratio of April 1 SWE to water year-to-date precipitation. The purple boundaries indicate the North American snow ecoregions Level III generated by the US Environmental Protection Agency (US EPA, 2015). These ecoregions include the Cascades, Idaho Batholiths, Intermountain West, Rockies, Sierra Nevada, and Wasatch and Uinta Mountains.

2.3 Synthetic Forecasts

In this study, synthetic forecasts are used to more clearly understand the role of forecast errors on economic value. We recognize that true forecasts have irregular error structures, which are difficult to interpret. To help interpret the relationship between forecast errors and PEV in true forecast systems, we introduce systematic errors into both the mean and standard deviation of observed AMJJ volumes. 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). The choice to set the mean of synthetic forecasts equal to observations and the standard deviation to inter-annual variability ensures the synthetic forecasts reflect key characteristics of the observed system. Aligning the mean with observations maintains comparability, while using inter-annual variability captures the system's inherent uncertainty. This design is crucial for studying irregular error structures, as it realistically represents the scale and variability of true forecasts.

By mirroring these properties, the synthetic experiments provide a controlled yet representative framework for analyzing how irregular error structures impact forecast value.

The observations are modified by applying a percent change to the mean, followed by a percent change to their standard deviation (Fig. 4a). An ensemble of forecast members is then generated, normally distributed around the modified mean and standard deviation. The varying spread of ensemble members reflects different potential hydrologic futures, allowing us to evaluate the performance of the forecast systems not only in terms of a single prediction but across a wide range of possible outcomes. Additionally, if the errors result in negative values, we truncate the range of the forecast to be greater than or equal to zero to avoid negative forecasts. In Fig. 4b, two synthetic forecasts are presented: one with a 50% increase in both the mean and standard deviation, represented by the blue line and ribbon, and another with a 50% decrease, represented by the red line and ribbon. These lines illustrate the ensemble spread of possible synthetic forecast based on the modified statistics. For comparison, the black dotted line and ribbon show the ensemble spread derived from the original observations and their standard deviation (i.e., interannual variability), serving as a reference point for evaluating deviations in the forecasts. Additionally, the white circle and triangle denote the original mean and standard deviation of the observations, respectively, offering a baseline to assess how the synthetic adjustments impact the overall distribution.

https://doi.org/10.5194/egusphere-2024-4046
Preprint. Discussion started: 14 January 2025

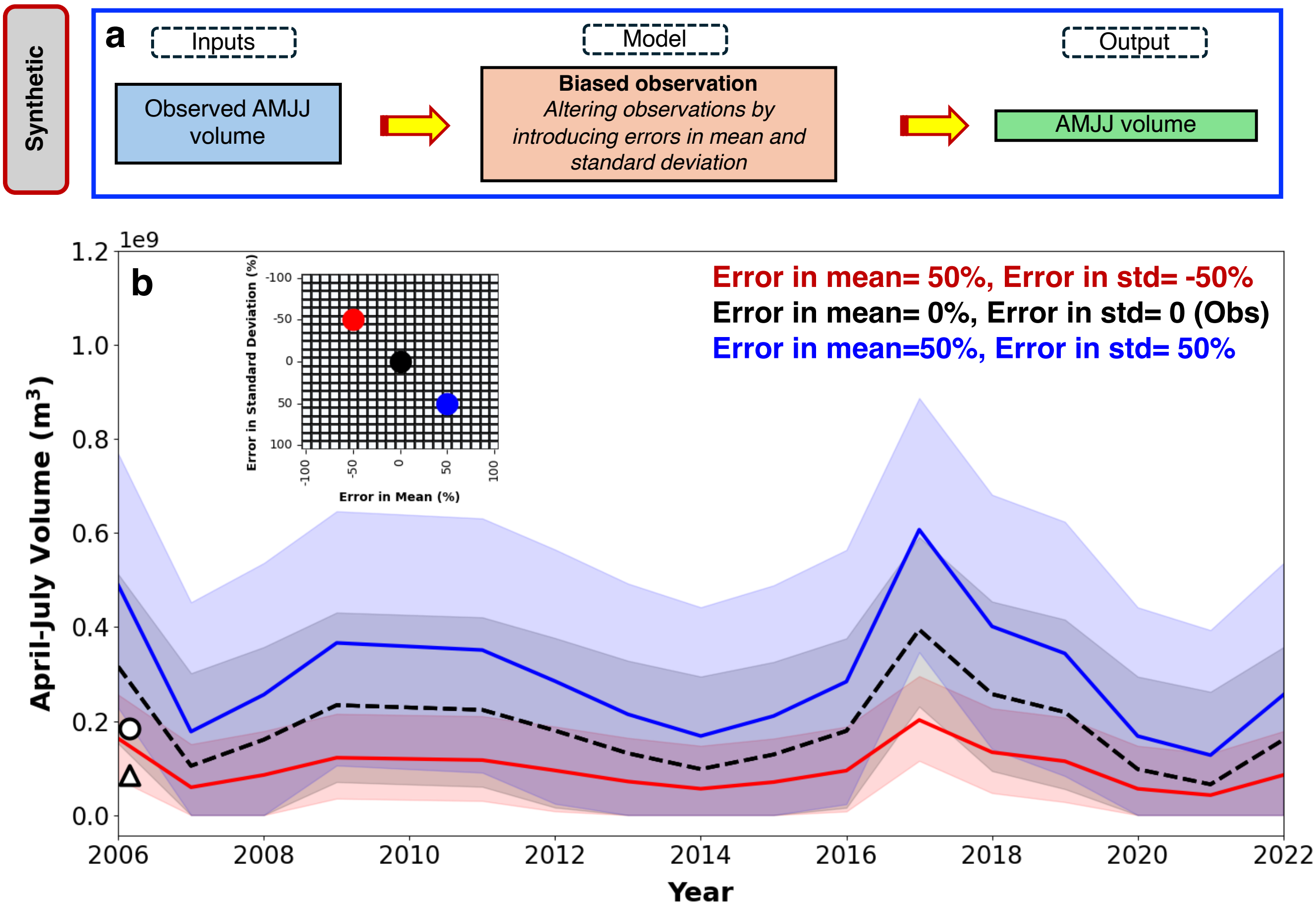


Figure 4: Schematic of the model workflow used to generate synthetic forecasts. (a) Illustration of two synthetic forecasts with ensemble spread in AMJJ volumes: one with a 50% increase in both the mean and standard deviation, represented by the blue line and ribbon, and another with a 50% decrease, represented by the red line and ribbon (b). The black dotted line and ribbon show the ensemble spread derived from the original observations and their standard deviation (i.e., interannual variability), whereas the white circle and triangle show the original mean and standard deviation of the observations, respectively. These forecasts correspond to different error structures shown by an inset grid.

2.4 True Forecasts

A schematic of model workflows of three true forecast systems is provided in Fig. 5 – two designed for this study and one used operationally. The two designed true forecast systems use the Ensemble Streamflow Prediction (ESP) framework. The first is a process-based hydrologic model (WRF-Hydro – WRFH; Gochis et al., 2020), which simulates streamflow evolution based on physical processes like snowmelt, soil moisture, and runoff (Fig. 5a). The second is a deep-learning model (LSTM; Hochreiter and Schmidhuber, 1997), which leverages historical patterns from the data (Fig. 5b). 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). 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. A detailed description of the ESP methodology is provided in Sect. 2.4.1, and the implementation of both models, including input data, model architecture, calibration/training, and forecast generation, is discussed in Sect. 2.4.2 and 2.4.3.

In addition, we also use Natural Resources Conservation Services statistical forecasts (now "NRCS") operational forecasts over the study watersheds to benchmark true forecasts. These forecasts were chosen since they are methodologically consistent across all study regions and easily accessible for a larger number of basins and years. The NRCS employs a Principal Component Regression model. This model is usually modified to retain the principal components (Garen, 1992; Lehner et al., 2017) and uses predictors like SWE, accumulated precipitation from SNOTEL, and antecedent streamflow from USGS to predict AMJJ volumes (Fig. 5c).

All true forecasts include five forecasted exceedance probabilities at 90, 70, 50, 30, and 10%. To clarify, 90% means there is a 90% chance that the observed AMJJ volumes will exceed this forecast value and a 10% chance that it will be less than this forecast value. These probabilities are based on the multiple ensemble members in all true forecasts. In order to make all forecasts comparable, the same five probabilities of exceedance were obtained from both true and synthetic forecasts. True forecast systems often deviate from idealized assumptions, exhibiting non-normal error distributions and varying variances due to the influence of dynamic, unpredictable factors and system-specific behaviors. This phenomenon is demonstrated in Fig. A3, where an exposition of these irregular error structures is presented through time-series analyses of AMJJ volumes. These time series illustrate how interannual fluctuations in volumes reveal underlying heteroscedasticity, skewness, and other deviations from standard statistical norms.

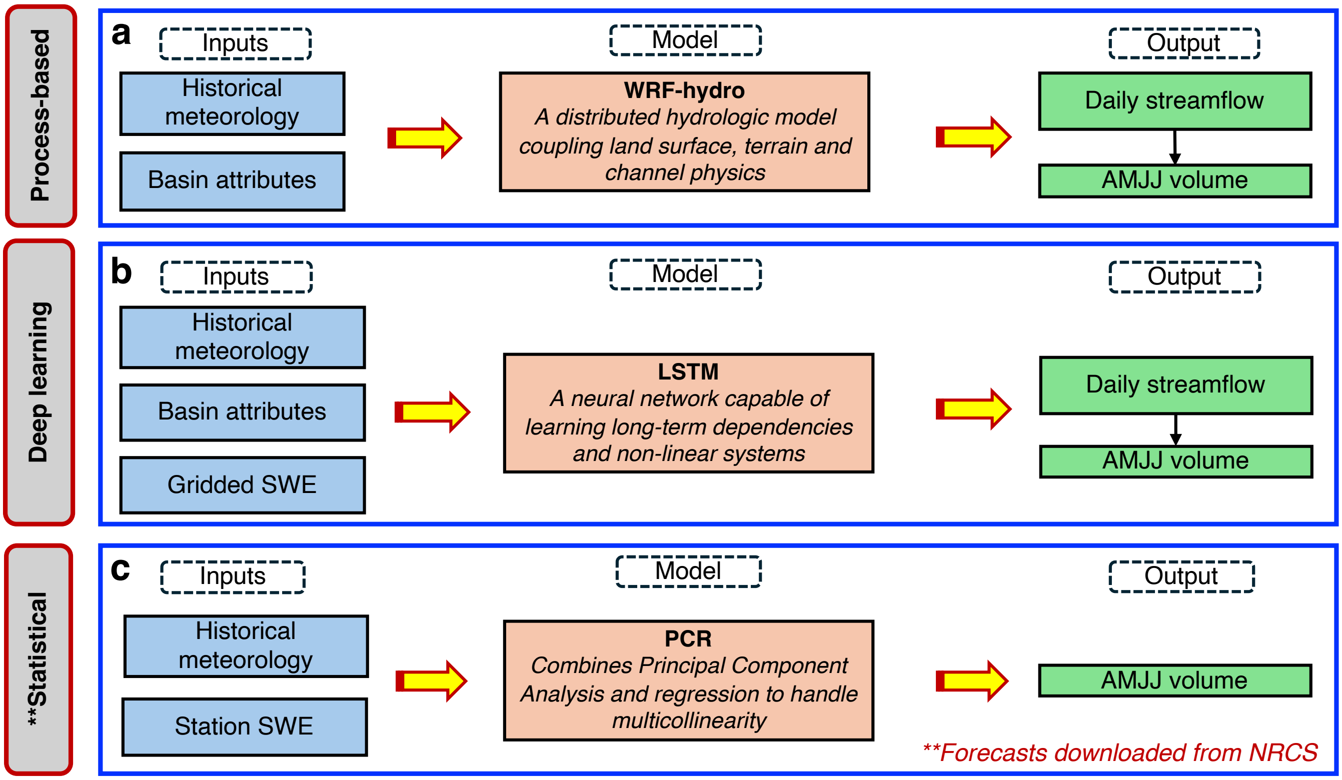


Figure 5: Schematic of model workflows used to generate true forecasts, including the inputs, model type, and outputs. (a) shows the workflow for the process-based hydrologic model, WRF-Hydro, (b) for the deep learning model, LSTM, and (c) for the NRCS statistical forecasts.

2.4.1 Ensemble Streamflow Predictions (ESPs)

In general, ESP forecasts generated on April 1 (i.e., forecast date) hold significant operational importance. This is because April 1 historically serves as a surrogate for the timing of peak SWE conditions and provides near-maximum predictive information (Livneh & Badger, 2020; Pagano et al., 2004). In this study, April 1 as a forecast date is closely tied to forecast skill and serves as an optimal point for calculating forecast value. However, depending on the region and the context of decision-making, users may choose a different forecast date that better aligns with their needs and associated forecast skill. The ESP simulation begins at the start of the water year (October 1), utilizing true meteorological forcings to initialize the model's initial conditions on April 1. Using these initial conditions on April 1 and meteorological forcings from historical years, an ensemble of streamflow traces is produced in the forecast period (April-July) as a function of the current hydroclimatic state and historical weather conditions (Day, 1985; Troin et al., 2021).

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. For example, a forecast generated on April 1 (illustrated in Fig. 6) uses observed meteorology up to that date, with the model's initial conditions preserved, and then generates streamflow traces based on meteorological forcings from historical years for the remainder of the forecast period.

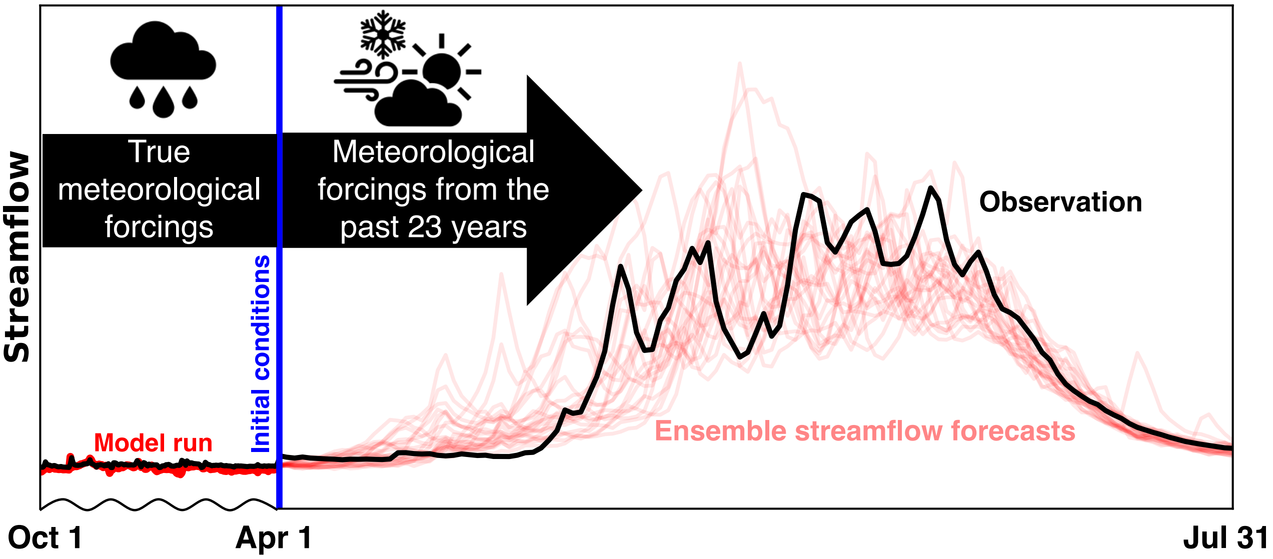


Figure 6: Illustration of an ESP forecast issued on April 1. The thick red line on the left depicts the model run before the forecast date using 'true' meteorological forcings starting from October 1. Using the model's initial conditions on April 1 (shown in blue) and historical meteorological forcings from the past 23 years, ensemble streamflow forecasts are generated (shown with faint red lines). Data are from Johnson Creek, ID, USGS basin 13313000, for the forecast year 2011. The broken x-axis shown here is not uniform and represents the ESP conceptually (Modi et al., 2024).

2.4.2 Implementation of WRF-hydro in an ESP framework

WRFH model architecture

WRFH is a distributed hydrologic model architecture designed to facilitate the coupling of hydrologic models with atmospheric models through improved representations of terrestrial hydrologic processes associated with spatial redistribution of surface, subsurface, and channel waters across the land surface (Gochis et al., 2020). At its modeling core, WRFH uses the Noah-MP land surface model, an improved version of the baseline Noah land surface model (Ek et al., 2003; Niu et al., 2011), that offers multi-parameterization through several vegetation, snow, radiation transfer, runoff and groundwater schemes. We use the National Water Model (NWM) scheme configuration developed and managed by NOAA to generate short-to-medium-range streamflow forecasts over the 2.7 million stream locations nationwide (Cosgrove et al., 2024). We only match the physics permutations used in the NWM configuration and not the routing configuration used in the operational NWM. We rely on a channel network that uses a default channel structure and is generated using Hydrosheds Digital Elevation Model data (Lehner et al., 2008). WRFH is set up on a 1 km horizontal grid spacing, simulating lateral water redistribution on the surface and shallow sub-surface on a 100 m grid spacing. The model is run hourly, with model outputs aggregated daily for analysis purposes. A description of WRFH model parameters and calibration is provided in Appendix A1.

WRFH model inputs

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). The Noah-MP land surface model is parametrized using surface albedo, leaf area index and green fraction from the Moderate Resolution Imaging Spectrometer (Myneni et al., 2015), land-use/land-cover from the United States Department of Agriculture – National Agricultural Statistics Service (CropScape - NASS CDL Program, 2019), soil type from State Soil Geographic (STATSGO), maximum snow albedo and soil temperature from the WRF Preprocessing System data page managed by UCAR (WRF Preprocessing System (WPS) Geographical Static Data, 2019). Daily streamflow estimates from the USGS's National Water Information System (USGS NWIS) are obtained for the USGS stream gages corresponding to the basin outlets that are used to calibrate the model and described below.

WRFH forecast generation

We generate WRFH ESP forecasts on April 1 for WY2006-2022 before (now WRFH_{DEF}) and after calibration (now WRFH_{CAL}). These forecasts leverage historical meteorological data from all available years WY1983-WY2022 except the forecast year by using them as inputs to WRFH. 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. An ensemble of streamflow traces is produced in the forecast period using these memory states on the forecast date and historical meteorological forcings. The forecasted daily streamflow is further cumulated to AMJJ volume and used for analysis.

2.4.3 Implementation of LSTM in an ESP framework

LSTM model architecture

This study adopts a model architecture similar to Kratzert et al. (2019), as followed by Modi et al. (2024) (now "M24"), which has been shown to simulate and forecast streamflow well for basins with minimal anthropogenic influence. This M24 setup only includes hyperparameters – externally set values that govern the training process - not model parameters or inputs. This list of hyperparameters is briefly outlined and explained in Table 2. Using the M24 setup, the LSTM includes a single hidden layer comprising 256 units, where units act as computational units through which data flows, and the hidden layer is responsible for learning the intricate structures in the data. Additionally, the hidden layer is configured with a dropout rate of 0.4, which involves randomly dropping neurons during training to mitigate overfitting. The input sequence length used is 270 days, which specifies the number of preceding time steps fed into the LSTM to produce streamflow on a given day. A description of LSTM training is provided in Appendix A2.

Table 2: The LSTM hyperparameters used in this study (adapted from Kratzert et al. (2019) and Modi et al. (2024)).

Parameter	Description	Selected Value
Number of hidden layers	The number of stacked LSTM layers in the model	1
Number of units	The number of memory cells in each LSTM layer that determine the capacity to learn from the data	256
Input sequence length	The length of preceding time steps fed into the LSTM	270
Batch size	The number of training samples used in one iteration	2000
Dropout rate	The fraction of the units to drop during training to prevent overfitting	0.4
Number of epochs	The number of times the entire training dataset is passed through the model	40
Optimizer	The algorithm used to minimize the loss function	Adam
Learning rate	The step size used by the optimization algorithm to update the model weights	0.001

LSTM model inputs

The training inputs for the LSTM model (as detailed in Table A2) include meteorological forcings from the AORC (Fall et al., 2023), which are aggregated daily and spatially averaged across each basin using 1 km grid cells and identical to the WRFH inputs. These forcings consist of precipitation, average wind speed, 2 m average air temperature, incoming longwave and shortwave radiation, near-surface air pressure, and vapor pressure. In addition to these meteorological forcings, static predictors are included, consisting of basin attributes from the GAGES-II dataset, which remain constant over time and are selected to mirror those utilized in the CAMELS dataset, following the work of Arsenault et al. (2022) and Kratzert et al. (2019). We obtain daily snow information from the gridded snow dataset developed at the University of Arizona (Broxton et al., 2019; Zeng et al., 2018 – now UA), spatially averaged for each basin from 1/16-degree grids. Lastly, daily streamflow estimates from the USGS's National Water Information System (USGS NWIS) are obtained for the USGS streamgages corresponding to the basin outlets.

LSTM forecast generation

We generate LSTM ESP forecasts on April 1 for WY2006-2022, excluding years used in training, using model parameters from fully-trained settings. These forecasts leverage historical meteorological data and snow information from all available years WY1983-WY2022 except the forecast year. For ESP forecasts on April 1, the LSTM simulation begins at the start of the water year, i.e., October 1, using true meteorological forcings and snowpack information to obtain LSTM's memory states on the forecast date. During the forecast period, the historical meteorological data is used similarly to process-based models. However, special treatment is applied to snowpack information, integrating known snowpack information on the forecast date and assumptions about snow evolution after the forecast date as a way to boost the representation of hydrologic memory that is commensurate with the physical hydrological system. We adopt the "$\mathrm{ESP_{RetroSWE}}$" forecast experiment from Modi et al. (2024), which integrates the known SWE information on the forecast date (from the forecast year) with explicit accumulation and ablation rates after the forecast data from individual historical years. More information on the design and performance of "$\mathrm{ESP_{RetroSWE}}$" is provided by Modi et al. (2024). The forecasted daily streamflow is further cumulated to AMJJ seasonal volume and used for analysis.

2.5 Performance metrics

We employed four fundamental performance metrics to calibrate/train the models and evaluate forecast accuracy, drawing from those widely adopted to quantify streamflow accuracy. The Nash Sutcliffe Efficiency (NSE) was used to quantify streamflow prediction accuracy of the different models. The NSE ranges from negative infinity to 1, with 1 indicating perfect agreement between the simulated and observed values, and values closer to 0 indicating poorer performance. The Normalized Root Mean Square Error (NRMSE, in %) was used to analyze the skill of simulated AMJJ streamflow volume against the corresponding observed streamflow volumes. The RMSE was normalized by the median of observed streamflow volumes, and values closer to 0 indicate better performance. The correlation assesses the agreement in patterns between the simulations and observations, with values ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation). The ratio of standard deviation compares the spread between the simulations and observations to assess whether the simulations capture the correct level of variability in the observations. A ratio of standard deviation of 1 indicates the simulations have captured the correct level of variability. 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.

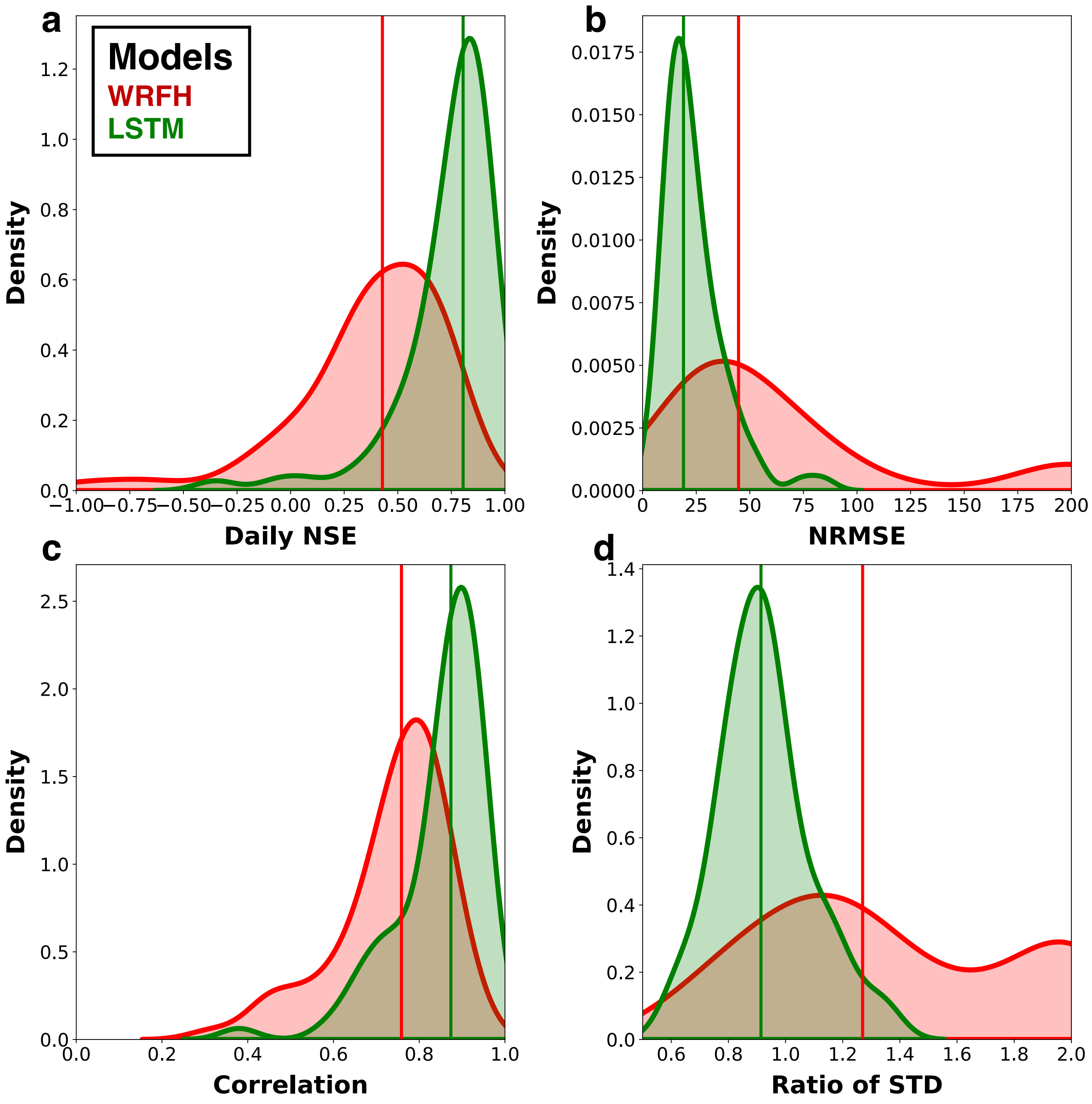
3 Results

We first compare the historical model performance from the WRFH and LSTM models with respect to the observations (Sect. 3.1). In Section 3.2, we analyze how errors in mean and standard deviation impact the forecast skill and value for synthetic (i.e., imposed errors on observations) and true forecasts (i.e., estimated with respect to the observations). In section 3.3, we examine the relationship between forecast skill and value from different forecast systems, with different severities of drought and the impact of categorical variables, particularly on forecast value.

3.1 Historical model performance of our designed true forecast systems

We assess the performance of our designed true forecast systems using historical data to ensure their effectiveness in accurately simulating streamflow. 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). LSTM model consistently outperformed the WRFH model across all metrics, with statistically significant improvements. For example, LSTM showed a median NSE and NRMSE of 0.80 and 20%, whereas WRFH showed 0.42 and 45%, respectively. 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. LSTM also showed a reasonable ratio of standard deviation of 0.95, whereas WRFH showed 1.25. These results suggest that the LSTM models, particularly LSTM, perform much better in simulating streamflow than the WRFH models. The WRFH and LSTM showed satisfactory utility in simulating daily and seasonal streamflow and were chosen for further comparison to analyze the skill-value relationship for different model architectures. To underscore the importance of model calibration and training, we compare the performance of the models before and after calibration/training. In general, we observe improvements across all metrics for both models (additional details can be found in Appendix A3).

https://doi.org/10.5194/egusphere-2024-4046
Preprint. Discussion started: 14 January 2025



465 **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.**

3.2 Forecast skill and value are affected by errors in mean and standard deviation

In section 3.2.1, we first analyze synthetic forecasts to gain insights into their skill and value with respect to the errors in mean and standard deviation. In section 3.2.2, we examine true forecasts, quantifying the errors in mean and standard deviation, and evaluate their skill and value (Sect. 3.2.2). Finally, we overlap skill and value from true forecasts with those from synthetic forecasts to diagnose and interpret how errors in mean and standard deviation impact forecast skill and value. 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).

3.2.1 Synthetic forecasts

Figures 8a and 8b illustrate the sensitivity of forecast skill and value to errors in both mean and standard deviation across drought years. In Fig. 8a, a lower number indicates better forecast skill, meaning darker shades (close to purple) represent worse skill, whereas lighter shades (close to yellow) indicate good skill. The optimal forecast skill (close to zero) occurs particularly around errors in the mean between -20% and 20% and errors in the standard deviation of -100% and -50%. It is important to note that a standard deviation of 0 indicates how closely the forecasted variability aligns with the historical interannual variability. As errors in mean or standard deviation increase beyond these ranges, forecast skill worsens. In Fig. 8b, a higher number indicates a greater value, meaning darker shades (close to purple) represent a low value, whereas lighter shades (close to yellow) indicate a greater value. The optimal forecast value (closer to 0.9) is observed with an error in mean between -20% and 20% and an error in standard deviation between -100% and 0%. A key observation is that a greater forecast value extends further into positive errors in the mean compared to negative errors, resulting in a symmetric forecast skill around mean errors but an asymmetric forecast value.

We present four synthetic forecasts (Fig. 8 c-f) to demonstrate how forecast skill and value are impacted by systematic errors in mean and standard deviation in case of a categorical decision. In each plot, 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}. 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. 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. This comparison reveals why forecast skill remains symmetric around errors in mean while forecast value is distinctly asymmetric. This highlights the different sensitivities of skill and value to errors in mean and standard deviation, likely due to the interplay of categorical measures, where forecast value responds differently than forecast skill.

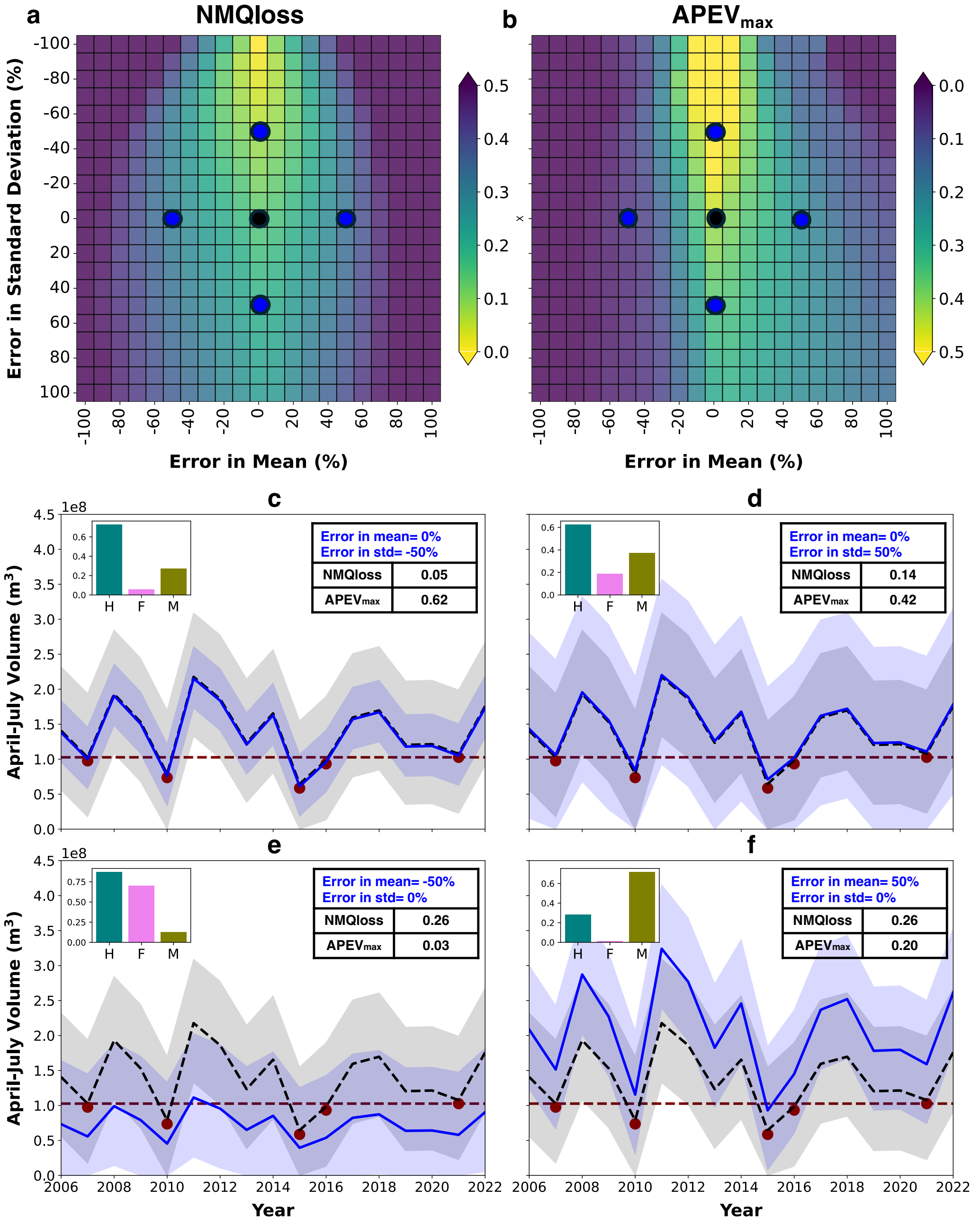


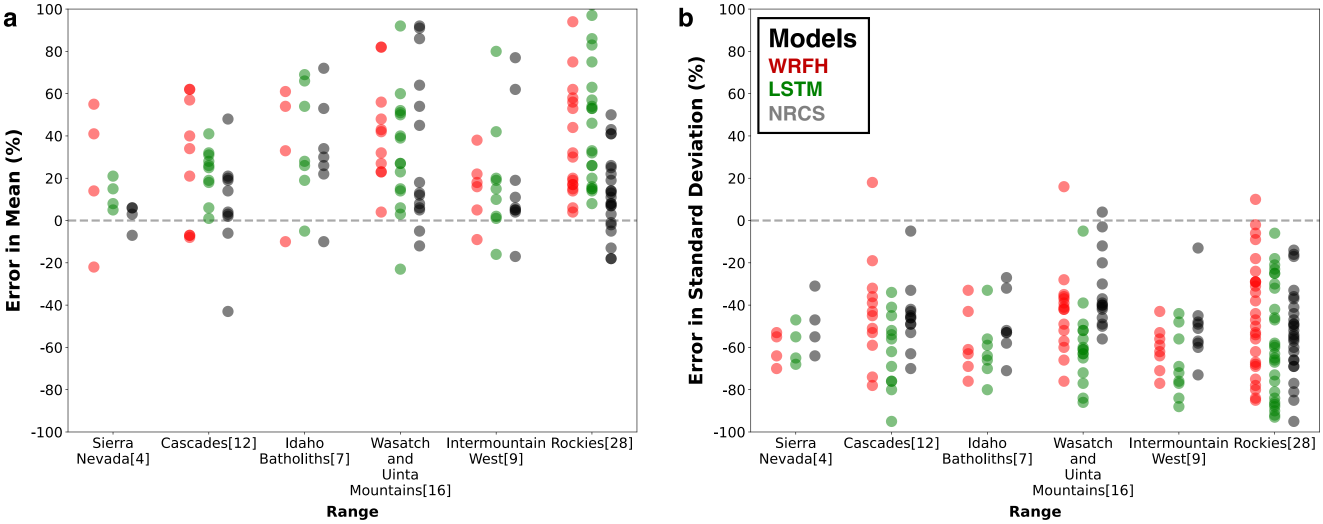
Figure 8: Sensitivity of quantile loss (forecast skill) and APEV$_{max}$ (forecast value) to errors in mean and standard deviation for synthetic forecasts. The background heatmaps (a and b) represent synthetic forecasts, with lower values showing better forecast skill (closer to yellow) and higher values better forecast value (closer to yellow). We also illustrate four synthetic forecasts (shown in blue) corresponding to different errors in mean and standard deviation (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.

https://doi.org/10.5194/egusphere-2024-4046
Preprint. Discussion started: 14 January 2025

3.2.2 True forecasts

Error in mean and standard deviation

Figure 9 illustrates the errors in mean and standard deviation for all true forecast systems across 76 basins. Across all models, 510 there is a consistent trend of overprediction in mean during drought years (Fig. 9a), with a standard deviation in forecasts lower than interannual variability from historical records (Fig. 9b). The degree of overprediction is generally higher in the Wasatch and Unita Mountains and the Rockies, while it is smaller in the Sierra Nevada, Cascades, Idaho Batholiths, and the Intermountain West. This is likely because the limited precipitation and snow observations in high-elevation regions introduce uncertainty in interpolated precipitation values (Vuille et al., 2014), which are assimilated into the model inputs (i.e., AORC). 515 An intercomparison of the errors in mean across the models reveal significant differences. The median error in the mean is 55% for WRFH, 30% for the LSTM, and 14% for the NRCS model. LSTM shows lower mean errors than WRFH$_{CAL}$, aligning with historical performance trends, while NRCS performs best, exhibiting the smallest errors in the mean as observed in Fig. 7. In contrast to overprediction of the mean, these models mostly show a standard deviation that is lower than interannual variability during WY2006-2022, as indicated by the negative errors in standard deviation (Fig. 9b). These results are 520 consistent with the trends observed in the synthetic forecasts (Fig. 8), where higher forecast skill and value were associated with negative errors in standard deviation. This understanding of errors in mean and standard deviation underscores the importance of capturing both mean state and variability for improving forecast performance and value, particularly in complex mountainous regions like the Rockies, where observational limitations pose challenges.



525 **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.**

Forecast skill

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. The background heatmaps represent the median skill from synthetic forecasts across basins, while the scatter points represent true forecast systems based on the estimated errors with respect to the observation during drought years. Each dot in Fig. 10 represents a basin with colors showing the median skill and value only during drought years. We overlap true forecasts over synthetic forecasts to systematically analyze and understand the role of irregular error structures in true forecast systems on the forecast skill. WRFH and LSTM show good correspondence when compared to the synthetic forecasts (i.e., colors match well between the points and heatmap), based on the estimated RMADs of 30% and 23%, respectively. Notably, NRCS shows the highest consistency and robustness, with a RMAD of 20%, closely aligning with the synthetic forecasts. The scatter points' distribution across each heatmap highlights the sensitivities of the forecast skill to errors in mean and standard deviation for the different forecast systems. Overall, this approach highlights the importance of considering errors in mean and standard deviation when diagnosing true forecast skill. It offers valuable insights into the reliability and robustness of forecasts in real-world scenarios, emphasizing how different systems perform under varying conditions of uncertainty.

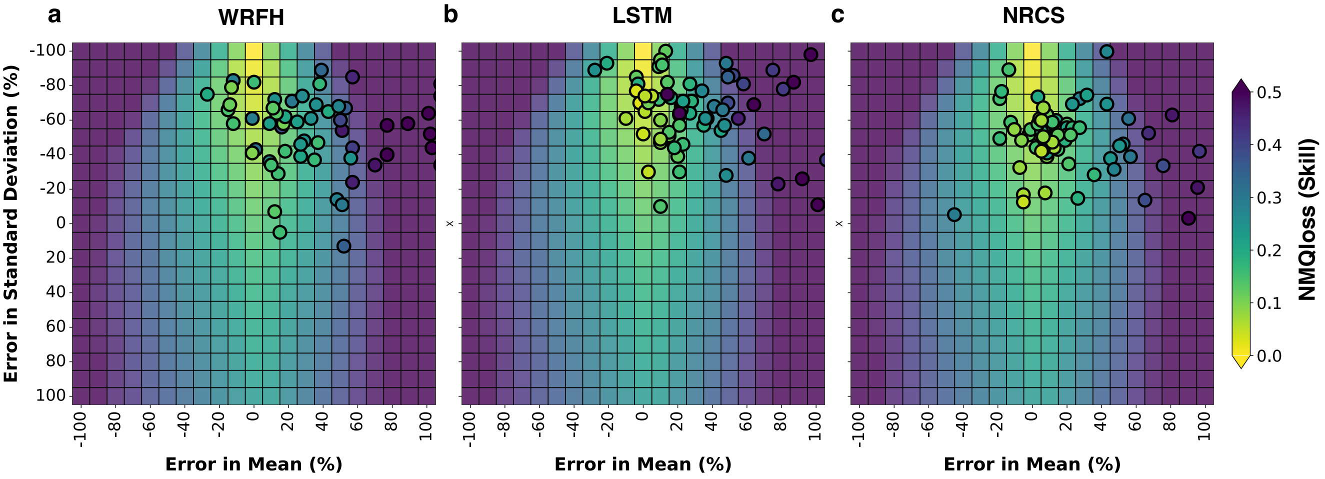


Figure 10: Comparison of skill between synthetic and true forecast systems to mean and standard deviation. Normalized mean quantile loss (NMQloss) of three forecast systems (WRFH, LSTM, and NRCS) represented as scatter points (each point represents a basin), indicating the true skill during drought years between WY 2006 and 2022. The background heatmaps represent the sensitivity of skill to errors in mean and standard deviation for synthetic forecasts. RMAD for true forecast systems from the optimal scenario are 30%, 23%, and 20% for WRFH, LSTM, and NRCS, respectively.

Forecast value

Figure 11 is similar to that in Figure 10; however, it focuses on APEV$_{max}$ rather than NMQloss. Despite the good correspondence observed in forecast skill (Fig. 10), all true forecast systems demonstrate poor correspondence in value when compared to synthetic forecasts. This can be seen by the significant difference in the colors of points and heatmaps. This results in estimated RMAD for WRFH, LSTM, and NRCS to 100%, 81%, and 91%, respectively, dramatically different from the deviations in skill. These large deviations show that errors in mean and standard deviation do not effectively explain the variations in the forecast value between true and synthetic forecasts. None of the true forecast systems were able to consistently capture forecast value, as seen from our comparison with synthetic forecasts. The distribution of scatter points across each heatmap further emphasizes that $APEV_{max}$, unlike NMQloss, is not a simple function of errors in mean and standard deviation or, in broad terms, forecast skill.

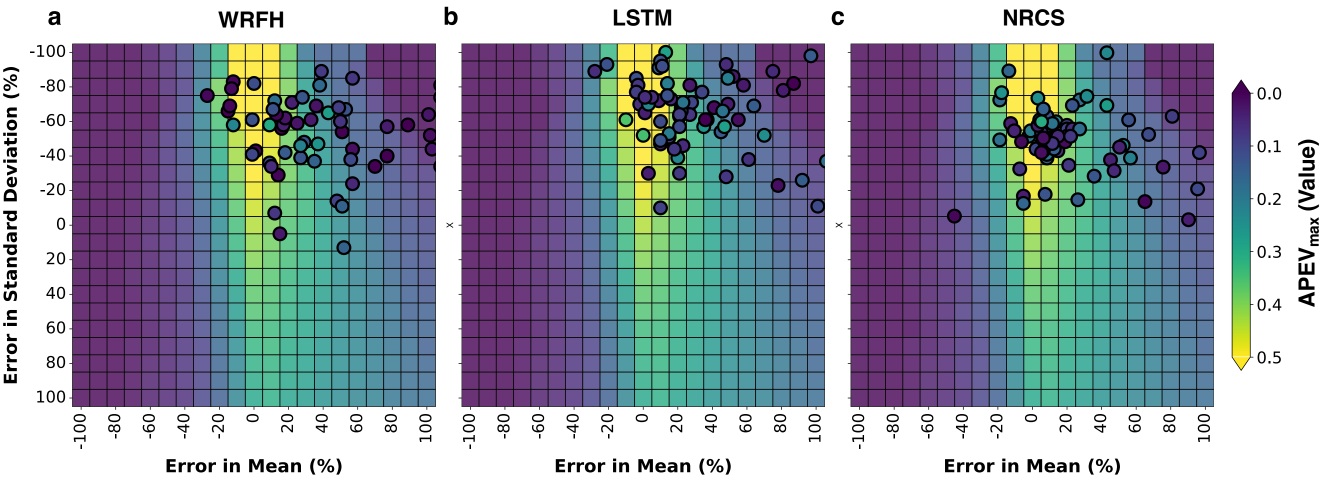


Figure 11: Comparison of value between synthetic and true forecast systems to the errors in mean and standard deviation. Area under PEV_{max} curve ($APEV_{max}$) of three forecast systems (WRFH, LSTM, and NRCS) represented as scatter points (each point represents a basin), indicating the true value during drought years between WY 2006 and 2022. The background heatmaps represent the sensitivity of $APEV_{max}$ to errors in mean and standard deviation for synthetic forecasts. RMAD for true forecast systems from the optimal scenario are 100%, 81%, and 91% for WRFH, LSTM, and NRCS, respectively.

3.3 Relationship between skill and value

3.3.1 Comparison between synthetic and true forecasts

We use the overlap between synthetic and true forecast systems from Figs. 10 and 11 to explore their skill-value relationship. Figure 12a compares the skill (NMQloss) and value ($APEV_{max}$) of the synthetic forecasts (i.e., grids in the heatmap) that overlapped with the true forecast systems (i.e., scatter points) based on errors in mean and standard deviation. Similarly, Figure 12b shows the skill and value of the true forecast systems. Both scatter plots show the relationship between NMQloss (forecast skill) and $APEV_{max}$ (forecast value) for three true forecast systems (WRFH, LSTM, and NRCS), with each point corresponding to a different basin. The dashed lines in the plots represent fitted exponential curves, highlighting the general trend that as skill increases (i.e., as NMQloss decreases), the value also improves (i.e., $APEV_{max}$ increases). The optimal skill and value are obtained at coordinate (0,1), where skill declines along the X-axis and value increases along the Y-axis. For synthetic forecasts, this trend is more pronounced, with high correlation values (≦0.65) across all models, indicating a strong negative relationship between NMQloss and $APEV_{max}$ across the entire range of NMQloss. In contrast, for the true forecasts, the relationship between NMQloss and $APEV_{max}$ weakens (r≦0.38) and becomes more variable suggesting that good forecast skill does not always translate to good forecast value (Turner et al., 2017). These plots collectively demonstrate that while NMQloss and $APEV_{max}$ are related, their relationship is complex, particularly in true forecast systems. 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.

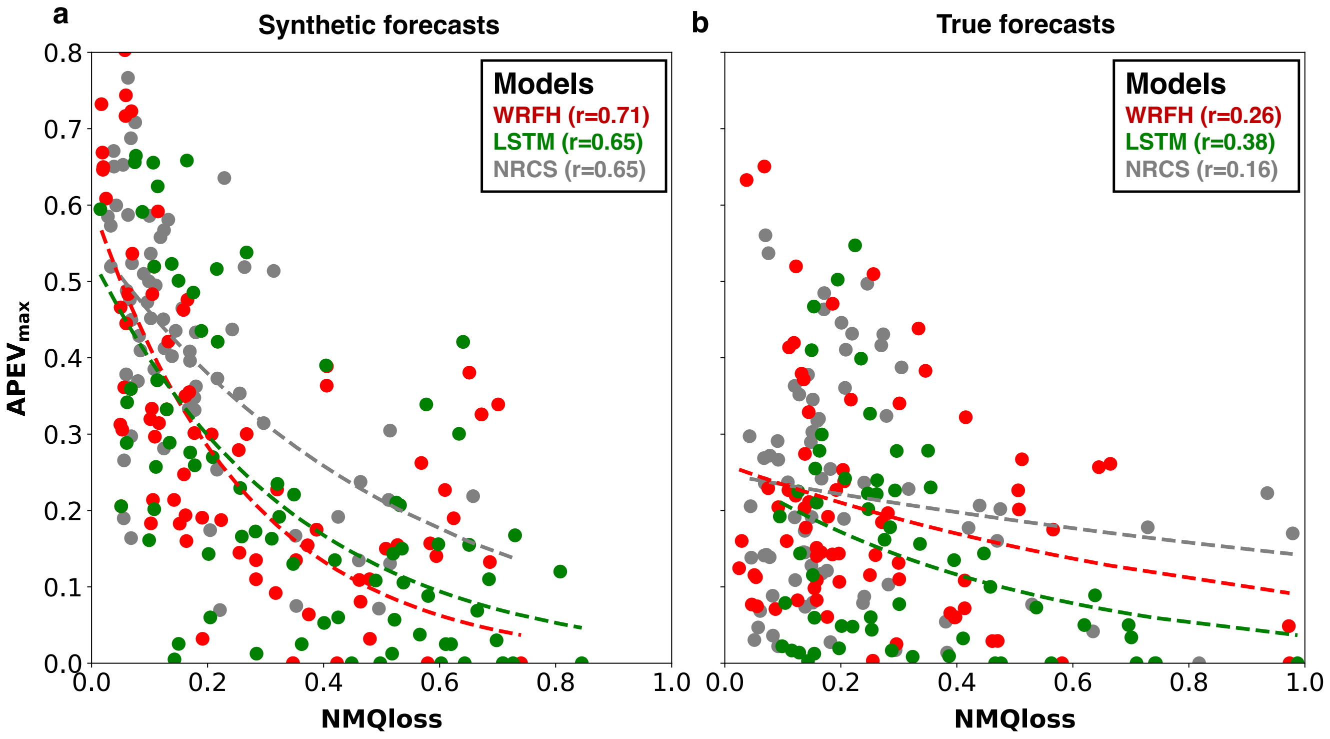


Figure 12: Scatter plots depicting the relationship between skill (NMQloss) and value (APEVmax) for synthetic and true forecast systems. The points in (a) and (b) represent the synthetic forecast (the grid of the heatmap) that overlap with true forecast systems (scatter point) in Figures 4.10 and 4.11.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

3.3.2 Skill-Value relationship monotonically changes with the severity of drought

Figure 13 illustrates the relationship between NMQloss and $APEV_{max}$ for three drought scenarios related to different severities. This includes three scenarios: AMJJ volume less than the 35th percentile (P_{35}), less than the 25th percentile (P_{25} used consistently in earlier analyses), and less than the 15th percentile (P_{15}), represented by green, orange, and red colors, respectively. Importantly, these scenarios are not independent of one another, as events identified below P_{35} also encompass those below

https://doi.org/10.5194/egusphere-2024-4046
Preprint. Discussion started: 14 January 2025

P_{15} and P_{25}. The top density plot shows the distribution of NMQloss across all true forecast systems and basins, showing generally wide distributions with median values around 0.20. The right density plot represents $APEV_{max}$, which shows a consistent increase in median values from 0.12 to 0.20 as the drought severity decreases (i.e., from P_{15} to P_{35}). This widening 595 of distributions suggests that the estimated skill and value for drought scenarios that are not limited to extremely dry events (i.e., P_{35}) tend to improve, i.e., higher accuracy and better economic benefit. Hence, the relationship changes monotonically with drought severity. Therefore, the decrease in forecast value is likely attributable to the increase in forecast error, as predictive models increasingly struggle in simulating progressively more extreme drought events (Chaney et al., 2015).

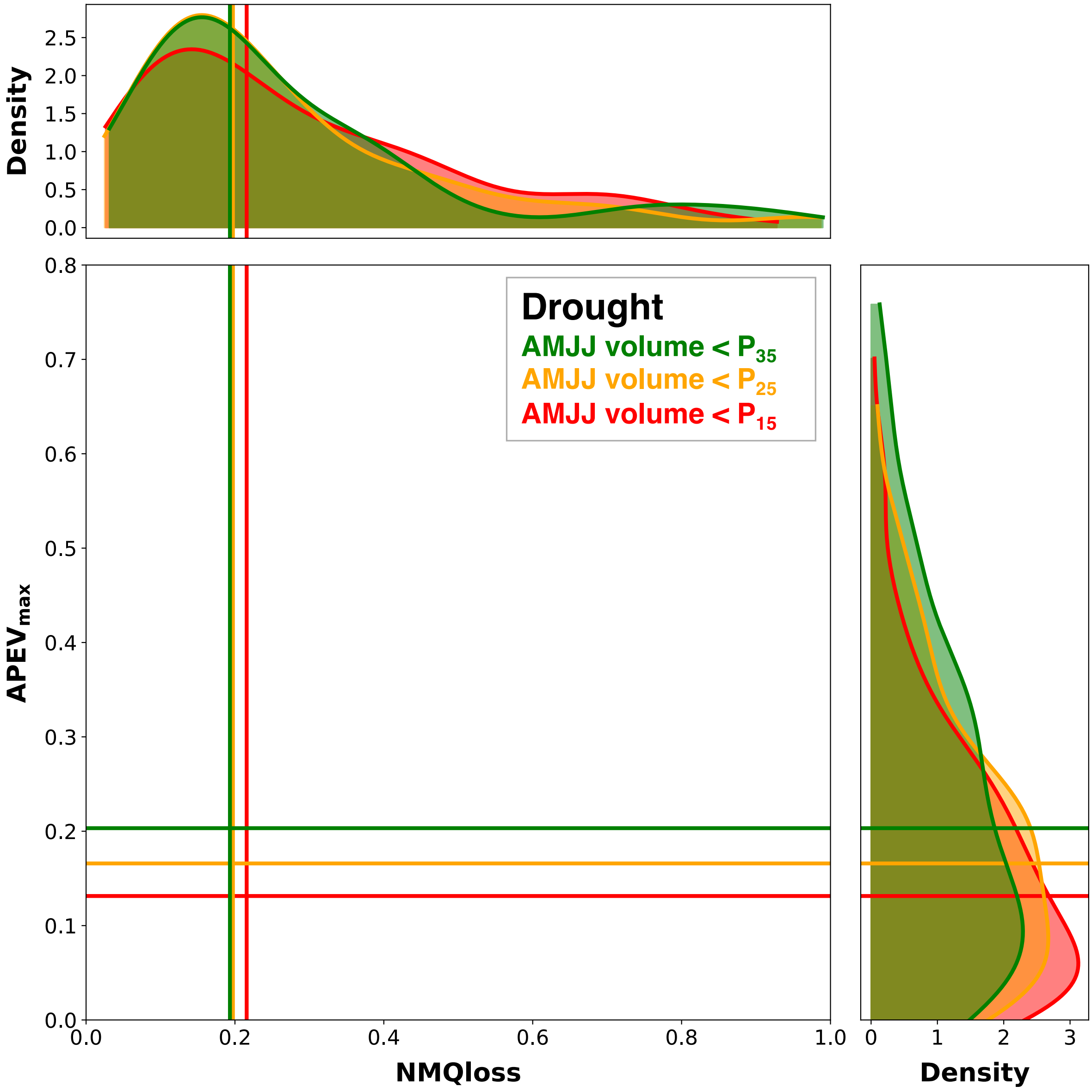


Figure 13: Relationship between NMQloss and APEV$_{max}$ shown for three drought scenarios related to different severities. These drought severities are represented by AMJJ volume being less than 35th percentile (P_{35} - green), 25th percentile (P_{25} - orange), and 15th percentile (P_{15} - red). The top density plot shows the distribution of NMQloss across all forecast systems and basins, whereas the right-side density plot displays the distribution of APEV$_{max}$.

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

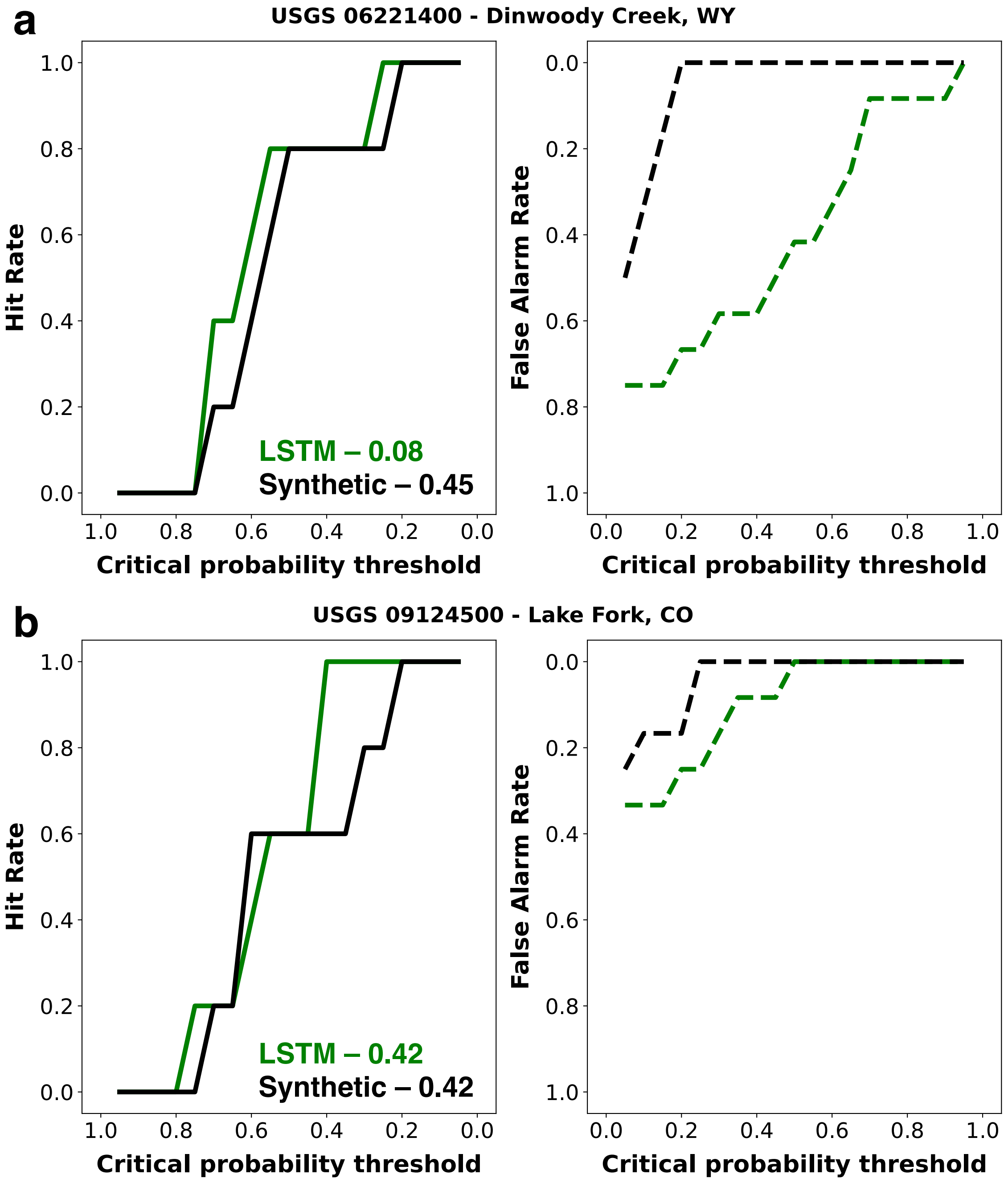
In decision-making, a high hit rate ensures timely actions for critical events like drought, while a low false alarm rate prevents unnecessary responses and maintains trust in the forecast system. Balancing these metrics is crucial for forecast value, as it determines the forecast's ability to support efficient and reliable decision-making. We analyze two critical components of APEV$_{max}$: the Hit Rate and False Alarm Rate (Fig. 14). This analysis focuses on two distinct basins, Dinwoody Creek, WY (Fig. 14a) and Lake Fork, CO (Fig. 14b), across various critical probability thresholds (τ) – minimum probability at which a drought event is deemed likely enough to trigger an action. The left plots for each basin show the Hit Rate, while the right plots depict the False Alarm Rate. 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.

In the case of Dinwoody Creek, both synthetic and true forecasts demonstrate a similar pattern where, as the critical probability threshold (τ) decreases, the Hit Rate generally increases, eventually reaching a maximum of 1 (Fig. 14a - left). The value of 1 suggests that both forecasts effectively identify all drought events (below P_{25} between WY2006 and 2022) when the threshold becomes less strict. 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). However, as the threshold decreases, the False Alarm Rates for both forecasts diverge significantly before converging at maximum rates of 0.5 and 0.75 for the synthetic and true forecasts, respectively. This divergence results in a notable difference in APEV$_{max}$ values: 0.45 for the synthetic forecast and 0.08 for the true forecast.

In the case of Lake Fork, a similar trend is observed for the Hit Rate. As the critical probability threshold decreases, both the synthetic and true forecasts consistently detect more drought events as the threshold becomes less strict (Fig. 14b - left). However, the behavior of the False Alarm Rate differs from that in Dinwoody Creek. Here, both forecasts exhibit a gradual increase in the False Alarm Rate as the threshold decreases, but they converge more closely at maximum rates of 0.25 and 0.32 for the synthetic and true forecasts, respectively. This convergence results in similar APEV$_{max}$ values for both forecasts, each approximately 0.42.

Overall, these analyses highlight how the balance between Hit and False Alarm Rate impacts APEV$_{max}$ in different basins. While Dinwoody Creek shows a clear discrepancy in economic value between synthetic and true forecasts due to their divergent False Alarm Rates, Lake Fork displays a more aligned relationship, with both forecasts yielding similar APEV$_{max}$. These differences exist because of irregular error structures that are better captured in categorical measures than skill.

https://doi.org/10.5194/egusphere-2024-4046
Preprint. Discussion started: 14 January 2025



635 **Figure 14: Attribution of Hit Rate and False Alarm Rate across varying critical probability thresholds (τ). Two basins are shown: Dinwoody Creek, WY (top panels) and Lake Fork, CO (bottom panels). The left panels show the Hit Rate as a function of the critical probability threshold (τ - minimum probability at which a drought event is deemed likely enough to trigger an action) for the LSTM forecast (green) and its corresponding synthetic forecast (black). The right panels depict the False Alarm Rate. The values indicate the APEV$_{max}$ corresponding to each forecast.**

3.3.4 Value is largely explainable by hit and false alarm rates

Figure 15 illustrates the forecast value of three true forecast systems with respect to hit and false alarm rates. Unlike Figures 10 and 11, which analyzed errors in mean and standard deviation, this figure focuses on understanding the variability in the
640 value with respect to hit and false alarm rates. The background heatmaps represent the median value from synthetic forecasts across basins, while the scatter points represent the median value from each forecast system. This comparison was performed across 76 basins during drought years (below P_{25}) between WY2006 and 2022. Unlike Fig. 11, WRFH and LSTM show better correspondence of value when compared to the synthetic forecasts, based on the estimated RMADs of 78% and 70%, respectively. The estimated deviations are still higher primarily resulting from differences in smaller magnitude of forecast
645 value. Notably, NRCS shows the highest consistency and robustness, with a RMAD of 61%, closely aligning with the synthetic forecasts.

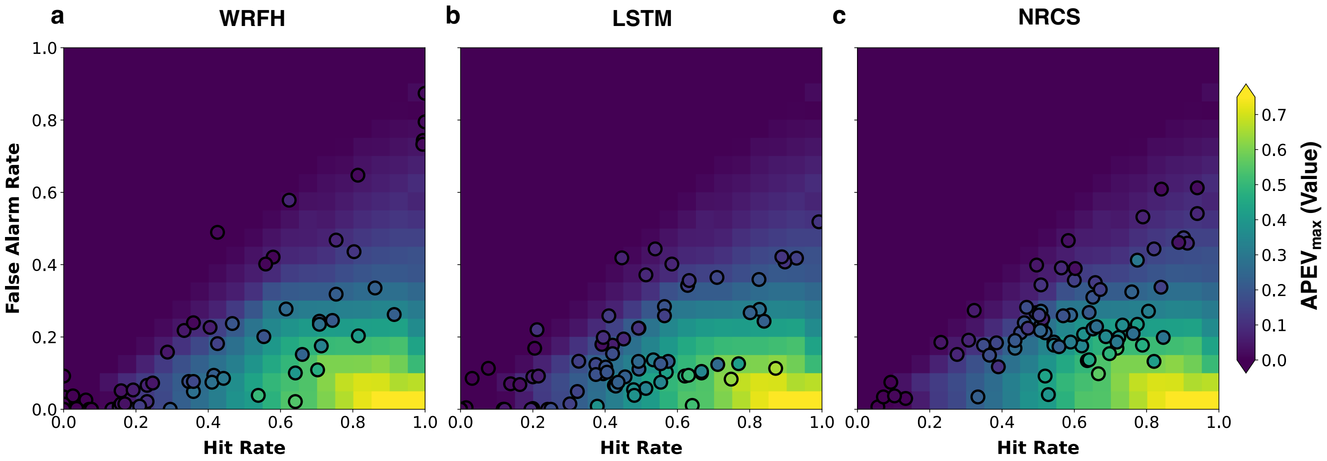


**Figure 15: APEV$_{max}$ of three forecast systems (WRFH, LSTM, and NRCS) represented as scatter points (each point represents a basin), indicating the actual value during drought years between WY 2006 and 2022. The background heatmaps represent the
650 sensitivity of APEV$_{max}$ to hit and false alarm rates for synthetic forecasts. RMAD for true forecast systems from the optimal scenario are 78%, 70%, and 61% for WRFH, LSTM, and NRCS, respectively.**

Discussion

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
655 observations, the value represents the economic benefits derived from utilizing those forecasts in the decision-making process. We use the relatively simple PEV metric, based on a cost-loss model, to evaluate how forecast skill in 76 unmanaged snow-dominated basins translates into value, assuming a hypothetical group of decision-makers. Our analysis demonstrated that skill and value are not always aligned in a straightforward manner attributed to the inherent quality of forecasting systems in unmanaged basins. To better understand the relationship between skill and value in the unmanaged basins from true forecasts,

we compare these true forecasts with synthetic forecasts – created by imposing systematic errors on observed streamflow volumes (Fig. 4). Conversely, the true forecast systems include a process-based hydrologic model (WRF-Hydro), a deep learning model (LSTM), and operational forecasts from the Natural Resources Conservation Service (NRCS).

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, were 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). This indicates that good forecast skill does not always equate to high forecast value, especially under real-world conditions. We also found that in drought scenarios limited to extremely dry events, forecast value tends to decrease due to the heightened economic impact of these events, showing a monotonic relationship with skill (Fig. 13).

Lastly, we found that categorical measures, such as the hit and false alarm rates, better explained the discrepancies in forecast value between synthetic and true forecast systems than skill metric used in the study (Fig. 14). This was confirmed by showing the correspondence of forecast value between synthetic and true forecasts, which was largely driven by categorical measures like hit and false alarm rates (Fig. 15). 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.

However, PEV assumes risk-neutral decision-makers and is limited to binary decision contexts, which may oversimplify real-world decision-making challenges (Laugesen et al., 2023). In water management, decisions often involve continuous or multi-categorical variables, such as balancing water supply needs, hydropower generation, and flood control, which PEV does not fully capture (Laugesen et al., 2023; Portele et al., 2021). While more advanced and flexible metrics like the Relative Utility Value (RUV; Laugesen et al., 2023) offer improved decision-making capabilities by incorporating user-specific utility functions, we opted for PEV due to its simplicity and broad operational applicability. RUV provides granular insights into forecast value across different decision thresholds but introduces additional complexities that are unique to each user, including their decision-making preferences, risk tolerance, and operational priorities. RUV uses the same inputs as PEV. However, RUV allows the economic model, damage function, and risk aversion to be explicitly specified (Laugesen et al., 2023). One of the important benefits of RUV is that it uses the whole probabilistic forecast and does not need a conversion to a categorical forecast like PEV (Laugesen et al., 2023). PEV's straightforward interpretation and widespread usage in hydrologic and meteorological applications made it more suitable for our evaluation without introducing unnecessary complexities. The results from this study raise an important question about whether the categorical nature of the events and the experimental nature of PEV are indeed driving the observed outcomes. This potential alignment may suggest that categorical error measures are performing better simply because they match the structure of our experimental design. To clarify this, further consideration is needed to understand whether this relationship reflects a true advantage of categorical measures or is an artifact of the setup, i.e., a comparison with the synthetic forecasts generated by imposing regular error structures. By testing alternative error measures like RUV that are not categorical and adjusting the experimental design, we can better assess whether the effectiveness of the forecasts is truly a function of forecast skill or simply due to the structure of the experiment. Such additional analysis will help confirm or refute the notion that categorical measures work better only because they align more closely with how events and costs are defined in this model.

There are several limitations to the probabilistic forecasts used in this study. First, the datasets used for generating these forecasts typically have their own limitations, such as the absence of common standards for intercomparison, a lack of uncertainty estimates for assessing data reliability, and a lack of characterization of human intervention (Addor et al., 2020). In the case of LSTM-ESP forecasts, the use of only a single deep learning model (LSTM) is a limitation, which could be replaced by alternative neural networks (Cho et al., 2014; Vaswani et al., 2017) or physics-guided architectures (Feng et al., 2022b, a; Hoedt et al., 2021) to improve forecast performance. Additional limitations, as discussed by Modi et al. (2024), include the need to test different hyperparameters, extend the training period, and explore the use of other snowpack treatments that may improve the model's performance. For WRFH-ESP forecasts, biases in initial hydrologic conditions, which arise due to lack of knowledge and incomplete process representation (DeChant and Moradkhani, 2011), and parameter uncertainty potentially resulting from ill-constrained calibration (Arheimer et al., 2020; Hirpa et al., 2015; Wood et al., 2016) contribute to forecast biases.

We also recognize that a comparison with operational ESP forecasts generated by the River Forecast Centers might be more appropriate for this study. However, due to the limited availability of operational ESP forecasts (starting in 2015) for our study basins, as well as inconsistent methodologies across regions, we chose to use the NRCS forecasts. Importantly, it should be noted that the differences in forecast volumes between NRCS and operational ESP forecasts are minor in the context of the overall forecast uncertainty (Lukas and Payton, 2020).

Conclusions

This study explored how the skill of seasonal streamflow forecasts translates into economic value for decision-making in unmanaged basins across the western US. We used synthetic forecasts to systematically analyze the skill and value of true forecasts produced by process-based (WRFH), deep-learning (LSTM) models, and operational forecasts from NRCS. 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.

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. Our comparisons between synthetic and true forecasts revealed that forecast skill across the basins was more sensitive to errors in mean and standard deviation. However, these errors do not adequately explain the variations in forecast value. This is primarily due to the irregular model error structures, which impact categorical measures such as hit and false alarm rates, causing high forecast skill to not necessarily result in high forecast value. 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.

The analysis also reveals a clear relationship between drought severity and skill-value relationship. Models consistently struggle to predict severe drought events, and forecast value worsens monotonically with drought severity. We conclude the study by demonstrating that categorical error measures, such as the hit and false alarm rates, largely explained the forecast value. 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.

Appendix A

A1 WRFH model parameters and calibration

The WRFH has several tunable parameters associated with soil properties, the surface and subsurface routing schemes, baseflow and groundwater schemes, snow schemes and the channel configuration (Cuntz et al., 2016; Lahmers et al., 2021). We use a calibration approach associated with the NWM scheme configuration following Lahmers et al. (2021) and Cosgrove et al. (2024), that selects calibration parameters based on previous sensitivity studies (Cuntz et al., 2016; Mendoza et al., 2015), model developer surveys, and a WRF-Hydro parameter sensitivity study (further described in Lahmers et al. 2021). These parameters are distributed (distinct to each grid), and the calibration is performed on the basis of either scalar multipliers (multiplying a scalar value from the calibration range with the actual values as shown in Table 1) or simply replacing the actual values. The scalar multipliers ensure the original model parameters are spatially coherent and physically consistent with a priori catchment properties (e.g., Gupta et al., 2008, 2009) whereas the replacement ensures that parameters are constant throughout the entire domain. The model parameters tuned for this analysis are mentioned in Table 1, including the calibration range, initial values, adjustment type, parameter description, and units.

Table A1: WRFH Calibration parameters, including their calibration range, initial values, adjustment type, parameter description, and units.

Parameter	Minimum	Maximum	Initial	Type	Description	Units
Soil Parameters						
BEXP	0.4	1.9	1	Multiplier	Pore size distribution index	Dimensionless
SMCMAX	0.8	1.2	1	Multiplier	Saturation soil moisture content (i.e., porosity)	Volumetric fraction
DKSAT	0.2	10	1	Multiplier	Saturated hydraulic conductivity	m s^{-1}
RSURFEXP	1	6	5	Replace	Soil evaporation resistance exponent	Dimensionless
Runoff parameters						
REFKDT	0.1	4	1	Replace	Surface runoff parameter; REFKDT is a tuneable parameter that significantly impacts surface infiltration and hence the partitioning of total runoff into surface and subsurface runoff. Increasing REFKDT decreases surface runoff	Unitless
SLOPE	0	1	0.3	Replace	Linear scaling of “openness” of bottom drainage boundary	0-1

RETDEPRTFAC	0.1	20000	1	Replace	Multiplier on retention depth limit	Unitless
LKSATFAC	10	10000	1000	Replace	Multiplier on lateral hydraulic conductivity (controls anisotropy between vertical and lateral conductivity)	Unitless
Groundwater parameters						
ZMAX	10	250	50	Replace	Maximum groundwater bucket depth	mm
EXPON	1	8	3	Replace	Exponent controlling rate of bucket drainage as a function of depth	Dimensionless
Vegetation parameters						
CWPVT	0.5	2	1	Multiplier	Canopy wind parameter for canopy wind profile formulation	m^{-1}
VCMX25	0.6	1.4	1	Multiplier	Maximum carboxylation at 25^{O}C	$\mu molm^{-2}s^{-1}$
MP	0.6	1.4	1	Multiplier	Slope of Ball-Berry conductance relationship	Unitless
Snow parameters						
MFSNO	0.25	2	1	Multiplier	Melt factor for snow depletion curve; larger value yields a smaller snow cover fraction for the same snow height	Dimensionless

A total of 14 model parameters were calibrated with an iterative Dynamically Dimensioned Search approach (Tolson and Shoemaker, 2007). This algorithm was developed for computationally expensive optimization problems such as distributed watershed model calibration, which automatically scales the search strategy in model parameter space based on the user-specified maximum iterations (Tolson and Shoemaker, 2007). In the initial iterations, the algorithm searches globally, and as the procedure approaches the maximum number of iterations, the search transitions from a global to local search, making it computationally efficient and finds equally good solutions as compared to the dominant Shuffled Complex Evolution algorithm (Tolson and Shoemaker, 2007). In this study, the model is cycled over the calibration period 250 times to minimize an objective cost function based on the works of Cosgrove et al. (2024) and Lahmers et al. (2021). It is important to note that we restrict the iterations to 250 due to limited computing resources. However, in an ideal scenario, such as an operational context, this number could scale up to thousands of iterations, depending on the complexity of the physical processes in the region. A 5-year calibration period for each basin was selected based on the maximum standard deviation of streamflow between WY1986-2005. This ensures calibration periods are selected based on the first, basin's hydrologic conditions that are responsible for its water balance simulations, and the second, distinct climate years that allow for consideration of the broad effects of non-stationarity (Myers et al., 2021). A 5-year calibration period is short but has been adopted in earlier model implementations attributable to the limitations of computational resources (Cosgrove et al., 2024; Lahmers et al., 2021). The objective cost function is a weighted Nash Sutcliffe Efficiency (NSEwt; Equation 3) consisting of equal parts NSE (Nash and Sutcliffe, 1970) and NSE calculated for the log of the discharge ($\mathrm{NSE_{log}}$) using daily streamflow observations (Cosgrove et al., 2024; Lahmers et al., 2021).

$$NSEwt = \frac{1}{2}\left(2 - \frac{\sum_{t=1}^{T}\left(Q_{obs,t} - Q_{sim,t}\right)^2}{\sum_{t=1}^{T}\left(Q_{obs,t} - Q_{sim,t}\right)^2} - \frac{\sum_{t=1}^{T}(\log\left(Q_{obs,t}\right) - \log\left(Q_{sim,t}\right))^2}{\sum_{t=1}^{T}(\log\left(Q_{obs,t}\right) - \log\left(Q_{sim,t}\right))^2}\right) \tag{3}$$

A2 LSTM model training

The LSTM training process, as illustrated in Fig. A1, was adapted from Modi et al. (2024), who provide a more comprehensive exposition. It begins by initializing weights and biases using the Xavier uniform distribution (Glorot and Bengio, 2010). During each iteration, a random batch of 2000 samples is drawn from the training data to make predictions. The model is trained regionally, using training data from 664 basins across the CONUS from WY1983-2000. Each sample consists of a streamflow observation on a given day (the dependent variable) and the input sequence of the preceding 270 days, creating a "sequence-to-value" prediction. Since streamflow on any given day is dependent on the preceding 270 days, batches are randomly selected across basins without requiring chronological order (Kratzert et al., 2018). Static basin attributes alongside meteorological forcings are included as inputs to inform model of basin characteristics. During each iteration, the predictors (X) pass through the model's weights (w) and biases (b) to produce streamflow predictions (y_{sim}), and the error (or loss) is computed relative to the observations (y_{obs}). The model parameters are then updated through back-propagation.

To account for varying hydroclimatic conditions across basins, the training loss function is a basin average Nash Sutcliffe Efficiency (NSE), which normalizes the mean squared error for each basin using streamflow variance (Kratzert et al., 2019). This prevents large, humid basins from dominating the loss function. Unlike process-based models where parameters are updated after each complete model run, LSTM parameters are updated after each epoch – where an epoch represents one full pass of the training data. For example, if there are 100,000 training samples and a batch size of 2000, one epoch would consist of 50 iterations (100,000/2000). In this study, 40 epochs were used for training with a single seed and the Adam optimizer, which offers better efficiency than Stochastic Gradient Descent (Ruder, 2016). Multiple seeds were not tested, as the performance impact was minimal (Kratzert et al., 2019).

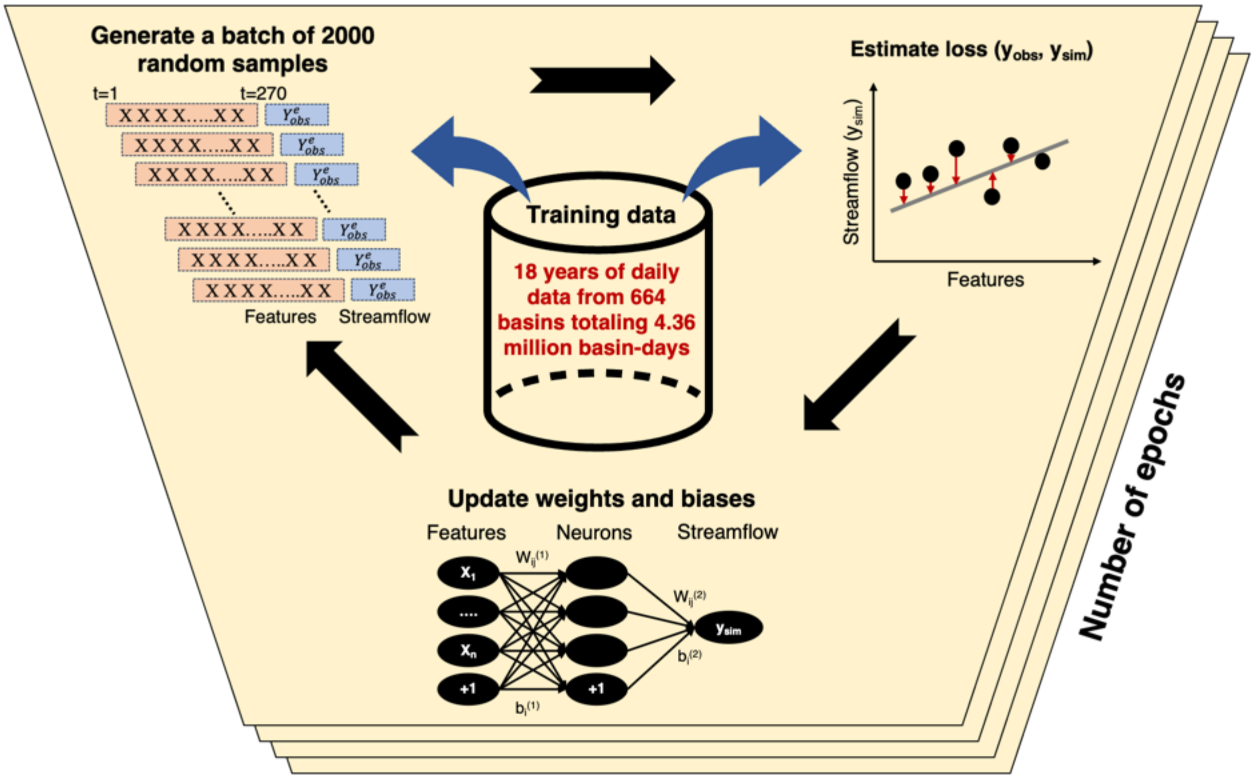


Figure A1 - Schematic of LSTM model training for each iteration within an epoch. In each iteration, 2000 independent random samples are drawn from 18 years of daily data from 664 basins totaling 4.36 million basin-days. Each sample consists of 270 days, i.e., input sequence length, of preceding predictors (X) and one target observation (y_{obs}). The loss is computed between observed discharge (y_{obs}) and the network's prediction (y_{sim}). The model parameters, including weights (w_1...w_m) and biases (b_1…b_m), are updated after every iteration. Epoch refers to the complete passing of the entire training dataset through the model algorithm once. The weights and biases are model parameters, whereas the batch size, input sequence length, and number of epochs are the hyperparameters (Modi et al., 2024).

Table A2 - Training predictors for LSTM models. It consists of meteorological forcings (source: AORC), static basin attributes (source: GAGES-II), and snow data (source: UA) with streamflow data (source: USGS) as the predictand. The asterisk indicates that the predictor was only included in one of the two trained LSTM models.

CATEGORY	NAME	DESCRIPTION
Static	PPTAVG_BASIN	Mean annual precipitation (mm)
	PET	Mean annual potential evapotranspiration (mm)
	T_AVG_BASIN	Average annual air temperature (°C)
	SNOW_PCT_PRECIP	Snow percent of total precipitation estimate
	WDMAX_BASIN	Watershed average of monthly max. number of days of measurable precipitation
	WDMIN_BASIN	Watershed average of monthly min. number of days of measurable precipitation

	PRECIP_SEAS_IND	Precipitation seasonality index (Markham, 1970; Dingman, 2002). Index of how much annual precipitation falls seasonally (high values) or spread out over the year (low values).
	RUNAVE7100	Mean annual total runoff (mm)
	RE	Runoff efficiency = PPTAVG_BASIN/RUNAVE7100
	ELEV_MAX_BASIN	Maximum watershed elevation (m)
	ELEV_MIN_BASIN	Minimum watershed elevation (m)
	DRAIN_SQKM	Watershed drainage area (km^2)
	SLOPE_PCT	Mean watershed slope (%)
	FORESTNLCD06	Watershed percent forest (%)
	PLANTNLCD06	Watershed percent planted/cultivated
	PNV_BAS_PCT	Percentage of the watershed covered by the dominant potential natural vegetation
	ROCKDEPAVE	Average value of total soil thickness examined (in)
	AWCAVE	Average value for the range of available water capacity for the soil layer
	CLAYAVE	Average value of clay content (%)
	SILTAVE	Average value of silt content (%)
	SANDAVE	Average value of sand content (%)
	PERMAVE	Average permeability (in/hr)
	KFACT_UP	Average K-factor for the uppermost soil horizon in each soil component. K-factor is an erodibility factor which quantifies the susceptibility of soil particles to detachment and movement by water.
Meteorological	PRCP	Average daily precipitation (mm/day)
	WIND	Average wind speed (m/s)
	TAS	2 m daily average air temperature (^{o}C)
	SRAD	Incoming shortwave solar radiation (W/m^2)
	LRAD	Incoming longwave solar radiation (W/m^2)
	PRES	Near-Surface Air pressure (Pa)
	VP	Near-Surface Vapor Pressure (Pa)
*Snow	SWE	Average Snow Water Equivalent (mm)
Streamflow	SF	Average daily streamflow (mm/day)

A3 Historical performance evaluation of our designed true forecast systems before and after calibration/training.

As shown in Fig. A3, for WRFH, the improvements were minimal across most metrics before (WRFH$_{DEF}$) and after calibration (WRFH$_{CAL}$), except for the variability (ratio of standard deviation) that improved from 1.65 to 1.25. With LSTM, major improvements were seen with the median daily NSE, improving from 0.58 to 0.77. In general, the improvements across all metrics for both models underscore the importance of model calibration and training, as seen with LSTM$_{FINAL}$ and WRFH$_{CAL}$ (Fig. A3).

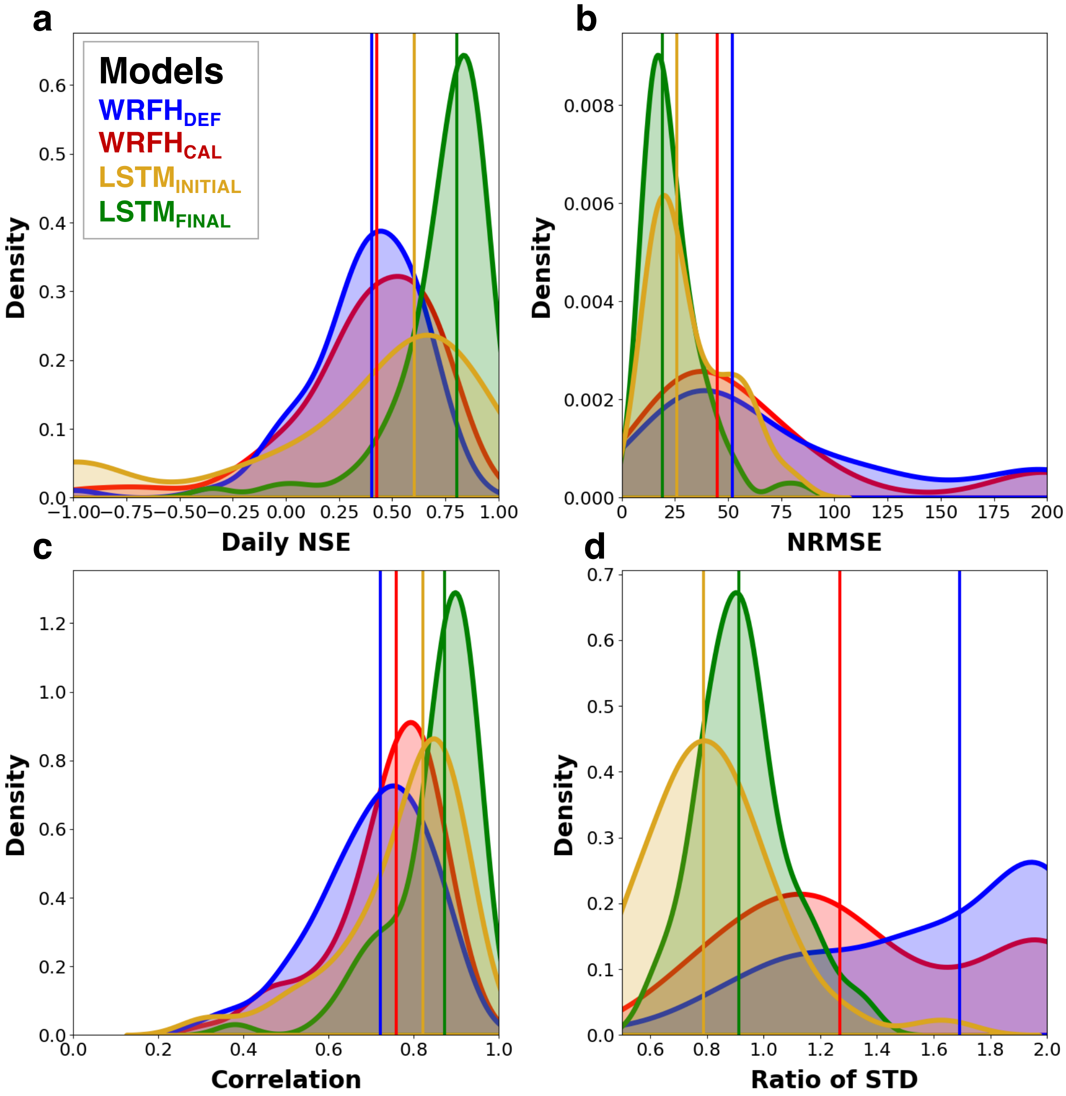


Figure A2: Historical model performance of true forecast systems. (a) Daily NSE, (b) NRMSE of the total April-July streamflow volumes, (c) daily correlation, and (d) Ratio of the standard deviation against observations for WRFH (default and calibrated) and LSTM (initial and final) models. Comparison shown for the 76 basins during the testing period, WY2001-2010.

A4 Exposition of irregular error structures in true forecasts

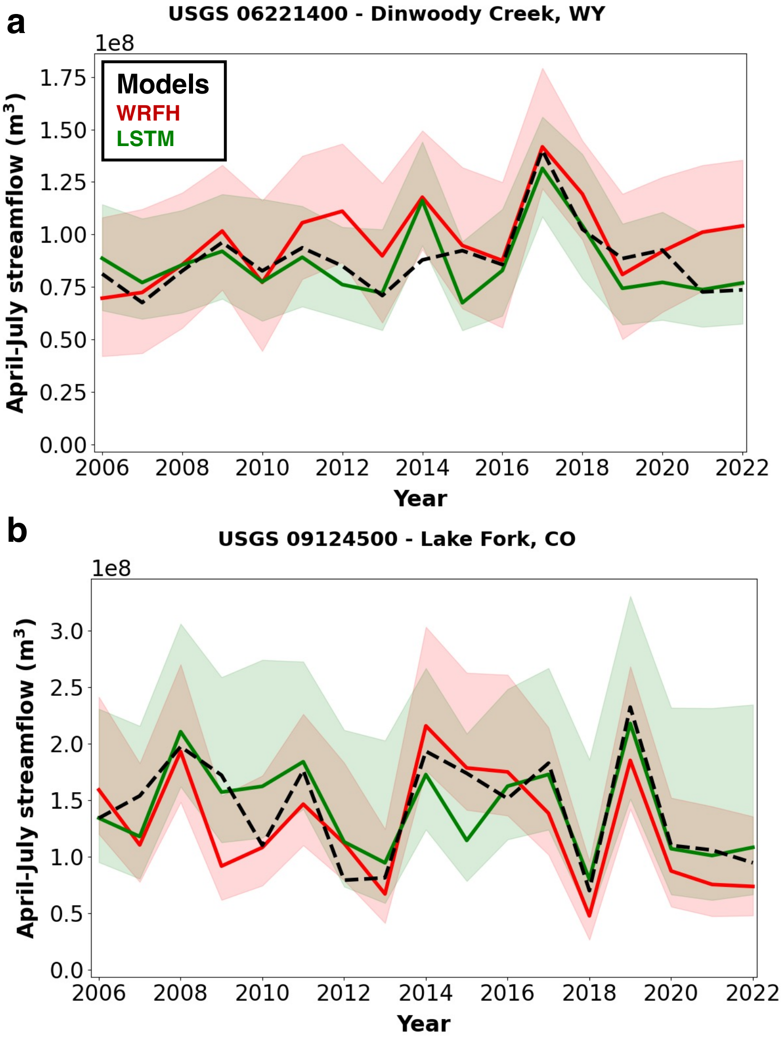


Figure A3: April-July streamflow volume from two true forecast systems (WRFH and LSTM) in WY2006-2022 at Dinwoody Creek (USGS 06221400) and Lake Fork (USGS 09124500).

Code and data availability

All data products used in the analysis are publicly available. A total of 664 GAGES-II basins are selected following screening criteria that ensure minimal upstream regulation and continuous data availability for at least 30 years. The meteorological forcings, basin attributes, snow and streamflow data are obtained from AORC (Fall et al., 2023), GAGES-II (U.S. Geological Survey, 2023), UA (Broxton et al., 2019a) and the US Geological Survey streamflow gages (United States Geological Survey, 2024) respectively. NRCS forecast data and SNOTEL snowpack observations are downloaded from the National Water and Climate Center portal (United States et al., 2024). The Modi & Livneh (2024) data set provides the source code, training data, and model runs for the LSTM model used in this research. The code for the WRF-hydro model (V5.2) is available online (McCreight et al., 2021).

Author contributions

Conceptualization: PM and BL; methodology, validation, formal analysis, investigation, and writing (review and editing): all authors; software: PM and BL; data curation: PM and BL; writing (original draft preparation): PM and BL; visualization: PM and BL; supervision, project administration, and funding acquisition: all authors. All authors have read and agreed to the published version of the paper.

Competing interests

The contact author has declared that none of the authors has any competing interests.

Acknowledgments

We acknowledge funding support from the NOAA grant # NA20OAR4310420 Identifying Alternatives to Snow-based Streamflow Predictions to Advance Future Drought Predictability and NSF grant BCS # 2009922 Water-Mediated Coupling of Natural-Human Systems: Drought and Water Allocation Across Spatial Scales. We gratefully acknowledge constructive and insightful suggestions from two anonymous reviewers, which improved the clarity of the original manuscript. We are extremely grateful for all the computing resources provided by the University of Colorado Boulder Research Computing.

References

Abaza, M., Anctil, F., Fortin, V., and Turcotte, R.: A Comparison of the Canadian Global and Regional Meteorological Ensemble Prediction Systems for Short-Term Hydrological Forecasting, Monthly Weather Review, 141, 3462–3476, https://doi.org/10.1175/MWR-D-12-00206.1, 2013.

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, Hydrol. Earth Syst. Sci., 21, 5293–5313, https://doi.org/10.5194/hess-21-5293-2017, 2017.

Addor, N., Do, H. X., Alvarez-Garreton, C., Coxon, G., Fowler, K., and Mendoza, P. A.: Large-sample hydrology: recent progress, guidelines for new datasets and grand challenges, Hydrological Sciences Journal, 65, 712–725, https://doi.org/10.1080/02626667.2019.1683182, 2020.

Amlung, M., Yurasek, A., McCarty, K. N., MacKillop, J., and Murphy, J. G.: Area under the curve as a novel metric of behavioral economic demand for alcohol., Experimental and Clinical Psychopharmacology, 23, 168–175, https://doi.org/10.1037/pha0000014, 2015.

Arheimer, B., Pimentel, R., Isberg, K., Crochemore, L., Andersson, J. C. M., Hasan, A., and Pineda, L.: Global catchment modelling using World-Wide HYPE (WWH), open data, and stepwise parameter estimation, Hydrol. Earth Syst. Sci., 24, 535–559, https://doi.org/10.5194/hess-24-535-2020, 2020.

Arsenault, R., Martel, J.-L., Brunet, F., Brissette, F., and Mai, J.: Continuous streamflow prediction in ungauged basins: Long Short-Term Memory Neural Networks clearly outperform hydrological models, Hydrometeorology/Modelling approaches, https://doi.org/10.5194/hess-2022-295, 2022.

Broxton, P., Zeng, X., and Dawson, N.: Daily 4 km Gridded SWE and Snow Depth from Assimilated In-Situ and Modeled Data over the Conterminous US, Version 1, https://doi.org/10.5067/0GGPB220EX6A, 2019a.

Broxton, P. D., Leeuwen, W. J. D., and Biederman, J. A.: Improving Snow Water Equivalent Maps With Machine Learning of Snow Survey and Lidar Measurements, Water Resour. Res., 55, 3739–3757, https://doi.org/10.1029/2018WR024146, 2019b.

Chaney, N. W., Herman, J. D., Reed, P. M., and Wood, E. F.: Flood and drought hydrologic monitoring: the role of model parameter uncertainty, Hydrol. Earth Syst. Sci., 19, 3239–3251, https://doi.org/10.5194/hess-19-3239-2015, 2015.

Cho, K., van Merrienboer, B., Bahdanau, D., and Bengio, Y.: On the Properties of Neural Machine Translation: Encoder-Decoder Approaches, https://doi.org/10.48550/ARXIV.1409.1259, 2014.

Clark, M. P., Schaefli, B., Schymanski, S. J., Samaniego, L., Luce, C. H., Jackson, B. M., Freer, J. E., Arnold, J. R., Moore, R. D., Istanbulluoglu, E., and Ceola, S.: Improving the theoretical underpinnings of process-based hydrologic models, Water Resources Research, 52, 2350–2365, https://doi.org/10.1002/2015WR017910, 2016.

Cosgrove, B., Gochis, D., Flowers, T., Dugger, A., Ogden, F., Graziano, T., Clark, E., Cabell, R., Casiday, N., Cui, Z., Eicher, K., Fall, G., Feng, X., Fitzgerald, K., Frazier, N., George, C., Gibbs, R., Hernandez, L., Johnson, D., Jones, R., Karsten, L., Kefelegn, H., Kitzmiller, D., Lee, H., Liu, Y., Mashriqui, H., Mattern, D., McCluskey, A., McCreight, J. L., McDaniel, R., Midekisa, A., Newman, A., Pan, L., Pham, C., RafieeiNasab, A., Rasmussen, R., Read, L., Rezaeianzadeh, M., Salas, F., Sang, D., Sampson, K., Schneider, T., Shi, Q., Sood, G., Wood, A., Wu, W., Yates, D., Yu, W., and Zhang, Y.: NOAA's National Water Model: Advancing operational hydrology through continental-scale modeling, J American Water Resour Assoc, 60, 247–272, https://doi.org/10.1111/1752-1688.13184, 2024.

Crochemore, L., Ramos, M.-H., Pappenberger, F., Andel, S. J. V., and Wood, A. W.: An Experiment on Risk-Based Decision-Making in Water Management Using Monthly Probabilistic Forecasts, Bulletin of the American Meteorological Society, 97, 541–551, https://doi.org/10.1175/BAMS-D-14-00270.1, 2016.

Cuntz, M., Mai, J., Samaniego, L., Clark, M., Wulfmeyer, V., Branch, O., Attinger, S., and Thober, S.: The impact of standard and hard-coded parameters on the hydrologic fluxes in the Noah-MP land surface model: HARD-CODED PARAMETERS IN NOAH-MP, J. Geophys. Res. Atmos., 121, 10,676-10,700, https://doi.org/10.1002/2016JD025097, 2016.

Day, G. N.: Extended Streamflow Forecasting Using NWSRFS, Journal of Water Resources Planning and Management, 111, 157–170, https://doi.org/10.1061/(ASCE)0733-9496(1985)111:2(157), 1985.

DeChant, C. M. and Moradkhani, H.: Improving the characterization of initial condition for ensemble streamflow prediction using data assimilation, Hydrol. Earth Syst. Sci., 15, 3399–3410, https://doi.org/10.5194/hess-15-3399-2011, 2011.

Demargne, J., Wu, L., Regonda, S. K., Brown, J. D., Lee, H., He, M., Seo, D.-J., Hartman, R., Herr, H. D., Fresch, M., Schaake, J., and Zhu, Y.: The Science of NOAA's Operational Hydrologic Ensemble Forecast Service, Bulletin of the American Meteorological Society, 95, 79–98, https://doi.org/10.1175/BAMS-D-12-00081.1, 2014.

Water Supply Forecast Rodeo: Forecast Stage: https://www.drivendata.org/competitions/259/reclamation-water-supply-forecast/, last access: 9 April 2024.

Ek, M. B., Mitchell, K. E., Lin, Y., Rogers, E., Grunmann, P., Koren, V., Gayno, G., and Tarpley, J. D.: Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model, J. Geophys. Res., 108, 8851, https://doi.org/10.1029/2002JD003296, 2003.

GAGES-II: geospatial attributes of gages for evaluating streamflow.: https://water.usgs.gov/GIS/metadata/usgswrd/XML/gagesII_Sept2011.xml, last access: 15 April 2021.

Falcone, J. A., Carlisle, D. M., Wolock, D. M., and Meador, M. R.: GAGES: A stream gage database for evaluating natural and altered flow conditions in the conterminous United States, Ecology, 91, 621–621, https://doi.org/10.1890/09-0889.1, 2010.

Fall, G., Kitzmiller, D., Pavlovic, S., Zhang, Z., Patrick, N., St. Laurent, M., Trypaluk, C., Wu, W., and Miller, D.: The Office of Water Prediction's Analysis of Record for Calibration, version 1.1: Dataset description and precipitation evaluation, J American Water Resour Assoc, 59, 1246–1272, https://doi.org/10.1111/1752-1688.13143, 2023.

Feng, D., Liu, J., Lawson, K., and Shen, C.: Differentiable, Learnable, Regionalized Process-Based Models With Multiphysical Outputs can Approach State-Of-The-Art Hydrologic Prediction Accuracy, Water Resources Research, 58, e2022WR032404, https://doi.org/10.1029/2022WR032404, 2022a.

Feng, D., Beck, H., Lawson, K., and Shen, C.: The suitability of differentiable, learnable hydrologic models for ungauged regions and climate change impact assessment, Catchment hydrology/Modelling approaches, https://doi.org/10.5194/hess-2022-245, 2022b.

Ficchì, A., Raso, L., Dorchies, D., Pianosi, F., Malaterre, P.-O., Van Overloop, P.-J., and Jay-Allemand, M.: Optimal Operation of the Multireservoir System in the Seine River Basin Using Deterministic and Ensemble Forecasts, J. Water Resour. Plann. Manage., 142, 05015005, https://doi.org/10.1061/(ASCE)WR.1943-5452.0000571, 2016.

Fleming, S. W., Garen, D. C., Goodbody, A. G., McCarthy, C. S., and Landers, L. C.: Assessing the new Natural Resources Conservation Service water supply forecast model for the American West: A challenging test of explainable, automated, ensemble artificial intelligence, Journal of Hydrology, 602, 126782, https://doi.org/10.1016/j.jhydrol.2021.126782, 2021.

Garen, D. C.: Improved Techniques in Regression-Based Streamflow Volume Forecasting, Journal of Water Resources Planning and Management, 118, 654–670, https://doi.org/10.1061/(ASCE)0733-9496(1992)118:6(654), 1992.

Glorot, X. and Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, 249–256, 2010.

Gochis, D. J., Barlage, M., Cabell, R., Casali, M., Dugger, A., FitzGerald, K., McAllister, M., McCreight, J., RafieeiNasab, A., Read, L., Sampson, K., Yates, D., and Zhang, Y.: The WRF-Hydro® modeling system technical description, (Version 5.1.1)., NCAR Technical Note, 108, 2020.

Gupta, H. V., Wagener, T., and Liu, Y.: Reconciling theory with observations: elements of a diagnostic approach to model evaluation, Hydrol. Process., 22, 3802–3813, https://doi.org/10.1002/hyp.6989, 2008.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of Hydrology, 377, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Hamlet, A. F., Huppert, D., and Lettenmaier, D. P.: Economic Value of Long-Lead Streamflow Forecasts for Columbia River Hydropower, J. Water Resour. Plann. Manage., 128, 91–101, https://doi.org/10.1061/(ASCE)0733-9496(2002)128:2(91), 2002.

Hirpa, F. A., Fagbemi, K., Afiesimam, E., Shuaib, H., and Salamon, P.: Saving Lives: Ensemble-Based Early Warnings in Developing Nations, in: Handbook of Hydrometeorological Ensemble Forecasting, edited by: Duan, Q., Pappenberger, F., Thielen, J., Wood, A., Cloke, H. L., and Schaake, J. C., Springer Berlin Heidelberg, Berlin, Heidelberg, 1–22, https://doi.org/10.1007/978-3-642-40457-3_43-1, 2015.

Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, Neural Computation, 9, 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735, 1997.

Hoedt, P.-J., Kratzert, F., Klotz, D., Halmich, C., Holzleitner, M., Nearing, G., Hochreiter, S., and Klambauer, G.: MC-LSTM: Mass-Conserving LSTM, https://doi.org/10.48550/arXiv.2101.05186, 10 June 2021.

Jolliffe, I. T. and Stephenson, D.: Forecast verification: a practitioner's guide in atmospheric science, J. Wiley, Chichester, 2003.

Kaune, A., Chowdhury, F., Werner, M., and Bennett, J.: The benefit of using an ensemble of seasonal streamflow forecasts in water allocation decisions, Hydrol. Earth Syst. Sci., 24, 3851–3870, https://doi.org/10.5194/hess-24-3851-2020, 2020.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M.: Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks, Hydrol. Earth Syst. Sci., 22, 6005–6022, https://doi.org/10.5194/hess-22-6005-2018, 2018.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S.: Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning, Water Resour. Res., 55, 11344–11354, https://doi.org/10.1029/2019WR026065, 2019.

Lahmers, T. M., Hazenberg, P., Gupta, H., Castro, C., Gochis, D., Dugger, A., Yates, D., Read, L., Karsten, L., and Wang, Y.-H.: Evaluation of NOAA National Water Model Parameter Calibration in Semi-Arid Environments Prone to Channel Infiltration, Journal of Hydrometeorology, https://doi.org/10.1175/JHM-D-20-0198.1, 2021.

Laugesen, R., Thyer, M., McInerney, D., and Kavetski, D.: Flexible forecast value metric suitable for a wide range of decisions: application using probabilistic subseasonal streamflow forecasts, Hydrol. Earth Syst. Sci., 27, 873–893, https://doi.org/10.5194/hess-27-873-2023, 2023.

Lehner, B., Verdin, K., and Jarvis, A.: New Global Hydrography Derived From Spaceborne Elevation Data, Eos Trans. AGU, 89, 93, https://doi.org/10.1029/2008EO100001, 2008.

Lehner, F., Wood, A. W., Llewellyn, D., Blatchford, D. B., Goodbody, A. G., and Pappenberger, F.: Mitigating the Impacts of Climate Nonstationarity on Seasonal Streamflow Predictability in the U.S. Southwest, Geophysical Research Letters, 44, 12,208-12,217, https://doi.org/10.1002/2017GL076043, 2017.

Li, D., Wrzesien, M. L., Durand, M., Adam, J., and Lettenmaier, D. P.: How much runoff originates as snow in the western United States, and how will that change in the future?: Western U.S. Snowmelt-Derived Runoff, Geophys. Res. Lett., 44, 6163–6172, https://doi.org/10.1002/2017GL073551, 2017.

Livneh, B. and Badger, A. M.: Drought less predictable under declining future snowpack, Nat. Clim. Chang., 10, 452–458, https://doi.org/10.1038/s41558-020-0754-8, 2020.

Lukas, J. and Payton, E.: Colorado River Basin Climate and Hydrology: State of the Science, https://doi.org/10.25810/3HCV-W477, 2020.

Maurer, E. P. and Lettenmaier, D. P.: Potential Effects of Long-Lead Hydrologic Predictability on Missouri River Main-Stem Reservoirs*, J. Climate, 17, 174–186, https://doi.org/10.1175/1520-0442(2004)017<0174:PEOLHP>2.0.CO;2, 2004.

McCreight, J., FitzGerald, K., Cabell, R., Fersch, B., Donaldwj, Dugger, A., Laurareads, Dan, McAllister, M., Logankarsten, Dunlap, R., Nels, Fanfarillo, A., Arezoorn, Mattern, D., Barlage, M., Champham, Wrf-Hydro-Nwm-Bot, Prasanth Valayamkunnath, TimLahmers, and Aheldmyer: NCAR/wrf_hydro_nwm_public: WRF-Hydro® v5.2.0, , https://doi.org/10.5281/ZENODO.4479912, 2021.

Mendoza, P. A., Clark, M. P., Barlage, M., Rajagopalan, B., Samaniego, L., Abramowitz, G., and Gupta, H.: Are we unnecessarily constraining the agility of complex process-based models?, Water Resour. Res., 51, 716–728, https://doi.org/10.1002/2014WR015820, 2015.

Modi, P. and Livneh, B.: Long Short Term Memory simulations and code for 664 basins in the Ensemble Streamflow Prediction framework (LSTM-ESP), , https://doi.org/10.5281/ZENODO.14213154, 2024.

Modi, P. A., Small, E. S., Kasprzyk, J., and Livneh, B.: Investigating the role of snow water equivalent on streamflow predictability during drought (In review), Journal of Hydrometeorology, 2021.

Modi, P. A., Jennings, K. S., Kasprzyk, J. R., Small, E. E., Wobus, C. W., and Livneh, B.: Using Deep Learning in Ensemble Streamflow Forecasting: Exploring the Predictive Value of Explicit Snowpack Information, https://doi.org/10.22541/essoar.172222576.62134567/v1, 29 July 2024.

Mosavi, A., Ozturk, P., and Chau, K.: Flood Prediction Using Machine Learning Models: Literature Review, Water, 10, 1536, https://doi.org/10.3390/w10111536, 2018.

Murphy, A. H.: What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting, 1993.

Myers, D. T., Ficklin, D. L., Robeson, S. M., Neupane, R. P., Botero-Acosta, A., and Avellaneda, P. M.: Choosing an arbitrary calibration period for hydrologic models: How much does it influence water balance simulations?, Hydrological Processes, 35, https://doi.org/10.1002/hyp.14045, 2021.

Myneni, R., Knyazikhin, Y., and Park, T.: MOD15A2H MODIS/Terra Leaf Area Index/FPAR 8-Day L4 Global 500m SIN Grid V006, https://doi.org/10.5067/MODIS/MOD15A2H.006, 2015.

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T., and Duan, Q.: Development of a large-sample watershed-scale hydrometeorological dataset for the contiguous USA: dataset characteristics and assessment of regional variability in hydrologic model performance, Catchment hydrology/Modelling approaches, https://doi.org/10.5194/hessd-11-5599-2014, 2014.

Niu, G.-Y., Yang, Z.-L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., Kumar, A., Manning, K., Niyogi, D., Rosero, E., Tewari, M., and Xia, Y.: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements, J. Geophys. Res., 116, D12109, https://doi.org/10.1029/2010JD015139, 2011.

Pagano, T., Garen, D., and Sorooshian, S.: Evaluation of Official Western U.S. Seasonal Water Supply Outlooks, 1922–2002, Journal of Hydrometeorology, 5, 896–909, https://doi.org/10.1175/1525-7541(2004)005<0896:EOOWUS>2.0.CO;2, 2004.

Pagano, T., Wood, A., Werner, K., and Tama-Sweet, R.: Western U.S. Water Supply Forecasting: A Tradition Evolves, Eos Trans. AGU, 95, 28–29, https://doi.org/10.1002/2014EO030007, 2014.

Portele, T. C., Lorenz, C., Dibrani, B., Laux, P., Bliefernicht, J., and Kunstmann, H.: Seasonal forecasts offer economic benefit for hydrological decision making in semi-arid regions, Sci Rep, 11, 10581, https://doi.org/10.1038/s41598-021-89564-y, 2021.

Rheinheimer, D. E., Bales, R. C., Oroza, C. A., Lund, J. R., and Viers, J. H.: Valuing year-to-go hydrologic forecast improvements for a peaking hydropower system in the Sierra Nevada: VALUING HYDROLOGIC FORECASTS FOR HYDROPOWER, Water Resour. Res., 52, 3815–3828, https://doi.org/10.1002/2015WR018295, 2016.

Richardson, D. S.: Skill and relative economic value of the ECMWF ensemble prediction system, Quart J Royal Meteoro Soc, 126, 649–667, https://doi.org/10.1002/qj.49712656313, 2000.

Ruder, S.: An overview of gradient descent optimization algorithms, https://doi.org/10.48550/ARXIV.1609.04747, 2016.

Slack, J. R. and Landwehr, J. M.: Hydro-climatic data network: a US Geological Survey streamflow data set for the United States for the study of climate variations, 1874–1988. USGS Open-File Report 92-129, US Geological Survey, 1992.

Svoboda, M., LeComte, D., Hayes, M., Heim, R., Gleason, K., Angel, J., Rippey, B., Tinker, R., Palecki, M., Stooksbury, D., Miskus, D., and Stephens, S.: THE DROUGHT MONITOR, Bull. Amer. Meteor. Soc., 83, 1181–1190, https://doi.org/10.1175/1520-0477-83.8.1181, 2002.

Thiboult, A., Anctil, F., and Ramos, M. H.: How does the quantification of uncertainties affect the quality and value of flood early warning systems?, Journal of Hydrology, 551, 365–373, https://doi.org/10.1016/j.jhydrol.2017.05.014, 2017.

Tolson, B. A. and Shoemaker, C. A.: Dynamically dimensioned search algorithm for computationally efficient watershed model calibration: DYNAMICALLY DIMENSIONED SEARCH ALGORITHM, Water Resour. Res., 43, https://doi.org/10.1029/2005WR004723, 2007.

Troin, M., Arsenault, R., Wood, A. W., Brissette, F., and Martel, J.-L.: Generating ensemble streamflow forecasts:A review of methods and approaches over the past 40 years, Water Resources Research, n/a, e2020WR028392, https://doi.org/10.1029/2020WR028392, 2021.

Turner, S. W. D., Bennett, J. C., Robertson, D. E., and Galelli, S.: Complex relationship between seasonal streamflow forecast skill and value in reservoir operations, Hydrol. Earth Syst. Sci., 21, 4841–4859, https://doi.org/10.5194/hess-21-4841-2017, 2017.

https://doi.org/10.5194/egusphere-2024-4046
Preprint. Discussion started: 14 January 2025

WRF Preprocessing System (WPS) Geographical Static Data: https://www2.mmm.ucar.edu/wrf/users/download/get_sources_wps_geog.html, last access: 27 July 2019.

United States, US Department of Agriculture, National Resource Conservation Service, and National Water and Climate Center: Air and Water Database., 2024.

United States Geological Survey: USGS Water Data for the Nation, https://doi.org/10.5066/F7P55KJN, 2024.

U.S. Geological Survey: GAGES-II: Geospatial Attributes of Gages for Evaluating Streamflow, https://doi.org/10.5066/P96CPHOT, 2023.

CropScape - NASS CDL Program: https://nassgeodata.gmu.edu/CropScape/, last access: 15 December 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I.: Attention Is All You Need, https://doi.org/10.48550/ARXIV.1706.03762, 2017.

Verkade, J. S., Brown, J. D., Davids, F., Reggiani, P., and Weerts, A. H.: Estimating predictive hydrological uncertainty by dressing deterministic and ensemble forecasts; a comparison, with application to Meuse and Rhine, Journal of Hydrology, 555, 257–277, https://doi.org/10.1016/j.jhydrol.2017.10.024, 2017.

Watts, G., Christierson, B. V., Hannaford, J., and Lonsdale, K.: Testing the resilience of water supply systems to long droughts, Journal of Hydrology, 414–415, 255–267, https://doi.org/10.1016/j.jhydrol.2011.10.038, 2012.

Wilks, D. S.: A skill score based on economic value for probability forecasts, Meteorological Applications, 8, 209–219, https://doi.org/10.1017/S1350482701002092, 2001.

Wood, A. W. and Lettenmaier, D. P.: A Test Bed for New Seasonal Hydrologic Forecasting Approaches in the Western United States, Bull. Amer. Meteor. Soc., 87, 1699–1712, https://doi.org/10.1175/BAMS-87-12-1699, 2006.

Wood, A. W., Hopson, T., Newman, A., Brekke, L., Arnold, J., and Clark, M.: Quantifying Streamflow Forecast Skill Elasticity to Initial Condition and Climate Prediction Skill, Journal of Hydrometeorology, 17, 651–668, https://doi.org/10.1175/JHM-D-14-0213.1, 2016.

Wood, E. F., Schubert, S. D., Wood, A. W., Peters-Lidard, C. D., Mo, K. C., Mariotti, A., and Pulwarty, R. S.: Prospects for Advancing Drought Understanding, Monitoring, and Prediction, Journal of Hydrometeorology, 16, 1636–1657, https://doi.org/10.1175/JHM-D-14-0164.1, 2015.

Zeng, X., Broxton, P., and Dawson, N.: Snowpack Change From 1982 to 2016 Over Conterminous United States, Geophysical Research Letters, 45, https://doi.org/10.1029/2018GL079621, 2018.