Water Resources Research

RESEARCH ARTICLE
10.1029/2024WR039324

**Special Collection:**
A Continuing Special Collection for Translational Water Research: Looking for Theory-Practice Synthesis for Actionable Outcomes

# Synthetic Ensemble Forecasts: Operations-Based Evaluation and Inter-Model Comparison for Reservoir Systems Across California

**Zachary P. Brodeur[1], William Taylor[2], Jonathan D. Herman[2], and Scott Steinschneider[1]**

[1]Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY, USA, [2]Department of Civil and Environmental Engineering, University of California, Davis, CA, USA

**Key Points:**

- We develop and evaluate the reservoir operations-based performance of established and novel synthetic hydrologic ensemble forecast methods
- Synthetic forecasts are evaluated via ensemble verification techniques applied to reservoir storage and release series
- We assess the spatial generalizability of synthetic forecast approaches through a multisite numerical experiment

**Supporting Information:**

Supporting Information may be found in the online version of this article.

**Correspondence to:**

Z. P. Brodeur,
zpb4@cornell.edu

**Citation:**

Brodeur, Z. P., Taylor, W., Herman, J. D., & Steinschneider, S. (2025). Synthetic ensemble forecasts: Operations-based evaluation and inter-model comparison for reservoir systems across California. *Water Resources Research*, *61*, e2024WR039324. https://doi.org/10.1029/2024WR039324

Received 30 OCT 2024
Accepted 27 MAY 2025

**Author Contributions:**

**Conceptualization:** Zachary P. Brodeur, William Taylor, Jonathan D. Herman, Scott Steinschneider
**Data curation:** Zachary P. Brodeur
**Formal analysis:** Zachary P. Brodeur, William Taylor, Jonathan D. Herman, Scott Steinschneider

© 2025. The Author(s).
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

## Abstract

Synthetic ensemble forecasts are an important tool for testing the robustness of forecast-informed reservoir operations (FIRO). These forecasts are statistically generated to mimic the skill of hindcasts derived from operational ensemble forecasting systems, but they can be created for time periods when hindcast data are unavailable, allowing for a more comprehensive evaluation of FIRO policies. Nevertheless, it remains unclear how to determine whether a candidate synthetic ensemble forecasting approach is sufficiently representative of its real-world counterpart to support FIRO policy evaluation. This highlights a need for formal *fit-for-purpose* validation frameworks to advance synthetic forecasting as a generalizable risk analysis strategy. We address this research gap by first introducing a novel operations-based validation framework, where reservoir storage and release simulations under a FIRO policy are compared when forced with a single ensemble hindcast and many different synthetic ensembles. We evaluate the suitability of synthetic forecasts based on formal probabilistic verification of the operational outcomes. Second, we develop a new synthetic ensemble forecasting algorithm and compare it to a previous algorithm using this validation framework across a set of stylized, hydrologically diverse reservoir systems in California. Results reveal clear differences in operational suitability, with the new method consistently outperforming the previous one. These findings demonstrate the promise of the newer synthetic forecasting approach as a generalizable tool for FIRO policy evaluation and robustness testing. They also underscore the value of the proposed validation framework in benchmarking and guiding future improvements in synthetic forecast development.

## 1. Introduction

Forecast-informed reservoir operations (FIRO) has emerged as a promising approach for managing large dams in the United States and globally (Jasperse et al., 2020; Nohara et al., 2020; Porter et al., 2018; Ralph et al., 2022, 2023; Woodside et al., 2022). FIRO is defined by the American Meteorological Society (AMS) as “a reservoir-operations strategy that better informs decisions to retain or release water by integrating additional flexibility in operation policies and rules with enhanced monitoring and improved weather and hydrological forecasts” (AMS, 2025). The strategy has the potential to enhance flood risk reduction while also improving water supply, ecosystem, hydropower, and other objectives, as has been long recognized in academic research (e.g., Faber & Stedinger, 2001; Giuliani et al., 2019; Hejazi et al., 2014). Recent implementations of FIRO have leveraged advances in forecast accuracy (Badrinath et al., 2023; Lavers et al., 2020; Sukovich et al., 2014), operational hydrologic ensemble forecasting systems (Abbaszadeh et al., 2020; Demargne et al., 2014; Troin et al., 2021), and a growing socio-political and institutional focus on climate resilience to foster robust and adaptative water system management frameworks (Blum & Miller, 2019; Brown et al., 2015; Di Baldassarre et al., 2019; Thaler, 2021). Nevertheless, a critical challenge remains when trying to promote FIRO adoption among practitioners and institutions that prioritize flood risk management: How can we ensure that FIRO strategies are robust across a broad spectrum of possible flood events, forecasts, and patterns of forecast error? Or more precisely, how can we demonstrate that FIRO strategies consistently maintain or improve reservoir performance under a broad range of forecast scenarios—including extreme events and substantial forecast errors—relative to status quo operations?

To address this challenge, recent work has advanced synthetic forecasts as a key tool for testing the robustness of FIRO strategies (Brodeur & Steinschneider, 2021; Brodeur et al., 2024). Synthetic forecasts are generated using statistical methods to produce many plausible forecast sequences that emulate the forecast attributes of the modeled system, such as skill and uncertainty across lead times, but rely solely on observed streamflow or climate as input. As a result, they can generate forecasts for any event where observations are available, including those

**Funding acquisition:** Scott Steinschneider
**Investigation:** Zachary P. Brodeur
**Methodology:** Zachary P. Brodeur, William Taylor, Jonathan D. Herman, Scott Steinschneider
**Project administration:** Scott Steinschneider
**Resources:** Jonathan D. Herman, Scott Steinschneider
**Software:** Zachary P. Brodeur, William Taylor, Jonathan D. Herman
**Supervision:** Scott Steinschneider
**Validation:** Zachary P. Brodeur, Scott Steinschneider
**Visualization:** Zachary P. Brodeur, Scott Steinschneider
**Writing – original draft:** Zachary P. Brodeur
**Writing – review & editing:** Zachary P. Brodeur, William Taylor, Jonathan D. Herman, Scott Steinschneider

drawn from extended historical records, stochastic weather or streamflow generators, or climate change projections. In contrast, current hindcasting capabilities typically yield a single ensemble forecast sequence spanning only the past 30–40 years (Guan et al., 2022). Synthetic forecasts overcome this limitation by enabling plausible forecast generation across a much broader and more diverse set of hydrologic scenarios, including extreme and low-probability events. This expanded capability supports more comprehensive stress-testing of FIRO policies and aligns with feedback from practitioners, who emphasize the need to assess policy performance under extreme and uncertain conditions.

Although the concept of synthetic forecasts is not new (Grygier et al., 1989; Lamontagne & Stedinger, 2018; Lettenmaier, 1984), recent FIRO efforts have highlighted the need for methods that can emulate forecast behavior across multiple sites and lead times at daily or sub-daily timescales. Further, there is a need for synthetic forecasts that accurately capture the probabilistic characteristics of ensemble forecasts, which have emerged as a key requirement for state-of-the-art FIRO strategies (Delaney et al., 2020; Jasperse et al., 2020; Ralph et al., 2022, 2023). Brodeur et al. (2024) addressed this need by developing a novel statistical approach to generate synthetic ensemble forecasts that mimic the hindcast attributes of the Hydrologic Ensemble Forecast Service (HEFS) (Demargne et al., 2014) for one watershed in California. HEFS is a key hydrologic forecasting tool used in recent FIRO viability assessments led by the U.S. Army Corps of Engineers (USACE) (Jasperse et al., 2020; Ralph et al., 2022, 2023), though other hydrologic ensemble forecasting systems, including those utilizing advanced machine learning, are also available or in development (Troin et al., 2021).

Brodeur et al. (2024) found that their synthetic ensemble forecasting model captured the attributes of HEFS ensemble forecasts reasonably well using forecast ensemble verification techniques (Wilks, 2019). However, their analysis also included a preliminary evaluation of how these synthetic forecasts performed when used to drive a FIRO control policy (the Ensemble Forecast Operations (EFO) model (Delaney et al., 2020)). While the resulting reservoir outcomes (e.g., storage and releases) showed general qualitative agreement between synthetic and actual forecast ensembles, the analysis also revealed operational weaknesses in the synthetic forecasts that were not captured by traditional streamflow forecast verification statistics. These shortcomings in operations-based performance are particularly important to practitioners and can present a significant barrier to the adoption of synthetic forecasts. Although Brodeur et al. (2024) identified these challenges, they did not formally quantify them, highlighting a critical research gap: How can the operations-based performance of synthetic forecasting methods be rigorously evaluated to determine whether they are “good enough” for use in operational risk analysis?

In this study, we address this question by introducing a framework to evaluate whether synthetic forecast ensembles produce operational outcomes that preserve the statistical properties of water resources system simulations forced with hindcast data. Our approach tests whether synthetic forecasts are fit for purpose by validating model outcomes on simulated reservoir storage or release series that were not the target during calibration but are more relevant for real-world applications (see Shabestanipour et al. (2023) and Stedinger and Taylor (1982) for an analogous approach in stochastic hydrology). This approach contrasts with traditional verification methods, which focus primarily on how well the statistical properties of streamflow hindcasts are reproduced (Grygier et al., 1989; Lamontagne & Stedinger, 2018; Lettenmaier, 1984; Rougé, 2021). While standard statistical verification remains important, fit-for-purpose validation provides a complementary lens, ensuring that the synthetic data capture the characteristics that matter most in practice.

In the proposed framework, we generate many synthetic forecast ensembles, which are then used to drive a risk-based reservoir simulation model for the period when hindcasts are available. The resulting ensemble of storage and release series are compared to the single series of storage and releases generated from the reservoir model driven with the original hindcast data. If the hindcast-driven storage or release series behaves like a random draw from the synthetic ensembles, it suggests statistical consistency in operational outcomes between the actual and synthetic forecasts. This can be tested with well-established techniques from the ensemble verification literature (Wilks, 2019).

In this study, we demonstrate the proposed framework by evaluating the performance of synthetic ensemble streamflow forecasts generated using two different algorithms: the semi-parametric approach in Brodeur et al. (2024) and a new non-parametric method introduced in this work (which forms a secondary contribution of this study). These methods are tested across a wide range of hydrologic regimes in California (semi-arid south, snowmelt dominated north-central Sierra Nevada, rain-dominated Coastal Range) using stylized reservoir

19447973, 2025, 6, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024WR039324 by Test, Wiley Online Library on [02/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

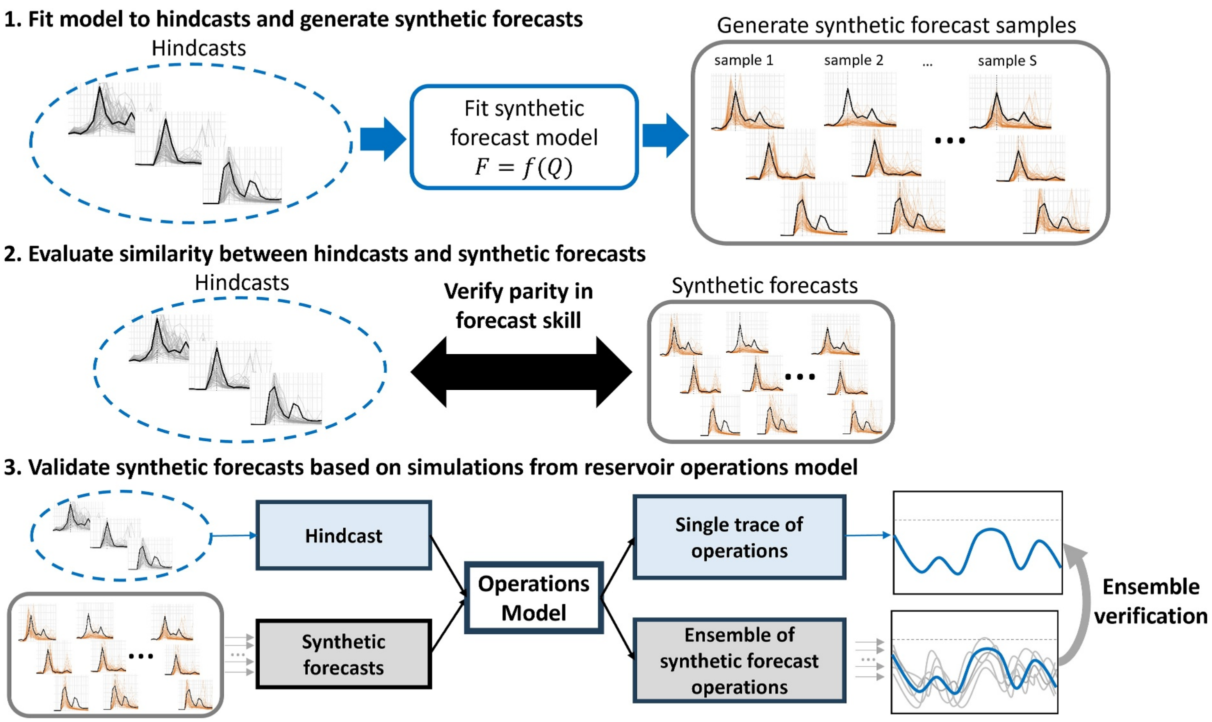
1. Fit model to hindcasts and generate synthetic forecasts
Hindcasts
Fit synthetic forecast model $F = f(Q)$
Generate synthetic forecast samples
sample 1
sample 2
...
sample S
2. Evaluate similarity between hindcasts and synthetic forecasts
Hindcasts
Verify parity in forecast skill
Synthetic forecasts
3. Validate synthetic forecasts based on simulations from reservoir operations model
Hindcast
Synthetic forecasts
Operations Model
Single trace of operations
Ensemble of synthetic forecast operations
Ensemble verification

**Figure 1.** Conceptual diagram of experimental design.

systems with multiple configurations (e.g., different capacities and maximum safe releases). The use of these stylized models—designed to represent a range of system characteristics rather than a detailed simulation of any one reservoir—enables us to assess the generalizability of synthetic forecasting algorithms across distinct hydrologic and system contexts, which are known to significantly affect FIRO policy outcomes (Taylor et al., 2024). Ultimately, this study aims to provide a framework for evaluating the suitability of synthetic forecasting methods for use in operational risk analysis, with the broader goal of supporting their adoption in practice.

## 2. Data

We utilize HEFS ensemble hindcasts and full natural flow observations from the California/Nevada River Forecast Center (CNRFC) at five locations in California: Oroville Dam (site id: ORDC1), New Bullards Bar Dam (site id: NBBC1), New Hogan Lake (site id: NHGC1), Lake Mendocino (site id: LAMC1), and Prado Dam (site id: ADOC1). These locations are all part of previous or ongoing FIRO studies being led by the U.S. Army Corps of Engineers (Jasperse et al., 2020; Ralph et al., 2022; Talbot et al., 2023; Woodside et al., 2022). Our analysis focuses solely on the inflow forecasts for these reservoirs, excluding downstream control points (see experimental design in Section 3.4 for more detail). The hindcast period differs slightly between sites, but the period of 1990-10-01 to 2019-08-15 is common to all and is used for this analysis. The hindcasts are generated from the NOAA/NCEP GEFS version 12 (Guan et al., 2022) through a frozen version of the HEFS model (Demargne et al., 2014) and differ from archived operational forecasts that reflect evolving skill, effects of regulation, and forecaster-in-the-loop inputs (CNRFC, 2022). The ensemble hindcasts are generated daily at 12:00 GMT at hourly scales out to 15 days, and the number of ensemble members varies between 39 and 41 depending on the site. We aggregate to daily forecasts by averaging over 24-hr periods in the forecasts and observations.

## 3. Methods

The experimental design presented in this study offers a comprehensive framework for evaluating synthetic ensemble forecasts for reservoir operations risk analysis, assuming the availability of a suitable hindcast data set. The approach comprises three main steps, as illustrated in Figure 1. Step 1 involves fitting a synthetic forecast model and generating many samples of synthetic ensemble forecasts. The synthetic forecast model (a) infers forecast skill characteristics from a hindcast data set, relying solely on observed streamflow series ($Q$) as input;

19447973, 2025, 6, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024WR039324 by Test, Wiley Online Library on [02/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

and (b) uses this inferred skill to produce fully stochastic simulations of plausible forecast sequences corresponding to $Q$. The implementation of this step is described in Section 3.1, where we introduce two synthetic ensemble forecasting algorithms: the method in Brodeur et al. (2024) (Section 3.1.1) and a novel alternative (Section 3.1.2).

Step 2 entails formally verifying the synthetic forecasts against the actual streamflow hindcasts to ensure that key skill characteristics are preserved. We developed a comprehensive verification methodology in prior work (Brodeur et al., 2024), which we briefly revisit in the Results.

Step 3 consists of validating the synthetic forecasts using a reservoir simulation model with forecast-informed operations. We run this model using both the original hindcast sample and the ensemble of synthetic forecast samples. This yields a single set of operational outcomes (e.g., storage and release series) for the original hindcast, which is then compared against the ensemble of operational outcomes from the synthetic forecasts. This validation process is described in detail in Sections 3.2–3.3 and constitutes the primary fit-for-purpose evaluation that underpins this study's core contribution. To demonstrate the generalizability of this approach, we apply it across various reservoir configurations and hydrologic regimes (see Section 3.4).

## 3.1. Synthetic Ensemble Forecast Generators

Brodeur et al. (2024) introduced a synthetic ensemble forecasting approach that is unique in its ability to generate multiple, plausible ensemble forecast realizations for any observed data within or outside the hindcast period. While other studies (Cassagnole et al., 2020; Rougé et al., 2023) have developed alternative synthetic ensemble forecasting methods, their focus was on generating synthetic forecasts with different skill properties within the hindcast period, and they cannot produce forecasts outside the hindcast period (i.e., for pre-hindcast or synthetic observations). These limitations reduce their utility for testing operational robustness and adaptability to climate change. In the subsections below, we first briefly describe the approach in Brodeur et al. (2024) before presenting a second, novel synthetic ensemble forecasting method, both of which can address these limitations.

### 3.1.1. Synthetic Ensemble Forecasts: syn-M1

Here, we provide a short synopsis of the synthetic ensemble forecasting algorithm detailed in Brodeur et al. (2024), hereafter referred to as "syn-M1." The approach employs a multivariate probabilistic model to simulate errors between the observations and hindcasts across multiple sites, lead times, and ensemble members. It addresses the complexity of these errors—characterized by cross-correlation, auto-correlation, non-normality, and heteroscedasticity—using a multivariate generalized autoregressive conditional heteroscedastic (GARCH) model. This model incorporates both parametric components (e.g., a model for heteroscedasticity, a skew exponential power distribution for non-normality, and a vector autoregressive model for persistence) and non-parametric components (e.g., an empirical copula and k-Nearest Neighbor (kNN) sampling for cross-correlation). To preserve seasonal characteristics of forecast errors, the model is fit separately for each month. Synthetic forecasts can be generated for any period with observed inflow time series for each forecast site, but computation demands restrict the number of sites, leads, and ensemble members that can be generated simultaneously. The syn-M1 model serves as a benchmark against which we compare a new synthetic forecasting algorithm, described next.

### 3.1.2. Synthetic Ensemble Forecasts: syn-M2

Our second synthetic ensemble forecasting algorithm, "syn-M2," is a non-parametric approach composed of two main steps: (a) kNN sampling of ensemble forecasts from the hindcast period and (b) scaling of those resampled ensemble forecasts (Figure 2). The degree of scaling is limited to preserve skill and stability across lead times.

First, we define two subsets of the data called the hindcast period and the generation period. In the hindcast period, we have both observations and ensemble $H$-day ahead hindcasts (e.g., $H = 15$). The generation period contains all observations of streamflow for which synthetic forecasts are required. The hindcast and generation periods can overlap, and the generation period can extend before or after the hindcast period.

For any given streamflow observation in the generation period, $Q_t^{\text{gen}}$, we define the $H$-day hydrograph on and following day $t$, $Q_{t:t+H}^{\text{gen}}$. In step 1, we calculate the Euclidean distance between $Q_{t:t+H}^{\text{gen}}$ and each $H$-day hydrograph from the hindcast period, and a candidate hydrograph $Q_{t':t'+H}^{\text{hind}}$ is selected based on this distance vector using kNN

19447973, 2025, 6, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024WR039324 by Test, Wiley Online Library on [02/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

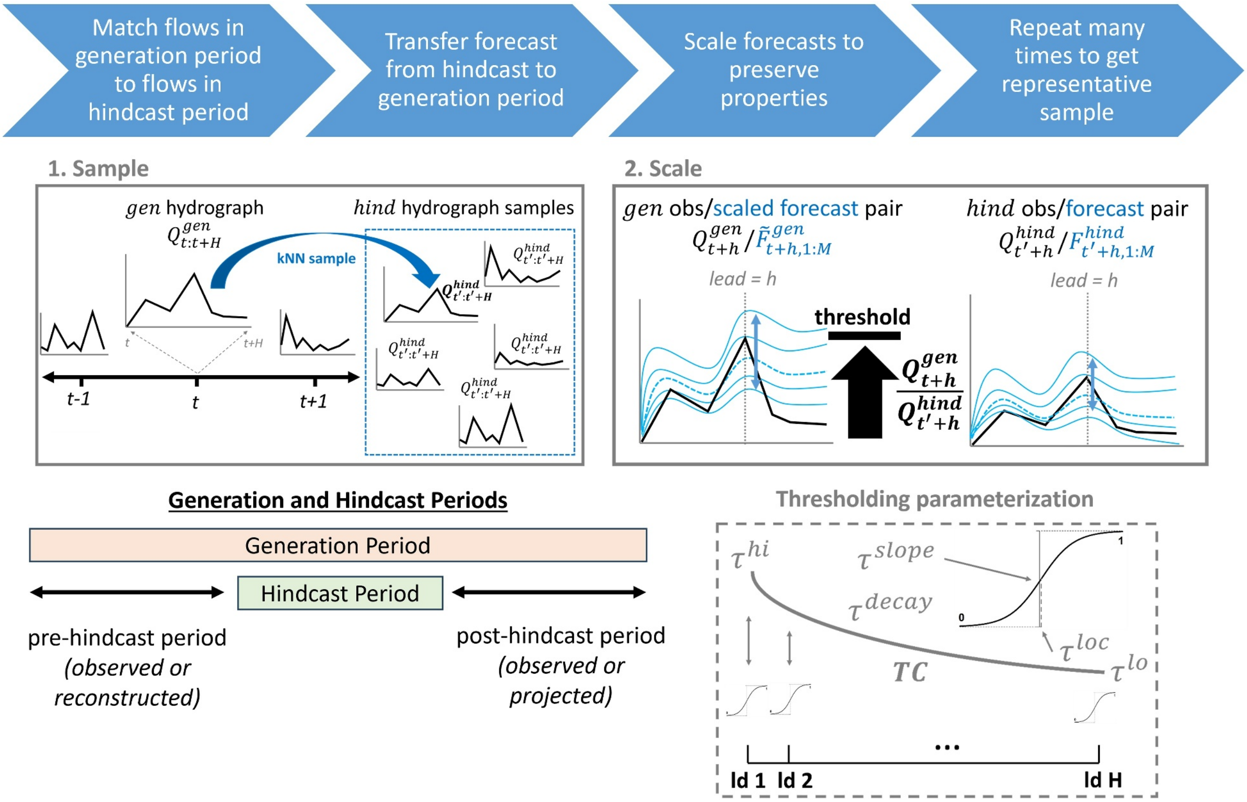
Match flows in generation period to flows in hindcast period
Transfer forecast from hindcast to generation period
Scale forecasts to preserve properties
Repeat many times to get representative sample
1. Sample
gen hydrograph
$Q^{gen}_{t:t+H}$
hind hydrograph samples
kNN sample
$Q^{hind}_{t':t'+H}$
t-1
t
t+1
t+H
2. Scale
gen obs/scaled forecast pair
$Q^{gen}_{t+h}/\tilde{F}^{gen}_{t+h,1:M}$
hind obs/forecast pair
$Q^{hind}_{t'+h}/F^{hind}_{t'+h,1:M}$
lead = h
threshold
$\frac{Q^{gen}_{t+h}}{Q^{hind}_{t'+h}}$
Generation and Hindcast Periods
Generation Period
Hindcast Period
pre-hindcast period (observed or reconstructed)
post-hindcast period (observed or projected)
Thresholding parameterization
$\tau^{hi}$
$\tau^{slope}$
$\tau^{decay}$
$\tau^{loc}$
$\tau^{lo}$
TC
1
0
ld 1
ld 2
...
ld H

**Figure 2.** Conceptual diagram of syn-M2 depicting the k-Nearest Neighbor sampling and scaling steps, where the scaling is limited by a threshold that varies across lead times. The bottom left shows a possible configuration of generation and hindcast periods, in which synthetic forecasts can be developed for the same period as the hindcast period, but also in years before or after that period (based on observations, reconstructions, or projections of flow).

sampling (we set $k = 30$ based on experimentation). Here, $t'$ is some day in the hindcast period. For periods of overlap between the hindcast and generation sets, we restrict the algorithm so that $t' \neq t$, that is, hindcast ensembles for an observation in the hindcast period cannot be resampled for that same observation in the generation period. The selected hydrograph $Q^{\text{hind}}_{t':t'+H}$ is associated with an ensemble forecast with $M$ members, $F^{\text{hind}}_{t':t'+H,1:M}$, which becomes the candidate ensemble forecast issued on day t in the generation period.

In step 2, we scale each of the $m$ members of the candidate ensemble $F^{\text{hind}}_{t':t'+H,1:M}$ by the ratio between $H$-day hydrographs in the generation and hindcast periods, calculated separately for each of the $H$ days, to produce the synthetic ensemble forecast at timestep $t$, $\tilde{F}^{\text{gen}}_{t:t+H,1:M}$:

$$\tilde{F}^{\text{gen}}_{t:t+H,1:M} = \left\{ \frac{Q^{\text{gen}}_{t+h}}{Q^{\text{hind}}_{t+h}} \times F^{\text{hind}}_{t+h,m} \middle| h \in \{1,\ldots,H\}, m \in \{1,\ldots,M\} \right\} \tag{1}$$

The scaling in Equation 1 maintains key attributes in the relationship between observations and forecast ensembles from the hindcast period, including the ratio of observations to the ensemble mean forecast and the ensemble coefficient of variation. In our analysis of the HEFS hindcasts discussed in Section 2, we find no evidence that these relationships change with the magnitude of flood events in the hindcast period (see Text S1 in Supporting Information S1).

The degree of scaling in Equation 1 is adjusted to manage two key challenges that arise in the procedure: (a) instability when observed flow values are near zero (small denominator in Equation 1), and (b) overestimation of forecast skill as lead times increase. The first challenge can lead to unrealistically large synthetic forecasts. The second issue arises when generating synthetic forecasts for $H$-day hydrographs $Q^{\text{gen}}_{t:t+H}$ in the generation period that contain very large floods toward the end of the $H$-day period, particularly if those floods are larger than any

19447973, 2025, 6, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024WR039324 by Test, Wiley Online Library on [02/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

observed flows in the hindcast period. Operational ensemble forecasts (e.g., HEFS) tend to underestimate large events at long lead times (Brown et al., 2014). However, when selecting the candidate $H$-day hydrograph from the hindcast period $Q^{\text{hind}}_{t':t'+H}$, it is more likely that the magnitude of flow toward the end of the $H$-day period in this hydrograph will be smaller than that in $Q^{\text{gen}}_{t:t+H}$, given the case described above. The resampled hindcasts $F^{\text{hind}}_{t':t'+H,1:M}$ associated with $Q^{\text{hind}}_{t':t'+H}$ will likely exhibit less underestimation at long leads than is desired, because the flows toward the end of the $H$-day period are smaller in $Q^{\text{hind}}_{t':t'+H}$ than in $Q^{\text{gen}}_{t:t+H}$. This results in synthetic forecasts $\tilde{F}^{\text{gen}}_{t:t+H,1:M}$ that are too skillful (i.e., not biased as low as the hindcasts) when the scaling procedure is applied. This issue is not as prominent at short leads, where hindcast ensembles do not exhibit the same degree of underestimation.

Both challenges above are driven by upward scaling (i.e., $\frac{Q^{\text{gen}}_{t+h}}{Q^{\text{hind}}_{t+h}}$ in Equation 1) that is too large. To address these challenges, we cap the degree of scaling using thresholds that vary by lead time. We define a threshold curve, $\text{TC}_{1:H}$, with three parameters that determine the upper scaling limit ($\tau^{\text{hi}}$), the lower scaling limit ($\tau^{\text{lo}}$), and the rate of decay ($\tau^{\text{decay}}$) between those two limits across lead times.

$$\text{decay}_h = \frac{e^{\tau^{\text{decay}}*(H-h+1)} - e^{\tau^{\text{decay}}}}{e^{2\tau^{\text{decay}}} - e^{\tau^{\text{decay}}}} \tag{2a}$$

$$\text{TC}_h = \left(\tau^{\text{hi}} - \tau^{\text{lo}}\right) * \frac{\text{decay}_h}{\max(\text{decay})} + \tau^{\text{lo}} \tag{2b}$$

$$\text{TC}_1 = \tau^{hi}; \text{TC}_H = \tau^{\text{lo}}; \tau^{\text{lo}} \geq 1; \tau^{hi} > \tau^{lo}$$

These parameters are constrained to ensure a monotonically decreasing curve across lead times. In addition, a sigmoid function with location ($\tau^{\text{loc}}$) and scale ($\tau^{\text{slope}}$) is used to modulate the degree of scaling based on the magnitude of the observed streamflow event in $Q^{\text{gen}}_{t:t+H}$ (shown here for a specific lead time $h$):

$$\text{TF}_h = \frac{1}{1 + e^{-\left(\tau^{\text{loc}} + \tau^{\text{slope}} \times \acute{Q}^{\text{gen}}_{t+h}\right)}} \tag{3a}$$

$$\acute{\text{T}}\text{C}_h = (\text{TC}_h - 1) \times \text{TF}_h + 1 \tag{3b}$$

Here, we first define a centered and scaled version of the logarithmic transform of observed flows in the generation set, $\acute{Q}^{\text{gen}}_{t+h}$, that are more symmetrically distributed than $Q^{\text{gen}}_{t+h}$ and have a constrained range. Values of $\acute{Q}^{\text{gen}}_{t+h}$ when $Q^{\text{gen}}_{t+h} = 0$ are re-calculated using the minimum non-zero value in the observational data set to avoid undefined values. The value $\acute{Q}^{\text{gen}}_{t+h}$ is then used as input to a sigmoid function to determine $\text{TF}_h$ (a value between 0 and 1), which adjusts downward the scaling threshold $\text{TC}_h$ to its final value $\acute{\text{TC}}_h$. This helps to constrain the scaling for very small observed flows in $Q^{\text{gen}}_{t:t+H}$. The scaling factor in Equation 1 is then limited to not exceed $\acute{\text{T}}\text{C}_h$, leading to the final scaling used to develop synthetic ensemble forecasts:

$$\tilde{F}^{\text{gen}}_{t:t+H,1:M} = \left\{ \max\left( \frac{Q^{\text{gen}}_{t+h}}{Q^{\text{hind}}_{t+h}}, \acute{\text{T}}\text{C}_h \right) \times F^{\text{hind}}_{t+h,m} \middle| h \in \{1, \dots, H\}, m \in \{1, \dots, M\} \right\} \tag{4}$$

The parameters of the threshold curve $\left\{\tau^{\text{hi}}, \tau^{\text{lo}}, \tau^{\text{decay}}, \tau^{\text{loc}}, \tau^{\text{slope}}\right\}$ are calibrated using 5-fold cross validation to maximize parity in ensemble skill (i.e., ensemble continuous ranked probability scores, CRPS; Wilks, 2019) between the synthetic ensemble forecasts and hindcasts for the top 1% of flows across lead times (see Text S2 in Supporting Information S1 for more detail). This cross-validation helps prevent information leakage between hindcast and generation periods during model evaluation. We address performance variability between random initializations by generating 5 random seeds of syn-M2 forecasts and choosing the seed that maximizes performance.

19447973, 2025, 6, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024WR039324 by Test, Wiley Online Library on [02/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The syn-M2 approach is significantly more efficient and scalable than syn-M1, allowing it to be readily applied to more complex multisite modeling scenarios. In addition, the optimizable approach allows syn-M2 to account for different skill dynamics across lead times, making it adaptable to hindcast sources other than the ones used in this study.

### 3.2. Risk-Based Operations Policy

Synthetic forecasts and hindcasts are used to drive a reservoir systems model with risk-based operating rules. We adapt the simulation-optimization framework from Taylor et al. (2024) that provides a simplified implementation of the EFO policy from Delaney et al. (2020). We provide a brief overview of the approach here but direct the reader to Taylor et al. (2024) and Text S3 in Supporting Information S1 for more detail.

The simulation model follows a basic mass balance relationship:

$$S_t = S_{t-1} + Q_t - R_t - \mathrm{Spill}_t \tag{5}$$

Where $S_t$ is storage, $Q_t$ is inflow, $R_t$ is the prescribed (or controlled) release, and $\mathrm{Spill}_t$ is an uncontrolled release over the spillway, all for day $t$. The risk-based policy makes prescribed flood control releases when cumulative ensemble inflow forecasts cause some proportion of storage forecasts to exceed a threshold (e.g., top of spillway crest). This proportion varies with lead time and is specified by a risk tolerance curve. For each lead time that exceeds the allowable risk (i.e., more forecast members exceed the storage threshold than are allowed by the risk tolerance curve), a release is calculated that reduces the risk below the curve. Whichever lead time yields the highest required release becomes the prescribed flood control release on day $t$. We refer to this lead time as the "release lead" for day $t$. Prescribed releases are then limited by a maximum safe release based on downstream channel capacity. Uncontrolled spills are activated when storage exceeds the spillway crest.

We adopt a parsimonious risk curve with five parameters, similar to the approach taken by Taylor et al. (2024), which allows us to capture a diverse range of shapes (see Figure S5 in Supporting Information S1). To optimize the parameters of this risk curve, we utilize the Differential Evolution algorithm (Storn & Price, 1997) with 10,000 function evaluations. As in Taylor et al. (2024), the simulation-optimization routine is executed over the HEFS hindcast period, using an aggregated objective function that rewards higher average storage while imposing significant penalties for spills and for exceeding maximum safe release values. As the focus of the experiment is flood control, the model does not include releases for water supply or environmental flows. We also do not include a flood pool, enabling the reservoir to reach full capacity between major inflow events. This represents the riskiest scenario for optimization, as the policy must effectively manage flood control risks during extreme events starting from or near full reservoir capacity.

### 3.3. Operations-Based Validation of Synthetic Forecasts

Using the policy in Section 3.2, we can simulate reservoir operations using the original ensemble hindcast, which will result in a single trace of operational outcomes (storages, releases). Similarly, we can simulate operations using many synthetically generated ensemble forecasts during the hindcast period, which will result in an ensemble of operational outcomes for each synthetic forecast method (syn-M1 and syn-M2). We can then use ensemble verification techniques to probabilistically evaluate the ensemble of synthetic operational outcomes against the actual forecast outcomes as outlined previously in Figure 1. We refer to this process as the "reservoir operations-based validation" of the synthetic forecasts, or simply "operations-based validation." The goal of this framework is to establish a formal statistical method for comparing two or more synthetic forecast algorithms based on their ability to replicate the operational performance characteristics of the original hindcast.

Several measures can be used to evaluate ensembles of synthetic operational outcomes against actual operational outcomes, including reliability, spread-error relationships, and accuracy (Wilks, 2019). For operations-based validation, we focus solely on the reliability of the operational ensembles, as our main interest lies in the distributional similarity between operations driven by synthetic forecasts and those driven by hindcasts. Other aspects of ensemble verification, such as spread-error relationships and accuracy, are more closely related to the predictive capabilities of the ensembles, which are not the primary focus of this work.

19447973, 2025, 6, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024WR039324 by Test, Wiley Online Library on [02/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

19447973, 2025, 6, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024WR039324 by Test, Wiley Online Library on [02/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

**Table 1**
*Baseline Reservoir Configurations*

| Reservoir | Capacity ($km^3$) | Safe release ($m^3/s$) | Ramping ($m^3/s/day$) |
|---|---|---|---|
| ORDC1 | 4.3468 | 4,248 | 874 |
| NBBC1 | 1.1977 | 1,950 | 816 |
| NHGC1 | 0.3910 | 269 | 55 |
| LAMC1 | 0.1437 | 113 | 113 |
| ADOC1 | 0.2133 | 283 | 283 |

*Note.* Note that the same ramping rate is applied to both positive and negative release changes.

We focus on three operational variables: reservoir storage, releases, and release leads (see description below). Both storage and release series are continuous variables bounded by operational constraints (i.e., reservoir capacity, maximum safe release). To evaluate ensemble reliability for these variables, we employ two techniques: (a) probability integral transforms (PIT), and (b) reliability diagrams. Both methods yield a 1:1 relationship for ideal ensembles when compared against the "observed" series (i.e., the storages and releases from the reservoir model driven by the original hindcast).

PIT plots (Huang & Zhao, 2022) evaluate the equiprobability of the observed series within the ensemble across multiple instances, determining if it behaves as a random draw from the ensemble. They assess whether the distribution of empirical non-exceedance probabilities of each observation evaluated under the ensemble is uniform. The PIT analysis can be summarized using the statistic $\pi_{\text{rel}}$, which is a normalized quantity between 0 and 1 that captures departures from the 1:1 relationship. A value of 0 indicates perfect reliability while values greater than 0 indicate decreasing performance (see Text S4 in Supporting Information S1 for more detail).

Reliability diagrams evaluate the ensemble's ability to capture extreme events (i.e., events exceeding a threshold) by framing the verification problem using binary outcomes (0 = non-occurrence, 1 = occurrence). These diagrams compare the predicted probability of events against the relative observed frequency of events across different bins of predicted probabilities. We only apply reliability diagrams to releases because there is no sensible threshold for storage in our experiments (by design, storage is often right below the spillway level; see Section 3.2). We summarize each reliability diagram using the reliability component of the Brier Score, $\text{BS}_{\text{rel}}$, to capture deviations from the 1:1 relationship. $\text{BS}_{\text{rel}}$ has a similar interpretation to $\pi_{\text{rel}}$, but differs slightly in that deviations from the 1:1 relationship are weighted by bin size (see Text S4 in Supporting Information S1 for more detail).

Release leads (i.e., the forecast lead time used to determine flood control releases; see Section 3.2 and Text S3 in Supporting Information S1), are integer variables between 1 and $H$ and require a different validation procedure. We compare the distribution of categorical release leads from the hindcast run to the distribution of release leads from all synthetic forecast driven simulations using a $\chi^2$ test (Teegavarapu et al., 2019; Wilks, 2019). We coarsen the data for this test to ensure sufficient sample sizes by binning the release leads into four categories (early leads: 1–3 days; moderate leads: 4–6 days; late leads: 7–9 days; and very late leads: 10–14 days). This analysis is unique because it captures how well the synthetic forecasts capture dynamics between the risk curve and forecast behavior across different lead times.

### 3.4. Numerical Experiment

We employ a multisite, stylized experimental design to evaluate synthetic forecast performance across various hydrologic regimes and reservoir configurations. Observed inflows and HEFS hindcasts are collected for five locations in California: Oroville Dam, New Bullards Bar Dam, New Hogan Lake, Lake Mendocino, and Prado Dam (see Section 2). We also gather basic reservoir attributes for each of these sites (Jasperse et al., 2020; Taylor et al., 2024; Woodside et al., 2022), including capacity, safe maximum release, and ramping rate (see Table 1). These attributes span a diverse range of system configurations related to the ratios of mean inflow to capacity and maximum inflow to safe maximum release, both of which Taylor et al. (2024) highlighted as key factors influencing reservoir policy performance.

For each site, we train and generate synthetic forecasts to the hindcast data for both methods (syn-M1 and syn-M2) using a 5-fold cross validation described in Text S5 in Supporting Information S1. That is, the synthetic forecasts used to drive all reservoir simulations shown in the results were developed using models trained to different portions of the data than the period for generation. This ensures that the performance comparison between syn-M1 and syn-M2 is based completely on out-of-sample data.

As part of the experimental design, we also modify reservoir attributes from their baseline values to test how synthetic forecasts perform for more highly constrained systems. Specifically, for each reservoir, we lower the maximum safe release (and ramping rates proportionally) until the optimized, hindcast-driven reservoir model can no longer avoid spills in the simulation over the hindcast period. This is referred to as the "highly constrained" scenario. We then set the maximum safe release and ramping rates halfway between those of the baseline scenario and the highly constrained scenario and term this the "moderately constrained" scenario. For each site and scenario, we optimize a new risk curve, creating a 5 × 3 matrix of sites and configurations (15 separate cases) that compose the multisite, stylized design. Then we simulate from the FIRO operations model described in Section 3.2 across all HEFS, syn-M1, and syn-M2 forecasts for each of the 15 cases and use the same operations-based validation framework from Section 3.3 to evaluate the performance of the synthetic forecasts.

## 4. Results

We first present the validation for synthetic forecasts at a single site, Lake Oroville (ORDC1), in detail before extending the analysis to other sites. Lake Oroville is a large, snowmelt fed system that is the largest reservoir in California's State Water Project.

### 4.1. Verification of Synthetic Ensemble Inflow Forecasts

Before reporting results for the operations-based validation of the synthetic forecasts, we first briefly compare the synthetic ensemble inflow forecasts to the inflow hindcasts. In this comparison, 100 samples of synthetic forecasts from both syn-M1 and syn-M2 were generated for the entire hindcast period at ORDC1, and performance measured using three metrics: cumulative rank histograms, binned spread-error (BSE) diagrams, and CRPS skill scores (see Brodeur et al. (2024) for further detail on these metrics). These same diagnostics were developed for the HEFS hindcasts, and both are shown in Figure 3. If results for the hindcasts (blue lines or points in Figure 3) are encapsulated within the synthetic ensemble results (orange or pink lines and boxplots), this indicates that the synthetic forecasts were able to reproduce the performance of the hindcasts. We show cumulative rank histograms and BSE diagrams for 1-day and 3-day leads and focus on the top 1% of flows for the cumulative rank histograms.

The cumulative rank histograms show that the HEFS hindcast is slightly biased low for large events at the 1-day lead, and this bias grows at the 3-day lead (Figures 3a–3d). This profile is captured well by syn-M2 but not by syn-M1, which exhibits too much downward bias in the forecast ensembles. A similar result is seen in the BSE diagrams, where syn-M1 exhibits too much forecast error in relation to spread, suggesting the forecast ensembles are underdispersed or biased (Figures 3e–3h).

Figures 3i and 3j shows the CRPS skill scores for the HEFS hindcasts and the 100 samples of synthetic forecasts across four selected lead times (1-, 3-, 5-, and 10-day) and two subsets of the data: the lower 99% of inflows (Figure 3i) and the upper 1% (Figure 3j). A CRPS skill score closer to one indicates better forecast accuracy compared to a reference climatological forecast. Both syn-M1 and syn-M2 slightly overestimate the CRPS skill score of HEFS for most flow events (Figure 3i), particularly at longer leads. However, for the largest events (Figure 3j), syn-M2 emulates HEFS skill and shows a clear advantage over syn-M1.

Across the range of ensemble forecast verification metrics, syn-M2 performs equal or better than syn-M1. Results are similar at the other four sites (see Text S6 in Supporting Information S1), although the degree of outperformance depends on the metric and site. Importantly though, these analyses do not provide sufficient evidence that either synthetic forecasting technique will lead to operational outcomes that are statistically similar to those simulated with the original HEFS hindcasts. This issue is addressed with operations-based validation, discussed next.

### 4.2. Operations-Based Validation

The operations-based validation for Oroville reservoir requires an optimized risk tolerance curve to guide FIRO operations (described in Section 3.2). Figure 4a shows the optimized curve, which accepts zero forecast risk out to 2 days, then smoothly transitions to a forecast risk threshold of 1.0 (no risk aversion) at day 9 and beyond. So, for example, the policy will make flood control releases if any ensemble member suggests storage will exceed the spillway crest at a 1–2 days lead, whereas at a 6-day lead, at least 40% of ensemble members need to exceed the spillway crest to initiate a flood control release. Figure 4b shows time series of simulated storage for 4 years between 1994 and 1997, forced with the HEFS hindcasts used to train the policy. These results indicate frequent, large forecast-based pre-releases ahead of large inflow events. Importantly, this FIRO policy (simulated with no water supply withdrawals) can maintain mean reservoir storage near full capacity without spills and can also

19447973, 2025, 6, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024WR039324 by Test, Wiley Online Library on [02/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

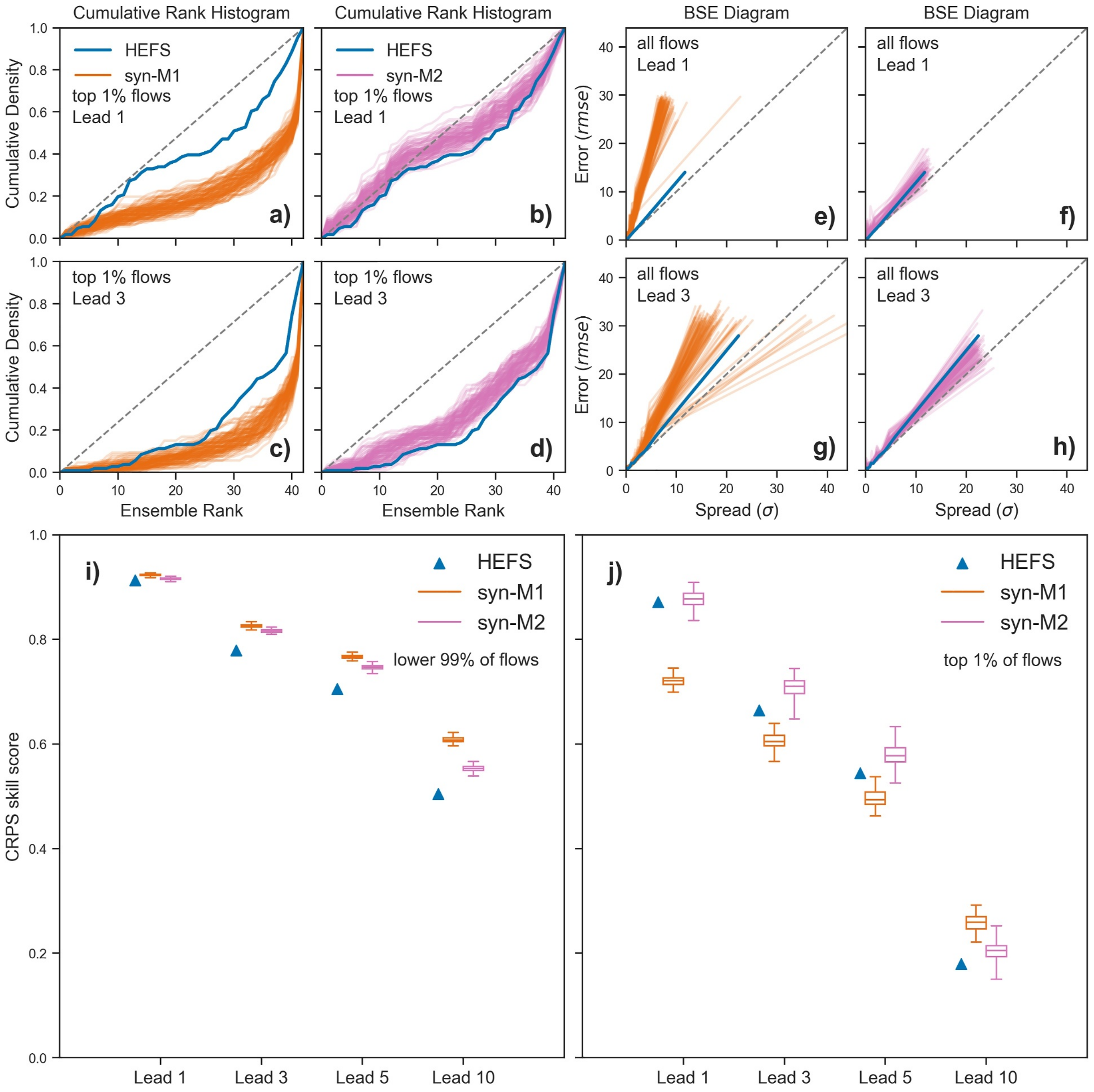

**Figure 3.** Ensemble verification statistics for ORDC1 across the Hydrologic Ensemble Forecast Service inflow hindcasts and 100 synthetic forecast samples, with all synthetic forecasts developed out of sample. (a–d) Cumulative rank histograms for specified percentile (based on observed inflows) and lead; (e–h) binned spread-error diagrams for specified lead; and (i and j) CRPS skill score for specified percentile, with a reference ensemble forecast based on the 30-member ensemble of observed values (1990–2019) for each day of the water year.

decrease the 99th percentile release ($R_{99}$) from 3,373.4 m$^3$/s under baseline (no forecast) operations to 1,169.7 m$^3$/s. These findings mirror the FIRO water supply and flood risk benefits demonstrated in earlier studies (Delaney et al., 2020; Jasperse et al., 2020; Taylor et al., 2024).

Figures 4c–4h shows time series of simulated storage, release, and release leads derived from this policy during a 30-day period around the flood of record in January 1997. Results include the policy outcomes forced with the HEFS hindcasts, as well as 100 samples of synthetic forecasts from syn-M1 (c,e,g) and syn-M2 (d,f,h). Using the HEFS hindcasts (blue lines), the policy initiates pre-releases around 7 days prior to the flood peak (i.e., around 12-26-1996), drawing down the reservoir's storage to make room for forecasted inflows (Figures 4c and 4d). These releases ramp up at shorter leads, with very large pre-releases around 4 days ahead of the event (Figures 4e and 4f). The release leads (Figures 4g and 4h) also show a smooth decline toward shorter leads as the event

19447973, 2025, 6, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024WR039324 by Test, Wiley Online Library on [02/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

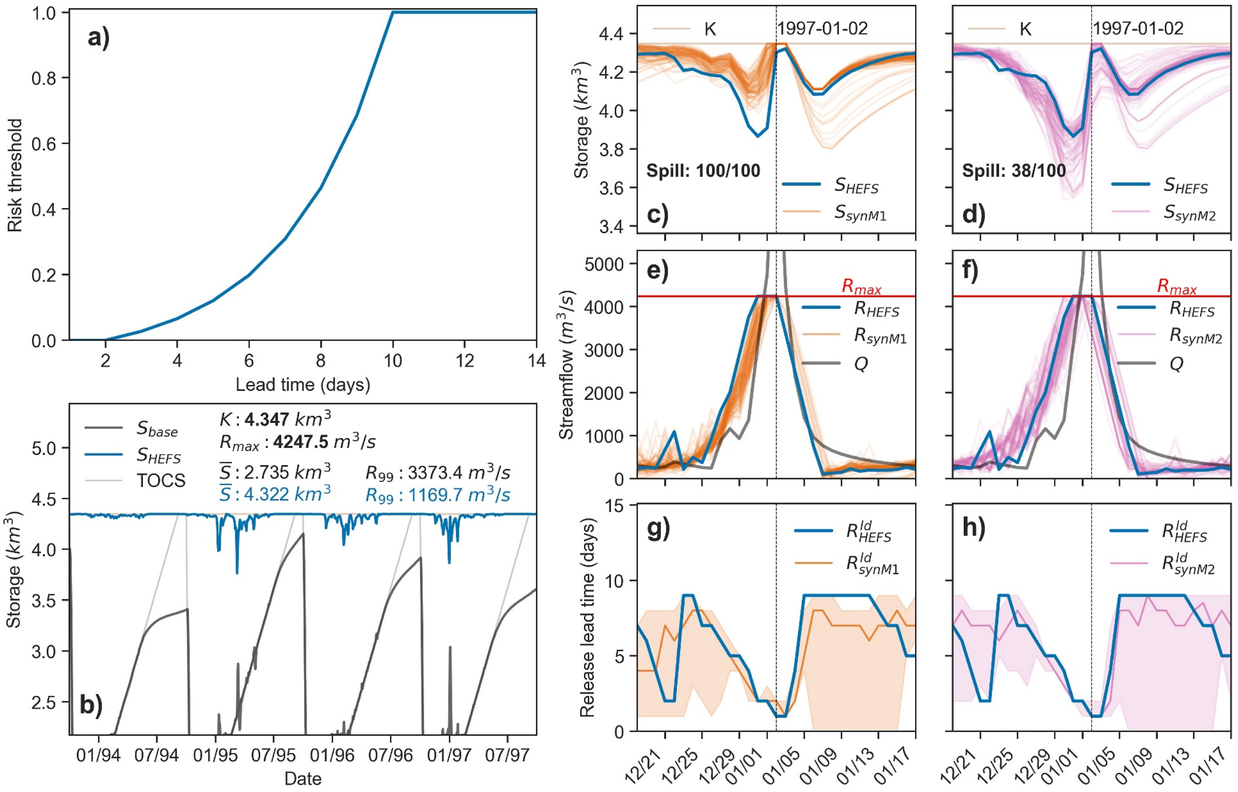

**Figure 4.** (a) Calibrated risk curve for ORDC1. (b) Time series of storage based on the calibrated risk curve with Hydrologic Ensemble Forecast Service (HEFS) hindcasts (blue) compared to baseline, no-forecast operations (black) for a wet 4-year period. The baseline, no-forecast policy is taken from Taylor et al. (2024). Statistics shown include mean storage ($\overline{S}$) and the 99th percentile release ($R_{99}$), calculated across the entire hindcast period. (c–(h) Comparison between HEFS, syn-M1 (orange), and syn-M2 (pink) storage ($S$) and release ($R$) dynamics for a 30-day window around the 1997-01-02 flood. (c–(d) Storage time series; (e–(f) Release and reservoir inflow (Q); and (g–(h) policy release lead time ($R^{\mathrm{ld}}$). The solid orange and pink lines (light shading) are the median (90% bounds) of the synthetic ensemble for (g) and (h) respectively.

approaches. When the flood peak occurs on 01-02-1997, the evacuated space is quickly refilled to full storage without triggering emergency spillway usage.

The HEFS-driven FIRO policy, optimized using the HEFS hindcast sequence, successfully manages to draw down storage just enough to avoid spilling during the 1997 flood. However, this is not the case for many of the synthetic forecast samples, to which the policy was not trained. In fact, all 100 synthetic forecasts samples from syn-M1 and 38% of samples from syn-M2 resulted in spills (Figures 4c and 4d). A key question, though, is whether these outcomes are plausible. This hinges on whether the synthetic forecasts correctly capture the uncertainty present in the hindcasts. Reservoir operations based on synthetic forecasts should closely align with those driven with HEFS hindcasts, albeit with greater variability. This is especially true in cases with high forecast uncertainty, such as flood events caused by atmospheric rivers, where small errors in predicted landfall location can result in large errors in forecasted streamflow (Ralph et al., 2019). Thus, one way to evaluate the plausibility of synthetic forecast performance is to determine whether operational outcomes driven by HEFS hindcasts fall within the distribution of operational outcomes generated by the synthetic forecasts.

By this measure, syn-M2 outperforms syn-M1, especially for storage and release series (Figures 4c–4f). Reservoir storage driven by HEFS hindcasts is drawn down below the minimum storage sequence from syn-M1 prior to the flood, but it is contained within the ensemble of storage sequences under syn-M2 (Figures 4c and 4d). These improvements extend to the release sequences, particularly for the 7 days preceding the flood peak (Figures 4e and 4f). These results suggest that operational outcomes driven by HEFS hindcasts could plausibly represent a random draw from the distribution of outcomes under syn-M2, but this seems unlikely for syn-M1. Notably, both synthetic forecast methods have few or no members that capture the drawdown on 12-22-1996, likely due to an inaccurately forecasted inflow peak (distinct from the 01-02-1997 flood) at a 2-day lead in HEFS. The challenge

19447973, 2025, 6, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024WR039324 by Test, Wiley Online Library on [02/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

19447973, 2025, 6, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024WR039324 by Test, Wiley Online Library on [02/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

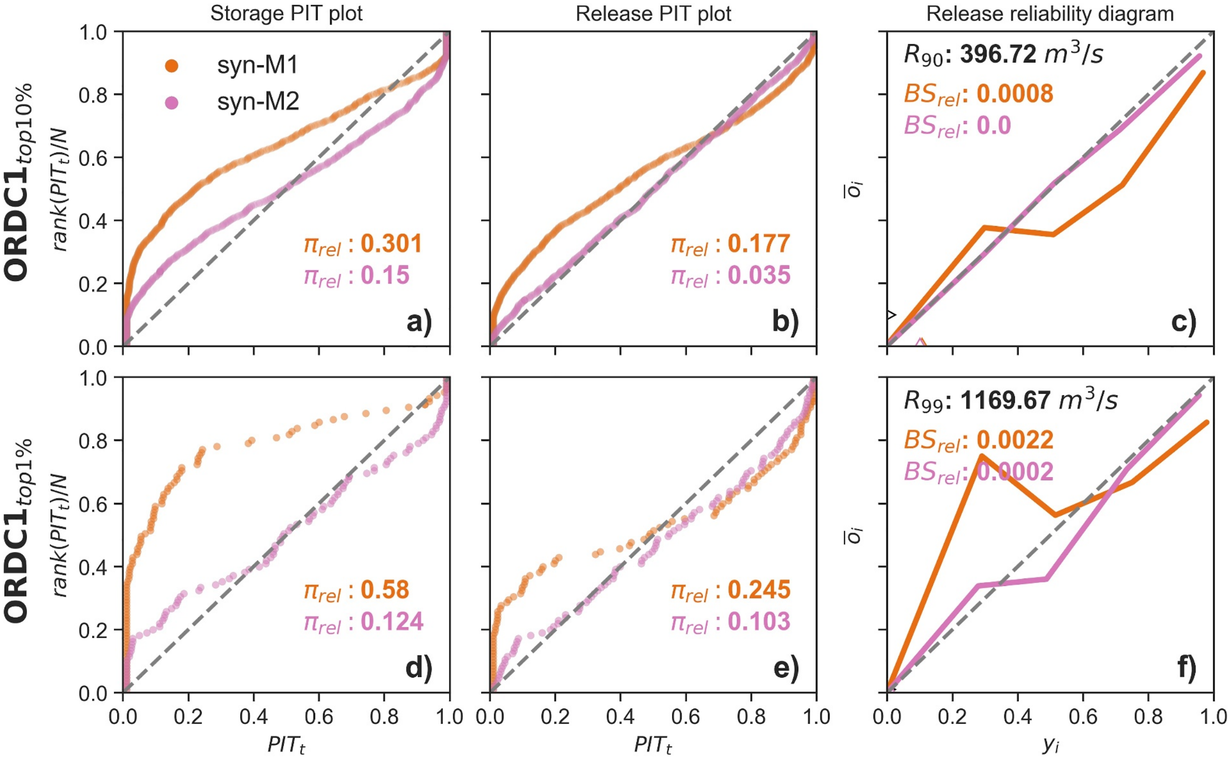

**Figure 5.** Ensemble verification of release and storage ensembles from syn-M1 (orange) and syn-M2 (purple) for ORDC1. (a–c) Verification of storage and release ensembles for the top 10% of inflows: (a) Storage probability integral transforms plot with reliability metric ($\pi_{\mathrm{rel}}$, closer to zero is better); (b) same as (a) but for releases; (c) reliability diagram for 90th percentile release, with Brier Score metric ($\mathrm{BS}_{\mathrm{rel}}$, closer to zero is better). Triangles on the $x$ and $y$ axes show the unconditional probabilities of occurrence for synthetic forecasts and hindcasts. (d–f) As in (a–c) but for the top 1% of inflows for (d–e) and 99th percentile release for (f).

of emulating false positive forecasts was highlighted in Brodeur et al. (2024) and continues to pose difficulties for both synthetic forecast methods.

Although the comparisons in Figure 4 are informative and consistent across sites (see Text S7 in Supporting Information S1), they are specific to an individual event and do not formally quantify the differences in performance across syn-M1 and syn-M2. This is addressed in Figures 5 and 6, which show results for the formal probabilistic ensemble verification approaches described in Section 3.3 for Lake Oroville. Similar plots for the other four sites are presented in Texts S8 and S9 in Supporting Information S1, and mirror the outcomes shown in Figures 5 and 6 below.

We assess the synthetic forecast storage and release outcomes in terms of PIT diagrams and reliability diagrams (releases only; storage does not have a meaningful "extreme" threshold), where a 1:1 relationship constitutes ideal performance (Figure 5). These diagrams are constructed using days containing the top 10% of inflows ($n \approx 1000$; Figures 5a–5c) and the top 1% of inflows ($n \approx 100$; Figures 5d–5f), although similar results are obtained if based on all days in the hindcast period (not shown). In the PIT diagrams for storage and releases, syn-M1 and syn-M2 performance degrades slightly from the top 10% to the top 1% of inflows. Further, both methods perform worse in the storage PIT diagram, reflecting the effect of a high degree of autocorrelation in the storage time series. Nonetheless, syn-M2 clearly outperforms syn-M1 in this analysis for both inflow regimes and approaches near ideal reliability for releases in the top 10% of inflows. This general finding applies to the release reliability diagrams as well, where syn-M2 outperforms syn-M1, particularly in the top 1% of inflows. The $\pi_{\mathrm{rel}}$ and $\mathrm{BS}_{\mathrm{rel}}$ metrics associated with the PIT and reliability diagrams also indicate numerically better performance of syn-M2 over syn-M1 (lower values are better).

Figure 6 focuses on the release lead variable for ORDC1, showing histograms of the frequency with which different lead times drive release decisions in the FIRO-based policy. Like Figure 5, these results are presented for the top 10% and 1% of inflows. Figure 6 also presents the results of $\chi^2$ tests comparing the frequency of release

19447973, 2025, 6, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024WR039324 by Test, Wiley Online Library on [02/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

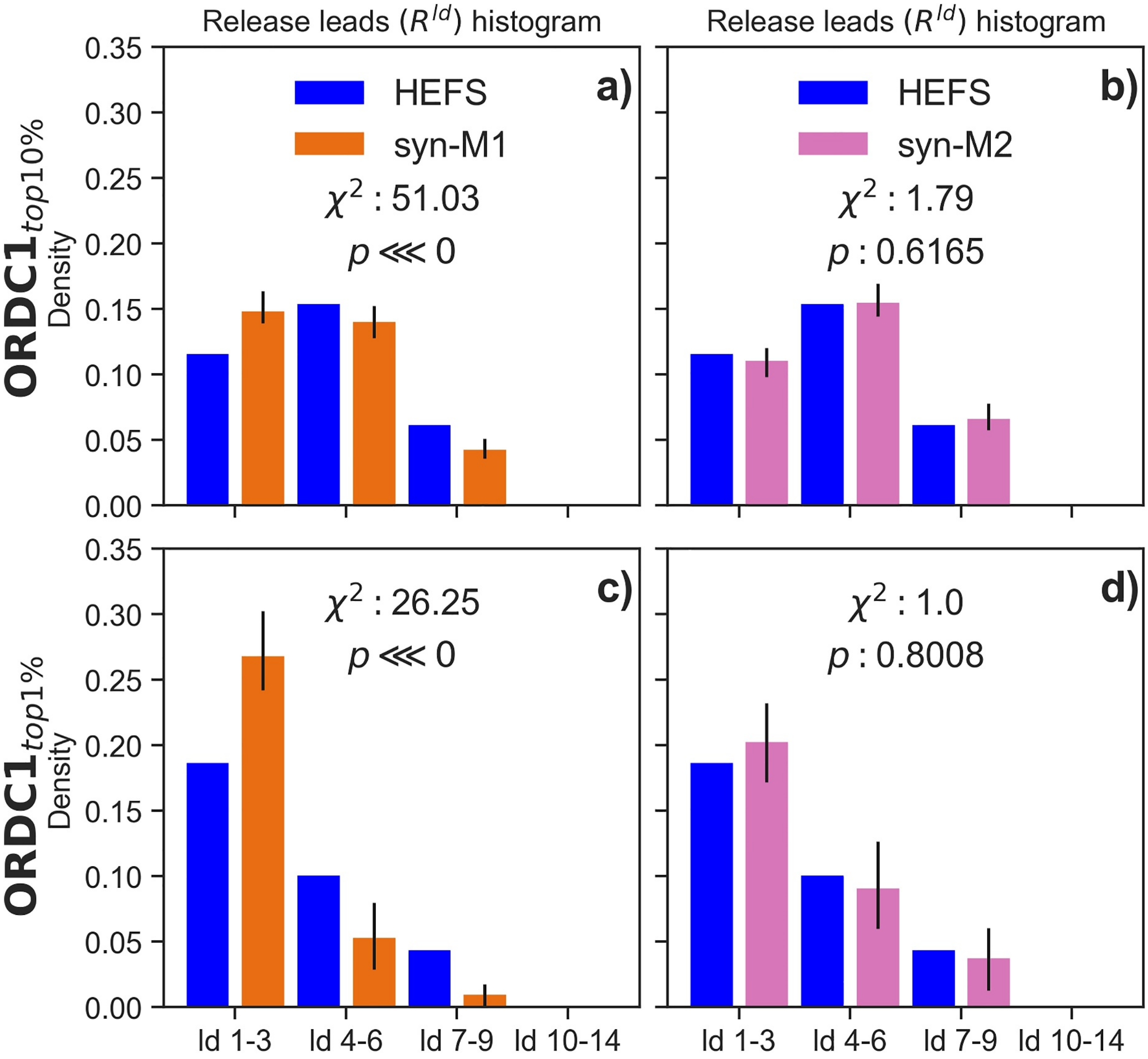

**Figure 6.** (a–b) Release lead histogram comparison between Hydrologic Ensemble Forecast Service and 100 aggregated samples of syn-M1 and syn-M2 for the top 10% of inflows. For synthetic forecasts, bar height is the median while error bars show range across 100 samples. (c–d) As in panels (a–b) but for the top 1% of inflows.

leads in four different bins (early, moderate, late, and very late leads) between the policy driven by HEFS hindcasts and synthetic forecasts. Recall that for the $\chi^2$ test, lower $\chi^2$ values (higher $p$-values) indicate an inability to reject the null hypothesis that the hindcast and synthetic forecast release leads come from the same distribution. Results for the other sites are shown in Text S9 in Supporting Information S1.

Release leads resulting from the HEFS hindcasts (blue bars in Figure 6) show policy decisions that frequently rely on the short to medium lead forecasts (1–3 days and 4–6 days) to drive releases, occasionally rely on late lead forecasts, and never use very late forecasts. Both synthetic forecasting algorithms (syn-M1 and syn-M2) show release leads that follow a similar pattern to HEFS. Here, syn-M2 slightly outperforms syn-M1 in replicating the HEFS operational outcomes. The median release leads from syn-M2 (bar heights in Figure 6) are closer to the HEFS median values than syn-M1, and the syn-M2 spread (error bars) better contain the HEFS median values compared to syn-M1. For both the top 10% and 1% of inflows, the much lower $\chi^2$ values (higher $p$-values) for syn-M2 show that its release lead distribution is statistically indistinguishable from HEFS with high confidence.

In Figure 7, we combine four metrics from Figures 5 and 6 into a single polar plot that summarizes performance for the two synthetic forecast methods (syn-M1 and syn-M2) across storage ($S_{\pi_{\mathrm{rel}}}$), release ($R_{\pi_{\mathrm{rel}}}, R_{\mathrm{BS}_{\mathrm{rel}}}$), and release lead ($R^{\mathrm{ld}}_{\chi^2}$) outcomes. These metrics are normalized so that values on any axis that are closer to zero for one

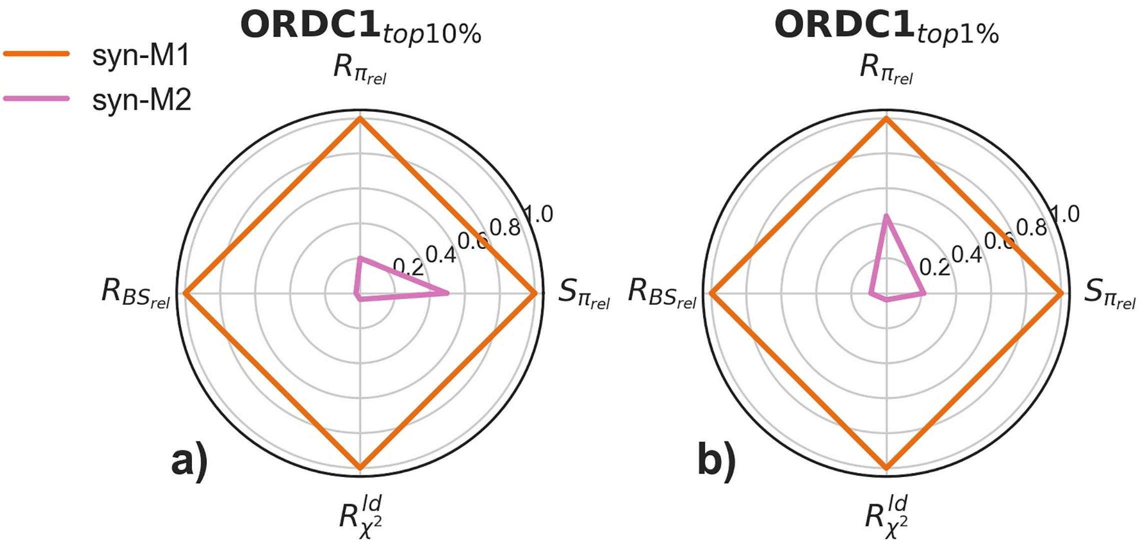

**Figure 7.** Polar plot comparison of release probability integral transforms (PIT) reliability ($R_{\pi_{\mathrm{rel}}}$), storage PIT reliability ($S_{\pi_{\mathrm{rel}}}$), release Brier Score reliability ($R_{\mathrm{BS}_{\mathrm{rel}}}$), and the release lead $\chi^2$ metric ($R^{\mathrm{ld}}_{\chi^2}$) for synthetic forecast ensembles. Performance is shown separately for the (a) top 10% and (b) top 1% percent of inflows. $R_{\pi_{\mathrm{rel}}}$, $S_{\pi_{\mathrm{rel}}}$, and $R_{\mathrm{BS}_{\mathrm{rel}}}$ are equal to the metrics shown in Figure 5, while $R^{\mathrm{ld}}_{\chi^2}$ is equal to the metric in Figure 6. For $R_{\mathrm{BS}_{\mathrm{rel}}}$, we use the (a) 90th and (b) 99th percentile of non-zero releases to define events. All metrics are zero-referenced (closer to zero is better performance) and scaled by the worse performing synthetic forecasting method.

method indicate better performance compared to the other method. Figure 7 highlights that for Lake Oroville, syn-M2 outperforms syn-M1 in all four metrics by a significant degree for days that include the top 10% and 1% of inflows. This summary provides a concise comparison between the two synthetic forecast models, which will facilitate easier comparisons across multiple sites simultaneously (described next).

### 4.3. Multisite Operations-Based Validation With Increasing System Constraints

We optimized risk tolerance curves for each of the five locations, incorporating site-specific reservoir configurations (see Table 1) and HEFS inflow forecasts, and repeated the optimization under three distinct constraint scenarios: baseline, moderately constrained, and highly constrained. Figure 8 shows the resulting risk tolerance curves for each site and scenario. Under the baseline configuration, the curves exhibit significant diversity in shape, reflecting the complex interactions between hydrology and reservoir configurations across different locations. As constraints increase—specifically, with lower maximum safe releases and ramping rates - the curves become progressively more risk-averse and begin to converge toward similar profiles under the highly constrained configuration. We use this diverse set of policies to evaluate synthetic forecast performance across differing hydrologic conditions and forecast lead times.

Figure 9 presents polar plots summarizing syn-M1 and syn-M2 performance across sites and constraint scenarios using metrics for storage ($S_{\pi_{\mathrm{rel}}}$), release ($R_{\pi_{\mathrm{rel}}}, R_{\mathrm{BS}_{\mathrm{rel}}}$), and release lead ($R^{\mathrm{ld}}_{\chi^2}$) outcomes, similar to Figure 7. Again,

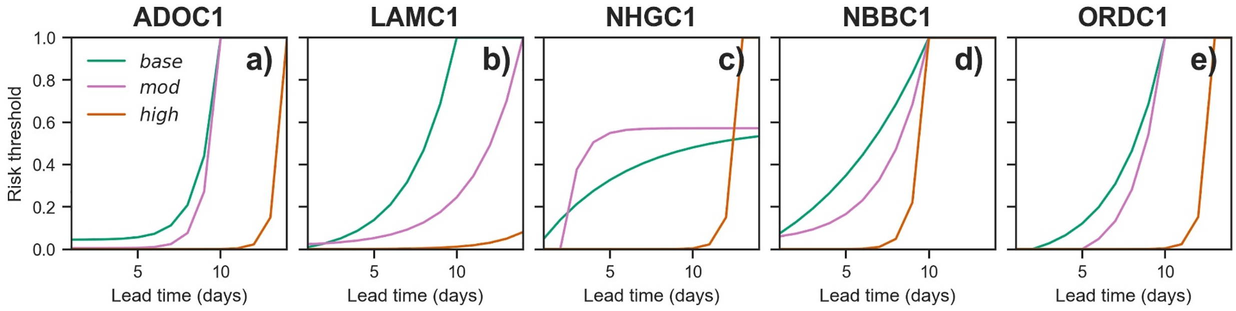

**Figure 8.** Optimized risk curves across sites for baseline constrained (base), moderately constrained (mod), and highly constrained (high) cases (described in Section 3.4).

19447973, 2025, 6, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024WR039324 by Test, Wiley Online Library on [02/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

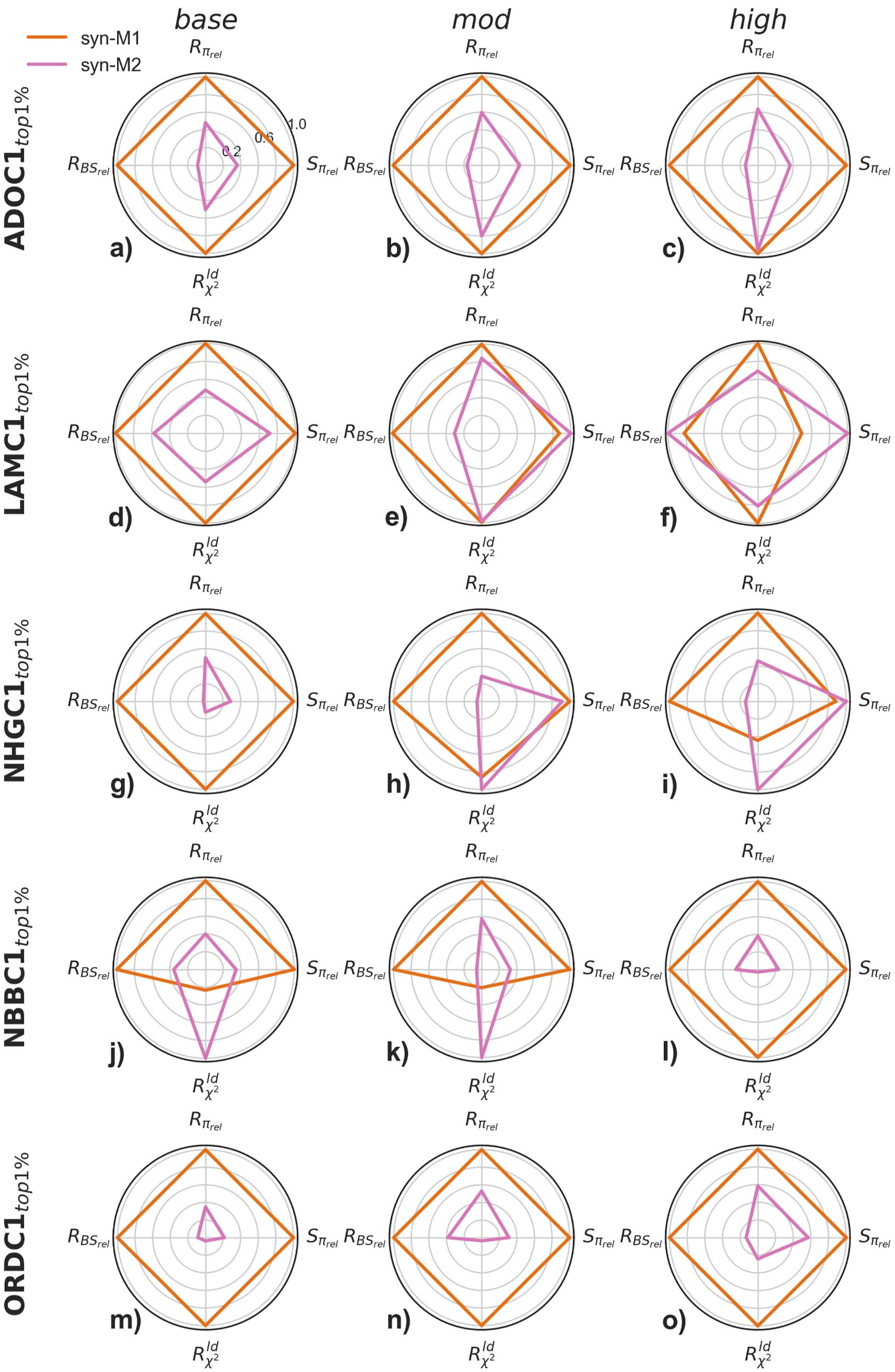

**Figure 9.** Multisite polar plot comparison (as in Figure 7) for baseline constrained (base), moderately constrained (mod), and highly constrained (high) cases for the top 1% of inflows.

19447973, 2025, 6, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024WR039324 by Test, Wiley Online Library on [02/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

results focus on the top 1% of inflows, though similar patterns are observed for the top 10% (see Text S10 in Supporting Information S1). Results in Figure 9 show that syn-M2 generally outperforms syn-M1 across sites and constraint scenarios, albeit with a few exceptions. For example, at New Bullards Bar under the baseline and moderately constrained scenario, syn-M1 more accurately represents the distribution of release lead times ($R^{\mathrm{ld}}_{\chi^2}$) compared to syn-M2 (Figures 9j and 9k). However, these examples often result from instances where both methods exhibit a high degree of statistical parity with HEFS, and the absolute differences in performance between them is small (see Text S9 in Supporting Information S1). Additionally, a more detailed analysis of reservoir operations at specific sites reveals serious challenges for syn-M1, particularly at locations with flashy hydrology like Prado Dam and New Hogan Lake (see Text S7 in Supporting Information S1). Overall, these results indicate that while syn-M2 does not completely dominate syn-M1 on all operations-based validation metrics across sites and scenarios, it consistently outperforms syn-M1 on balance.

## 5. Discussion and Conclusions

This study contributes an operations-based validation framework for synthetic ensemble forecasts to evaluate their suitability for robustness testing of forecast-informed reservoir polices. The framework employs probabilistic ensemble verification tools to evaluate the operational outcomes of synthetic forecasts generated by FIRO policies optimized using available hindcast data (Delaney et al., 2020; Taylor et al., 2024). Unlike traditional synthetic forecast verification, which compares synthetic flow forecasts to hindcasts of streamflow, operations-based validation more directly assesses whether synthetic forecasts are fit for their intended purpose. Additionally, this work introduces a new synthetic ensemble forecasting algorithm (syn-M2), which serves as an alternative to the synthetic ensemble forecasting algorithm (syn-M1) developed by Brodeur et al. (2024). The two algorithms were compared through operations-based validation across multiple locations in California, each characterized by different hydrologic and system attributes, to evaluate the generalizability of both methods.

The results indicate that syn-M2 outperforms syn-M1 in both forecast verification and operations-based validation. At certain locations, such as Lake Oroville, syn-M2 consistently excelled across all measures of storage, release, and release lead outcomes. The synthetic forecast ensembles approached near-ideal behavior, even during rare flood events, indicating that syn-M2 produces operational outcomes that are difficult to statistically distinguish from those generated by HEFS hindcasts. At other locations, some trade-offs emerged between syn-M2 and syn-M1 across operational variables (storage, release, release lead), and both algorithms deviated somewhat from the hindcast behavior. Still, syn-M2 consistently demonstrated superior performance to syn-M1 on balance, and it exhibited minimal sensitivity to the hydrologic and system attributes of the locations where it was applied, underscoring its generalizability.

Ultimately, the operations-based validation presented in this study aims to differentiate between various synthetic forecasting approaches, quantify improvements among them, and determine whether the synthetic forecasts are "good enough" for operational use. We contend that the suite of operations-based validation tools developed here provides multiple lines of evidence to address these questions and enhances the role of synthetic forecasts as a practical risk analysis tool for FIRO. Given its strong performance in this study, we propose that future work should consider the syn-M2 algorithm for future FIRO robustness testing. In addition to its computational efficiency, syn-M2 is also readily extensible to multisite applications and designed to preserve spatial correlations within its sampling framework. Thus, it is likely well-suited for the multisite analyses anticipated in both research initiatives (Taylor et al., 2024) and real-world FIRO implementations (Talbot et al., 2023), although future research is needed to fully test this potential. Additionally, while this study focused on California watersheds, the strong generalizability observed for syn-M2 suggests that this method could be applied effectively in other regions.

Future efforts to expand forecast-informed operations in reservoir systems should begin using synthetic forecasting methods like syn-M2, or other approaches validated as fit for purpose, to ensure that operating policies are robust across a wide range of potential flood events, forecasts, and patterns of forecast error. While prior literature has examined the robustness of operating policies through synthetic forecasts in research contexts (Lamontagne & Stedinger, 2018; Nayak et al., 2018; Rougé, 2021), this tool has not yet been implemented in practice. The results of this study demonstrated that policies trained on hindcasts can lead to undesirable outcomes, such as spills, when confronted with plausible alternative forecast sequences. Although the simulations conducted were based on a deliberately unlikely scenario without flood pools—intended to test the synthetic forecasts under high-risk

19447973, 2025, 6, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024WR039324 by Test, Wiley Online Library on [02/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

19447973, 2025, 6, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024WR039324 by Test, Wiley Online Library on [02/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

conditions - our findings nonetheless underscore the need for comprehensive robustness testing to ensure FIRO policies function effectively in real-world settings.

If used in practice, synthetic forecasting should be integrated into the early stages of FIRO viability assessments where feasible. A collaborative approach between research and operations will not only broaden the potential applications of synthetic forecasts but will also uncover new challenges for researchers to address. For example, our recent interactions with practitioners revealed a need for hourly synthetic forecasts that reflect the uncertainty inherent in operational forecasts, incorporating human adjustments rather than relying solely on frozen model hindcasts, as these are the forecasts used by reservoir operators in practice. Furthermore, this work focused solely on the performance of synthetic ensemble forecasts within a single policy optimization method, specifically the EFO approach of Delaney et al. (2020). Future research should investigate the operational performance of synthetic forecasts under alternative policy frameworks, such as those outlined in recent water control manual updates (Folsom Dam; USACE, 2019), which may highlight different attributes of the synthetic ensembles.

Synthetic forecasts also hold promise to address key research questions. For instance, FIRO may be an effective climate adaptation strategy, especially as machine learning bolsters the skill of conventional dynamical forecasting methods (Lam et al., 2022). However, few if any studies to date have been able to explore this potential. Synthetic forecasts developed with simulated "observations" under future climate change scenarios (Thyer et al., 2024) and with modified attributes of forecast skill provide a unique capability in this context (Lamontagne & Stedinger, 2018; Rougé et al., 2023). These methods can advance the science of climate-adaptive FIRO policies and enable testing of these policies under varying levels of forecasting skill improvement. However, applying these methods would require techniques to address structural biases between observed and simulated flows in the historical record, as such biases could influence the calibration of synthetic forecasting models. Additionally, methods to account for changing forecast-error structures—such as those arising from shifts in hydrologic regimes (e.g., from snow-dominated to rain-driven systems)—will be critical for robust future applications.

Synthetic forecasts can also be generated at longer lead times (e.g., sub-seasonal to seasonal), enabling testing of water supply focused operations that are influenced by non-flood hydrometeorology, such as droughts. Additionally, hybrid approaches that integrate short- and long-lead forecasting—for example, FIRO paired with Managed Aquifer Recharge (FIRO-MAR; CADWR, 2019)—stand to benefit from synthetic forecasts that span multiple timescales. Synthetic ensemble forecasts, applied to both streamflow and climate variables (see Brodeur & Steinschneider, 2021), also hold promise for broader forecast-informed system management, including hydropower scheduling and renewable energy integration in systems reliant on variable, non-dispatchable renewables.

Despite the promising outcomes of this work, several important limitations require discussion. Our analysis relied on approximately 30 years of high-quality hindcasts generated from frozen versions of the climate and hydrologic models used to support HEFS. Such long, consistent, and high-quality hindcast data sets may not be available in other regions, posing challenges for broader applications of synthetic forecasting methods. In previous work (Brodeur & Steinschneider, 2021), we found that synthetic forecasting performance begins to degrade with fewer than 15 years of hindcast data and is significantly compromised with less than 5 years. Although this study employs a 5-fold cross-validation approach - effectively using about 24 years of data per training fold - this still does not address the performance limitations associated with shorter data sets.

Additionally, in many regions, only archived operational forecasts are available, rather than hindcasts from a frozen model version. These archived data sets often reflect the effects of iterative model upgrades and human-in-the-loop adjustments. As a result, synthetic forecasts derived from them may resample heterogeneous skill characteristics, complicating their use for consistent operations-based testing. These issues underscore the challenges of applying synthetic forecasting in the absence of full-length, high-quality hindcasts like those used in this study. Future work is needed to address these challenges. Additionally, these issues highlight the value of high-quality hindcasts where they do exist and provide strong rationale for increased investment in their development and maintenance.

19447973, 2025, 6, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024WR039324 by Test, Wiley Online Library on [02/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

## Data Availability Statement

All code and data for the synthetic ensemble forecast models (syn-M1 and syn-M2) are available in Brodeur (2025a, 2025b) while code and data for forecast verification and the multisite reservoir simulation model are available in Brodeur (2025c).

**Acknowledgments**

This work was supported by the U.S. Army Corps of Engineers Engineer Research and Development Center Forecast Informed Reservoir program (Award USACE W912HZ1920023) and the U.S. National Science Foundation (award no. 2205239). We thank Brett Whitin from the CNRFC for provisioning the HEFS hindcasts and observations for this study.

## References

Abbaszadeh, P., Gavahi, K., & Moradkhani, H. (2020). Multivariate remotely sensed and in-situ data assimilation for enhancing community WRF-Hydro model forecasting. *Advances in Water Resources*, *145*, 103721. https://doi.org/10.1016/j.advwatres.2020.103721

American Meteorological Society. (2025). *Forecast-informed reservoir operations*. AMS Glossary of Meteorology. Retrieved from https://glossary.ametsoc.org/wiki/Forecast-informed_reservoir_operations

Badrinath, A., Delle Monache, L., Hayatbini, N., Chapman, W., Cannon, F., & Ralph, M. (2023). Improving precipitation forecasts with convolutional neural networks. *Weather and Forecasting*. https://doi.org/10.1175/WAF-D-22-0002.1

Blum, A. G., & Miller, A. (2019). Opportunities for forecast-informed water resources management in the United States. *Bulletin of the American Meteorological Society*, *100*(10), 2087–2090. https://doi.org/10.1175/BAMS-D-18-0313.1

Brodeur, Z. (2025a). zpb4/Synthetic-Forecast-v1-FIRO-DISES: May 20, 2025 release (v1.0.0) [R; GitHub]. Zenodo. https://doi.org/10.5281/zenodo.15477173

Brodeur, Z. (2025b). zpb4/Synthetic-Forecast-v2-FIRO-DISES: May 20, 2025 release (v1.0.0) [R; GitHub]. Zenodo. https://doi.org/10.5281/zenodo.15477332

Brodeur, Z. (2025c). zpb4/FIRO_syn-forecast_ops-validation: May 20, 2025 Release (v1.0.0) [Python; GitHub]. Zenodo. https://doi.org/10.5281/zenodo.15477575

Brodeur, Z. P., Delaney, C., Whitin, B., & Steinschneider, S. (2024). Synthetic forecast ensembles for evaluating forecast informed reservoir operations. *Water Resources Research*, *60*(2), e2023WR034898. https://doi.org/10.1029/2023WR034898

Brodeur, Z. P., & Steinschneider, S. (2021). A multivariate approach to generate synthetic short-to-medium range hydro-meteorological forecasts across locations, variables, and lead times. *Water Resources Research*, *57*(6). https://doi.org/10.1029/2020WR029453

Brown, C. M., Lund, J. R., Cai, X., Reed, P. M., Zagona, E. A., Ostfeld, A., et al. (2015). The future of water resources systems analysis: Toward a scientific framework for sustainable water management. *Water Resources Research*, *51*(8), 6110–6124. https://doi.org/10.1002/2015WR017114

Brown, J. D., He, M., Regonda, S., Wu, L., Lee, H., & Seo, D.-J. (2014). Verification of temperature, precipitation, and streamflow forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS): 2. Streamflow verification. *Journal of Hydrology*, *519*, 2847–2868. https://doi.org/10.1016/j.jhydrol.2014.05.030

CADWR. (2019). *Flood-MAR research and data development plan: Priority actions to expand implementation of effective and efficient flood-MAR projects in California*. California Department of Water Resources. Retrieved from https://water.ca.gov/-/media/DWR-Website/Web-Pages/Programs/Flood-Management/Flood-MAR/Flood-MAR-RDD-Plan_a_y_19.pdf?la=en&hash=F87030022485CF01BC6E9D6AE4C134818A4FCB0E

Cassagnole, M., Ramos, M.-H., Zalachori, I., Thirel, G., Garçon, R., Gailhard, J., & Ouillon, T. (2020). Impact of the quality of hydrological forecasts on the management and revenue of hydroelectric reservoirs—A conceptual approach. *Hydrology and Earth System Sciences Discussions*, *25*(2), 1033–1052. https://doi.org/10.5194/hess-2020-410

CNRFC. (2022). *Ensemble forecasting with the Hydrologic Ensemble Forecast Service (HEFS)*. California-Nevada River Forecast Center.

Delaney, C. J., Robert, H., Hartman, R., Mendoza, J., Dettinger, M. D., Delle Monache, L., et al. (2020). Forecast informed reservoir operations using ensemble streamflow predictions for a multipurpose reservoir in northern California. *Water Resources Research*, *56*(9), 26604. https://doi.org/10.1029/2019wr026604

Demargne, J., Wu, L., Wu, L., Satish, K. R., Regonda, S., James, D. B., et al. (2014). The science of NOAA's operational hydrologic ensemble forecast Service. *Bulletin of the American Meteorological Society*, *95*(1), 79–98. https://doi.org/10.1175/bams-d-12-00081.1

Di Baldassarre, G., Sivapalan, M., Rusca, M., Cudennec, C., Garcia, M., Kreibich, H., et al. (2019). Sociohydrology: Scientific challenges in addressing the sustainable development goals. *Water Resources Research*, *55*(8), 6327–6355. https://doi.org/10.1029/2018WR023901

Faber, B. A., & Stedinger, J. R. (2001). Reservoir optimization using sampling SDP with ensemble streamflow prediction (ESP) forecasts. *Journal of Hydrology*, *249*(1–4), 113–133. https://doi.org/10.1016/S0022-1694(01)00419-X

Giuliani, M., Zaniolo, M., Castelletti, A., Davoli, G., & Block, P. (2019). Detecting the state of the climate system via artificial intelligence to improve seasonal forecasts and inform reservoir operations. *Water Resources Research*, *55*(11), 9133–9147. https://doi.org/10.1029/2019WR025035

Grygier, J. C., Stedinger, J. R., & Yin, H. (1989). A generalized maintenance of variance extension procedure for extending correlated series. *Water Resources Research*, *25*(3), 345–349. https://doi.org/10.1029/WR025i003p00345

Guan, H., Zhu, Y., Sinsky, E., Fu, B., Li, W., Zhou, X., et al. (2022). GEFSv12 reforecast dataset for supporting subseasonal and hydrometeorological applications. *Monthly Weather Review*, *150*(3), 647–665. https://doi.org/10.1175/MWR-D-21-0245.1

Hejazi, M. I., Cai, X., Yuan, X., Liang, X.-Z., & Kumar, P. (2014). Incorporating reanalysis-based short-term forecasts from a regional climate model in an irrigation scheduling optimization problem. *Journal of Water Resources Planning and Management*, *140*(5), 699–713. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000365

Huang, Z., & Zhao, T. (2022). Predictive performance of ensemble hydroclimatic forecasts: Verification metrics, diagnostic plots and forecast attributes. *WIREs Water*, *9*(2), e1580. https://doi.org/10.1002/wat2.1580

Jasperse, J., Ralph, F. M., Anderson, M., Brekke, L., Malasavage, N., Dettinger, M. D., et al. (2020). *Lake Mendocino forecast informed reservoir operations: Final viability assessment*. UC San Diego. Retrieved from https://escholarship.org/uc/item/3b63q04n

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., et al. (2022). GraphCast: Learning skillful medium-range global weather forecasting (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2212.12794

Lamontagne, J. R., & Stedinger, J. R. (2018). Generating synthetic streamflow forecasts with specified precision. *Journal of Water Resources Planning and Management*, *144*(4), 04018007. https://doi.org/10.1061/(asce)wr.1943-5452.0000915

Lavers, D. A., Ralph, F. M., Richardson, D. S., & Pappenberger, F. (2020). Improved forecasts of atmospheric rivers through systematic reconnaissance, better modelling, and insights on conversion of rain to flooding. *Communications Earth & Environment*, *1*(1), 39. https://doi.org/10.1038/s43247-020-00042-1

Lettenmaier, D. P. (1984). Synthetic streamflow forecast generation. *Journal of Hydraulic Engineering*, *110*(3), 277–289. https://doi.org/10.1061/(ASCE)0733-9429(1984)110:3(277)

Nayak, M. A., Herman, J. D., & Steinschneider, S. (2018). Balancing flood risk and water supply in California: Policy search integrating short-term forecast ensembles with conjunctive use. *Water Resources Research*, *54*(10), 7557–7576. https://doi.org/10.1029/2018WR023177

Nohara, D., Takemon, Y., & Sumi, T. (2020). Real-time flood management and preparedness: Lessons from floods across the western Japan in 2018. In P. Gourbesville & G. Caignaert (Eds.), *Advances in hydroinformatics* (pp. 287–304). Springer. https://doi.org/10.1007/978-981-15-5436-0_22

Porter, J., Day, G., Schaake, J. C., & Wang, L. (2018). New York City's operations support tool: Utilizing hydrologic forecasts for water supply management. In Q. Duan, F. Pappenberger, J. Thielen, A. Wood, H. L. Cloke, & J. C. Schaake (Eds.), *Handbook of hydrometeorological ensemble forecasting* (pp. 1–42). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-40457-3_56-1

Ralph, F. M., Hutchinson, A., Anderson, M., Fairbank, T., Forbis, J., Haynes, A., et al. (2023). *Prado dam forecast informed reservoir operations: Final viability assessment*. UC San Diego. Retrieved from https://cw3e.ucsd.edu/FIRO_docs/Prado/FIRO_Prado_FVA.pdf

Ralph, F. M., James, J., Leahigh, J., Anderson, M., Forbis, J., Haynes, A., et al. (2022). *Yuba-feather forecast informed reservoir operations: Preliminary viability assessment*. UC San Diego. Retrieved from https://cw3e.ucsd.edu/FIRO_docs/Yuba-Feather_PVA.pdf

Ralph, F. M., Rutz, J. J., Cordeira, J. M., Dettinger, M., Anderson, M., Reynolds, D., et al. (2019). A scale to characterize the strength and impacts of atmospheric rivers. *Bulletin of the American Meteorological Society*, *100*(2), 269–289. https://doi.org/10.1175/BAMS-D-18-0023.1

Rougé, C. (2021). Generating families of synthetic forecasts of different skills from an existing forecast product. https://doi.org/10.5194/egusphere-egu21-12367

Rougé, C., Peñuela, A., & Pianosi, F. (2023). Forecast families: A new method to systematically evaluate the benefits of improving the skill of an existing forecast. *Journal of Water Resources Planning and Management*, *149*(5). https://doi.org/10.1061/jwrmd5.wreng-5934

Shabestanipour, G., Brodeur, Z., Farmer, W. H., Steinschneider, S., Vogel, R. M., & Lamontagne, J. R. (2023). Stochastic watershed model ensembles for long-range planning: Verification and validation. *Water Resources Research*, *59*(2), e2022WR032201. https://doi.org/10.1029/2022WR032201

Stedinger, J. R., & Taylor, M. R. (1982). Synthetic streamflow generation: 1. Model verification and validation. *Water Resources Research*, *18*(4), 909–918. https://doi.org/10.1029/WR018i004p00909

Storn, R., & Price, K. (1997). Differential Evolution: A simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, *11*(4), 341–359. https://doi.org/10.1023/A:1008202821328

Sukovich, E. M., Ralph, F. M., Barthold, F. E., Reynolds, D. W., & Novak, D. R. (2014). Extreme quantitative precipitation forecast performance at the weather prediction center from 2001 to 2011. *Weather and Forecasting*, *29*(4), 894–911. https://doi.org/10.1175/WAF-D-13-00061.1

Talbot, C., Forbis, J., & Ralph, M. (2023). Phase III of forecast-informed reservoir operations in the USACE: National expansion pathfinder. *ASCE Inspire*, 970–978. https://doi.org/10.1061/9780784485163.111

Taylor, W., Brodeur, Z. P., Steinschneider, S., Kucharski, J., & Herman, J. D. (2024). Variability, attributes, and drivers of optimal forecast-informed reservoir operating policies for water supply and flood control in California. *Journal of Water Resources Planning and Management*, *150*(10), 05024010. https://doi.org/10.1061/JWRMD5.WRENG-6471

Teegavarapu, R. S. V., Salas, J. D., & Stedinger, J. R., & American Society of Civil Engineers, & Environmental and Water Resources Institute (U. S.). (2019). *Statistical analysis of hydrologic variables: Methods and applications*. American Society of Civil Engineers.

Thaler, T. (2021). Social justice in socio-hydrology—How we can integrate the two different perspectives. *Hydrological Sciences Journal*, *66*(10), 1503–1512. https://doi.org/10.1080/02626667.2021.1950916

Thyer, M., Gupta, H., Westra, S., McInerney, D., Maier, H. R., Kavetski, D., et al. (2024). Virtual Hydrological Laboratories: Developing the next generation of conceptual models to support decision making under change. *Water Resources Research*, *60*(4), e2022WR034234. https://doi.org/10.1029/2022wr034234

Troin, M., Arsenault, R., Wood, A. W., Brissette, F., & Martel, J. (2021). Generating ensemble streamflow forecasts: A review of methods and approaches over the past 40 years. *Water Resources Research*, *57*(7). https://doi.org/10.1029/2020WR028392

USACE. (2019). *Folsom dam modification project water control manual update: Final supplemental environmental assessment/environmental impact report*. No. SCH # 2012102034 (p. 218). U.S. Army Corps of Engineers. Retrieved from https://www.govinfo.gov/content/pkg/GOVPUB-D103-PURL-gpo133465/pdf/GOVPUB-D103-PURL-gpo133465.pdf

Wilks, D. S. (2019). *Statistical methods in the atmospheric sciences* (4th ed.). Elsevier.

Woodside, G. D., Hutchinson, A. S., Ralph, F. M., Talbot, C., Hartman, R., & Delaney, C. (2022). Increasing stormwater capture and recharge using forecast informed reservoir operations, Prado dam. *Groundwater*, *60*(5), 634–640. https://doi.org/10.1111/gwat.13162

19447973, 2025, 6, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024WR039324 by Test, Wiley Online Library on [02/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License