Water Resources Research

RESEARCH ARTICLE
10.1029/2023WR034898

Synthetic Forecast Ensembles for Evaluating Forecast Informed Reservoir Operations

Zachary P. Brodeur[1], Chris Delaney[2], Brett Whitin[3], and Scott Steinschneider[1]

[1]Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY, USA, [2]Center for Western Weather and Water Extremes (CW3E), Scripps Institute of Oceanography, University of California-San Diego, La Jolla, CA, USA, [3]NOAA/NWS California-Nevada River Forecast Center (CNRFC), Sacramento, CA, USA

Key Points:

- We develop a method to produce synthetic hydrologic ensemble forecasts based on hindcast ensembles
- Synthetic and actual forecast ensembles exhibit good agreement based on ensemble forecast verification statistics
- Synthetic forecasts can extend robustness testing of forecast informed reservoir operations to better estimate out-of-sample performance

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

Received 15 MAR 2023
Accepted 17 JAN 2024

Author Contributions:

Conceptualization: Zachary P. Brodeur, Chris Delaney, Brett Whitin, Scott Steinschneider
Data curation: Chris Delaney, Brett Whitin
Formal analysis: Zachary P. Brodeur, Chris Delaney, Brett Whitin, Scott Steinschneider
Funding acquisition: Scott Steinschneider
Investigation: Zachary P. Brodeur, Chris Delaney, Scott Steinschneider
Methodology: Zachary P. Brodeur, Chris Delaney, Brett Whitin, Scott Steinschneider

Abstract Forecast informed reservoir operations (FIRO) is an important advance in water management, but the design and testing of FIRO policies is limited by relatively short (10–35 year) hydro-meteorological hindcasts. We present a novel, multisite model for synthetic forecast ensembles to overcome this limitation. This model utilizes parametric and non-parametric procedures to capture complex forecast errors and maintain correlation between variables, lead times, locations, and ensemble members. After being fit to data from the hindcast period, this model can generate synthetic forecast ensembles in any period with observations. We demonstrate the approach in a case study of the FIRO-based Ensemble Forecast Operations (EFO) control policy for the Lake Mendocino—Russian River basin, which conditions release decisions on ensemble forecasts from the Hydrologic Ensemble Forecast System (HEFS). We explore two generation strategies: (a) simulation of synthetic forecasts of meteorology to force HEFS; and (b) simulation of synthetic HEFS streamflow forecasts directly. We evaluate the synthetic forecasts using ensemble verification techniques and event-based validation, finding good agreement with the actual ensemble forecasts. We then evaluate EFO policy performance using synthetic and actual forecasts over the hindcast period (1985–2010) and synthetic forecasts only over the pre-hindcast period (1948–1984). Results show that the synthetic forecasts highlight important failure modes of the EFO policy under plausible forecast ensembles, but improvements are still needed to fully capture FIRO policy behavior under the actual forecast ensembles. Overall, the methodology advances a novel way to test FIRO policy robustness, which is key to building institutional support for FIRO.

1. Introduction

Forecast informed reservoir operations (FIRO) is a flexible reservoir operations strategy that leverages hydro-meteorological forecasts to better inform release decisions (AMS, 2020). Recent assessments of FIRO using state-of-the-art hydro-meteorological medium-range (1–14 day) forecasts have demonstrated substantial gains in water supply, water quality, and environmental flow performance while maintaining appropriate safety margins for flood risk (Jasperse et al., 2020). These results highlight FIRO as a soft water path (Gleick, 2002) toward preserving water supplies while mitigating flood risk, both under current conditions and under intensifying extremes projected under climate change (Aghakouchak et al., 2020; IPCC, 2021; Swain et al., 2018).

Forecast uncertainty has long been a concern for forecast informed operating policies (Rayner et al., 2005; Todini, 2018; Whateley et al., 2014), slowing their uptake and implementation despite gradual but consistent improvement in forecast skill (Bauer et al., 2015; Blum & Miller, 2019; C. M. Brown et al., 2015). Forecast uncertainty can be incorporated into policy design using ensemble forecasting and a number of control policy methods, including stochastic dynamic programming (Côté & Leconte, 2016; Kim et al., 2007; Turner et al., 2017), stochastic model predictive control (Castelletti et al., 2023; Ficchi et al., 2016; Raso et al., 2014), and heuristic methods (Delaney et al., 2020; Nayak et al., 2018; Semmendinger et al., 2022). For approaches that calibrate a control policy to hindcast data (e.g., stochastic dynamic programming; heuristic methods), FIRO policies can be overfit to the specific forecast errors in the hindcast period. The performance evaluation of FIRO based operations is also dependent on those same hindcasts, regardless of method.

Commonly, medium-range hydrological hindcasts are forced with meteorological hindcasts that come either directly from a global forecast model or from a downscaled regional forecast model. The computational burden associated with these models often limits hindcasts to a single model run with one small set of perturbed members (Guan et al., 2019). These members are intended to represent hindcast uncertainty, but this representation is limited by the small ensemble size. This issue is most acute for forecast uncertainty around extreme events,

Resources: Chris Delaney, Brett Whitin, Scott Steinschneider
Software: Zachary P. Brodeur, Chris Delaney
Supervision: Scott Steinschneider
Validation: Zachary P. Brodeur, Chris Delaney, Brett Whitin, Scott Steinschneider
Visualization: Zachary P. Brodeur, Chris Delaney
Writing – original draft: Zachary P. Brodeur
Writing – review & editing: Zachary P. Brodeur, Chris Delaney, Brett Whitin, Scott Steinschneider

which are inherently limited in number. Furthermore, the observations required to initialize meteorological forecast models are only available from the satellite era (i.e., starting around 1979), limiting the timespan for hindcasts (Hartmann, 2016). Some advanced forecast models in operational use have even shorter hindcast periods (e.g., 1989—present for NCEP-GEFS v12, Guan et al., 2019). While data-driven streamflow forecasting can alleviate some of the computational burdens around hindcast development, they either are limited to using antecedent streamflow, basin conditions, and climatological weather sequences as inputs, thereby limiting skill for medium-range forecasting (Troin et al., 2021), or they use remotely sensed or climate model forecast information as input and therefore cannot extend prior to the satellite era (Nevo et al., 2022; Slater et al., 2023).

The inability of available hindcasts to fully capture forecast uncertainty complicates FIRO policy adoption in regions like northern California, where a majority of water is delivered in a small number of large storms per year (Corringham et al., 2019; Dettinger, 2013; Dettinger et al., 2011; Hanak et al., 2011; Jasperse et al., 2020; Ralph et al., 2020). In this setting, large over-forecasts of a single storm might cause an excessive drawdown of valuable water supply, while large under-forecasts might lead to spills, structural damage, or downstream flooding. A reliable FIRO policy should be robust to these forecast errors and needs to accommodate a range of antecedent conditions and system states that are possible preceding an event, as well as a diversity of plausible forecast progressions leading up to an event. If only one hindcast progression for past extreme events is used to evaluate a FIRO policy, there is considerable risk that historic performance will not generalize well to future new forecasts and future extreme events.

Synthetic forecasts provide a means to reduce this risk (Alemu et al., 2011; Brodeur & Steinschneider, 2021; Lamontagne & Stedinger, 2018; Lettenmaier, 1984; Nayak et al., 2018; Raso et al., 2014; Turner et al., 2017). Synthetic forecasts emulate the behavior of hindcasts using statistical models of forecast error. By combining many forecast error simulations with observed data, FIRO policies can be evaluated over a large variety of ensemble forecast samples to evaluate their reliability (Nayak et al., 2018). In addition, synthetic forecasts can generate plausible forecasts during periods with observations but without hindcasts. This substantially increases the available sample size for FIRO policy calibration, validation, and testing, including a more diverse set of extreme events.

However, generating synthetic hydrologic forecasts is a "deceptively challenging problem" (Lamontagne & Stedinger, 2018), even for a single value forecast at a specified lead time. Forecast errors (and hydrologic model errors more generally) tend to be autocorrelated, heteroscedastic, and non-Gaussian (McInerny et al., 2017; Schoups & Vrugt, 2010), and they can be correlated with the observations (Lamontagne & Stedinger, 2018). This problem becomes far more complex when the forecast data contain multiple lead times and sites, because error dependencies exist both endogenously (e.g., autocorrelation) and exogenously (e.g., between errors at different lead times and locations).

Recently, Brodeur and Steinschneider (2021) developed a multivariate synthetic forecast methodology to model these types of complex error structures and demonstrated its effectiveness in replicating the uncertainty in medium-range ensemble mean forecasts. However, emerging FIRO policies (Corringham et al., 2019; Delaney et al., 2020; Jasperse et al., 2020) operate not on the ensemble mean, but rather on the full forecast ensemble. Some recent studies have explored synthetic generation of ensemble forecasts to evaluate the value of improved forecast information (Bertoni et al., 2021; Cassagnole et al., 2021; Rougè et al., 2023). However, Bertoni et al. (2021) only evaluated seasonal timescales, where the distribution of forecast errors is simpler to model. Cassagnole et al. (2021) and Rougè et al. (2023) synthetically generated daily ensemble streamflow forecasts at multiple lead times designed to mimic the behavior of an operational forecasting system. However, the approaches in both of these studies require that the synthetic forecasts match the temporal evolution of the available hindcast and be restricted to the historical period when hindcasts are available. In addition, Cassagnole et al. (2021) assumes (without clear justification) that forecasts are log-normally distributed around observed flows, and only evaluated reservoir performance using the ensemble mean of their synthetic forecasts. To date, there are no synthetic forecast approaches that can (a) emulate the nuanced behavior of modern daily medium-range hydrologic ensemble forecast systems (HEFSs), while also being able to (b) generate novel forecasting sequences with different timing of events compared to the hindcast, and (c) generate synthetic forecasts for extreme events outside of the hindcast period. The work in Brodeur and Steinschneider (2021) can achieve the latter two points, but not the former (i.e., not for ensemble forecasts).

This work builds upon Brodeur and Steinschneider (2021) to develop a novel multivariate procedure to generate synthetic hydrologic ensemble forecasts. The method emulates complex forecast error relationships within and between sites and lead times, captures correlations between ensemble members, and can generate alternative

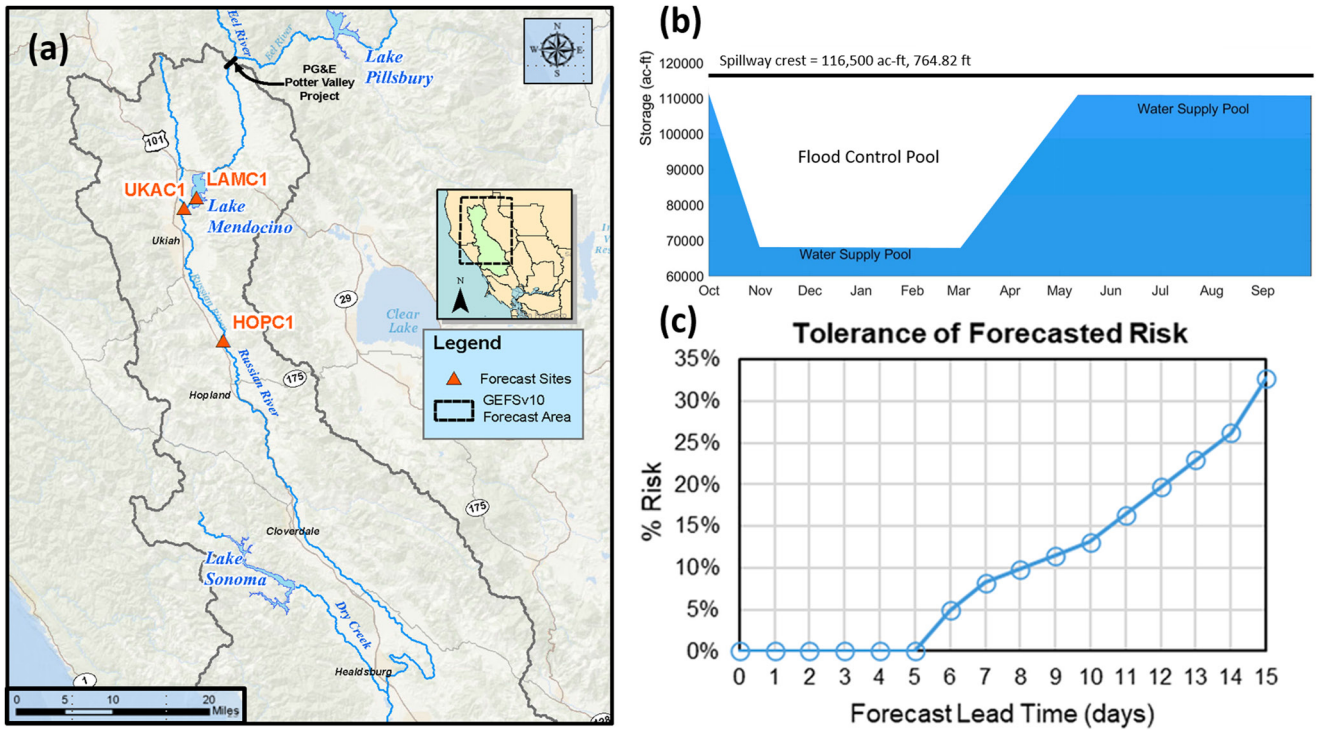

Figure 1. (a) Geographical area of study (Russian River watershed). The three forecast sites (LAMC1, UKAC1, HOPC1) are indicated in red and the 1° grid cell of the GEFSv10 forecast centered at 39°N 123°W is shown in the inset. Note: The sub-watershed boundaries for the three forecast sites are all contained within the single GEFSv10 1° grid cell. (b) Water supply and flood control pool at Lake Mendocino across the year, along with level of spillway crest. (c) The risk tolerance curve used to determine flood control releases.

forecasting sequences around events both in and outside of the hindcast period. We demonstrate this approach in a case study of the Russian River—Lake Mendocino system using the recently developed Ensemble Forecast Operations (EFO) policy being tested as part of the multi-agency FIRO initiative in the western US (Jasperse et al., 2020). For both calibration and operations, the EFO model ingests hydrologic ensemble forecasts from the National Weather Service (NWS) HEFS, which is forced by ensemble mean meteorological data from the NOAA/NWS Global Ensemble Forecast System (GEFS).

We compare two approaches to generate synthetic forecasts. In the first, we fit a model directly to the HEFS ensemble output, while in the second, we generate synthetic GEFS meteorological inputs that are used to force HEFS to create hydrologic ensemble forecasts. We apply ensemble verification techniques (Wilks, 2019) to evaluate how well the synthetic ensemble forecasts mimic the actual ensemble forecasts, and we investigate simulated ensemble behavior around extreme inflow events. We also evaluate the synthetic forecasts by using them as inputs to the EFO model over the available hindcast period (1985–2010), thereby validating the synthetic forecasts in an operational context (Stedinger & Taylor, 1982). Finally, we use the synthetic forecasts to explore how FIRO-based policies calibrated to the hindcasts perform during the pre-hindcast period of 1948–1984. Ultimately, the assessments performed in this work seek to advance synthetic forecasts as a tool for building trust in the continued development and spread of FIRO for sustainable water management.

2. FIRO Background and Case Study

2.1. Lake Mendocino, HEFS, and the EFO Policy

The Russian River watershed in northwestern California is a large watershed (3,850 km^2) with headwaters that drain into Lake Mendocino, a 116,500 acre-foot reservoir that is operated for flood control, water supply, and hydropower, and includes a water supply pool ranging between 68,400 and 111,000 acre-feet throughout the year (see Figure 1). The watershed sits on California's northwest coast and receives the majority of its annual inflow from a small number of atmospheric rivers in the cold season (October-March). Hydrologic forecasts at 1–14 day

lead times in the Russian River are developed through the NWS HEFS (Demargne et al., 2014). This hydrologic forecasting system is forced by NCEP GEFS ensemble mean forecasts of maximum and minimum temperature and precipitation. The HEFS utilizes the Sacramento Soil Moisture Accounting (SAC-SMA) hydrologic model to convert these weather forecasts into forecasted hydrologic response at key inflow points of interest. HEFS produces an internal ensemble of perturbed meteorological data (mean areal temperature (MAT)/precipitation, MAT/mean areal precipitation (MAP)) from the GEFS ensemble mean forecast, creating an ensemble of forecast members that are less biased and more reliable than the raw GEFS ensemble (Demargne et al., 2014).

The EFO policy developed by Delaney et al. (2020) for Lake Mendocino is a heuristic control policy that uses pre-defined rules to process ensemble output from HEFS, thereby conditioning releases on both expected inflow and uncertainty associated with the full ensemble. We provide a brief overview of this policy here, but direct the reader to Delaney et al. (2020) for more detail. The model used to simulate storage (S_t) in Lake Mendocino is based on a simple water balance calculation:

$$S_t = S_{t-1} + Q_t - E_t - L_t - R_t^{\text{Spill}} - R_t^{\text{Cntr}} \tag{1}$$

$$R_t^{\text{Ctrl}} = \max\left(\max\left(R_t^{\text{Flood}}, R_t^{\text{WS}}\right), R_t^{\text{Emgc}}\right) \tag{2}$$

Here, Q_t is the daily reservoir inflow, E_t is evaporative loss, and L_t represents water diversions upstream. Releases from the system are separated into controlled releases (R_t^{Ctrl}) and uncontrolled spills (R_t^{Spill}). Spills are triggered when storage exceeds 116,500 acre-feet (the spillway crest) and reach a maximum of 47,300 cfs when water levels reach the top of the dam (153,700 acre-feet). Controlled releases are calculated after accounting for changes in storage due to spill and are based on flood control releases (R_t^{Flood}), water supply releases (R_t^{WS}), and controlled emergency releases (R_t^{Emgc}).

Under the EFO policy, flood control releases (R_t^{Flood}) are made if ensemble streamflow forecasts suggest there is a risk that storage will exceed 111,000 acre-feet (the top of the water supply pool in the dry season; Figure 1b). These releases are based on the ensemble forecast through a risk tolerance curve (Figure 1c), which determines the allowable fraction of ensemble members that can exceed the 111,000 acre-feet storage limit before releases must be increased to mitigate flood risk. For each lead time, the model simulates the ending storage at that lead time for each of the ensemble forecast members, assuming no flood control releases are made. It then calculates the percentage of ensemble members that lead to storage exceeding 111,000 acre-feet. If that percentage exceeds the risk tolerance curve for that lead time, proposed flood control releases R_t^{Flood} are calculated that would reduce that percentage below the risk tolerance curve. These proposed values are then constrained not to exceed maximum release limits, which consider maximum flood control release regulations, the capacity of the controlled outlet, and the downstream flow limit of 8,000 cfs at the Hopland site (see Figure 1 and Text S1 in Supporting Information S1). This is repeated separately for each lead time, and the largest flood control release calculated across lead times is selected as the final value of R_t^{Flood}.

For Lake Mendocino, the risk tolerance curve was calibrated separately by lead time out to 14 days. As shown in Figure 1c, there is no risk appetite at short leads (0% risk threshold for lead times of 1–5 days), but allowable risk increases gradually with lead time as forecasts become more uncertain. Note that under the EFO policy, flood releases can cause storage to drop into the water supply pool in anticipation of a flood. The EFO policy can also be adapted into a perfect forecast operations (PFO) policy simulating whether perfect inflow information over the next 14 days would lead to storage exceeding 111,000 acre-feet, and if so, increasing flood control releases R_t^{Flood} to ensure storage does not exceed this threshold. In this case, there is only a single forecast member (the perfect inflow forecast), and the risk tolerance is fixed at 0% for all lead times.

To define the final controlled release R_t^{Ctrl}, the policy first takes the maximum between R_t^{Flood} and designated water supply releases (R_t^{WS}), which vary throughout the year and are set to meet downstream agricultural and domestic water demands and maintain minimum in-stream environmental flows in the Upper Russian River (Sonoma Water, 2016). The final controlled release is then set to the maximum of that result and R_t^{Emgc}, which specifies emergency releases conditioned on storage, as specified in the water control manual for Lake Mendocino (USACE, 2003). Additional detail on water balance and release terms $\left(E_t, L_t, R_t^{\text{WS}}, R_t^{\text{Emgc}}, R_t^{\text{Spill}}\right)$, as well as additional system model constraints (e.g., maximum release rates), can be found in Text S1 in Supporting Information S1.

2.2. Data

Synthetic forecasts in this work are based on a 61-member HEFS ensemble hindcast at three locations in the Russian River Basin: Lake Mendocino (LAMC1) inflows and downstream local flows at the Ukiah (UKAC1) and Hopland (HOPC1) forecast sites (red triangles in Figure 1). Hereafter we drop the "1" postscripts on the location names for convenience. The two downstream forecasts are processed by the EFO to constrain releases to prevent downstream flooding. Because the LAMC forecasts are the primary determinant of EFO release decisions, we focus on LAMC metrics in the Results (Section 4).

The HEFS hindcast, composed of 1–14 day lead time forecasts at an hourly resolution and initialized daily between 1 October 1985 to 30 September 2010, are obtained from the California-Nevada River Forecast Center (CNRFC, 2022). The forecasts are aggregated to daily mean values between 12:00 and 11:59 GMT. For observations we use California Data Exchange Center daily full natural flow data for each of the gauges, where downstream flows at HOPC and UKAC are adjusted to represent only local flow contributions by subtracting out upstream contributions. Any negative values in the observed full natural flow record are corrected to zero.

We use forecast values of maximum/minimum temperature (TMAX/TMIN) and precipitation (PRECIP) from the NCEP GEFS v2 (Hamill et al., 2013; NOAA PSL, 2022) hindcast data set between 1985 and 2010, taken from the 1° grid box centered at 39°N and 123°W that overlaps most of the Russian River watershed (Figure 1a) and all of the forecasted sub-watershed boundaries of interest in this study. These data are also provided by the CNRFC and include ensemble mean forecasts of the three variables at 6-hourly increments out to a 14-day lead time. For observational data, we use 6-hourly gauge-based observations of MAT and MAP at the LAMC, HOPC, and UKAC locations. In addition to the gauge-based estimates at each location, we use the spatially weighted MAT and MAP for the entire 1° grid cell as the basis for calculating and simulating forecast errors based on the GEFS output.

3. Methods

We develop two approaches to create synthetic hydrologic ensemble forecasts (Figure 2). The first approach, referred to as "syn-HEFS," involves directly fitting a synthetic forecast model to the HEFS ensemble output. This model requires only the HEFS output and the hydrologic observations for model fitting. The second approach, called "syn-GEFS," targets the GEFS meteorological forcing data that drives HEFS. This approach is fit to the GEFS TMAX, TMIN, and PRECIP ensemble mean forecast data. Synthetic meteorological forecasts are then processed through the HEFS model to generate an ensemble of hydrologic forecasts. Because each synthetic meteorological forecast must be processed through HEFS and is 6-hourly, this method is more computationally intensive than syn-HEFS.

Both approaches can produce M synthetic forecast ensembles, each with E ensemble members (here, we set $E = 61$, the same as in the HEFS hindcast). Synthetic forecast data can be generated for any period with hydrological (syn-HEFS) or meteorological (syn-GEFS) observations and can be directly input into the EFO model to evaluate FIRO policies. Both approaches build from the methods developed in Brodeur and Steinschneider (2021), which can produce M synthetic forecasts, each for the ensemble mean (i.e., $E = 1$). Therefore, we first provide an overview of the methodology in Brodeur and Steinschneider (2021) common to both syn-HEFS and syn-GEFS (Section 3.1), and then highlight components of syn-HEFS (Section 3.2) and syn-GEFS (Section 3.3) that differ from that original approach and expand their capacity to model synthetic forecast ensembles with multiple members.

3.1. Synthetic Forecasts for a Single Trace

In Brodeur and Steinschneider (2021), we developed a multivariate approach to model and simulate complex forecast errors from a single forecast trace, the ensemble mean forecast. This approach considered a single observational timeseries (O_t) of length n over an available hindcast period and l forecast time series ($F_{t,l}$) of the same length associated with lead times $l = 1,\ldots,L$. Each of the l forecasts will have an associated error time series ($e_{t,l}$), defined as the difference between the observation and the forecast ($O_t - F_{t,l}$). If observations and forecasts are available at multiple locations or for multiple hydrometeorological variables (e.g., precipitation and temperature), then the error time series across those locations and variables can be concatenated across all lead times. This

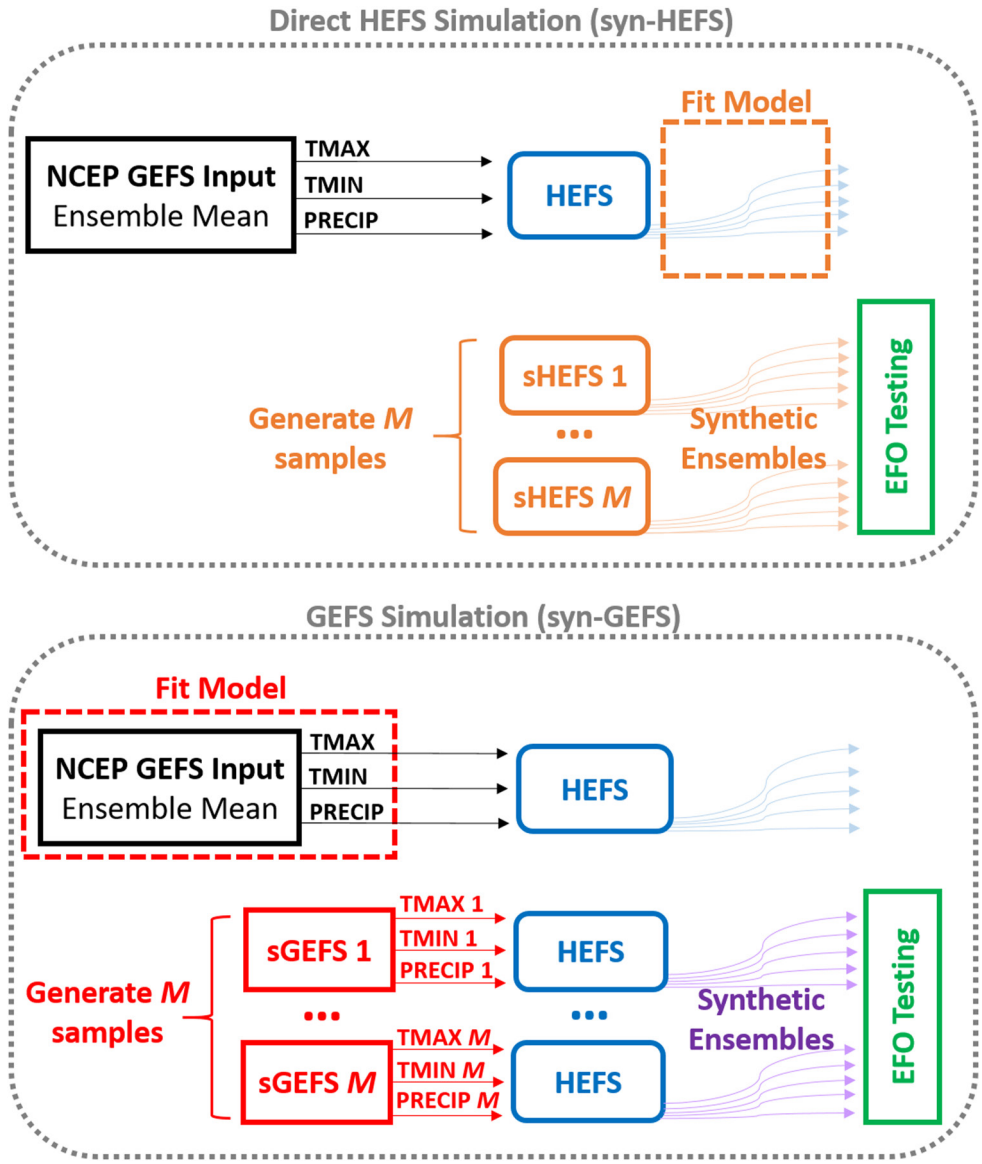

Figure 2. Conceptual diagram of two proposed synthetic ensemble forecast generation approaches, "syn-HEFS" and "syn-GEFS." Orange (purple) coloration will be used through the remainder of the manuscript in reference to syn-HEFS (syn-GEFS) output.

produces error time series $e_{t,k}$, with $k = 1,\ldots,K$ and K equal to the product of the number of lead times, locations, and variables.

The central section of Figure 3 shows the primary components of the synthetic forecast procedure in Brodeur and Steinschneider (2021) that operate on the forecast errors $e_{t,k}$, with some minor modifications from that original study. To produce a set of independent, random deviates for each of the K forecast time series, the first three steps of the procedure are: (a) standardize the errors to remove heteroscedasticity; (b) remove auto- and cross-correlation both within and between time series; and (c) fit suitable distributions to enable sampling of new random deviates. The order of steps above reflects a change to our previous work (steps 1–2 are reversed in Brodeur & Steinschneider, 2021), because we found this change led to better replication of forecast variance.

In step 1, we standardize the raw forecast errors from the hindcast period using a linear model between the standard deviation ($\sigma_{t,k}$) for each time series of forecast errors and the observations.

$$\sigma_{t,k} = \gamma_{0,k} + \gamma_{1,k} O_t \tag{3a}$$

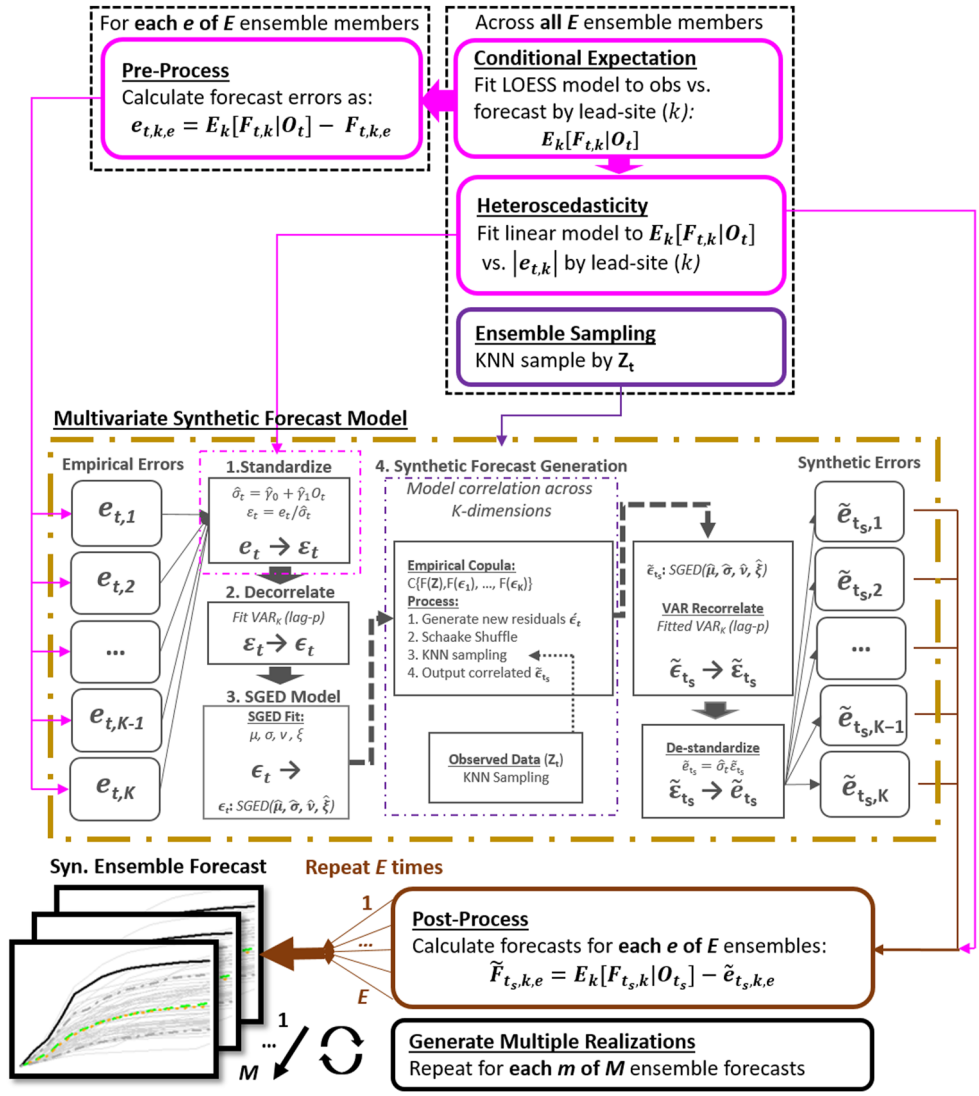

Figure 3. Syn-HEFS conceptual diagram. Model elements inside the gold dashed box/dark gray text represent the multivariate synthetic forecast procedure developed in Brodeur and Steinschneider (2021), with some modifications (i.e. Section 3.1, steps 1–4). Elements outside the gold dashed box highlight innovations in this study used to extend the model to the generation of *M* synthetic forecast ensembles, each with *E* members.

$$\epsilon_{t,k} = \frac{e_{t,k}}{\sigma_{t,k}} \tag{3b}$$

This model is estimated using an ordinary least squares linear regression between the observations and absolute errors ($|e_{t,k}|$), which serve as a proxy for $\sigma_{t,k}$. We constrain both the intercept ($\gamma_{0,k}$) and slope parameters ($\gamma_{1,k}$) to be greater than 0. In addition, we constrain the intercept to be no lower than the mean standard deviation estimate of the lowest decile of the data, since very low intercept values can produce extremely large and unrealistic standardized errors at low observed flow values.

In step 2, a vector autoregressive (VAR) model is used to remove auto- and cross-correlation within and between the *K* standardized error time series $\epsilon_{t,1:K}$. The VAR model is fit using the LASSO penalized "BigVAR" package in R (Nicholson et al., 2020, 2021) with a maximum lag value of 3.

In step 3, the standardized and uncorrelated residuals of the VAR model ($\epsilon_{t,k}$) from the hindcast period are fit to a skew generalized error distribution (SGED), which is well suited for non-Gaussian, fat tailed, and skewed

distributions (Schoups & Vrugt, 2010; Wurtz et al., 2020). There are four parameters of the SGED distribution that are fit to capture the mean, standard deviation, skew, and kurtosis of the data.

At the conclusion of steps 1–3, the VAR model residuals $\epsilon_{t,k}$ may still exhibit dependencies across the K lead times, locations, and variables, because these dependencies are not always fully captured by a low-order, linear model. The residuals $\epsilon_{t,k}$ are often also correlated with the observations themselves (Lamontagne & Stedinger, 2018). To capture these correlations in new samples, we employ an empirical copula and k-Nearest-Neighbor (kNN) sampling procedure based on the rank correlation between the hindcast period observations and $\epsilon_{t,k}$ (step 4). In brief, this procedure independently samples new random deviates $\acute{\epsilon}_{t,k}$ from the fitted SGED distributions for the length of the hindcast period, and then reorders these samples using a Schaake Shuffle (Clark et al., 2004) based on the order of the empirical residuals $\epsilon_{t,k}$ from the hindcast period. We denote these reordered residuals $\tilde{\epsilon}_{t,k}$. The reordering correlates the vector of SGED samples $\tilde{\epsilon}_{t,k}$ across its K dimensions (lead times, sites, variables), and also correlates these residuals with the time series of observations O_t from the hindcast period. Then for any time step t_s in the historical record (pre-, during, or post-hindcast period) with an observation O_{t_s}, we can select $\tilde{\epsilon}_{t,k}$ to form our synthetic forecast across the K dimensions (lead times, sites, variables) via kNN sampling:

1. Find the k observations in the hindcast period that are most similar to the observation at simulation time t_s.
2. Randomly select one of those hindcast period observations (O_t).
3. For that selected observation from the hindcast period, select the vector of SGED residuals $\tilde{\epsilon}_{t,k}$ that have been associated with that observation via the Schaake Shuffle. These residuals are now associated with simulation time t_s ($\tilde{\epsilon}_{t_s,k}$).

The steps above ensure that standardized residuals $\tilde{\epsilon}_{t_s,k}$ used to develop new synthetic forecasts maintain appropriate cross-correlations between lead times, sites, and variables, and also maintain correlations between the residuals and the observations. As in Brodeur and Steinschneider (2021), we note that O_t and O_{t_s} can be replaced by any plausible sampling criterion Z_t and Z_{t_s}.

New sets of raw forecast errors $\tilde{e}_{t_s,k}$ are generated by reversing steps 1–3: the new sets of synthetic forecast random deviates $\tilde{\epsilon}_{t_s,k}$ are recorrelated via simulation from the fitted VAR model to become $\tilde{\varepsilon}_{t_s,k}$, and then de-standardized via the normalization model to become $\tilde{e}_{t_s,k}$.

3.2. "syn-HEFS" Ensemble Generation

The methodology above works well for simulating a single value forecast (e.g., the ensemble mean) across K dimensions (i.e., the product of lead times, locations, and variables), where K is at least an order of magnitude smaller than the length of the record to ensure precise model estimation. However, individual ensemble members exhibit more noise than the ensemble mean, and preliminary experimentation showed that synthetic forecasts of ensemble members with the approach in Section 3.1 were unstable, especially around extreme events. In addition, in order to generate synthetic forecast ensembles, the number of dimensions K must increase significantly to accommodate multiple ensemble members (E). This can render the approach computationally intractable for even moderately sized values of E (e.g., $K > 1{,}000$ for 14 lead times, 3 sites, and $E = 30$ ensemble members). Yet, if the algorithm in Section 3.1 is applied separately for each ensemble member, it may fail to capture clustering across the ensemble (Wilks, 2019 and references therein).

Therefore, we modify the approach in Section 3.1 in three ways to both stabilize synthetic forecasts of individual ensemble members around extremes and to capture inter-ensemble correlation, all while controlling computational complexity. The underlying strategy is to model certain properties of each forecast ensemble member separately and other properties globally (i.e., across all ensemble members). Hereafter, we use the subscript e to reference specific ensemble members (e.g., $e_{t,k,e}$ is the forecast error for ensemble member e of E). The goal of the synthetic forecast generator is to develop multiple (M) stochastic realizations of an ensemble of forecast traces (each realization having E members), thereby sampling from the space of plausible ensembles rather than being limited to a single ensemble with E members (as is the case with the current HEFS forecasting system).

The first major innovation consists of data pre-processing when calculating the empirical forecast errors ($e_{t,k,e}$). Building from the work in Lamontagne and Stedinger (2018), we define the raw forecast errors not as the difference between forecasts and observations, but rather as the difference between forecasts and the conditional expectation of the forecasts given the value of the observation (Figure 3, "Pre-Process").

$$e_{t,k,e} = E_k\left[F_{t,k,e}|O_t\right] - F_{t,k,e} \quad (4)$$

This change helps accommodate the fact that hydro-meteorological forecasts, even in their raw deterministic form (e.g., the raw GEFS precipitation forecasts), exhibit some degree of Type-II conditional bias, with the most important implication being their tendency to underestimate extreme events (J. D. Brown et al., 2012). Hydrological model calibration and meteorological ensemble post-processing further exacerbate this problem, leading to underestimation bias around large observed inflow events (J. D. Brown et al., 2012; Demargne et al., 2014; Vogel, 2017). We found that explicit modeling of this bias as a pre-processing step enhanced representation of the Type-II conditional bias, stabilized model output, and enabled better modeling of persistence in the forecasts, particularly across lead times.

To model this bias, we estimate a conditional expectation $E_k[F_{t,k,e}|O_t]$ globally (i.e., using all of the ensemble members) using a locally weighted polynomial regression (LOESS) between the observations and the HEFS ensemble median (Figure 3, "Conditional Expectation"). The conditional expectation must be fit to each of the K dimensions separately, since the expected forecast for a particular observational value will differ by site and lead-time.

The second innovation utilizes the conditional expectation model and the raw errors across all ensemble members for a specific site and lead time ($e_{t,k}$) to construct a global linear model for heteroscedasticity as described in Equations 3a and 3b. Specifically, we aggregate $e_{t,k,e}$ to estimate $\sigma_{t,k}$ by first taking the absolute value of the raw errors across all ensemble members, then calculating the ensemble mean: $\hat{\sigma}_{t,k} = \frac{1}{E}\sum_{e=1}^{E}|e_{t,k,e}|$. This $\hat{\sigma}_{t,k}$ value replaces $\sigma_{t,k}$ in Equation 3a while O_t is replaced by $E_k[F_{t,k,e}|O_t]$ to estimate the heteroscedastic linear model parameters $\gamma_{0,k}$ and $\gamma_{1,k}$ globally (Figure 3, "Heteroscedasticity").

The third major innovation is to adjust the kNN sampling procedure in order to preserve the correlation structure between ensemble members (Figure 3, "Ensemble Sampling"). Specifically, at each time step in the synthetic forecast generation, the kNN-sampled observation used to select Schaake Shuffled random deviates $\tilde{\epsilon}_{t_s,k,e}$ is common across all ensemble members, ensuring that the forecast errors selected at each time step will maintain the empirical correlation structure across ensemble members. However, the remainder of the synthetic forecasting procedure (gold box in Figure 3) is still applied individually for each ensemble member. That is, the synthetic errors for each ensemble member are generated independently from each other, with the exception that the correlation in the random deviates $\tilde{\epsilon}_{t_s,k,e}$ across ensemble members is preserved by the global kNN-sampling scheme.

Finally, we use the cumulative flows over the forecast horizon of 14 days to define our sampling criterion Z_t ($Z_t = \sum O_{t:t+l}\,; l = 14$) instead of the single day observations O_t. This procedure ensures that forecast error properties that accompany large, sustained inflow events are better captured in the synthetic forecasts. In cases with multiple locations, one can define Z_t based on the location with the largest inflows (LAMC in our case study), although other approaches could be tested as well.

Once raw forecast errors $\tilde{\epsilon}_{t_s,k,e}$ are generated for each ensemble member, they must be subtracted from the conditional expectation $E_k\left[F_{t_s,k,e}|O_{t_s}\right]$ to recover a synthetic forecast value $\tilde{F}_{t_s,k,e}$ (Figure 3, "Post-Process"). This entire process can then be repeated to generate M synthetic forecast ensembles, each with E members (Figure 3, "Generate Multiple Realizations").

3.3. "syn-GEFS" Ensemble Generation

The synthetic generation of meteorological input data largely follows the procedure described in Section 3.1 for a single trace, because the "syn-GEFS" strategy uses synthetic forecasts of the GEFS ensemble mean to force the HEFS model (which generates its own hydrologic ensemble; see Section 2.1). However, we note a few modifications here. First, observations of TMAX, TMIN, and PRECIP are based on gauge derived MAT/MAP estimates. Biases between TMAX and TMIN forecasts from GEFS and MAT observations are removed using a monthly bias correction factor. For PRECIP, we employ the same conditional mean approach as described in Section 3.2, because precipitation forecasts are conditionally biased low around large events at longer lead times. Finally, the observations used for the kNN sampling of synthetic forecast errors are based not on a single observation at time t, but rather a vector of observations consisting of gauge-based 6-hourly precipitation estimates for day, t and $t-1$ across the 3 locations (LAMC, HOPC, UKAC). This yielded an observational vector of length 24 (3 locations, 2

time steps, and 4 6-hourly values in each day), which we reduced to a 3-dimensional vector via principal component analysis for use in the kNN resampling. M synthetic forecasts of the GEFS ensemble mean were supplied to the CNRFC, where each of the $m = 1,\ldots,M$ forecast traces were run through the HEFS model to develop M separate hydrologic ensembles with E members each.

In comparison to syn-HEFS, syn-GEFS is more computationally expensive. In this work, the syn-HEFS procedure was fit and used to simulate 100 63-year ensemble synthetic forecast samples within 24 hr on a small high performance computing resource (22 node cluster with dual 20-Core Intel Xeon Gold 5218R CPUs 2.1 GHz, 192.0 GB of RAM). In contrast, the syn-GEFS procedure requires 2–3 days for the generation of a single sample, mainly because syn-GEFS must utilize the existing HEFS architecture in hindcast mode, which is not optimized for the generation of many samples. It is likely that the efficiency of syn-GEFS could be improved if parallelized versions of HEFS were developed for synthetic forecasting.

3.4. Synthetic Forecast Evaluation

The literature on forecast ensemble verification (Wilks, 2019 and references therein) highlights three important characteristics of a well-calibrated forecast: (a) the ensemble should be an unbiased predictor of the observations; (b) the ensemble should consist of "equally likely" members, where observations fall uniformly across the ensemble; and (c) the spread of the ensemble should be reflective of forecast uncertainty, that is, accurate forecasts should be associated with tightly bounded ensembles and vice versa.

To measure these properties for the original HEFS forecast ensemble, we employ binned spread-error (BSE) diagrams, cumulative rank histograms, and ensemble continuous ranked probability scores (eCRPS) (Wilks, 2019). In brief, BSE diagrams assess the relationship between the ensemble spread and the ensemble mean forecast error across quantiles of the ensemble spread. Well calibrated ensemble forecasts will show a 1:1 relationship between spread and skill. Cumulative rank histograms assess the "consistency" of the ensemble forecast. For ideal ensemble forecasts, each observation appears as a random draw from the ensemble and would yield a flat (i.e., uniform) rank histogram and a linear cumulative rank histogram. Departures from this ideal condition often result from issues of aggregate forecast bias or under/over-dispersion. Finally, eCRPS measures the skill or "accuracy" of ensemble forecasts. It rewards the concentration of probability density around the observation and so is very sensitive to the central tendency of the ensemble, but it is also sensitive to the sharpness of the ensemble (i.e., the ensemble spread). Unlike the two previous metrics that are (typically) employed in aggregate, the eCRPS is a single score calculated for each forecasted event. A more detailed explanation of these measures and their interpretation can be found in Text S2 in Supporting Information S1.

To evaluate how well the synthetic ensemble forecasts match the original HEFS forecasts, we assess how well the BSE diagrams, cumulative rank histograms, and eCRPS from the synthetic forecast ensembles match those of the original forecast ensemble. That is, we seek synthetic forecasts that demonstrate parity with the actual forecasts, where "good" synthetic forecasts mimic both the strengths and weaknesses of the original forecast. We focus this analysis on a subset of lead times (1-day, 3-day, 5-day, and 10-day) that are operationally relevant but also exhibit some appreciable skill over climatology (DeFlorio et al., 2018; DeHaan et al., 2021).

We also apply a two-sample DTS hypothesis test (Dowd, 2020, 2023) between the HEFS and synthetic forecast cumulative rank histograms and eCRPS distributions to diagnosis their similarity. In this test, the null hypothesis is that the HEFS and synthetic forecast verification metrics come from the same distribution, and we deem the results "statistically indistinguishable" if the null hypothesis cannot be rejected. For the cumulative rank histograms, we conduct the DTS test separately between HEFS and each of the synthetic forecast samples at the $p = 0.1$ significance level. In contrast, we apply the DTS test ($p = 0.1$) directly to the continuous (but unequal in size) distributions of eCRPS for HEFS versus the eCRPS values pooled across all samples from the synthetic forecasts.

We further validate the synthetic forecasts by assessing how well they perform during extreme events and whether they are "fit for purpose" from an operational perspective (Stedinger & Taylor, 1982). In particular, we compare the actual and synthetic ensemble forecast mean and variance of cumulative flow totals at different lead times prior to major extreme events. In addition, we compare EFO operations during the hindcast period when forced with the actual and synthetic ensemble forecasts, with a focus on system dynamics during major floods.

Finally, we evaluate how well EFO operations perform using the synthetic forecasts during the pre-hindcast period (1948–1984), especially during extreme events that were not present in the hindcast period (1985–2010).

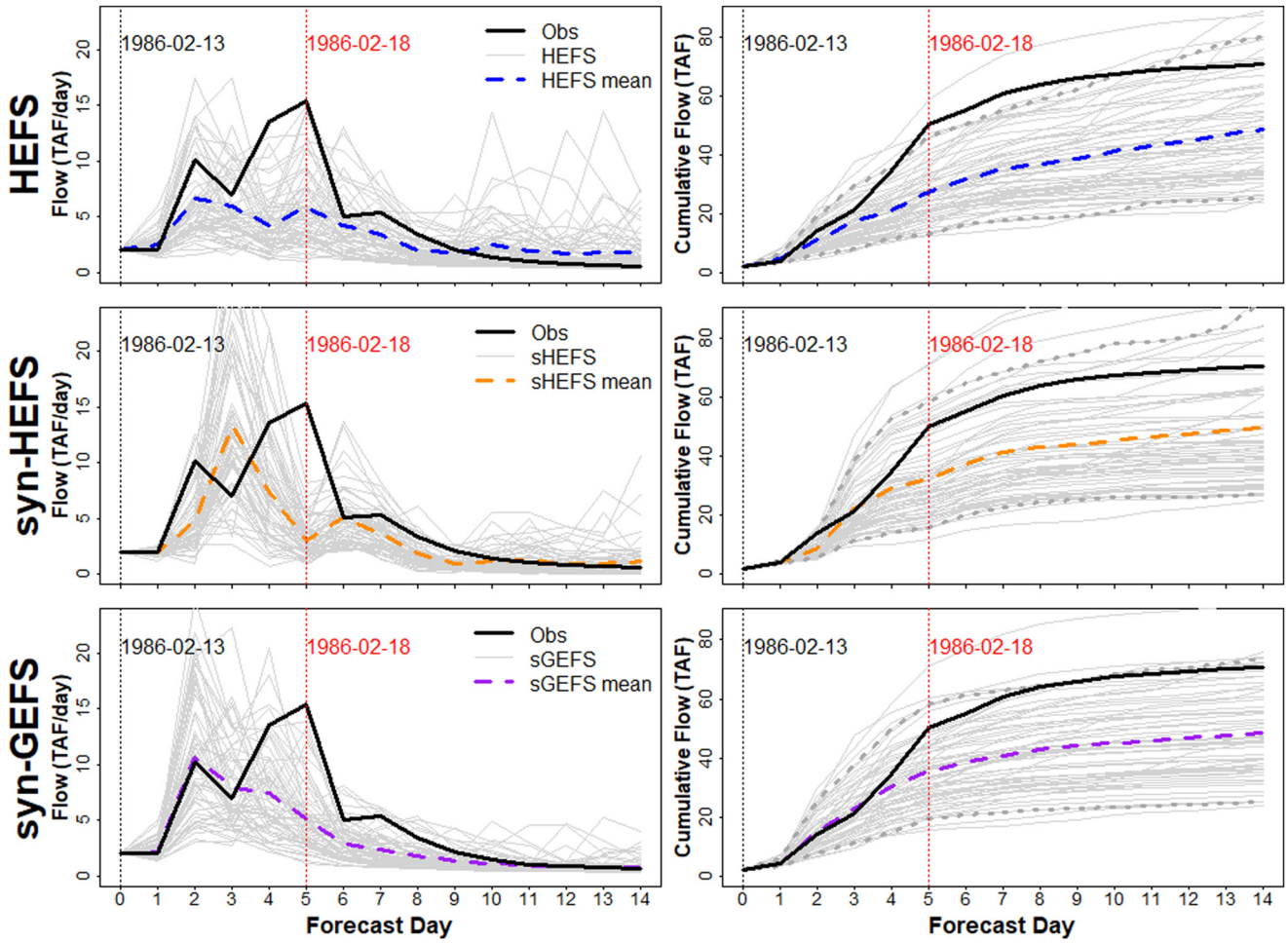

Figure 4. Examples of 61-member ensemble forecasts (light gray lines) for the extreme inflow event on 18 February 1986, issued on 13 February 1986 (5-day lead), where each panel shows the full 14-day forecast time period. The top row shows the Hydrologic Ensemble Forecast System forecast, while the middle (bottom) row shows one sample of an ensemble forecast from syn-HEFS (syn-GEFS). The left column shows daily inflow forecasts while the right column shows cumulative inflow forecasts. The ensemble means for each panel are indicated by the thick dashed lines, and gray dotted lines show the 90th percentile bounds for the cumulative ensemble inflows.

4. Results

4.1. Synthetic Forecast Verification

We first illustrate the general behavior of the original HEFS ensemble forecast and both "syn-HEFS" and "syn-GEFS," using one extreme inflow event on 18 February 1986 as an example (Figure 4). Here, and in all subsequent figures, we focus on forecasts at the LAMC location. In all panels of Figure 4, we show 1–14 day lead forecasts, with the February 18 extreme inflow event occurring 5 days after the forecast issue date. The cumulative forecasts in the right column represent the forecast progressively summed across lead times, which is the format processed by EFO.

Figure 4 illustrates how synthetic ensemble forecasts from syn-HEFS and syn-GEFS broadly capture similar attributes to those of the HEFS hindcasts. However, due to the stochastic nature of the algorithms, any one sample (i.e., any single ensemble forecast) from syn-HEFS and syn-GEFS will differ from each other and from HEFS for individual events. For instance, in the left column of Figure 4 at a 2-day lead (15 February 1986), the HEFS hindcast underpredicts the observed flow on average, with a limited number of ensemble members extending above the observation. After the 2-day lead, HEFS forecasted flows steadily decline. In contrast, syn-HEFS has a more severe underprediction at a 2-day lead, but peaks at a 3-day lead well above the observations with substantial spread. Syn-GEFS shows qualitatively similar behaviors to HEFS, but is approximately unbiased at a 2-day lead. At a 5-day lead when the peak observation occurs, all three models are biased below the observation,

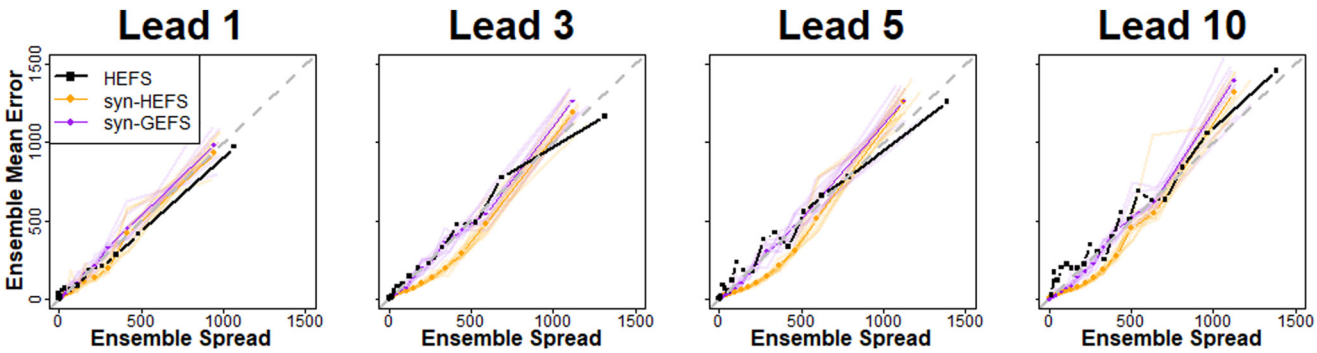

Figure 5. Ensemble binned spread-error diagrams showing ensemble mean error versus ensemble spread at the LAMC location across four selected lead times for observations in the rainy season (December-March). The 20 plotted points for each forecast are based on 20 equally spaced data bins derived from sorted forecast ensemble spreads. Light orange/purple lines show performance for 10 samples of synthetic forecast ensembles, with median performance shown by the darker line. The 1:1 line (gray dashed) indicates an ideal forecast.

but HEFS and syn-GEFS capture the peak in the upper tail of the ensemble spread while syn-HEFS does not. At longer leads (9–14 days), both HEFS and syn-HEFS show more variability than syn-GEFS, which manifests prominently in the cumulative forecasts at these longer leads. Additional synthetic ensembles for the 1986 flood and other extreme events are shown in Figures S3.1–3.5 in Supporting Information S1 and show how different synthetic samples can differ from each other and from HEFS for individual events. Altogether, these results illustrate the kind of inter-sample variability that can be achieved with the two synthetic forecast models for individual events, as well as the manner in which forecast behaviors at different lead times propagate into the cumulative ensemble behavior.

To determine whether the synthetic forecasts are properly replicating the behavior of HEFS, we need to compare properties of the ensembles across many events, and for many sample ensembles from the synthetic algorithms. We first examine how synthetic forecast performance compares to HEFS in an ensemble BSE diagram (Figure 5), focusing on observations in the middle of the wet season (December-March). At shorter leads (lead 1 and 3), the HEFS forecast exhibits well calibrated behavior across bins, which is reflected in both synthetic forecast models. HEFS forecasts are also well calibrated at longer leads (lead 5 and 10), although there is more variability in the spread-error relationship in the lower bins (i.e., lower left) that stabilizes at the higher bins (i.e., upper right). The syn-HEFS is biased low (i.e., over-dispersed) for smaller ensemble spread bins, but approaches HEFS for moderate spread values, while the spread-error relationship is more stable under syn-GEFS and aligns well with HEFS. Notably, neither synthetic forecast method reaches the HEFS ensemble mean error or spread at the highest quantiles, particularly at long lead times. Overall, however, these results suggest that the synthetic forecasts approximate the spread-error calibration relationship of HEFS reasonably well.

Next, we examine cumulative rank histograms and the distribution of eCRPS for HEFS and for multiple samples from both synthetic algorithms, again focusing on observations in December-March. Recall that the cumulative rank histograms show the ensemble forecast consistency (equi-probability of members), while the distribution of eCRPS emphasizes forecast accuracy. We note that an eCRPS score of 0 indicates a perfect forecast, while larger values indicate lower skill (see Text S2 in Supporting Information S1). We stratify the cumulative rank histograms and eCRPS distributions for observations less than the 90th percentile (i.e., small and average flows) and above the 90th percentile (i.e., high flows), because we are particularly interested in forecast behavior for the largest events. We note that stratifying rank histograms across the distribution of observations can obscure absolute measures of forecast skill (Bellier et al., 2017; Siegert et al., 2012), but this is not a concern here because we are mainly interested in the comparison of ensemble behavior between HEFS and the synthetic forecasts.

For observations less than the 90th percentile, the cumulative rank histograms show that the HEFS forecast ensemble underpredicts the observations at shorter leads but smoothly progresses toward a well calibrated forecast at long leads (Figure 6, first row). The syn-GEFS forecasts generally mimic this behavior, although with less underprediction bias at shorter leads (1–5 days), while the syn-HEFS forecasts deviate more from HEFS at 1–3 day leads before trending toward the HEFS behavior at 5 and 10 day leads. The distribution of eCRPS for observations less than the 90th percentile (Figure 6, third row) shows that forecast accuracy in HEFS degrades with lead time (as expected), and that the level of degradation is well matched in both syn-HEFS and syn-GEFS.

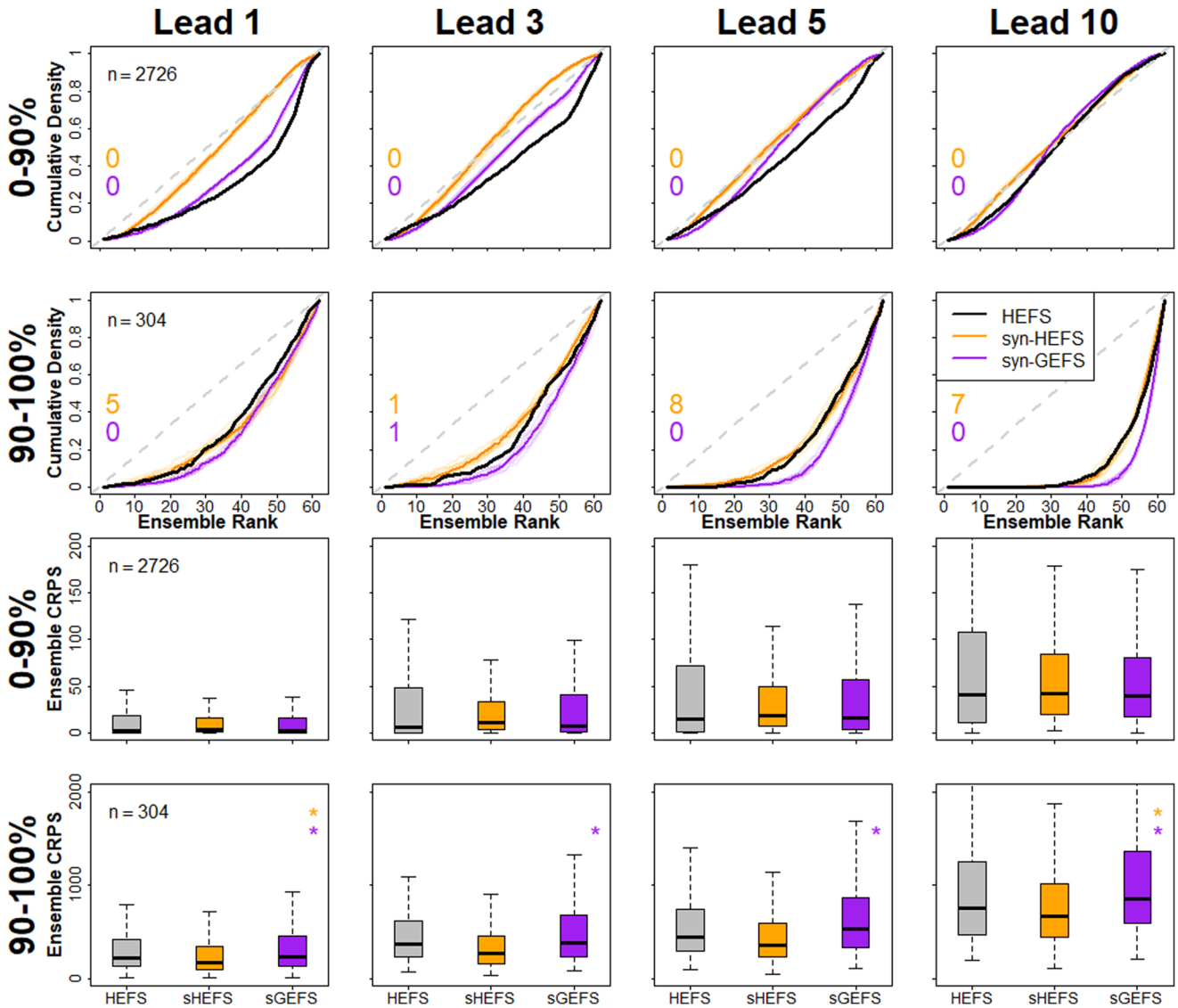

Figure 6. Top 2 rows: Cumulative rank histograms for Hydrologic Ensemble Forecast System (HEFS), syn-HEFS, and syn-GEFS at the LAMC location across 4 selected lead times in December-March (rainy season), shown separately for observations in the lower 90th percentile (0%–90% row) and the upper tenth percentile (90%–100% row). The dashed gray lines indicated the ideal 1:1 relationship (flat rank histogram) and the "*n*" value shows the number of observations in each subset. 10 samples of the syn-HEFS (syn-GEFS) are shown in light orange (purple), while the darker line shows the median value of the 10 samples. The orange (purple) numbers at left of plots indicate the number of syn-HEFS (syn-GEFS) samples that are statistically indistinguishable from HEFS via a DTS test ($p = 0.1$). Bottom 2 rows: The distribution of ensemble CRPS (ensemble continuous ranked probability scores) for 10 samples from each synthetic generation method compared to the single sample from HEFS. Asterisks denote synthetic forecast distributions that are statistically indistinguishable from HEFS ($p = 0.1$) via a DTS test.

Though results for both syn-HEFS and syn-GEFS are qualitatively similar to HEFS in many cases, especially the eCRPS distributions, they are statistically different than the corresponding distributions under HEFS according to the DTS test.

For observations above the 90th percentile (which are of most interest), the cumulative rank histograms suggest a high degree of parity between HEFS and both synthetic forecast algorithms (Figure 6, second row). All models show significant underprediction bias that worsens with lead time, although syn-GEFS exaggerates this under-prediction bias compared to HEFS at the longest leads. Similarly, the distribution of eCRPS for large flow events (Figure 6, fourth row) is comparable between HEFS and both synthetic forecast methods across lead times, although accuracy is slightly lower (higher) for syn-GEFS (syn-HEFS) at long leads. The DTS test confirms a higher degree of parity between HEFS and syn-HEFS compared to syn-GEFS for the rank histograms, but greater similarity between HEFS and syn-GEFS for the eCRPS distributions. Overall, though, the cumulative

rank histograms and eCRPS distributions suggest both synthetic forecast methods produce ensembles that mimic the behavior of the HEFS hindcasts for the largest observations.

We conduct two additional comparisons to further evaluate our synthetic forecasts against HEFS. First, we confirm the reliability of ensemble-mean meteorological forecasts from syn-GEFS in Text S4 in Supporting Information S1. Second, we compare syn-HEFS to a "reference" version of the syn-HEFS model (syn-HEFS-ref), based on the original algorithm in Brodeur and Steinschneider (2021). This comparison is used to illustrate the importance of the modifications that were developed in this study in Section 3.2, specifically the global estimation of the conditional expectation, heteroscedasticity, and inter-ensemble kNN sampling procedures. This comparison is described in Text S5 in Supporting Information S1, and demonstrates the importance of these algorithmic developments to accurate emulation of HEFS ensemble forecasts.

4.2. Event Specific Forecast Validation

In this section, we focus on whether the synthetic forecasts capture important ensemble attributes of HEFS for rare extreme events. We focus on the syn-HEFS forecasts owing to the computational ease of creating many forecast samples, but present similar results for syn-GEFS in Text S6 in Supporting Information S1, albeit for a smaller number of samples.

Figure 7 shows the ensemble mean and variance of cumulative flow totals at different lead times prior to the 10 largest extreme events in the hindcast period. For reference, Figure 4 visualizes the ensemble mean and spread of cumulative forecasted flows at one lead time (5-day) and for one extreme event (the 1986 flood). For each lead time and event in Figure 7, we show a distribution of 100 different ensemble means and variances based on 100 synthetic ensemble forecast samples from syn-HEFS. To ease interpretation, all values in Figure 7 are shown as a ratio to the cumulative ensemble mean and variance from the HEFS hindcast, such that values above (below) unity (green line in Figure 7) indicate ensemble means and variances that are greater than (lower than) that of the HEFS forecast for a given event and lead time. We also show the observed cumulative flow for each event and lead time in the ensemble mean panel (black triangles), also shown as a fraction of the HEFS ensemble mean. We note that both the ensemble mean and variance of cumulative flows are important attributes of the forecast in terms of how the EFO policy makes flood releases based on the risk tolerance curve (see Section 2.1).

Before examining the synthetic forecasts, we first note that the HEFS forecasts tend to underpredict the observations, and this underestimation grows with lead time (Figure 7, top row). There is, however, significant variability in the degree of ensemble mean underestimation across events and lead times, especially beyond the 1-day lead. For instance, the observations for 17 February 2004 are greater than 3.5 times the cumulative forecast ensemble mean value for HEFS for both the 3- and 5-day lead, but this is not the case at the 1- and 10-day leads.

Boxplots in Figure 7 that straddle the value of 1 suggest that the syn-HEFS synthetic forecasts are able to reproduce the same ensemble mean and variance of the HEFS hindcast for a given event and lead time. Overall, the synthetic forecasts are mostly able to capture the HEFS ensemble mean for each event and lead time within the range of 100 synthetic ensembles (Figure 7, top row). However, for any given event and lead time, the central tendency in the ensemble mean of the synthetic forecasts does not necessarily align with HEFS, and can be either above or below the hindcast. Since the HEFS hindcast is a single realization of a forecast ensemble, the central tendency of the synthetic forecasts cumulative ensemble mean should not necessarily align with the HEFS for each event. Still, there are a few events at different lead times where no synthetic ensemble sample is able to capture the HEFS ensemble mean behavior. For instance, during the extreme event on 17 February 2004 at a 3-day and 5-day lead, all synthetic ensemble forecast samples produce cumulative means that exceed that of the HEFS and are closer to the observed cumulative flow. That is, the syn-HEFS algorithm cannot produce the degree of underprediction for this extreme that was seen in HEFS.

For the cumulative ensemble variance (Figure 7, bottom row), there is substantially more variability in the syn-HEFS synthetic forecast distribution than the cumulative ensemble mean, but the overall comparison against HEFS is similar. For different events and lead times, the synthetic forecasts are able to capture the ensemble variance of HEFS within the range seen for 100 samples, with only a few exceptions. The synthetic cumulative ensemble variances are not always centered on the HEFS value for any given event and lead time, exhibiting both upward and downard biases, although the synthetic ensemble variances tend to become more biased downward with lead time on average. We note that the results for syn-GEFS are broadly similar (Figure S6a in Supporting Information S1), though the much smaller sample size (10 vs. 100 samples) makes meaningful comparisons difficult.

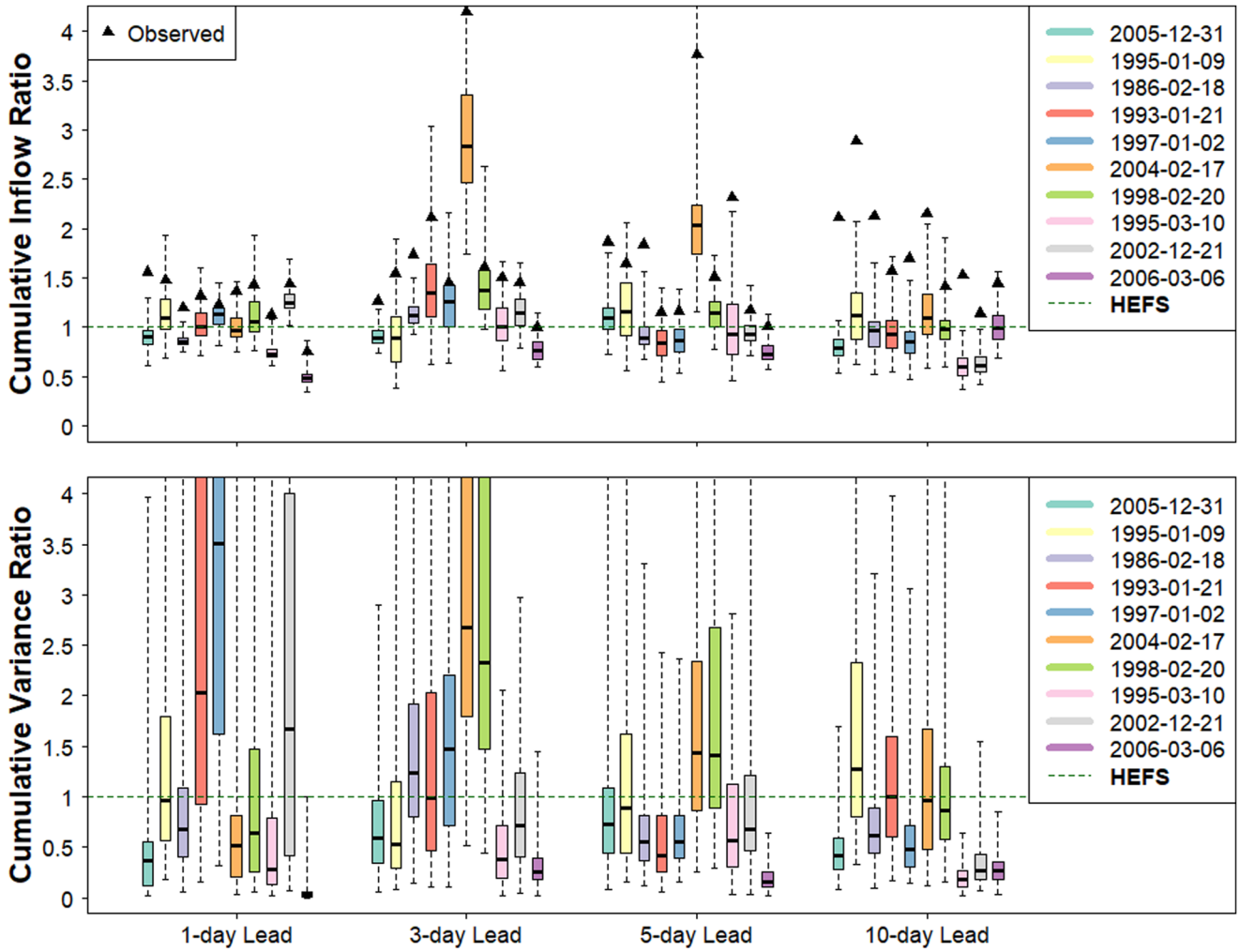

Figure 7. Analysis of 10 largest extreme events in the hindcast record for syn-HEFS at the LAMC location and four lead times. Top: Boxplots (black triangles) show the ratio of the cumulative ensemble mean of synthetic forecasts (cumulative observed flows) to the cumulative ensemble mean of the Hydrologic Ensemble Forecast System (HEFS) hindcast for each event and lead time. Each event has a single HEFS forecast cumulative ensemble mean, a single observation, and 100 cumulative ensemble means from random synthetic forecast samples. Bottom: As in top, but for cumulative ensemble variance. No "observed" value displayed because the observation has no variance. The boxplot whiskers in both panels extend to the extremes of the distribution.

Finally, we generated results for the top 100 events in a manner very similar to that of Figure 7 (see Figure S6b in Supporting Information S1), but with some modifications to account for event aggregation. Overall, the cumulative ensemble mean trends across 100 events are similar to those in Figure 7, although the syn-HEFS sample ensembles are more centered around the HEFS value, due to aggregation across multiple events. In addition, the analysis of syn-HEFS for the top 100 events highlights a tendency toward underprediction of ensemble variance as lead time increases. Nonetheless, both the HEFS cumulative ensemble mean and cumulative ensemble variance values fall within the interquartile range of the synthetic samples in a majority of cases across lead times, indicating that inter-sample variability in the synthetic model is sufficient to capture these cumulative statistics of the HEFS output. A similar result is seen for the top 100 events and syn-GEFS (Figure S6c in Supporting Information S1).

4.3. EFO Operational Validation

We further evaluate the synthetic forecasts based on how they influence FIRO-based operations under the EFO policy for Lake Mendocino, as compared to operations based on the HEFS hindcast. As in Delaney et al. (2020), we allow the EFO model to use the entire reservoir storage across all months to fully explore performance at the system boundaries, and we also include a PFO modeled operational trace for comparison. In the descriptions below, we use

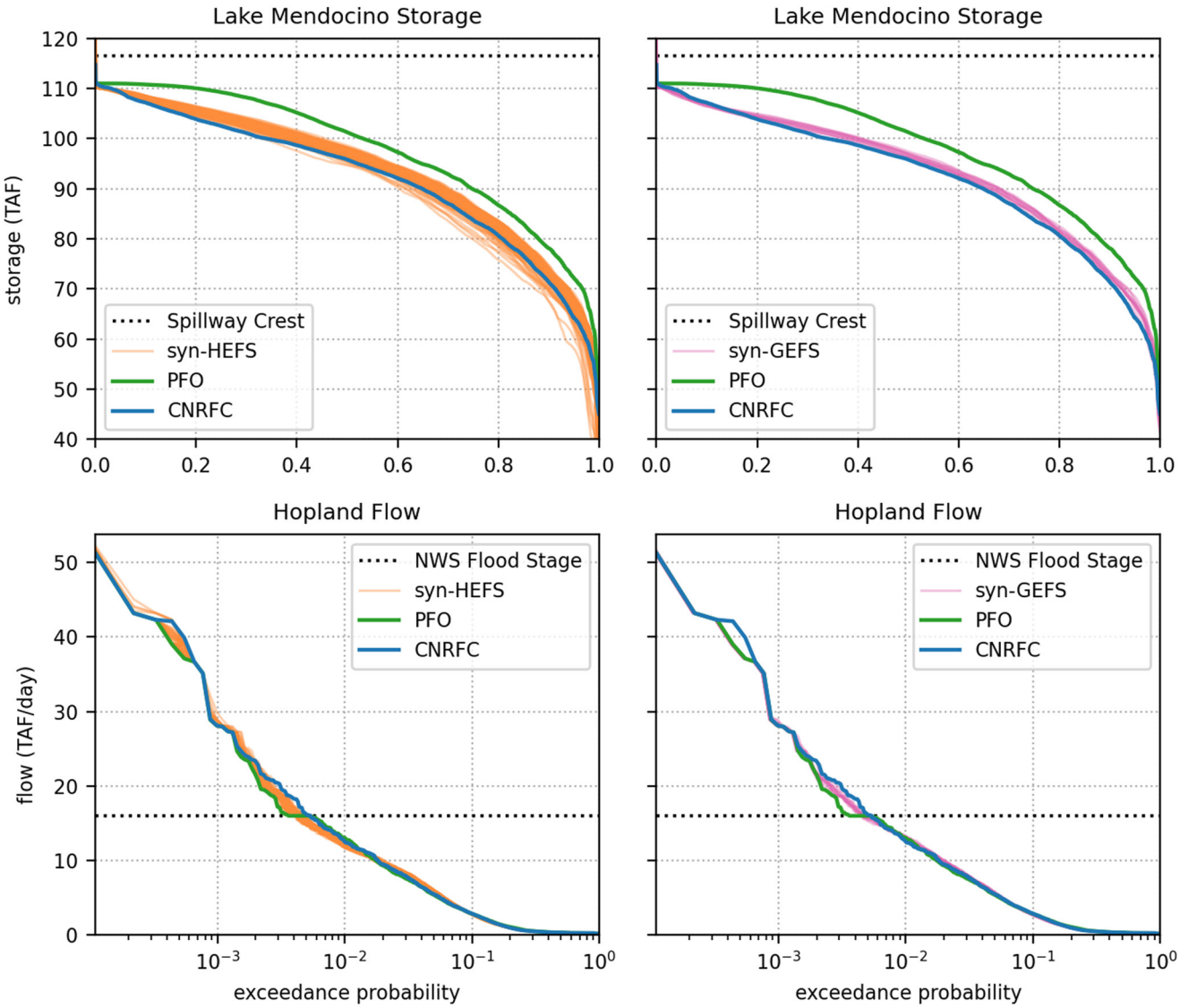

Figure 8. Operational results for the two synthetic forecast methods, where green lines correspond to perfect forecast operations, blue lines to California-Nevada River Forecast Center, and orange (purple) lines to syn-HEFS (syn-GEFS) with 100 (10) samples. Top row: Lake Mendocino storage sorted by exceedance probability. Bottom row: Flows at the downstream Hopland site sorted by exceedance probability.

"HEFS" to refer to the single trace of EFO output forced by the actual HEFS hindcast ensemble, PFO to refer to the single trace of EFO output forced with perfect forecasts, and "syn-HEFS" and "syn-GEFS" to refer to an ensemble of M traces of EFO outputs forced by syn-HEFS ($M = 100$) and syn-GEFS ($M = 10$) ensembles, respectively.

Figure 8 shows aggregated attributes of EFO operations across the WY 1986–2010 hindcast period, including the distribution of storage and flow at Hopland. The storage plots (top row) show that both syn-HEFS and syn-GEFS exhibit a similar storage distribution to the HEFS sample, although they show a slight bias toward the PFO results. The syn-GEFS is biased slightly toward the PFO across all but the highest storage values while the syn-HEFS shows the greatest bias at moderate to high storages (0.1–0.5 exceedance probabilities). In addition, there is more variability in the syn-HEFS samples than the syn-GEFS samples (albeit with a substantially larger sample size for syn-HEFS), especially at lower storage values. The synthetic forecasts also produce realistic downstream flows at Hopland (bottom row), where again, the syn-HEFS shows more variability than syn-GEFS. Altogether, these results show good agreement between HEFS and synthetic forecast forced EFO runs, though storage biases toward PFO at certain quantiles suggest a tendency toward either too skillful or under-dispersed ensembles for both methods.

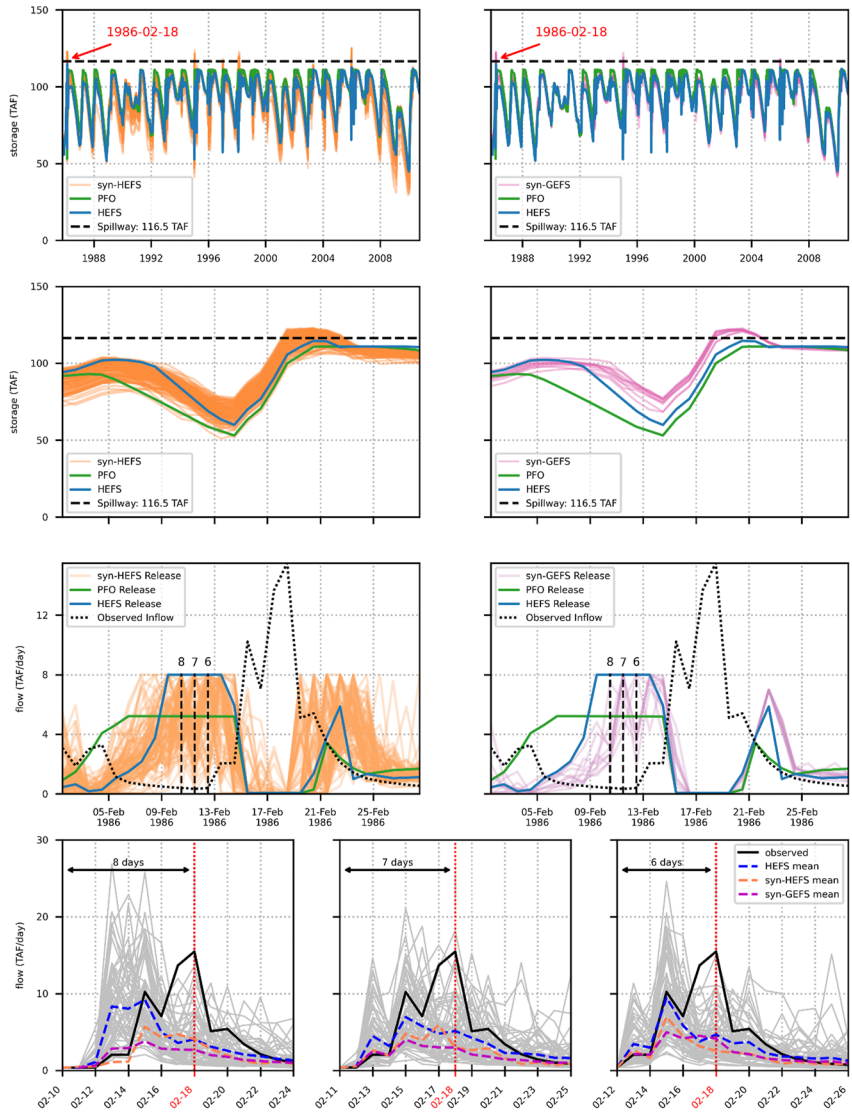

Figure 9. Extreme event Ensemble Forecast Operations performance example for 18 February 1986 event. Top row: Storage timeseries for the WY 1986–2010 hindcast period. Second row: Daily storage timeseries for the 1986 flood event, where 61% of syn-HEFS and 100% of syn-GEFS samples spill on 20 February 1986. Third row: Daily controlled release (R_t^{Ctrl}) decisions with observed inflow depicted as black dotted line in background. Note: 8 thousand acre-feet/day is the maximum allowable release to prevent flooding at Hopland. Bottom row: As in Figure 4, but for Hydrologic Ensemble Forecast System ensemble at selected lead times preceding 18 February 1986 inflow event. These lead times correspond to vertical dashed lines shown in the third row.

Figure 9 shows EFO operational dynamics across the hindcast period and emphasizes operations during individual extreme events. The timeseries plots (top row) show that both methods lead to storage traces that generally follow that of HEFS, with more variability around the HEFS seen for syn-HEFS than syn-GEFS (even considering the larger sample size for syn-HEFS). Unlike HEFS, operations forced with a subset of syn-HEFS and syn-GEFS ensembles both result in emergency spillway usage. For syn-GEFS, these spills are confined to three events, namely February 1986, January 1995, and December 2005, whereas syn-HEFS also produces spills in January 1997 and February 1998.

The remaining panels in Figure 9 show operational dynamics in more detail for one of these spill events (18 February 1986). The storage sequences (Figure 9, second row) under syn-HEFS and syn-GEFS start from a similar initial condition to the HEFS prior to the extreme inflow event. However, a majority (70%) of syn-HEFS samples and all syn-GEFS samples lead to spills during this event. One clear contributing factor is that the

Table 1
Summary of Extreme Events That Result in Spills for One or Both Synthetic Forecast Model Simulations

Date	Max. daily inflow	syn-HEFS spills	syn-GEFS spills
1986-02-18	15.3 TAF	70/100	10/10
1995-01-09	17.9 TAF	23/100	5/10
1997-01-02	11.3 TAF	2/100	0/10
1998-02-20	9.8 TAF	19/100	0/10
2005-12-31	19.5 TAF	95/100	5/10

Note. Maximum daily inflows are shown in thousand acre-feet (TAF). Also shown are the fraction of synthetic forecast samples (out of 100 for syn-HEFS and 10 for syn-GEFS) that lead to spills.

majority of synthetic samples do not result in the sustained maximum release commanded by HEFS between 9 February 1986 and 14 February 1986 (Figure 9, third row). While both syn-HEFS and syn-GEFS reproduce portions of this release sequence, they most often do not sustain the release for long enough or do not initiate the release early enough. These insufficient pre-releases appear to be the primary differentiator between samples that spill and those that do not (see Text S7 in Supporting Information S1). We note that no policy (HEFS, syn-HEFS, or syn-GEFS) achieves the earlier and deeper drawdown supported by PFO.

Further analysis of the original HEFS hindcast for this event (Figure 9, bottom row) shows a notable over-forecast prior to the main inflow event that is particularly prominent at 6–8 day lead times (i.e., forecasts initiated between 10 February 1986 and 12 February 1986). This suggests that for this event, the HEFS reacted to a shorter range overforecast when initiating this maximum release, rather than reacting to a correct forecast synchronized with the observed extreme inflow event. This type of overforecast represents a type-1 (false positive) error that may have resulted from a temporal error in the meteorological forecasts. This behavior is challenging to replicate in synthetic forecasts, because the model standardization procedure (Section 3.1, step 1) enforces a proportionality between observations and forecast errors, making large forecasts around very small observations exceedingly rare in generation. This type of preceding overforecast is present in 3 of the 4 other events that spill (1995, 1997, 2006), with the fourth event (1998) being a multi-peaked, clustered inflow event in which the synthetic forecast traces systematically draw down insufficiently after the first of two large inflow events (see Text S7 in Supporting Information S1). We note that for all events where syn-HEFS has some fraction of ensembles that lead to spills, there are other ensembles in the set of $M = 100$ that do not spill (see Table 1). This suggests that in the hindcast period, a subset of synthetic forecast ensembles do lead to operations with the same flood control benefits seen in the original HEFS hindcasts, but these benefits are not guaranteed.

4.4. Operational Usage—Forecast Extension

Finally, we highlight how synthetic forecasts can be used to create plausible forecast sequences around historical events of interest outside the hindcast record (e.g., pre-1985 for this version of HEFS). We generate only from syn-HEFS for this example, given its lower computational expense. Here, PFO operations can also be simulated for comparison, but not operations forced with HEFS hindcasts. We also note that the syn-HEFS forecasts in this pre-1985 period show verification results that align closely with syn-HEFS performance in the hindcast period (1985–2010; see Text S8 in Supporting Information S1), implying that the relationships between forecasts and the observations are maintained in this out-of-sample period.

Figure 10 shows syn-HEFS-forced operational dynamics across the pre-hindcast period, as well as during the December 1964 flood event, which is both the downstream flood of record for Hopland and the only event to historically result in emergency spillway use at Lake Mendocino. There is similar aggregate storage behavior in the pre-hindcast period (Figure 10a) to that observed in the hindcast period (Figure 8), where operations forced with synthetic forecasts have a lower overall storage distribution than PFO. In the timeseries panel (Figure 10c), syn-HEFS forced operations lead to spills in three cases (December 1955; December 1964; January 1970), at least in a fraction of the syn-HEFS samples. Operations under syn-HEFS during the flood of record (December 1964) lead to maximum allowable releases (8 thousand acre-feet/day) up to 7–8 days prior to the inflow event (Figure 10d), which helps avoid spills in a large majority of traces immediately following peak inflow (Figure 10b). This is likely because the syn-HEFS ensemble (bottom row) shows a relatively accurate forecast of the inflow event, even up to a week in advance (Figures 10e–10g). However, approximately one quarter of the traces spill on the subsequent, smaller inflow event on 7 January 1965 (Figure 10b), because storage in the reservoir was unable to sufficiently drawdown following the peak inflow a few days prior. A similar analysis is provided for the December 1955 and January 1970 events in Text S9 in Supporting Information S1.

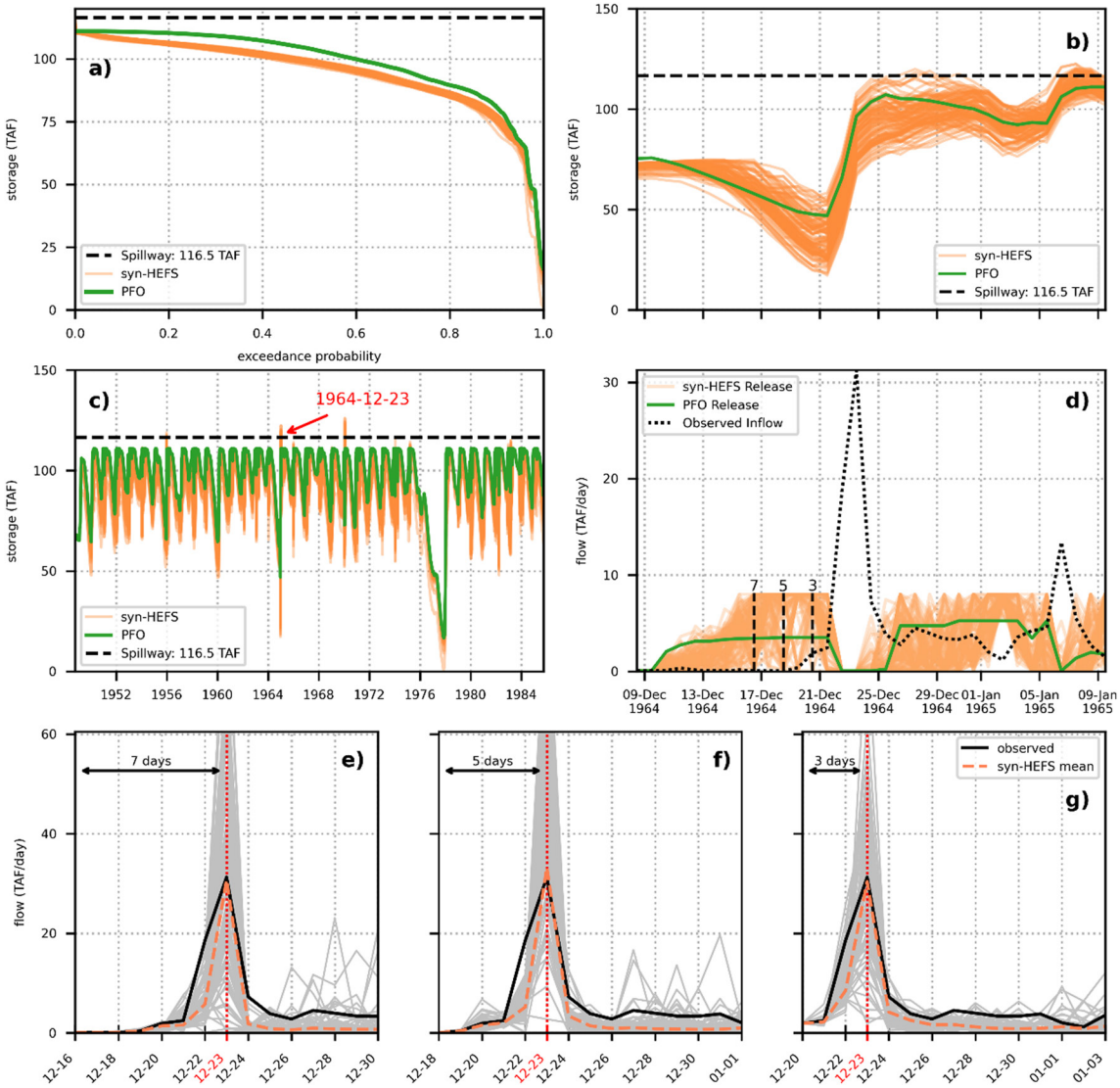

Figure 10. Forecast extension with syn-HEFS (orange) and perfect forecast operations (green) for the pre-hindcast period (WY1949–1985). (a) Storage exceedance and (c) time series plot for the entire pre-hindcast period. (b) Storage sequence and (d) controlled release (R_t^{Ctrl}) sequence prior to, during, and after the December 1964 flood. (e–g) syn-HEFS ensemble sample for December 1964 flood event at 7, 5, and 3-day leads from the maximum daily inflow.

5. Discussion and Conclusion

This study presented two novel methods to generate synthetic hydrological ensemble forecasts across multiple sites and lead times with ensemble sizes typical of current forecasting systems. In the first approach (syn-HEFS), we develop synthetic forecasts of the hydrologic ensemble directly, with innovations meant to capture inter-ensemble behavior. In the second approach (syn-GEFS), we develop synthetic forecasts of ensemble mean meteorological forcing, which are used to drive a process-based hydrologic ensemble forecast model. Both methods define an important contribution to the rapidly evolving area of forecast-informed reservoir operations, because they create new opportunities for testing the robustness of FIRO policies in a variety of contexts.

We demonstrate the utility of the synthetic forecast algorithms in a case study of the Russian River—Lake Mendocino system, using the HEFS model currently in operational use at NWS River Forecast Centers across the United States and the recently developed EFO FIRO-based operating policy (Delaney et al., 2020; Jasperse et al., 2020). This case study presents a real world and highly complex test bed for this approach, as the synthetic forecast model must replicate consistent forecast behavior across 3 sites, 14 lead times, and 61 ensemble members for full representation of system dynamics.

Evaluations of both synthetic forecast methods using common ensemble verification metrics showed a high degree of parity between the synthetic forecasts and HEFS, particularly in aggregate verification at the upper flow quantiles. The synthetic output emulated the calibration and skill of the HEFS forecasts well while preserving important biases. However, both synthetic methods struggled to replicate nuanced HEFS behavior for specific extreme events, highlighting some limitations of the modeling approach. At the granularity of individual events, the high degree of heterogeneity in HEFS forecast performance provides a unique simulation challenge. While HEFS forecast values in cumulative inflow statistics were often captured in the tails of the synthetic forecast sampling distribution, there were some tendencies to systematically underpredict the cumulative ensemble variance, particularly at longer forecast leads. The cumulative forecast integrates forecasts for each day leading up to an extreme event, which makes diagnosing the source of cumulative forecast discrepancies in the synthetic samples difficult. This is especially challenging at longer leads as the number of forecasted days (and associated errors) prior to the event increases. In addition, impactful type-1 errors (i.e., large HEFS forecasts when observed flows are very small) are difficult to replicate with the modeling structure presented. The addition of jitters to the conditional expectation in the synthetic forecast model might address these issues, since small increases in the conditional expectation when observed flows are very small will not only increase the mean of the forecasts but will also increase the variance through the heteroscedasticity model, allowing some ensemble members to reflect higher forecasted flows. This will be pursued in future work.

Importantly, we did not find evidence in this case study to suggest that the multi-step, computationally expensive syn-GEFS approach outperformed the more computationally efficient direct HEFS simulation (syn-HEFS) for large inflow events of greatest interest, at least for the SAC-SMA hydrologic model in HEFS used to converted synthetic meteorological forecasts to ensemble streamflow forecasts. We therefore conclude that syn-HEFS is likely sufficient for many types of FIRO risk analysis and may be preferable to syn-GEFS, especially when computation expense is a factor. Nonetheless, there are also scenarios where syn-GEFS may be preferred. For instance, syn-GEFS constricts the uncertainty modeling to the meteorological input space, which may enable easier generation of synthetic samples with manipulations to forecast skill.

The operational validation of the synthetic forecasts (Section 4.3) was generally consistent with the ensemble forecast verification and validation presented in Sections 4.1 and 4.2. Across many synthetic forecast samples, reservoir operations forced with synthetic forecasts closely aligned with the HEFS output. However, HEFS produced large drawdowns prior to individual extreme events that were sometimes, but not always, matched when using the synthetic forecasts. While some fraction of synthetic forecasts (particular from syn-HEFS) replicated the HEFS trace in avoiding spills for all major extreme events, these synthetic forecast traces did not always lead to the same magnitude of drawdown as HEFS. This is because nuanced behavior in HEFS was absent from most or all synthetic forecast samples. Given that the HEFS hindcast is only one realization of an underlying stochastic process, we expect that individual synthetic forecast samples should deviate from the hindcast, especially when filtered through an operational model. Still, if the synthetic forecast model was fully capturing HEFS behavior during extremes, one would expect at least some samples to cause similar magnitude drawdowns during major extremes.

The synthetic forecasts struggled with one nuance in the HEFS hindcasts in particular, which was large overforecasts of inflow during low flow periods that occurred several days prior to an extreme inflow event, particularly when those forecasts were issued at long (7–14 day) lead times. The EFO model leveraged this information in HEFS to implement a safe and conservative release sequence, but synthetic forecasts around these extremes resulted in smaller drawdowns and occasionally in spills. Some of this difference between the synthetic forecast and HEFS-driven results suggests room for improvement in the synthetic forecast model. However, it also suggests that the EFO policy might be vulnerable to plausible (but unobserved) HEFS forecasts in which large overforecasts several days prior to extreme events do not occur.

This vulnerability of EFO around key events in the hindcast period is an important finding, since it highlights a non-trival risk that the EFO model could lead to spills if subjected to plausible ensemble forecasts to which it was not trained. Even if spills under the synthetic forecast traces for some extreme events are caused by biases in the synthetic forecast model, the high degree of parity between the synthetic forecasts and HEFS during other extreme events suggests that some of the spills are likely due to plausible forecast variability that the EFO policy does not manage well. This was also demonstrated in the pre-hindcast period, where results showed that the current EFO policy has a near 25% chance of spilling during a cluster of two inflow events in WY 1965, even

though synthetic forecasts during the first (and much larger) inflow peak were accurate at a week ahead lead time. Overall, these initial operational findings highlight the promise of synthetic forecasts as a way to advance robust FIRO policy risk analysis, but they also emphasize the need for further improvement in the synthetic forecast algorithm to match nuanced HEFS behavior during extremes.

Finally, we posit that the utility of synthetic forecasts extends well beyond historical risk analysis of existing FIRO policies. For example, synthetic forecasts could be used in the calibration of operational models directly by substantially expanding the size and diversity of the available training data. This follows data augmentation approaches used to prevent overfitting in machine learning (Forestier et al., 2017; Oh et al., 2020) and could support robust (Giuliani et al., 2021) or "noisy" (Brodeur et al., 2020; Gupta et al., 2022) optimization/calibration techniques for FIRO policy design. Parameters of the synthetic forecast model could also be adjusted to produce synthetic forecasts of higher accuracy than the current generation of HEFS forecasts, allowing an investigation into the operational value of forecast improvements. Another area of interest is exploring FIRO behavior under non-stationary conditions, which to date has only been attempted with simple scaling approaches (Jasperse et al., 2020). While the present modeling architecture is not specifically designed to accommodate non-stationarity, one could generate synthetic forecasts from the model fit to the hindcast period but using projected flows under climate change as observations during simulation. This approach would enable investigations of FIRO as a strategy for climate change resiliency, and could be conducted assuming forecast skill remains the same under future climate conditions or that forecast skill changes (for better or worse) in the future.

Data Availability Statement

All code and data for the synthetic ensemble forecast model are available in Brodeur (2023a, 2023b) while code and data for the EFO model are in Delaney and Brodeur (2024a, 2024b).

Acknowledgments
This study was supported by the U.S. National Science Foundation, Grants 1803563 and 2205239. We would like to acknowledge Nathan Baskett (Sonoma Water) for help in developing Figure 1, Weiming Hu (James Madison University) for assistance in parallelizing the EFO model, and Robert Hartman (R. K. Hartman Consulting) for help in conceptualizing the experimental design.

References

Aghakouchak, A., Chiang, F., Huning, L. S., Love, C. A., Mallakpour, I., Mazdiyasni, O., et al. (2020). Climate extremes and compound hazards in a warming world. *Annual Review of Earth and Planetary Sciences*, *48*(1), 519–548. https://doi.org/10.1146/annurev-earth-071719-055228

Alemu, E. T., Palmer, R. N., Polebitski, A., & Meaker, B. (2011). Decision support system for optimizing reservoir operations using ensemble streamflow predictions. *Journal of Water Resources Planning and Management*, *137*(1), 72–82. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000088

American Meteorological Society. (2020). Forecast informed reservoir operations (FIRO) definition. Retrieved from https://glossary.ametsoc.org/wiki/Forecast-informed_reservoir_operations

Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, *525*(7567), 47–55. https://doi.org/10.1038/nature14956

Bellier, J., Zin, I., & Bontron, G. (2017). Sample stratification in verification of ensemble forecasts of continuous scalar variables: Potential benefits and pitfalls. *Monthly Weather Review*, *145*(9), 3529–3544. https://doi.org/10.1175/MWR-D-16-0487.1

Bertoni, F., Giuliani, M., Castelletti, A., & Reed, P. M. (2021). Designing with information feedbacks: Forecast informed reservoir sizing and operation. *Water Resources Research*, *57*(3), 1–19. https://doi.org/10.1029/2020WR028112

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. P. (2023a). FIRO_synthetic-forecast ensembles dataset v0 (Version 0) [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.7688974

Brodeur, Z. P. (2023b). FIRO_synthetic-forecast-ensembles: October 18, 2023 release (v1.0.1) [Software]. Zenodo. https://doi.org/10.5281/zenodo.10019063

Brodeur, Z. P., Herman, J. D., & Steinschneider, S. (2020). Bootstrap aggregation and cross-validation methods to reduce overfitting in reservoir control policy search. *Water Resources Research*, *55*(8), e2020WR027184. https://doi.org/10.1029/2020WR027184

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), 1–24. 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). Scientific framework for sustainable water management. *Water Resources Research*, *51*(8), 6110–6124. https://doi.org/10.1002/2015WR017114

Brown, J. D., Seo, D. J., & Du, J. (2012). Verification of precipitation forecasts from NCEP's short-range ensemble forecast (SREF) system with reference to ensemble streamflow prediction using lumped hydrologic models. *Journal of Hydrometeorology*, *13*(3), 808–836. https://doi.org/10.1175/JHM-D-11-036.1

Cassagnole, M., Ramos, M. H., Zalachori, I., Thirel, G., Garçon, R., Gailhard, J., & Ouillon, T. (2021). Impact of the quality of hydrological forecasts on the management and revenue of hydroelectric reservoirs-a conceptual approach. *Hydrology and Earth System Sciences*, *25*(2), 1033–1052. https://doi.org/10.5194/hess-25-1033-2021

Castelletti, A., Ficchì, A., Cominola, A., Segovia, P., Giuliani, M., Wu, W., et al. (2023). Model predictive control of water resources systems: A review and research agenda. *Annual Reviews in Control*, *55*, 442–465. https://doi.org/10.1016/j.arcontrol.2023.03.013

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B., & Wilby, R. (2004). The Schaake shuffle: A method for reconstructing space-time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology*, *5*(1), 243–262. https://doi.org/10.1175/1525-7541(2004)005<0243:TSSAMF>2.0.CO;2

Corringham, T. W., Martin Ralph, F., Gershunov, A., Cayan, D. R., & Talbot, C. A. (2019). Atmospheric rivers drive flood damages in the western United States. *Science Advances*, *5*(12), 1–8. https://doi.org/10.1126/sciadv.aax4631

Côté, P., & Leconte, R. (2016). Comparison of stochastic optimization algorithms for hydropower reservoir operation with ensemble streamflow prediction. *Journal of Water Resources Planning and Management*, *142*(2), 04015046. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000575

DeFlorio, M. J., Waliser, D. E., Guan, B., Lavers, D. A., Ralph, F. M., & Vitart, F. (2018). Global assessment of atmospheric river prediction skill. *Journal of Hydrometeorology*, *19*(2), 409–426. https://doi.org/10.1175/JHM-D-17-0135.1

DeHaan, L. L., Martin, A. C., Weihs, R. R., Delle Monache, L., & Ralph, F. M. (2021). Object-based verification of atmospheric river predictions in the Northeast Pacific. *Weather and Forecasting*, *36*(4), 1575–1587. https://doi.org/10.1175/WAF-D-20-0236.1

Delaney, C. J., & Brodeur, Z. P. (2024a). Lake Mendocino EFO model—Synthetic forecasts dataset: January 2, 2024 release (v1.0.0) [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.10453068

Delaney, C. J., & Brodeur, Z. P. (2024b). LkMendoEfoSynWrr: January 2, 2024 release (v1.0.1) [Software]. Zenodo. https://doi.org/10.5281/zenodo.10453090

Delaney, C. J., Hartman, R. K., Mendoza, J., Dettinger, M., Delle Monache, L., Jasperse, J., et al. (2020). Forecast informed reservoir operations using ensemble streamflow predictions for a multipurpose reservoir in northern California. *Water Resources Research*, *56*(9), e2019WR026604. https://doi.org/10.1029/2019WR026604

Demargne, J., Wu, L., Regonda, S. K., Brown, J. D., Lee, H., He, M., 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

Dettinger, M. D. (2013). Atmospheric rivers as drought busters on the U.S. West Coast. *Journal of Hydrometeorology*, *14*(6), 1721–1732. https://doi.org/10.1175/JHM-D-13-02.1

Dettinger, M. D., Ralph, F. M., Das, T., Neiman, P. J., & Cayan, D. R. (2011). Atmospheric rivers, floods and the water resources of California. *Water*, *3*(4), 445–478. https://doi.org/10.3390/w3020445

Dowd, C. (2020). A new ECDF two-sample test statistic. ArXiv:2007.01360 [Stat.ME]. https://doi.org/10.48550/arXiv.2007.01360

Dowd, C. (2023). twosamples: Fast permutation based two sample tests. R package version 2.0.1. Retrieved from https://cran.r-project.org/web/packages/twosamples/twosamples.pdf

Ficchi, A., Raso, L., Dorchies, D., Pianosi, F., Malaterre, P.-O., Van Overloop, P.-J., & Jay-Allemand, M. (2016). Optimal operation of the multireservoir system in the seine River Basin using deterministic and ensemble forecasts. *Journal of Water Resources Planning and Management*, *142*(1), 05015005. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000571

Forestier, G., Petitjean, F., Dau, H. A., Webb, G. I., & Keogh, E. (2017). Generating synthetic time series to augment sparse datasets. In *Proceedings of the IEEE International Conference on Data Mining, ICDM, 2017-November* (pp. 865–870). https://doi.org/10.1109/ICDM.2017.106

Giuliani, M., Lamontagne, J. R., Reed, P. M., & Castelletti, A. (2021). A state-of-the-art review of optimal reservoir control for managing conflicting demands in a changing world. *Water Resources Research*, *57*(12), e2021WR029927. https://doi.org/10.1029/2021WR029927

Gleick, P. H. (2002). Soft water paths. *Nature*, *418*(July), 373. https://doi.org/10.1038/418373a

Guan, H., Zhu, Y., Sinsky, E., Fu, B., Zhou, X., Li, W., et al. (2019). The NCEP GEFS-v12 reforecasts to support subseasonal and hydrometeorological applications. In *44th NOAA annual climate diagnostics and prediction workshop, (October)* (pp. 78–81).

Gupta, R. S., Steinschneider, S., & Reed, P. M. (2022). A multi-objective paleo-informed reconstruction of western US weather regimes over the past 600 years. *Climate Dynamics*, *60*(1–2), 339–358. https://doi.org/10.1007/s00382-022-06302-4

Hamill, T. M., Bates, G. T., Whitaker, J. S., Murray, D. R., Fiorino, M., Galarneau, T. J., et al. (2013). NOAA's second-generation global medium-range ensemble reforecast dataset. *Bulletin of the American Meteorological Society*, *94*(10), 1553–1565. https://doi.org/10.1175/BAMS-D-12-00014.1

Hanak, E., Lund, J., Dinar, A., Gray, B., Howitt, R., Mount, J., et al. (2011). Managing California's water. Retrieved from http://www.ppic.org/content/pubs/report/R_211EHR.pdf

Hartmann, D. L. (2016). *Global physical climatology* (2nd ed.). Elsevier.

IPCC. (2021). Climate change 2021: The physical science basis. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, et al. (Eds.), *Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change*. Cambridge University Press. https://doi.org/10.1017/9781009157896

Jasperse, J., Ralph, F. M., Anderson, M., Brekke, L. D., Malasavage, N., Dettinger, M. D., et al. (2020). Lake Mendocino forecast informed reservoir operations final viability assessment. In *Lake Mendocino FIRO steering committee, 28 December 2020, San Diego, CA*. Scripps Institute Center for Western Weather and Water Extremes. Retrieved from https://escholarship.org/uc/item/3b63q04n

Kim, Y.-O., Eum, H.-I., Lee, E.-G., & Ko, I. H. (2007). Optimizing operational policies of a Korean multireservoir system using sampling stochastic dynamic programming with ensemble streamflow prediction. *Journal of Water Resources Planning and Management*, *133*(1), 4–14. https://doi.org/10.1061/(ASCE)0733-9496(2007)133:1(4)

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

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)

McInerney, D., Thyer, M., Kavetski, D., Lerat, J., & Kuczera, G. (2017). Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors. *Water Resources Research*, *53*(3), 2199–2239. https://doi.org/10.1111/j.1752-1688.1969.tb04897.x

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

Nevo, S., Morin, E., Gerzi Rosenthal, A., Metzger, A., Barshai, C., Weitzner, D., et al. (2022). Flood forecasting with machine learning models in an operational framework. *Hydrology and Earth System Sciences*, *26*(15), 4013–4032. https://doi.org/10.5194/hess-26-4013-2022

Nicholson, W. B., Matteson, D. S., & Bien, J. (2021). BigVAR: Dimension reduction methods for multivariate time series. R Package version 1.0.6. Retrieved from https://cran.r-project.org/web/packages/BigVAR

Nicholson, W. B., Wilms, I., Bien, J., & Matteson, D. S. (2020). High dimensional forecasting via interpretable vector autoregression. *Journal of Machine Learning Research*, *21*, 1–52. Retrieved from https://www.jmlr.org/papers/volume21/19-777/19-777.pdf

NOAA/NWS California/Nevada River Forecast Center (CNRFC). (2022). Hydrologic ensemble forecast system (HEFS) streamflow forecast output, Lake Mendocino, Ukiah Junction, Hopland Junction, CA (LAMC1, UKAC1, HOPC1). Brett Whitin, P.E., CNRFC.

NOAA Physical Sciences Laboratory (NOAA PSL). (2022). NCEP global ensemble forecast system (GEFS) version 2 data archive. Brett Whitin, P.E., CA/NV RFC.

Oh, C., Han, S., & Jeong, J. (2020). Time-series data augmentation based on interpolation. *Procedia Computer Science*, *175*(2019), 64–71. https://doi.org/10.1016/j.procs.2020.07.012

Ralph, F. M., Cannon, F., Tallapragada, V., Davis, C. A., Doyle, J. D., Pappenberger, F., et al. (2020). West coast forecast challenges and development of atmospheric river reconnaissance. *Bulletin of the American Meteorological Society*, *101*(8), E1357–E1377. https://doi.org/10.1175/bams-d-19-0183.1

Raso, L., Schwanenberg, D., van de Giesen, N. C., & van Overloop, P. J. (2014). Short-term optimal operation of water systems using ensemble forecasts. *Advances in Water Resources*, *71*, 200–208. https://doi.org/10.1016/j.advwatres.2014.06.009

Rayner, S., Lach, D., & Ingram, H. (2005). Weather forecasts are for wimps: Why water resource managers do not use climate forecasts. *Climatic Change*, *69*(2–3), 197–227. https://doi.org/10.1007/s10584-005-3148-z

Rougè, 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), 04023015. https://doi.org/10.1061/jwrmd5.wreng-5934

Schoups, G., & Vrugt, J. A. (2010). A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. *Water Resources Research*, *46*(10), 1–17. https://doi.org/10.1029/2009WR008933

Semmendinger, K., Lee, D., Fry, L., & Steinschneider, S. (2022). Establishing opportunities and limitations of forecast use in the operational management of highly constrained multiobjective water systems. *Journal of Water Resources Planning and Management*, *148*(8), 04022044. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001585

Siegert, S., Bröcker, J., & Kantz, H. (2012). Rank histograms of stratified Monte Carlo ensembles. *Monthly Weather Review*, *140*(5), 1558–1571. https://doi.org/10.1175/MWR-D-11-00302.1

Slater, L. J., Arnal, L., Boucher, M.-A., Chang, A. Y.-Y., Moulds, S., Murphy, C., et al. (2023). Hybrid forecasting: Blending climate predictions with AI models. *Hydrology and Earth System Sciences*, *27*(9), 1865–1889. https://doi.org/10.5194/hess-27-1865-2023

Sonoma Water. (2016). *Fish habitat flows and water rights project draft environmental impact report*. Sonoma County Water Agency.

Stedinger, J. R., & Taylor, M. R. (1982). Synthetic streamflow generation: 1. Model verification and validation. *Water Resources Research*, *18*(4), 919–924. https://doi.org/10.1029/WR018i004p00919

Swain, D. L., Langenbrunner, B., Neelin, J. D., & Hall, A. (2018). Increasing precipitation volatility in twenty-first-century California. *Nature Climate Change*, *8*(5), 427–433. https://doi.org/10.1038/s41558-018-0140-y

Todini, E. (2018). Paradigmatic changes required in water resources management to benefit from probabilistic forecasts. *Water Security*, *3*, 9–17. https://doi.org/10.1016/j.wasec.2018.08.001

Troin, M., Arsenault, R., Wood, A. W., Brissette, F., & Martel, J.-L. (2021). Generating ensemble streamflow forecasts: A review of methods and approaches over the past 40 years. *Water Resources Research*, *57*(7), e2020WR028392. https://doi.org/10.1029/2020WR028392

Turner, S. W. D., Bennett, J. C., Robertson, D. E., & Galelli, S. (2017). Complex relationship between seasonal streamflow forecast skill and value in reservoir operations. *Hydrology and Earth System Sciences*, *21*(9), 4841–4859. https://doi.org/10.5194/hess-21-4841-2017

USACE. (2003). *United state Army Corps of Engineers, Coyote valley dam and Lake Mendocino, Russian river, California, water control manual: Appendix I to master water control manual Russian River basin, California*. U.S. Army Corps of Engineers, Sacramento District.

Vogel, R. M. (2017). Stochastic watershed models for hydrologic risk management. *Water Security*, *1*, 28–35. https://doi.org/10.1016/j.wasec.2017.06.001

Whateley, S., Palmer, R. N., & Brown, C. (2014). Seasonal hydroclimatic forecasts as innovations and the challenges of adoption by water managers. *Journal of Water Resources Planning and Management*, *141*(5), 04014071. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000466

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

Wurtz, D., Setz, T., Chalabi, Y., Boudt, C., Chausse, P., & Miklovac, M. (2020). fGarch: Rmetrics—Autoregressive conditional heteroskedastic modeling. R package version 3042.83.2. Retrieved from https://cran.r-project.org/web/packages/fGarch