Incorporating Mid-Term Temperature Predictions into Streamflow Forecasts and Operational Reservoir Projections in the Colorado River Basin

Erin Towler[1]; David Woodson[2]; Sarah Baker[3]; Ming Ge[4]; James Prairie[5]; Balaji Rajagopalan[6]; Seth Shanahan[7]; and Rebecca Smith[8]

Abstract: Skillful mid-term temperature predictions (up to five years out) offer a potential opportunity for water managers, especially in the Colorado River Basin (CRB), where streamflows are sensitive to temperature. The purpose of this paper is to develop and demonstrate a framework for how mid-term temperature predictions can be incorporated into streamflow forecasting and operational projections. The framework consists of three steps. First, 5-year average temperature predictions are obtained from two large ensemble climate model datasets. Second, hindcasts from the Ensemble Streamflow Predictions (ESP), an operationally used forecast method in the CRB, are post-processed using the 5-year average temperature predictions; specifically, a tercile-based block bootstrap resampling approach generates weighted streamflow ensembles called WeighESP. Third, ESP and WeighESP are run through an operational model, the Colorado River Mid-term Modeling System (CRMMS). Compared to ESP, WeighESP marginally improves streamflow forecast accuracy in the multi-year hindcasts up to five years out (i.e., years 1-5, 2-5, 2-4, and 2-3). The multi-year hindcasts show median annual root mean square error (RMSE) improvements between 437,000 and 771,000 m^3 (354 and 625 thousand acre-feet). Improvements in streamflow accuracy are more pronounced for the most recent hindcast run dates through 2016, partially due to ESP being run with climate time series data from 1981 to 2010. Next, CRMMS translates the streamflow forecasts into operational projections of end of calendar year (EOCY) pool elevations. WeighESP improves the accuracy of EOCY predictions, but mainly for longer leads of 3- and 4-years. For the 4-year lead, the median RMSE improves by 1.1 and 0.7 m (3.5 and 2.3 ft) for Lakes Powell and Mead, respectively. Although marginal improvements in pool elevation could be beneficial, not being realized until longer leads is a limitation. This study describes the need for better predictive tools at the mid-term timescale and underscores the importance of evaluating improvements in streamflow forecasts in decision-relevant terms. **DOI: 10.1061/(ASCE)WR.1943-5452.0001534.**

Introduction

Wood et al. (2020) provided an overview of streamflow forecasting in the Colorado River Basin (CRB), with a focus on the streamflow forecasts used to drive operational models for the US Bureau of Reclamation (Reclamation). Operational streamflow forecasts in the CRB are issued by the Colorado Basin River Forecasting Center (CBRFC), using the Ensemble Streamflow Prediction (ESP) method. ESP is a dynamic hydrological approach, initialized with basin-observed conditions, and run out with historical climate time series, to produce a probabilistic streamflow ensemble (Day 1985). The forecasted ESP streamflow traces are used to drive one of Reclamation's operational models, the Colorado River Mid-term Modeling System (CRMMS), formerly known as the Mid-Term Probabilistic Operations Model (MTOM). For a 5-year time horizon, CRMMS projects probabilistic monthly reservoir levels and releases, along with other water management variables. The first year of these risk-based projections is currently part of Reclamation's framework supporting stakeholder decision making in the CRB, and studies are under way that could result in the use of CRMMS projections for years two through five. Any improvements to ESP-based data that drive CRMMS would be welcome by the community and stakeholders in the basin, with potential benefits for management and planning.

In the CRB, there has been ongoing interest in quantifying the sensitivity of streamflow to temperature and precipitation. Woodhouse et al. (2016) showed that although cool season precipitation explains most of the Upper CRB streamflow variability, temperature exerts strong control under certain conditions. Further, temperature has been shown to be an influential control on runoff efficiency over past centuries (Woodhouse and Pederson 2018). Milly and Dunne (2020) reported a sensitivity of −9.3% streamflow per degree Celsius (C), and Udall and Overpeck (2017) found a similar decrease of 7% per degree C. On the other hand, Hoerling et al. (2019) attributed most of the streamflow decrease to changes

[1] Project Scientist, National Center for Atmospheric Research (NCAR), Boulder, CO 80307 (corresponding author). ORCID: https://orcid.org/0000-0002-1784-1346. Email: towler@ucar.edu; towlere@gmail.com

[2] Ph.D. Student, Dept. of Civil, Environmental, and Architectural Engineering, Univ. of Colorado at Boulder, Boulder, CO 80309. ORCID: https://orcid.org/0000-0001-9852-5355. Email: David.Woodson@colorado.edu

[3] Civil Engineer, Lower Colorado Basin, US Bureau of Reclamation, Boulder, CO 80301. Email: sabaker@usbr.gov

[4] Associate Scientist, National Center for Atmospheric Research (NCAR), P.O. Box 3000, Boulder, CO 80307. Email: mingge@ucar.edu

[5] Hydrologic Engineer, US Bureau of Reclamation, Boulder, CO 80301. Email: jprairie@usbr.gov

[6] Professor, Dept. of Civil, Environmental, and Architectural Engineering, Univ. of Colorado at Boulder, Boulder, CO 80309; Fellow, Cooperative Institute for Research in Environmental Sciences, Boulder, CO 80309. Email: balajir@colorado.edu

[7] Colorado River Programs Manager, Southern Nevada Water Authority, Las Vegas, NV 89106. Email: seth.shanahan@snwa.com

[8] Civil Engineer, US Bureau of Reclamation, Boulder, CO 80301. Email: rebeccasmith@usbr.gov

Note. This manuscript was submitted on March 24, 2021; approved on December 10, 2021; published online on February 9, 2022. Discussion period open until July 9, 2022; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Water Resources Planning and Management***, © ASCE, ISSN 0733-9496.

in precipitation and found that streamflow only decreases by 2.5% per degree C.

Given streamflow's sensitivity to climate, there have been efforts to improve streamflow forecasting in the CRB by incorporating future climate information. One approach has been to post-process the streamflow ensembles resulting from ESP. Werner et al. (2004) used an ESP member weighting scheme based on a climate index (e.g., the El Nino Southern Oscillation), which improved forecasting of spring runoff in three CRB sub-basins. Baker et al. (2021, 2019) demonstrated improvements over ESP using a k-nearest neighbors (k-nn) scheme, which weighed ESP based on 1-month and 3-month temperature and precipitation forecasts from the North American Multi-model Ensemble (NMME). Given that post-processing ESP has shown improvements on the seasonal time scale, there is a new opportunity to investigate potential improvements in mid-term prediction, or up to 5 years out.

The time horizon referred to as mid-term prediction in the water community overlaps with what is called "decadal prediction" in the climate community. Decadal climate prediction is a developing field that predicts potential climate variability and change on the timescale of 1 to 10 years in advance. These efforts use the same climate models and prescribed greenhouse gas forcings as their better-known counterparts, the multi-decadal to centennial climate change projections. Climate change projections have become fairly mainstream and well-known, such as from the Intergovernmental Panel on Climate Change (IPCC) Reports and their use in climate change impact studies. Climate change projections are uninitialized, i.e., they are free-running continuously through time. On the other hand, for decadal climate predictions, the climate model runs are initialized with observed current conditions. Although experimental, these predictions of the upcoming 1 to 10 years offer potential to capture decadal phenomenon that could contribute more skill than externally forced climate projections. Decadal climate predictions have been included in the experimental design in both phases 5 and 6 of the Coupled Model Intercomparison Project (CMIP5 and CMIP6). Global and regional evaluations have shown that decadal predictions have skill predicting temperature (Yeager et al. 2018). Precipitation is less skillful, although recent work suggests that using very large ensemble sizes (>70 members) can result in higher skill than previously found (Smith et al. 2019). Given the skill found on decadal time scales for temperature, research has begun to explore its potential usability. Towler et al. (2018) compared how decadal temperature predictions could be applied deterministically using an anomaly, versus probabilistically, which uses the likelihood of being in each tercile. Further, Towler and Yates (2020) assessed streamflow simulations using decadal temperature predictions in a process-based hydrological model versus an empirical/statistical approach; they found that including temperature improved the streamflow prediction from both approaches, but that there was substantial uncertainty without precipitation predictions. Temperature predictions have also been used to improve streamflow prediction for other timescales, such as seasonal forecasting (Lehner et al. 2017) and future climate change (Kiem et al. 2021). Henceforth, these climate predictions will be referred to as mid-term climate predictions, both for consistency with their application to water use, and to clarify the 5-year time horizon being explored.

The purpose of this paper is to develop and demonstrate a framework for how mid-term temperature predictions can be incorporated into streamflow forecasting as well as into operational reservoir projections. The framework has three steps. First, mid-term temperature predictions are obtained from two large ensemble climate datasets. Second, a method called "WeighESP" is developed to post-process ESP streamflows based on probabilistic temperature predictions. Third, streamflow ensembles are passed through CRMMS and the operational projections are evaluated within the context of current CRB operations and planning. This study provides two main contributions: first, the framework focuses on adding prediction skill for the mid-term planning horizon that aligns with an unmet societal need at this timescale (Sandgathe et al. 2020; Vera et al. 2010; Barsugli et al. 2009). Second, the framework provides a demonstration of how new streamflow forecasting methods can be tested and potentially implemented in operational practice, addressing a need to make research more useful to decision-making (NRC 2009).

Data

Climate Data

Observations

PRISM (Daly et al. 1997) is a gridded 4-km observational dataset of precipitation, minimum temperature (T_{min}), and maximum temperature (T_{max}). PRISM is available from 1980 to 2018. The average temperature (T_{avg}) was calculated as a function of T_{max} and T_{min}: $T_{avg} = 0.6T_{max} + 0.4T_{min}$ (Thornton et al. 1997). PRISM observations are used to assess the skill of the climate model datasets for a historical period.

Climate Model Datasets

This study examines two large ensemble climate datasets from NCAR's Community Earth System Model (CESM). The CESM decadal prediction large ensemble (DP) is utilized for the initialized prediction dataset (Yeager et al. 2018). This hindcast dataset is initialized annually on November 1st and is available for the time period between 1954 and 2017. Each annual initialization is run out for 122 months and includes 40 members (Yeager et al. 2018), where each member represents a unique climate trajectory or realization. Each ensemble member begins from a slightly different atmospheric state that is generated by randomly perturbing temperatures at round-off error levels. The ensemble spread results from internal climate variability. For this study, annual initializations from November 1980 through November of 2016 were used (for a total of 37 run dates, each with 40 members). For each run date, five-year hindcasts were used, i.e., for the first run date, November 1980 through October 1985 is used, and for the last run date, November of 2016 through October 2021 is used.

The DP is an experimental dataset, which is quite large since the annual initialization results in lead-based overlapping hindcasts. As such, it was decided to test a second large ensemble dataset: NCAR also has a corresponding uninitialized 40-member large climate projection ensemble (LE) for CESM (Kay et al. 2015). As with the DP, each LE ensemble member's trajectory results from small differences in initial conditions. For the LE, the initial condition temperature perturbation is imposed in 1920, and then members are run continuously forward in time. The LE uses the exact same model configuration and forcings as the DP, so their skill can be directly compared, offering insight on the value of the initialization for temperature predictions (see e.g., Goddard et al. 2012). The advantage of the LE is that it is similar to other Global Climate Model (GCM) data that is familiar to stakeholders in the CRB, and compared to the DP, is a smaller and more straightforward dataset to manipulate. For this study, for each of the 40 members of the LE, the continuous time series from November 1980 through October 2021 is used. Although not originally intended as a hindcast, the LE's continuous time series can be broken out and evaluated as such: for instance, November 1980 through October 1985 can be selected to correspond to the first DP run date, and so on.

Data Pre-Processing

Because climate models do not provide perfect representations of the Earth system, observational and model data need to be compared in terms of anomalies. All data is converted into anomalies using the 1981–2010 baseline period. This is straightforward for the LE projection dataset, since for long-running projection data, once it is past an initial spin-up period the climate model stabilizes and it is reasonable to assume that the model bias is constant in time (i.e., the LE is perturbed in 1920 but is only used for this study starting in November of 1980). However, for the DP, due to the annual initialization, the model bias is not constant in time, rather it is larger at times closer to initialization and decreases as the model approaches its preferred state (Meehl et al. 2014). As such, the anomalies need to be calculated as a function of their lead time. This is called "drift-correction" (Meehl et al. 2014) and is widely accepted and applied for decadal prediction datasets (e.g., Meehl and Teng 2012, 2014a, b; Towler et al. 2018). The protocol followed here to calculate the temperature prediction dataset anomalies is outlined by the US CLIVAR (CMIP–WGCM–WGSIP Decadal Climate Prediction Panel 2011).

Streamflow Data

Historical Streamflow

Reclamation provided monthly historical unregulated streamflow for the input sites used in CRMMS. As described in Baker (2019), there are 12 Upper CRB forecast locations in CRMMS (Fig. 1). Monthly flows were provided for 1980 to 2019. Unregulated flows were back calculated to simulate flows as if there were no reservoir regulation or depletion upstream of the forecast point, except for three transbasin diversions that are explicitly modeled in CRMMS (see Lukas et al. 2020 for details). Unregulated flows differ from naturalized flows, which represent the observed flows if no upstream reservoirs or diversions were present.

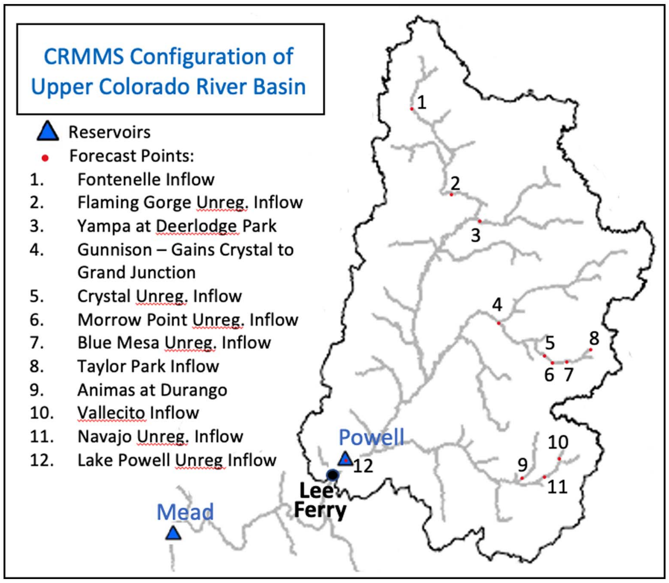

Fig. 1. Colorado River Mid-term Modeling System (CRMMS) configuration of the Upper Colorado River Basin, where dots represent approximate forecast locations, numbered from 1 to 12 and described in the top left table. Triangles represent reservoirs (only Lake Powell and Lake Mead are shown).

Ensemble Streamflow Predictions

ESP forecasts for the 12 Upper CRB locations were provided by the CBRFC. To create the forecasts, CBRFC uses a hydrologic model, the Sacramento Soil Moisture Accounting Model (SAC-SMA), which has been calibrated using historical conditions and temperature and precipitation traces from the 30-year climatological record (1981–2010).

Using the calibrated SAC-SMA, CBRFC issues monthly operational ESP forecasts; this study uses operational forecasts from 2012 to 2016. For each forecast, or so-called "run date", the model is run out 60 months. For each run date, the model is initialized to current basin conditions, and then run with temperature and precipitation time series from 1981 to 2010, creating a total of 30 ESP traces. Further, to have a longer record to evaluate forecast performance, CBRFC created reforecasts from 1981 to 2011. Reforecasts are run retrospectively, and unlike operational forecasts, they do not include short term forecasts of temperature and precipitation. These forecasts each only have 29 traces, reflecting the fact that the trace from the run date (i.e., year being forecast) is dropped; this avoids having a trace with perfect climate information.

This work uses ESP forecasts initialized with basin conditions every November from 1980 to 2016, for a total of 37 run dates. November was selected since it corresponds to the time when the DP hindcast dataset is initialized annually.

Climatology Streamflow Ensemble

To evaluate the streamflow forecasts, a reference forecast from climatology was derived for run dates from November 1980 through November 2016, i.e., 37 run dates. The climatology ensemble is comprised of historical unregulated streamflows. Similar to what was described in the ESP ensemble, to avoid using "perfect" climate information, for run dates from November 1980 through November 2009, the run date's observation is dropped, and the remaining 29 years of historical unregulated streamflows create the ensemble. For run dates from November 2010 through November 2016, all 30 years of the historical unregulated streamflows were part of the climatology ensemble.

Data Diagnostics and Mid-Term Climate Skill

Given the sensitivity of CRB flows to climate (e.g., Udall and Overpeck 2017; Woodhouse et al. 2016; Woodhouse and Pederson 2018; Milly and Dunne 2020; Hoerling et al. 2019), it is useful to look specifically at the relationship between climate and streamflow in the Upper CRB. Fig. 2 shows the relationship between the 5-year mean precipitation and temperature versus the naturalized streamflow at Lees Ferry, which is just downstream of the outflow for Lake Powell (Fig. 1). Naturalized streamflow is back calculated to have the effect of both upstream reservoir regulation and human induced depletions removed from the observed streamflow record. As expected, the 5-year average annual precipitation has the strongest correlation with 5-year average streamflow ($r = 0.95$). Although this is a very strong predictor for streamflow, the DP and LE do not show skill at predicting precipitation; Table 1 shows the anomaly correlation coefficient (ACC) between the large ensembles and observed climate for the 5-year mean. For precipitation, the ACC is 0.087 for the LE and –0.057 for the DP, indicating a lack of skill for this variable. Fig. 2 shows that T_{max} has a relatively high correlation with average streamflow ($r = –0.77$), followed by T_{avg} (–0.67), and T_{min} (–0.52). From Table 1, it can be seen that there is skill for both T_{min} (ACC = 0.76 for DP and 0.71 for LE) and T_{max} (0.54 for the DP and 0.48 for the LE). The skill of the DP is slightly higher than LE for the T_{min} and T_{max}.

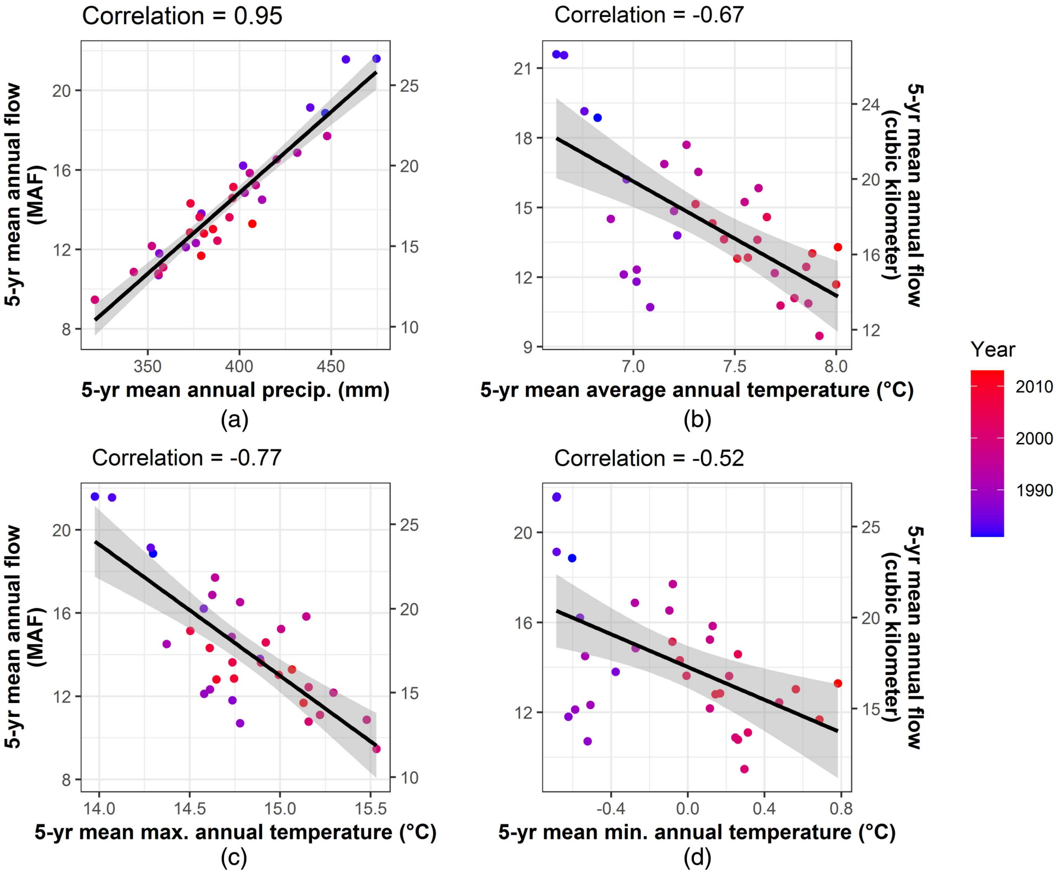

Fig. 2. Relationship between 5-year mean annual naturalized streamflow at Lees Ferry: (a) 5-year mean annual precipitation; (b) mean average annual temperature; (c) mean maximum annual temperature; and (d) mean minimum annual temperature from 1981 to 2018.

Methodology

Step1. Obtain Mid-Term Climate Prediction

LE and DP mid-term climate prediction ensembles are available as community resources for the research and practitioner community. Towler et al. (2018) discussed the two different formats familiar to practitioners that can be used to convey mid-term predictions: discrete and probabilistic. In this case, since the goal is to post-process streamflow ensembles, the probabilistic tercile-based approach was selected. This means that the mid-term temperature predictions are given in terms of the probability that the average temperature over the next 5 years will be below-normal, near-normal, and above-normal. This is similar to how seasonal forecasts are issued operationally [e.g., by NOAA's Climate Prediction Center and Columbia University's International Research Institute for Climate and Society (IRI)].

For each of the climate model datasets, T_{avg}, T_{min}, and T_{max} temperature variable anomalies were initially investigated, and the results of this paper focus on: (1) water year T_{avg} for the LE, (2) water year T_{avg} for the DP, (3) T_{min} winter (DJF) for the LE, and (4) Tm_{in} (DJF) for the DP. T_{avg} was selected since real-time forecasts from global climate centers are typically presented as averages (WMO 2020), making this a practical choice for potential utility. When considering T_{min} versus T_{max}, it was found that T_{min} had slightly higher skill than T_{max} (Table 1), but T_{max} had a higher correlation with streamflow than T_{min} (Fig. 2). However, since the Colorado River Basin is snow-dominated and T_{min} showed good skill (Table 1), winter T_{min} was also selected for demonstration. For each variable, the data was subset to include 5-year averages, averaged over the Upper CRB, starting in November for each of the 37 run dates (November 1980 to November 2016). November was selected for the DP, since that was the month of initialization. Although the LE is not a true hindcast, for consistency, the 5-year average beginning in November of the run dates was also used. Then for each run date, the 5-year average temperature anomaly for each ensemble member was calculated for both the LE and DP and compared to the 1981–2010 climatology.

Table 1. Anomaly correlation coefficient, ACC, for 5-year mean (i.e., multi-year forecast 1-5) minimum temperature (T_{min}), maximum temperature (T_{max}), and precipitation (Prec) for the decadal prediction large ensemble (DP) and large ensemble (LE) over the Upper Colorado River Basin

	ACC		
CESM	T_{min}	T_{max}	Prec
DP	0.76	0.54	−0.057
LE	0.71	0.48	0.087

The number of ensemble members that fell into the below-, near-, and above-normal terciles were counted. An example of this is shown for several run dates in Table 2. For example, for the DP hindcast initialized in November of 1980 and run out to October of 1985 (5 years), had 26 members whose 5-year average was below-normal, 11 members were near-normal, and 3 members were in the above-normal tercile. The LE projection over the same period was for 24 members to be in the below-normal tercile, 16 members in the near-normal tercile, and 0 in the above-normal tercile. Table 2 shows that regardless of the climate model, the counts for the first

Table 2. Number of ensemble members that were in each temperature tercile (below-, near-, and above- normal) for each run date of the decadal prediction large ensemble (DP) and large ensemble (LE), as well as what was observed using PRISM

	Run date		DP			LE			PRISM		
Index	Start	End	Below-	Near-	Above-	Below-	Near-	Above-	Below-	Near-	Above-
1	November 1980	October 1985	26	11	3	24	16	0	X	—	—
2	November 1981	October 1986	30	10	0	24	16	0	X	—	—
3	November 1982	October 1987	15	21	4	21	16	3	X	—	—
4	November 1983	October 1988	11	19	10	16	21	3	X	—	—
…	…	…	…	…	…	…	…	…	…	…	…
32	November 2011	October 2016	0	10	27	0	1	39	—	—	X
33	November 2012	October 2017	2	9	23	0	2	38	—	—	X
34	November 2013	October 2018	0	2	28	0	0	40	—	—	X
35	November 2014	October 2019	1	8	31	0	0	40	—	—	—

four run dates were skewed towards the below-normal temperature tercile, and the last four run date counts were skewed toward the above-normal tercile. This makes sense given the greenhouse gas forcing in the LE and DP models, consistent with the increasing trend in observed temperature over the basin. The final column shows the observed tercile, based on the 5-year temperature averages in PRISM. There were 35 five-year periods that could be validated with historical unregulated streamflows (run dates from November 1980 to November 2014, since the historical unregulated streamflows go through 2019) and 34 five-year periods that could be compared to PRISM observations (since the PRISM observations only go through 2018).

Step 2. Generate Weighted ESP Streamflow Ensembles

The WeighESP method generates an ensemble that is weighted to reflect the mid-term temperature tercile predictions obtained in Step 1. To review, for each run date, the ESP forecast is comprised of 29 (or 30) equally weighted streamflow traces, which are derived from temperature and precipitation time series from the climatological period. To describe how ESP is modified for WeighESP, the example of the November 1980 run date is used, where 29 traces are used (1981 is dropped). The first step is to bin each of the 29 ESP traces by their 5-year observed temperature average tercile, using 1981–2010 as the climatology (i.e., from the PRISM column of Table 2). The second step is to resample the 29 ESP traces, with replacement, based on the DP or LE predictions in Table 2. For example, for the November 1980 run date using the DP prediction, 65% (= 26/40) of the ESP streamflow traces are selected from the below-normal bin, ~27% (= 11/40) from the near-normal, and ~8% (= 3/40) from the above-normal. This generates a WeighESP sample of 100 members. The results for the Powell unregulated flows for all run dates can be seen in Fig. 3. For the November 1980 run date, WeighESP shifts the 5-year average distribution

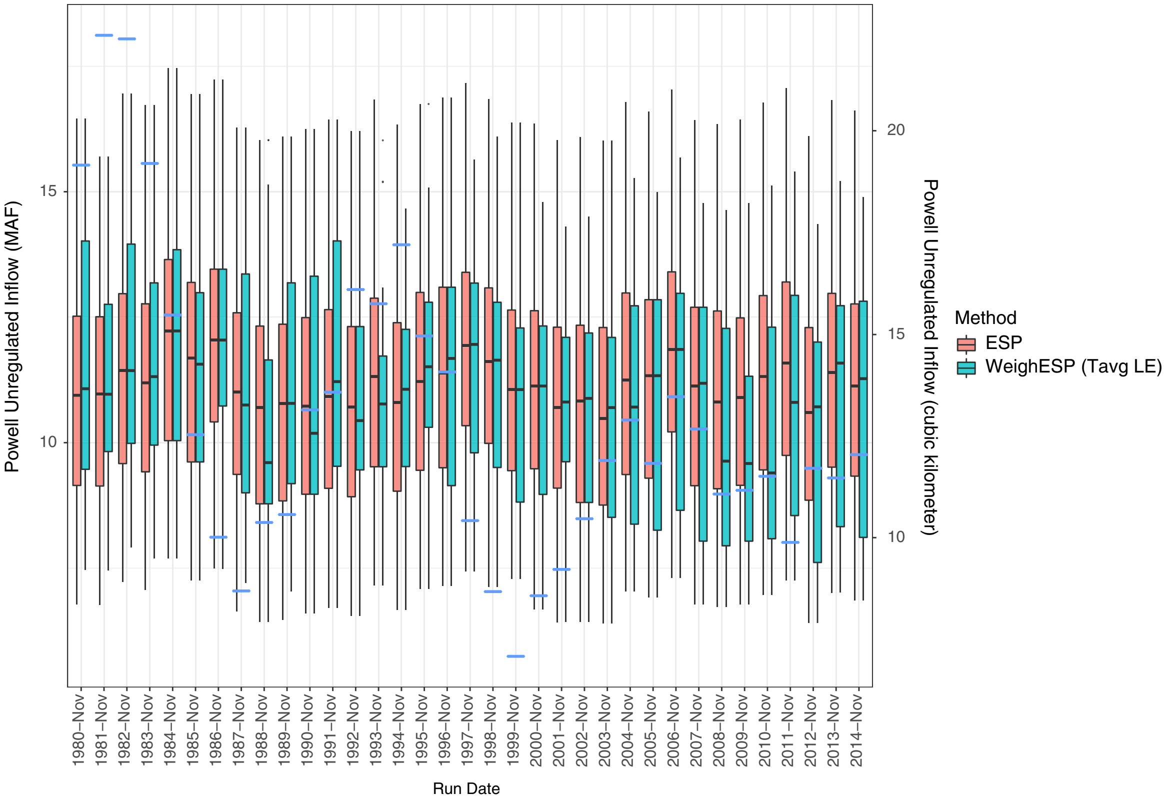

Fig. 3. Distribution of 5-year average Lake Powell unregulated inflows in million acre-feet (MAF) and cubic kilometers for ESP and WeighESP (boxplots) compared to historical data (horizontal dashes) for each run date.

towards higher flows, which was closer to the historical streamflow. However, there are also years where neither ESP or WeighESP distribution captures the observed flow (e.g., November 1981), highlighting the challenge of predicting streamflow at this timescale. This is repeated for all 12 Upper CRB forecast locations; it is noted that although the same weights are used for each run date (derived from Table 2), resampling occurs independently for each location, so different traces can be selected for each location.

The streamflow ensembles from ESP and WeighESP are evaluated using accuracy and skill metrics. To evaluate accuracy, root mean square error (RMSE) is calculated based on the error of each trace for each run date. Skill is calculated using the ranked probability skill score (RPSS; Wilks 1995). The RPSS evaluates forecast performance for multiple categories, which, in this case, are the streamflow terciles (below-, near-, and above-normal). The RPSS is calculated relative to the climatology ensemble using the library SpecsVerification in R.

Step 3. Evaluate Operational Projections

The ESP and WeighESP traces for the 12 Upper CRB forecast locations generated in Step 2 are used as inputs into CRMMS, Reclamation's mid-term operational probabilistic projection model. CRMMS was developed using the RiverWare software (Zagona et al. 2001) and is a rule-based water management model that simulates reservoir operations in the CRB. CRMMS results provide stakeholders with risk-based projections of monthly reservoir levels and basin conditions. In this study, the focus is on projections of pool elevations for Lakes Powell and Mead because they store the majority of water in the system [50 MAF (61.67 km^3) of the 60 MAF (74 km^3) of the system storage modeled in CRMMS] and are important for basin operations. End of year pool elevation projections are evaluated in terms of annual RMSE.

This study uses CRMMS in the context of the Colorado River Basin Operational Prediction Testbed (CRBOPT), which is a framework for assessing the skill of mid-term streamflow forecasts and associated CRMMS modeled operational projections in the CRB (Baker et al., forthcoming). The CRBOPT ingests Upper Basin streamflow forecasts and runs them through CRMMS. CRMMS solves for all modeled basin variables, including operations at twelve reservoirs, operating conditions, and water uses in the Lower Basin. The CRBOPT sets the Lower CRB intervening flows, Upper CRB tunnel diversions, and Lower Basin water use to historical values; this means that the only input being adjusted is the Upper CRB streamflow.

Results

Streamflow Ensembles

Although streamflow ensembles are generated for all 12 forecast points, results are shown for the most downstream point, the Lake Powell unregulated inflows (forecast location 12 in Fig. 1). The evaluation on the 5-year average is designated "1-5", i.e., for each trace, the monthly streamflows over the 60 months from the run date are averaged before they are evaluated. Evaluation metrics are also calculated on several additional multi-year and individual year forecasts. Multi-years include the aforementioned, 1-5, as well as 2-5, 2-4, and 2-3. Individual years included 1, 2, 3, 4, and 5; e.g., year 1 would include averaging streamflow over the first 12 months of the forecast (i.e. November through October of the following year).

Fig. 4 shows annual RMSE for Lake Powell unregulated inflows. Each boxplot is comprised of the RMSE for each run date; the run date sample size (n) depends on how many observational years were available to validate: n = 35 for 1-5 and 2-5, n = 36 for 2-4, n = 37 for 2-3. Fig. 4 compares climatology, ESP, and the four WeighESP hindcasts. The median annual RMSEs from the boxplots are shown in Table 3; lower RMSE values are better (RMSE = 0 indicates a perfect prediction). Across all forecast years, except for individual year 5, the median error from the climatology ensemble (i.e., the historical unregulated streamflows) is always higher (worse) than ESP and the WeighESP forecasts. This shows the importance of initial conditions in the streamflow forecasts through year 4. Further, the ESP median is always higher (worse) than the WeighESP multi-year forecasts (1-5, 2-5, 2-4, and 2-3). Water year T_{avg} (LE) had the lowest RMSE for three of the multi-year forecasts, and the second lowest RMSE for the remaining multi-year forecast (i.e., 1-5). In this study, improvements in multi-year streamflow forecasts might be expected, since WeighESP is conditioned on multi-year temperature forecasts. However, it is also interesting to look at the individual year forecasts. For year 1, when initial

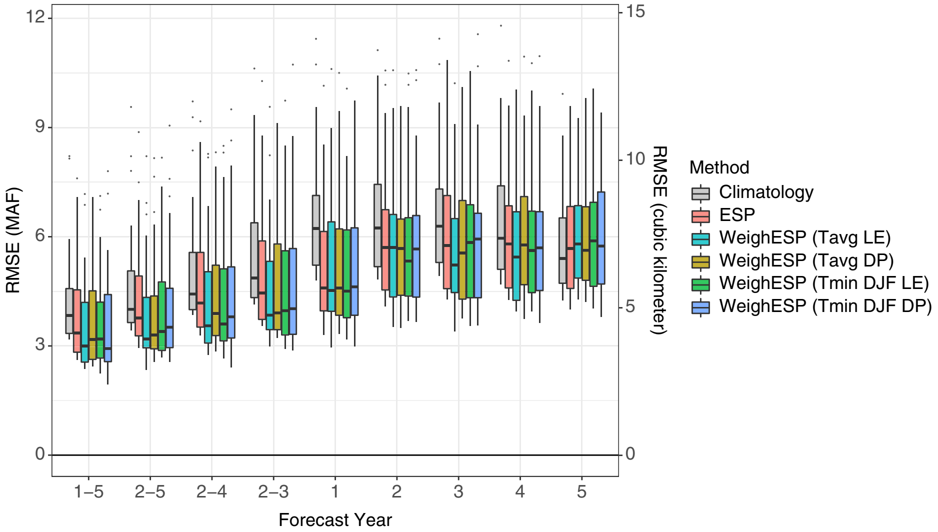

Fig. 4. For Lake Powell unregulated inflows, annual root mean squared error (RMSE) for climatology, ESP, and four WeighESP hindcasts [water year T_{avg} from the LE and DP and winter (DJF) T_{min} from the LE and DP] for traces averaged over multi-year and individual forecast years (using hindcast run dates from November 1980 to November 2016).

Table 3. For Lake Powell unregulated inflows, median annual root mean square error (RMSE) for climatology, ESP, and four WeighESP hindcasts in cubic kilometers (km^3)

	Median root mean square error (RMSE) (km^3)					
Forecast year	Climatology	ESP	T_{avg} (LE)	T_{avg} (DP)	$T_{min,DJF}$ (LE)	$T_{min,DJF}$ (DP)
1-5	4.72	4.14	3.70	3.92	3.93	**3.61**
2-5	4.94	4.64	**3.93**	4.08	4.19	4.33
2-4	5.46	5.15	**4.38**	4.79	4.44	4.69
2-3	6.01	5.49	**4.74**	4.82	4.89	4.96
1	7.68	5.66	5.57	5.67	**5.56**	5.70
2	7.69	7.04	7.03	7.00	**6.59**	6.99
3	7.76	7.11	**6.45**	6.85	7.20	7.32
4	7.35	7.16	**6.71**	7.11	6.93	7.02
5	**6.66**	7.00	7.16	6.94	7.26	7.09

Note: Bold shows the minimum for each forecast year.

Table 4. For Lake Powell unregulated inflows, difference between ESP and WeighESP (T_{avg} LE) median root mean square error (RMSE) in cubic kilometers (km^3) and million acre-feet (MAF)

Forecast year	Difference in median RMSE in km^3 (MAF)
1-5	0.437 (0.354)
2-5	0.716 (0.580)
2-4	0.771 (0.625)
2-3	0.749 (0.607)
1	0.083 (0.068)
2	0.005 (0.004)
3	0.662 (0.537)
4	0.444 (0.360)
5	−0.151 (−0.123)

conditions have the most impact, the ESP median is very similar to the WeighESP medians, though T_{min} DJF (LE) had the lowest RMSE. For the other individual years, ESP and the WeighESP medians are relatively similar. To summarize, Table 4 shows that taking the difference between ESP and WeighESP for T_{avg} (LE), offers an improvement of 436,652 and 770,925 m^3 [354 and 625 thousand acre-feet (KAF)] for multiyear forecasts, and improvements in individual year forecasts, except in year 5. The RMSE improvements for unregulated inflows to Lake Powell are more pronounced when looking at the most recent run dates, i.e., the run dates from November 2010 to November 2016; this is shown for the multi-year forecasts in Fig. 5. For run dates starting in November of 2010: n = 5 for 1-5 and 2-5, n = 6 for 2-4, n = 7 for 2-3. This shows that using equally weighted climate traces for 1981–2010, which is the current ESP practice, reduces the accuracy of the later run dates presumably because the traces do not include the more recent, warmer temperature time series.

Fig. 6 shows the RPSS results for Lake Powell unregulated inflows, using the climatology ensemble as a reference for ESP and WeighESP. The highest skill compared to climatology occurs in the forecast for individual year 1, reflecting the importance of the initial conditions. In year 1, ESP has the highest median, though the WeighESP medians are similar. Positive median skill is also seen in years 2-5, 2, 3, 4, and 5, but median skill is negative when averaging over other years (e.g. 1-5, 2-4, and 2-3). The RPSS is a probabilistic measure of skill, and the lack of skill found may indicate that the resulting trace ensembles are under confident or lacking discrimination.

Operational Projections

Based on the better median RMSEs from the LE-conditioned temperature predictions (Table 3), it was decided that the WeighESP streamflow ensembles conditioned from the water year T_{avg} (LE) would be run through CRMMS. To be more decision-relevant, CRMMS accuracy evaluation was performed differently than what was done directly on the unregulated streamflow. Specifically, RMSE was calculated on the December (end of calendar year, EOCY) pool elevation ensembles for each year of a given 5-year projection period and repeated for every 5-year block in the 1981–2017 hindcast (n = 31). EOCY elevations are important because calendar year-based operations are determined for the upcoming year based on elevation projections at the end of the current year.

Fig. 7 depicts an example 5-year CRMMS simulation that includes ESP and WeighESP traces initialized in November 2001 and run out through 2006. To assess skill across a time period including

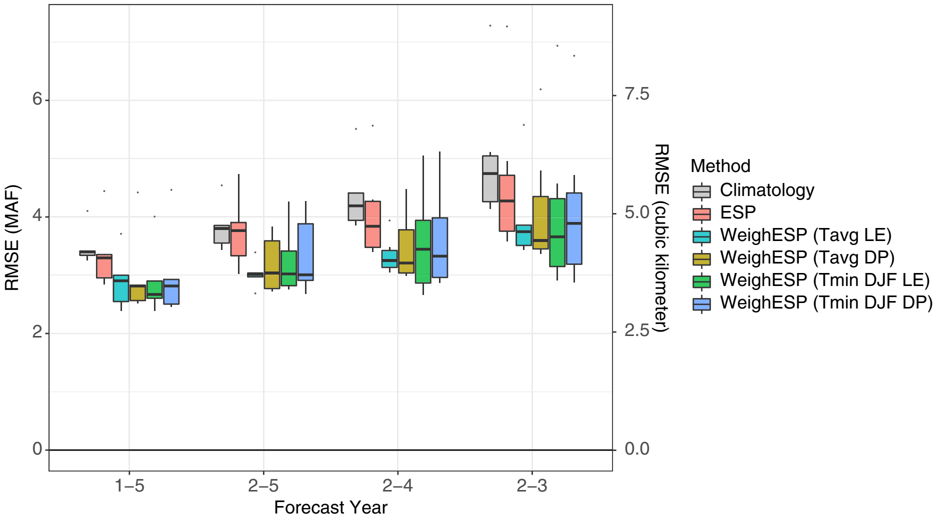

Fig. 5. For Lake Powell unregulated inflows, annual root mean squared error (RMSE) for climatology, ESP, and four WeighESP hindcasts [water year T_{avg} from the LE and DP and winter (DJF) T_{min} from the LE and DP] for traces averaged over the multi-year forecasts (using hindcast run dates from November 2010 to November 2016).

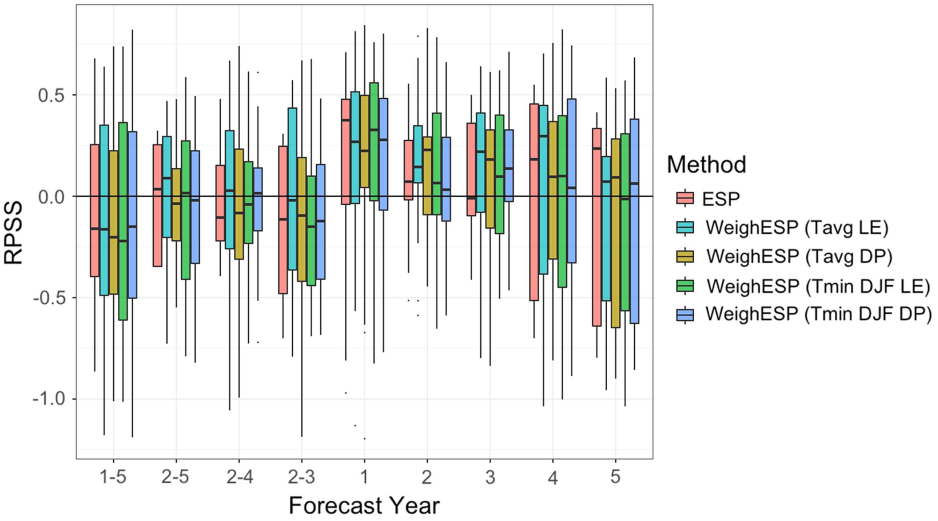

Fig. 6. For Lake Powell unregulated inflows, ranked probability skill score (RPSS) for ESP and four WeighESP hindcasts [water year T_{avg} from the LE and DP and winter (DJF) T_{min} from the LE and DP] as compared to climatology for traces averaged over the multi-year and individual forecast years (using hindcast run dates from November 1980 to November 2016).

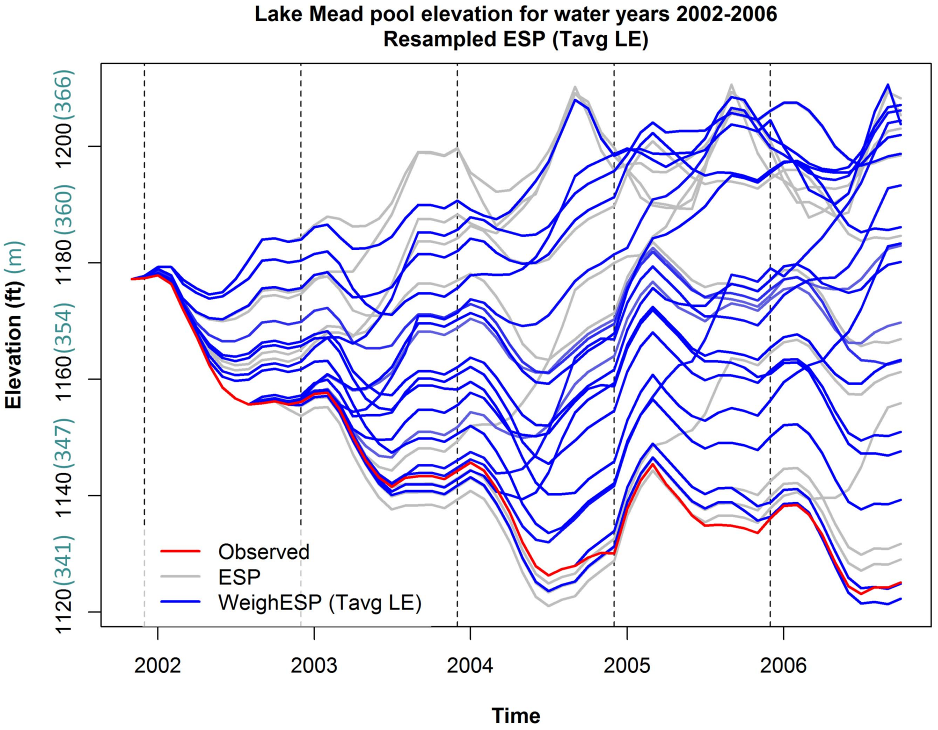

Fig. 7. CRMMS-simulated Lake Mead pool elevation spanning October 2001 to September 2006 from ESP (~30 traces), WeighESP (100 traces), and a quasi-observed historic time series. RMSE is calculated for each projection method on the end of calendar year ensemble spread, i.e., months 1, 13, 25, 37, and 49 (vertical lines).

two different operating guidelines (current operating guidelines did not take effect until 2008), a "quasi-observed" time series of reservoir elevations that approximate what would have happened under the current operating guidelines was generated (by providing the model perfect information about historical streamflow); this approximation is shown in Fig. 7. The vertical lines indicate EOCY points, in terms of number of months lead time, at which the RMSE of each ensemble was evaluated. This 'ensemble RMSE' represents the aggregate performance of all traces for a given projection method at each lead time and has precedent in Baker et al. (2021).

Fig. 8 shows the ensemble RMSE results for Lake Powell EOCY pool elevations from 31 different 5-year projection blocks in the 1981–2017 hindcast, separated by method and lead time. Since the ESP and WeighESP forecasts are initialized at the beginning of November, the first December pool elevation's RMSE is similar for both methods and quite low due to the 1-month lead time. Of greater interest are the subsequent lead times; the RMSE for lead times of 13- and 25-months are very similar for both forecast methods, but at 37- and 49-month leads (i.e., about 3- and 4-year leads) WeighESP has lower (better) median RMSE values

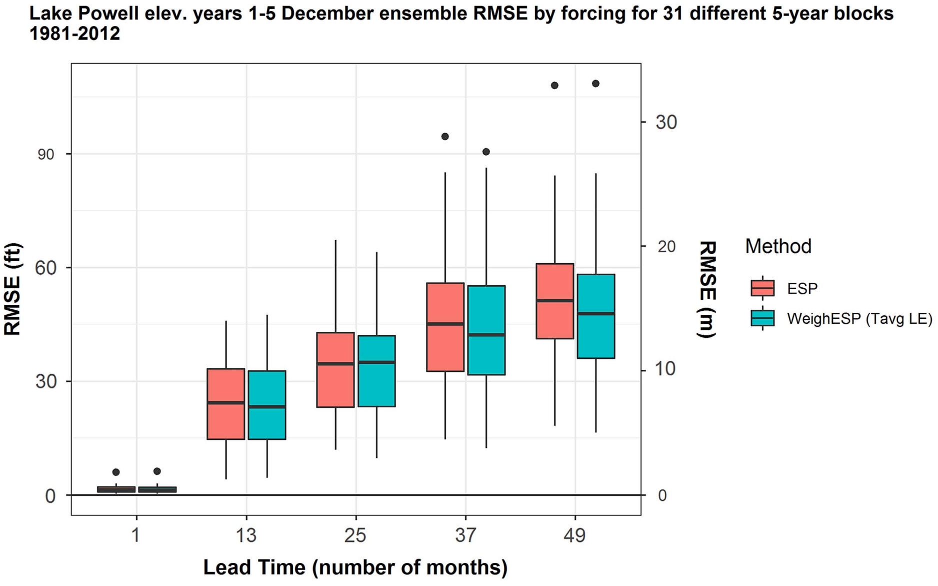

Fig. 8. Lake Powell pool elevation root mean square error (RMSE) for ESP and WeighESP hindcasts (T_{avg} LE), calculated on end of calendar year ensemble spread, i.e., RMSE at 1-, 13-, 25-, 37-, 49-month lead times for all 31 different 5-year blocks in the 1981–2017 hindcast (n = 31 RMSE values per boxplot).

Table 5. ESP and WeighESP (Tavg LE) hindcast RMSE spread in meters (m) and feet (ft) for a 49-month lead time. UW is upper whisker, Q75 is 75th percentile, Q50 is 50th percentile, Q25 is 25th percentile, and LW is lower whisker; whiskers are calculated as 1.5× the inner quartile range +/− the upper or lower quartile for the upper and lower whiskers, respectively

RMSE	Lake Powell		Lake Mead	
in m (ft)	ESP	WeighESP	ESP	WeighESP
UW	25.7 (84.2)	25.9 (84.9)	20.0 (65.6)	20.8 (68.4)
Q75	18.6 (61.0)	17.7 (58.2)	14.7 (48.3)	14.7 (48.1)
Q50	15.6 (51.3)	14.6 (47.8)	12.1 (39.7)	11.4 (37.4)
Q25	12.6 (41.3)	11.0 (36.1)	6.1 (20.1)	6.0 (19.7)
LW	5.5 (18.2)	5.0 (16.4)	3.5 (11.5)	3.1 (10.1)

than ESP. For example, at a 49-month lead time the median RMSE improves by ∼1.1 m (3.5 ft; Table 5). Results using blocks from only 2000 to 2017 show similar, slightly better results, with the median RMSEs performing marginally better for WeighESP for leads >13 months (Fig. 9).

Fig. 10 reveals similar results for the Lake Mead EOCY pool elevations for the 1981-2017 projection blocks. Again, lead month 1 shows low RMSE values, and ESP modestly outperforms WeighESP for 1-, 13-, and 25-month leads. However, for lead times of 37- and 49-months, WeighESP outperforms ESP, yielding reductions in hindcast median RMSE of –10.8% and –5.8%, respectively (Table 6). Results for 5-year projection blocks from the more recent period (2000–2017) are shown in Fig. 11, which show slightly different results. Here, WeighESP is similar or outperforms ESP at

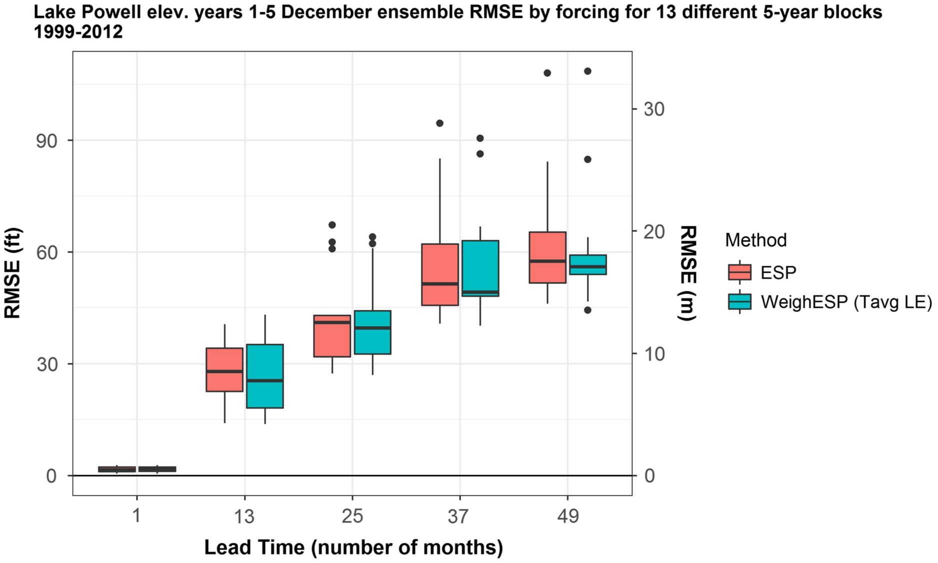

Fig. 9. Lake Powell pool elevation root mean square error (RMSE) for ESP and WeighESP (T_{avg} LE) hindcasts, calculated at the end of calendar year ensemble spread, i.e., RMSE at 1-, 13-, 25-, 37-, 49-month lead times for 5-year blocks in 2000–2017 hindcast (n = 13 RMSE values per boxplot).

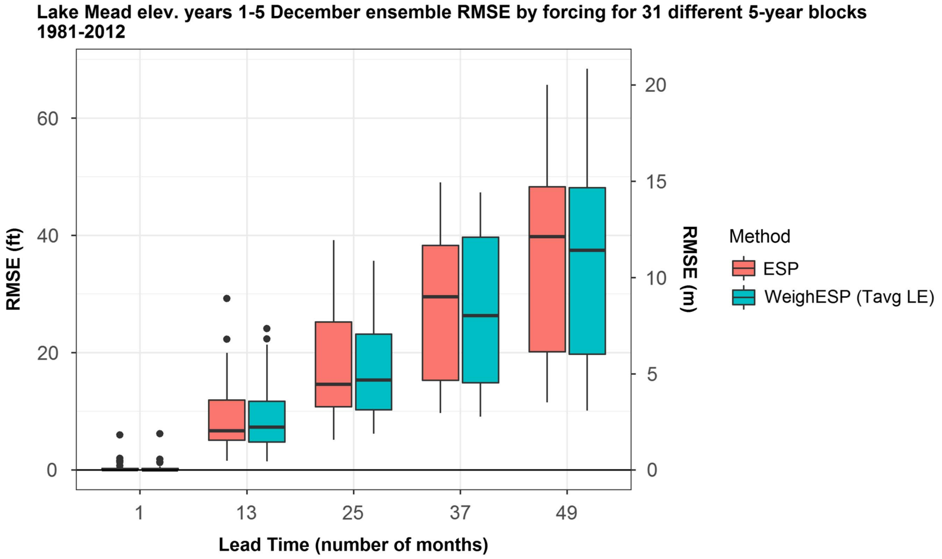

Fig. 10. Lake Mead pool elevation root mean square error (RMSE) for ESP and WeighESP (T_{avg} LE), calculated at the end of calendar year ensemble spread, i.e., RMSE at 1-, 13-, 25-, 37-, 49-month lead times for all 31 different 5-year blocks in 1981–2017 hindcast (n = 31 RMSE values per boxplot).

Table 6. Percent change in hindcast median annual root mean square error (RMSE) in meters (m) and feet (ft) for WeighESP (Tavg LE) relative to ESP

	Powell			Mead		
	Median RMSE in m (ft)			Median RMSE in m (ft)		
Lead time (months)	ESP	WeighESP	% change	ESP	WeighESP	% change
1	0.41 (1.33)	0.40 (1.32)	−0.75	0.01 (0.04)	0 (0)	N/A
13	7.41 (24.3)	7.07 (23.2)	−4.53	2.04 (6.69)	2.22 (7.29)	8.97
25	10.5 (34.6)	10.7 (35.0)	1.16	4.45 (14.6)	4.66 (15.3)	4.79
37	13.8 (45.2)	12.9 (42.2)	−6.64	8.99 (29.5)	8.02 (26.3)	−10.8
49	15.6 (51.3)	14.6 (47.8)	−6.82	12.1 (39.7)	11.4 (37.4)	−5.79

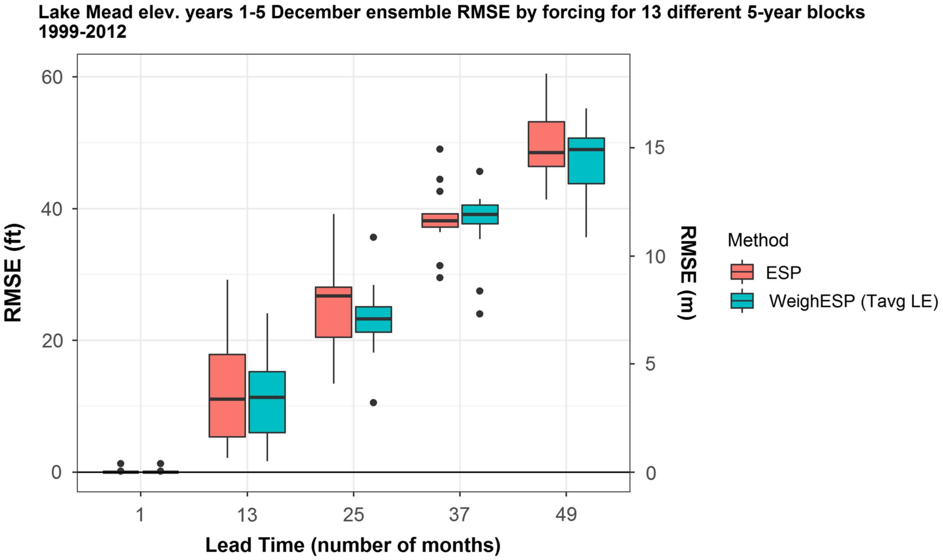

Fig. 11. Lake Mead pool elevation root mean square error (RMSE) for ESP and WeighESP (T_{avg} LE) hindcasts, calculated at the end of calendar year ensemble spread, i.e., RMSE at 1-, 13-, 25-, 37-, 49-month lead times for 5-year blocks in 2000–2017 hindcast (n = 13 RMSE values per boxplot).

earlier leads (13- and 25-months) but performs worse at longer leads. The differences between the results for Lake Powell and Mead could have to do with the fact that the two reservoirs operate in coordination with one another: Lake Mead inflows are modulated by guidelines governing releases from Lake Powell. Therefore, Lake Mead inflows are less directly affected by Lake Powell unregulated inflow variability (or the response can be delayed), and are more influenced by the Lake Powell pool elevation.

Discussion and Conclusions

Despite efforts to enhance operational streamflow forecasts, substantial improvements have not been realized (Welles et al. 2007; Pagano et al. 2004). Streamflow forecasts gain skill from two main sources: initial conditions, particularly at shorter lead times, and future climate forcings, which become increasingly important at longer lead times (Li et al. 2009; Wood et al. 2016). ESP is initialized with current basin conditions, but future climate information at seasonal or longer timescales is not typically incorporated (though CRB does use 5- to 10-day weather forecasts); rather, ESP is forced with equally weighted historical climate traces. Aiming to gain skill from future climate information, this study develops and demonstrates a simple way to post-process ESP traces so that the ensemble is weighed towards climate forcings that reflect the mid-term temperature predictions. However, one critical issue is that ESP is run with climate time series from 1981 to 2010, even for later run dates (i.e., November 2010–November 2016), which have experienced warmer temperatures. As such, updating the climate time series to include years since 2010 will allow for the usage of more recent years, providing additional variability to the streamflow traces (i.e., 40 traces rather than 30), and a new ESP benchmark for hindcast testing. At the time of this writing, ESP has included 35 traces since 2017, and will include 40 traces in the CBRFC's next calibration (2021). A next step could be to test the WeighESP methodology with additional traces. Another potential approach could be to pre-process ESP (Werner et al. 2004); for instance, ESP could be forced directly with new climate sequences that have been generated to reflect a given forecast (see Baker et al. 2021 and references therein). Nevertheless, tercile-based approaches, such as the WeighESP technique put forth in this study, are appealing in that they are simple, and have precedent in their use with seasonal forecasts (e.g., Towler et al. 2010). Other methods, such as k-nn approaches have also shown promise in the CRB (e.g., Baker et al. 2021). A parallel effort to incorporate mid-term temperature predictions using a random forest machine learning approach is also being pursued (Woodson et al. 2021).

Another avenue for increased predictability is to incorporate precipitation. Although the mid-term precipitation predictions are not yet skillful (Table 1), it should be noted that the resampling approach put forth here could be readily extended to a joint technique that includes precipitation and temperature (e.g., Briggs and Wilks 1996). As a first step, precipitation scenarios could be used to test the sensitivity. Another option is to examine the mid-term predictions for their ability to capture precipitation proxies, such as conducive circulation patterns or weather types (Towler et al. 2020) and is the subject of ongoing study.

Part of the study design was to determine the value of initialization of the DP versus the uninitialized LE. Although the DP showed slightly better correlation than the LE for temperature (Table 1), in terms of the conditioned streamflow, it was found that the results were quite similar, with the LE showing slightly better results than the DP. In short, the initialization did not add much, if any, value to the streamflow generation. Of greater importance for this probabilistic approach was having a large ensemble of temperature predictions. These findings are favorable for moving forward with the implementation of this approach for a few reasons:

1. The LE is a smaller, more straightforward dataset to work with than the DP given the annual initialization and lead-based overlapping hindcasts in the DP.
2. While the DP is initialized in November, the LE is a continuous time series; as such, future work could look at different initialization months of interest, as was done in Baker et al. (2021).
3. The LE is similar to other Global Climate Model (GCM) data that is familiar to stakeholders in the CRB. For example, bias-correcting the LE is the same as what is done for the CMIP3 and CMIP5 datasets. The DP requires a different technique, where the bias/drift-correction varies in time.
4. There is a new collection of multi-model large ensembles available (Deser et al. 2020), which could be explored to assess the impact of both initial conditions and model differences.

Although the Lake Powell unregulated inflows showed improvements for all the multiyear forecasts examined, the CRMMS operational projections mainly showed improvements at longer lead periods (approximately 3- and 4-year leads). This could be due to a cumulative effect, where there are small improvements in accuracy each month, but that are only realized at longer leads. It could also indicate that streamflow error reductions need to be larger to show up as improvements in the shorter lead operational projections. A key point is that streamflow forecasts need to be evaluated in a decision-relevant way, and that improvements in streamflow forecasts may not directly translate to improvements in operational projections. By working in close collaboration with Reclamation and other stakeholders, streamflow ensembles were evaluated using CRMMS, which established a pathway for testing and implementation. Given the importance of pool elevations to stakeholders in the region, even marginal improvements at longer leads shows promise for the technique and could be beneficial. Improving the error by several feet can matter, especially when the reservoirs are projected to be near threshold elevations that affect annual operations and water deliveries.

Data Availability Statement

All data and codes that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments

Erin Towler's research is made possible through the support of the Colorado River Climate and Hydrology Work Group and is funded by the Central Arizona Water Conservation District, Denver Water, California's Six Agency Committee, and Southern Nevada Water Authority. Erin Towler's material is based upon work supported by the National Center for Atmospheric Research (NCAR), which is a major facility sponsored by the National Science Foundation (NSF) under Cooperative Agreement No. 1852977. David Woodson's research is funded through USBR S&T #19264 and CU-Boulder Civil, Environmental, and Architectural Engineering. The information being offered herein represents the opinion of the author(s). It has not been formally disseminated by the Bureau of Reclamation. It does not represent and should not be construed to represent Reclamation's determination or policy.

References

Baker, S. A. 2019. "Development of sub-seasonal to seasonal watershed-scale hydroclimate forecast techniques to support water management." Ph.D. dissertation, Dept. of Civil, Environmental, and Architectural Engineering, Univ. of Colorado.

Baker, S. A., B. Rajagopalan, and A. W. Wood. 2021. "Enhancing ensemble seasonal streamflow forecasts in the Upper Colorado River Basin using multi-model climate forecasts." *J. Am. Water Resour. Assoc.* 57 (6): 906–922. https://doi.org/10.1111/1752-1688.12960.

Baker, S. A., A. W. Wood, B. Rajagopalan, J. Prairie, C. Jerla, E. Zagona, R. A. Butler, and R. Smith. Forthcoming. "The Colorado River

Basin operational prediction testbed: A framework for evaluating streamflow forecasts and reservoir operations." *J. Am. Water Resour. Assoc.*

Barsugli, J., C. Anderson, J. B. Smith, and J. M. Vogel. 2009. "Options for improving climate modeling to assist water utility planning for climate change." Accessed November 1, 2021. https://www.wucaonline.org/assets/pdf/pubs-whitepaper-120909.pdf.

Briggs, W. M., and D. S. Wilks. 1996. "Extension of the climate prediction center long-lead temperature and precipitation outlooks to general weather statistics." *J. Clim.* 9 (12): 3496–3504. https://doi.org/10.1175/1520-0442(1996)009<3496:EOTCPC>2.0.CO;2.

CMIP–WGCM–WGSIP (Coupled Model Intercomparison Project-Working Group on Coupled Modeling-Working Group on Subseasonal to Interdecadal Prediction) Decadal Climate Prediction Panel. 2011. "Data and bias correction for decadal climate predictions." Accessed January 29, 2022. https://www.wcrp-climate.org/decadal/references/DCPP_Bias_Correction.pdf.

Daly, C., G. Taylor, and W. Gibson. 1997. "The PRISM approach to mapping precipitation and temperature." In *Proc., 10th AMS Conf. on Applied Climatology*, 20–23. Boston: American Meteorological Society.

Day, G. 1985. "Extended streamflow forecasting using NWSRFS." *J. Water Resour. Plann. Manage.* 111 (2): 157–170. https://doi.org/10.1061/(ASCE)0733-9496(1985)111:2(157).

Deser, C., et al. 2020. "Insights from Earth system model initial-condition large ensembles and future prospects." *Nat. Clim. Change* 10 (4): 277–286. https://doi.org/10.1038/s41558-020-0731-2.

Goddard, L., A. Kumar, A. Solomon, D. Smith, G. Boer, P. Gonzalez, and T. Delworth. 2012. "A verification framework for interannual-to-decadal predictions experiments." *Clim. Dyn.* 40 (1–2): 245–272. https://doi.org/10.1007/s00382-012-1481-2.

Hoerling, M., J. Barsugli, B. Livneh, J. Eischeid, X. Quan, and A. Badger. 2019. "Causes for the century-long decline in Colorado River flow." *J. Clim.* 32 (23): 8181–8203. https://doi.org/10.1175/JCLI-D-19-0207.1.

Kay, J. E., et al. 2015. "The community earth system model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability." *Bull. Am. Meteorol. Soc.* 96 (8): 1333–1349. https://doi.org/10.1175/BAMS-D-13-00255.1.

Kiem, A. S., G. Kuczera, P. Kozarovski, L. Zhang, and G. Willgoose. 2021. "Stochastic generation of future hydroclimate using temperature as a climate change covariate." *Water Resour. Res.* 57 (2): 2020WR027331. https://doi.org/10.1029/2020WR027331.

Lehner, F., A. W. Wood, D. Llewellyn, D. B. Blatchford, A. G. Goodbody, and F. Pappenberger. 2017. "Mitigating the impacts of climate nonstationarity on seasonal streamflow predictability in the US southwest." *Geophys. Res. Lett.* 44 (24): 12–208. https://doi.org/10.1002/2017GL076043.

Li, H., L. Luo, E. F. Wood, and J. Schaake. 2009. "The role of initial conditions and forcing in seasonal hydrologic forecasting." *J. Geophys. Res. Atmos.* 114 (4): D04114. https://doi.org/10.1029/2008JD010969.

Lukas, J., E. Payton, J. Deems, I. Rangwala, and B. Duncan. 2020. "Observations—Hydrology." Chap. 5 in *Colorado River Basin climate and hydrology: State of the science*, edited by J. Lukas and E. Payton, 154–219. Boulder, CO: Univ. of Colorado Boulder.

Meehl, G. A., et al. 2014. "Decadal climate prediction: An update from the trenches." *Bull. Am. Meteorol. Soc.* 95 (2): 243–267, https://doi.org/10.1175/BAMS-D-12-00241.1.

Meehl, G. A., and H. Teng. 2012. "Case studies for initialized decadal hindcasts and predictions for the Pacific region." *Geophys. Res. Lett.* 39 (22): L22705. https://doi.org/10.1029/2012GL053423.

Meehl, G. A., and H. Teng. 2014a. "CMIP5 multi-model hindcasts for the mid-1970s shift and early 2000s hiatus and predictions for 2016–2035." *Geophys. Res. Lett.* 41 (5): 1711–1716. https://doi.org/10.1002/2014GL059256.

Meehl, G. A., and H. Teng. 2014b. "Regional precipitation simulations for the mid-1970s shift and early-2000s hiatus." *Geophys. Res. Lett.* 41 (21): 7658–7665. https://doi.org/10.1002/2014GL061778.

Milly, P. C. D., and K. A. Dunne. 2020. "Colorado River flow dwindles as warming-driven loss of reflective snow energizes evaporation." *Science* 367 (6483): 1252–1255. https://doi.org/10.1126/science.aay9187.

NRC (National Academies Press). 2009. "Informing decisions in a changing climate." In *Panel on strategies and methods for climate-related decision support*. Washington, DC: National Academies Press.

Pagano, T., D. Garen, and S. Sorooshian. 2004. "Evaluation of official western US seasonal water supply outlooks, 1922–2002." *J. Hydrometeorol.* 5 (5): 896–909. https://doi.org/10.1175/1525-7541(2004)005,0896:EOOWUS.2.0.CO;2.

Sandgathe, S., B. R. Brown, J. C. Carman, J. M. Infanti, B. Johnson, D. McCarren, and E. McIlvain. 2020. "Exploring the need for reliable decadal prediction." *Bull. Am. Meteorol. Soc.* 101 (2): E141–E145. https://doi.org/10.1175/BAMS-D-19-0248.1.

Smith, D. M., R. Eade, A. A. Scaife, L. Caron, G. Danabasoglu, T. M. Delsole, and T. Delworth. 2019. "Robust skill of decadal climate predictions." *NPJ Clim. Atmos. Sci.* 2 (1): 1–10 https://doi.org/10.1038/s41612-019-0071-y.

Thornton, P. E., S. W. Running, and M. A. White. 1997. "Generating surfaces of daily meteorological variables over large regions of complex terrain." *J. Hydrol.* 190 (3–4): 214–251. https://doi.org/10.1016/S0022-1694(96)03128-9.

Towler, E., D. Llewellyn, A. Prein, and E. Gilleland. 2020. "Extreme-value analysis for the characterization of extremes in water resources: A generalized workflow and case study on New Mexico monsoon precipitation." *Weather Clim. Extremes* 29 (Sep): 100260. https://doi.org/10.1016/j.wace.2020.100260.

Towler, E., D. PaiMazumder, and J. Done. 2018. "Towards the application of decadal climate predictions." *J. Appl. Meteorol. Climatol.* 57 (3): 555–568. https://doi.org/10.1175/JAMC-D-17-0113.1.

Towler, E., B. Rajagopalan, R. S. Summers, and D. Yates. 2010. "An approach for probabilistic forecasting of seasonal turbidity threshold exceedance." *Water Resour. Res.* 46 (Jun): W06511. https://doi.org/10.1029/2009WR007834.

Towler, E., and D. Yates. 2020. "Incorporating mid-term temperature predictions for water resources planning." *J. Appl. Meteorol. Climatol.* 60 (2): 171–183. https://doi.org/10.1175/JAMC-D-20-0134.1.

Udall, B., and J. Overpeck. 2017. "The twenty-first century Colorado River hot drought and implications for the future." *Water Resour. Res.* 53 (3): 2404–2418. https://doi.org/10.1002/2016WR019638.

Vera, C., et al. 2010. "Needs assessment for climate information on decadal timescales and longer." *Procedia Environ. Sci.* 1: 275–286. https://doi.org/10.1016/j.proenv.2010.09.017.

Welles, E., S. Sorooshian, G. Carter, and B. Olsen. 2007. "Hydro- logic verification: A call for action and collaboration." *Bull. Am. Meteorol. Soc.* 88 (4): 503–512. https://doi.org/10.1175/BAMS-88-4-503.

Werner, K., D. Brandon, M. Clark, and S. Gangopadhyay. 2004. "Climate index weighting schemes for NWS ESP-based seasonal volume forecasts." *J. Hydrometeorol.* 5 (6): 1076–1090. https://doi.org/10.1175/JHM-381.1.

Wilks, D. S. 1995. *Statistical methods in the atmospheric sciences*. New York: Elsevier.

WMO (World Meteorological Organization). 2020. "WMO lead centre for annual-to-decadal climate prediction." Accessed July 19, 2021. https://hadleyserver.metoffice.gov.uk/wmolc/.

Wood, A., L. Woelders, and J. Lukas. 2020. "Streamflow forecasting." Chap. 8 in *Colorado River Basin climate and hydrology: State of the science*, edited by J. Lukas and E. Payton, 287–333. Boulder, CO: Univ. of Colorado Boulder.

Wood, A. W., T. Hopson, A. Newman, L. Brekke, J. Arnold, and M. Clark. 2016. "Quantifying streamflow forecast skill elasticity to initial condition and climate prediction skill." *J. Hydrometeorol.* 17 (2): 651–668. https://doi.org/10.1175/JHM-D-14-0213.1.

Woodhouse, C. A., and G. T. Pederson. 2018. "Investigating runoff efficiency in Upper Colorado River streamflow over past centuries." *Water Resour. Res.* 54 (1): 286–300. https://doi.org/10.1002/2017WR021663.

Woodhouse, C. A., G. T. Pederson, K. Morino, S. A. McAfee, and G. J. McCabe. 2016. "Increasing influence of air temperature on upper Colorado River streamflow." *Geophys. Res. Lett.* 43 (5): 2174–2181. https://doi.org/10.1002/2015GL067613.

Woodson, D., B. Rajagopalan, S. Baker, R. Smith, J. Prairie, E. Towler, M. Ge, and E. Zagona. 2021. "Stochastic decadal projections of Colorado river streamflow and reservoir pool elevations conditioned on temperature projections." *Water Resour. Res.* 57 (12): e2021WR030936. https://doi.org/10.1029/2021WR030936.

Yeager, S. G., G. Danabasoglu, N. Rosenbloom, W. Strand, S. Bates, G. Meehl, and N. S. Lovenduski. 2018. "Predicting mid-term changes in the earth system: A large ensemble of initialized decadal prediction simulations using the community earth system model." *Bull. Am. Meteorol. Soc.* 99 (9): 1867–1886. https://doi.org/10.1175/BAMS-D-17-0098.1.

Zagona, E. A., T. J. Fulp, R. Shane, T. Magee, and H. M. Goranflo. 2001. "Riverware: A generalized tool for complex reservoir system modeling." *J. Am. Water Resour. Assoc.* 37 (4): 913–929. https://doi.org/10.1111/j.1752-1688.2001.tb05522.x.