Long-Lead Forecasting of Runoff Season Flows in the Colorado River Basin Using a Random Forest Approach

David Woodson[1]; Balaji Rajagopalan, F.ASCE[2]; and Edith Zagona[3]

Abstract: There is an increasing need for skillful runoff season (i.e., spring) streamflow forecasts that extend beyond a 12-month lead time for water resources management, especially under multiyear droughts and particularly in basins with highly variable streamflow, large storage capacity, proclivity to droughts, and many competing water users such as in the Colorado River Basin (CRB). Ensemble streamflow prediction (ESP) is a probabilistic prediction method widely used in hydrology, including at the National Oceanic and Atmospheric Administration (NOAA) Colorado Basin River Forecasting Center (CBRFC) to forecast flows that the Bureau of Reclamation uses in their water resources operational decision models. However, it tends toward climatology at 5-month and longer lead times, causing decreased skill, particularly in forecasts critical for management decisions. We developed a modeling approach for seasonal streamflow forecasts using a machine learning technique, random forest (RF), for runoff season flows (April 1–July 31 total) at the important gauge of Lees Ferry, Arizona, on the CRB. The model predictors include antecedent basin conditions, large-scale climate teleconnections, climate model projections of temperature and precipitation, and the mean ESP forecast from CBRFC. The RF model is fitted and validated separately for lead times spanning 0 to 18 months over the period 1983–2017. The performance of the RF model forecasts and CBRFC ESP forecasts are separately assessed against observed streamflows in a cross validation mode. Forecast performance was evaluated using metrics including relative bias, root mean square error, ranked probability skill score, and reliability. Measured by ranked probability skill score, RF outperforms a climatological benchmark at all lead times and outperforms CBRFC's ESP hindcasts for lead times spanning 6 to 18 months. For the 6- to 18-month lead times, the RF ensemble median had a root mean square error that was between ∼410- and ∼620-thousand acre-feet lower than that of the ESP ensemble median (i.e., RF reduced ensemble median RMSE by −9% to −12% relative to ESP). Reliability was comparable between RF and ESP. More skillful long-lead cross-validated forecasts using machine learning methods show promise for their use in real time forecasts and better informed and efficient water resources management; however, further testing in various decision models is needed to examine RF forecasts' downstream impacts on key water resources metrics like robustness, reliability, and vulnerability. **DOI: 10.1061/JWRMD5.WRENG-6167.** *© 2024 American Society of Civil Engineers.*

Introduction

In river basins that have highly variable interannual flows, that is, multiyear periods where streamflow is above or below average, there is a need for river flow forecasts that go beyond seasonal time scales to make efficient water resources management decisions. The need for long-lead water supply forecasts is heightened in basins with high storage capacity, semiarid climate, sustained dry periods, competing needs, and a large number of stakeholders—such as the Colorado River Basin (CRB). Anthropogenic climate change has also impacted the CRB and the ongoing 23-year-long drought that was caused not only by lower-than-average precipitation but also by warmer-than-average temperatures (Hoerling et al. 2019; Williams et al. 2020, 2022; Woodhouse et al. 2016; Xiao et al. 2018). The sensitivity of CRB streamflow to temperature is estimated between $-2.5\%°C^{-1}$ and $-14\%°C^{-1}$ (Hoerling et al. 2019; McCabe and Wolock 2007; Milly and Dunne 2020; Nowak et al. 2012; Udall and Overpeck 2017; Vano et al. 2012, 2014), and as the warming trend continues over the next decades, research has suggested that temperature-induced drying may offset any possible future precipitation increases, or worse, compound with precipitation decreases and result in even lower CRB flows or increased aridification (Milly and Dunne 2020; Udall and Overpeck 2017). Compounding these problems, the CRB is overallocated. The Colorado River's average annual flow over the last century is about 15 million acre-feet (MAF), but 16.5 MAF of Colorado River water is allocated per annum between the United States and Mexico, and evaporative losses are significant (Stern 2023). The earliest allocations were made during an anomalously wet pluvial period at the beginning of the 20th century (Gangopadhyay et al. 2022; Meko et al. 2007; Woodhouse et al. 2005).

The United States (US) Bureau of Reclamation ("Reclamation") manages water resources in the Western United States including the CRB. They generate ensemble forecasts of potential basin and reservoir conditions based on flow outlooks that project beyond a season up to a ∼2-year time period (Payton et al. 2020; Reclamation 2015, 2019). For example, Reclamation forecasts reservoir levels at Lakes Mead and Powell and these projections can influence

[1] Graduate Research Assistant, Dept. of Civil, Environmental, and Architectural Engineering, Univ. of Colorado Boulder, UCB 428, Boulder, CO 80309 (corresponding author). ORCID: https://orcid.org/0000-0001-9852-5355. Email: david.woodson@colorado.edu

[2] Professor, Dept. of Civil, Environmental, and Architectural Engineering, Univ. of Colorado Boulder, UCB 428, Boulder, CO 80309; Fellow, Cooperative Institute for Research in Environmental Sciences, Boulder, CO 80309. Email: balajir@colorado.edu

[3] Director, Center for Advanced Decision Support for Water and Environmental Systems, Univ. of Colorado Boulder, Boulder, CO 80309; Research Professor, Dept. of Civil, Environmental, and Architectural Engineering, Univ. of Colorado Boulder, UCB 428, Boulder, CO 80309. ORCID: https://orcid.org/0000-0003-1333-0589. Email: zagona@colorado.edu

Note. This manuscript was submitted on February 28, 2023; approved on November 8, 2023; published online on January 31, 2024. Discussion period open until June 30, 2024; 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.

decision making with respect to reservoir releases or conservation efforts. Notably, Reclamation's August 2022 24-month study suggested that Lake Mead would enter an unprecedented Level 2a shortage condition in the water year 2023, with a projected pool elevation of 1,048 ft on January 1, 2023 (Reclamation 2022). Lake Mead's water level did drop to 1,043′ in November 2022, shortly before an unforeseen, above-average winter precipitation season led to Lake Mead's water level rising to 1,065′ by August 2023. However, the Level 2a shortage condition had already been triggered and reduced 2023 water deliveries to Arizona, Nevada, and Mexico by 592,000, 25,000, and 104,000 acre-feet, respectively.

For projecting operations of the reservoirs, Reclamation uses forecasts generated by the National Oceanic and Atmospheric Administration's (NOAA) CRB River Forecasting Center (CBRFC). The CBRFC forecasts apply an Ensemble Streamflow Prediction (ESP) approach that involves initializing the Sacramento Soil Moisture Accounting (SAC-SMA) and SNOW-17 models with the current state of the basin conditions (e.g., stream and reservoir stages, soil moisture, and snowpack) and then forcing them with traces of observed precipitation and temperature subset from the historical record, resulting in a streamflow forecast of ensemble members derived from each meteorological year (Day 1985; Lukas and Payton 2020; Werner and Yeager 2013; Wood and Werner 2011). ESP forecasts are most skillful at seasonal (i.e., shorter) time scales for which persistence from the model's initial hydrologic conditions (IHC) is strong, but ESP is less skillful at longer lead times due to the decreasing predictive skill of basin IHCs (5–18 months). Hence, the ESP forecasts at longer lead times converge to historical average flows, i.e., climatology, after the persistence from IHCs is lost. Since the meteorological ensemble forcing contains both wet and dry extremes, the ESP forecasts do best when flows are in the median range but tend to overpredict during low-flow times and underpredict during high flow periods.

In the CRB, there has been an ongoing drought since the early 21st century ("the Millennium Drought" Hoerling et al. 2019; Salehabadi et al. 2022) that has resulted in the basin's reservoirs, including its two largest: Lakes Mead and Powell, to fall in 2022–2023 to their lowest pool elevation since Mead's filling in the 1930s and Powell's filling in the 1960s. Reclamation has been challenged to manage the reservoirs under the latest operating guidelines that have for the first time imposed shortages on downstream users and enacted creative new policies to avoid dropping the large reservoirs below their hydropower generating pool levels (Reclamation 2021; Smith et al. 2022). However, research has shown that Reclamation's 24-month study has a wet bias for lead times of 12- to 24-months ("year-2") due to the use of probabilities derived from a 30-year reference period spanning 1991–2020 for these lead times (Wang et al. 2022). Prior to fall 2021, the reference period was 1981–2010, which had average flows 9% higher than those during the Millennium Drought, leading to year-2 ensemble median projections up to ~7-MAF higher than observations during 2010–2021. At a 24-month lead time, ~70% of ensemble members had a wet bias and the ensemble median bias was ~1-MAF (Wang et al. 2022). The importance of these high-profile operational decisions and intense stakeholder scrutiny has raised the bar for the necessity of more skillful forecasting at lead times up to 2 years.

This pressing need for and importance of skillful forecasting in the CRB for both near-term operations and long-term planning has motivated extensive research and methodological advancement. For example, the United States Department of Agriculture (USDA) Natural Resource Conservation Service (NRCS) has a long history of issuing seasonal water supply forecasts for ~1,000 locations in the Western United States. NRCS monthly forecasts use principal component regression (and Z-score regression for daily forecasts) trained on predictors like snow-water equivalent (SWE), accumulated precipitation, and antecedent streamflow (Lukas and Payton 2020). Various studies have found added value from other covariates like soil moisture, temperature forecasts, and ocean teleconnections like El Niño Southern Oscillation (ENSO) (Harpold et al. 2017; Lehner et al. 2017; Rosenberg et al. 2011). NRCS water supply forecasts are generally issued January through April to predict the runoff season volume and have high skill due to the strong signal imparted by IHCs, but various machine learning approaches were shown to outperform the NRCS forecasts in three test basins (Fleming and Goodbody 2019). Other statistical ensemble models based on local polynomials and multi-models—using SWE and large climate variables from the ocean and atmosphere during winter have been shown to be quite skillful for modeling seasonal flows in the CRB and other basins in the Western United States (Bracken et al. 2010; Regonda et al. 2006a, b). Recently, Baker et al. (2021a) used North American multimodel ensemble forecasts of precipitation and temperature as well as antecedent streamflow with a K-nearest neighbor trace weighting approach to improve CBRFC ESP forecasts and found that the climate-conditioned forecasts performed better than ESP alone in predicting April, May, June, and July (AMJJ) runoff volume for forecasts made in winter and early spring. Finally, Zhao et al. (2021) used artificial neural networks and stepwise linear regression to hindcast Lees Ferry AMJJ streamflow for lead times up to 12 months using predictors including antecedent Pacific Sea surface temperatures, soil moisture, SWE, and precipitation: they found a correlation between predicted and observed streamflow of about 0.4 for a 12-month lead time.

Increasingly, researchers have sought to elicit skill at lead times of greater than 12 months. For example, Chikamoto et al. (2020) were able to produce skillful forecasts of CRB annual water supply up to two years into the future through ocean initialization of a decadal climate model. Similarly, Switanek and Troch (2011) generated skillful projections of the 10-year mean Lees Ferry flow using the preceding 10-year mean Atlantic multidecadal oscillation (AMO) and Pacific decadal oscillation (PDO). Studies have also found that incorporating climate model projections of temperature into streamflow forecasts through various statistical weighting approaches can improve performance at multiyear time scales (e.g., for lead times of 1- to 5-years; Towler et al. 2018, 2021; Towler and Yates 2021).

Recently, climate model projections have been used with nonparametric stochastic methods such as time series bootstrapping (Towler et al. 2021) and machine learning methods such as RF (Woodson et al. 2021) for decadal projections of streamflow in the CRB. These projections have been shown to translate skillfully into projections of water resources decision variables, especially in Woodson et al. (2021). Stochastic simulations to generate scenarios of flows for multi-decadal scale planning have been developed using time series methods such as wavelets and bootstrapping (Erkyihun et al. 2016; Rajagopalan et al. 2019), hidden Markov models (Bracken et al. 2016). Despite the growing interest in and need for decadal projections that span the gap between seasonal forecasts and multidecadal projections, there are little to no skillful year-2 forecasts available in the CRB.

To address this critical need, we developed a modeling framework to generate runoff season ensemble streamflow forecasts for the CRB at the Lees Ferry gauge–the most important gauge on the Colorado River through which passes 85%–90% of the average annual natural flow of the entire river, generated by snowmelt in the Upper CRB (UCRB) (Christensen et al. 2004). The framework uses a random forest (RF) technique, a machine learning approach, to model and test retrospective forecasts (i.e., hindcasts) of historical spring flows over the period 1983–2017 at various lead times spanning 0 to 18 months and compares them with the corresponding ESP forecasts from CBRFC. In the next section, Materials and Methods, we describe the modeling framework, the RF approach, and the datasets used as well as variable selection, model validation, and evaluation metrics. The results are subsequently described, concluding with a summary and reflections.

Materials and Methods

The modeling framework is intended to be employed in real time forecasting; hence, it is referred to as a forecasting model—much like other such models in literature—with the caveat that the validation of the model can only be done in a retrospective forecast (i.e., hindcast) mode, even when done in cross validation. Thus, we use forecast, retrospective forecast, and hindcast interchangeably in this paper.

Datasets

We used both historic observed information as well as future, simulated information in the modeling framework. Predictor selection was performed separately for each lead time and involved training a RF model for each year for each lead time on all of the predictors, then aggregating RF 'variable importance' and removing the variables that had a negative median variable importance over the entire 1983–2017 hindcast period, resulting in a suite of custom predictors for each lead time, described in Table 1.

The length of the training record for the RF model depends on the lead time since the predictors used for each lead time vary, as does the period of record for each predictor. RF hindcasts with lead times of 0, 1, 2, 3, and 4 months have a 35-year period of record for model training (1983–2017); while lead times of 6, 7, and 8 months use a 96-year record (1922–2017), and a 95-year record (1923–2017) is used for lead times of 12 and 18 months because an additional year must be dropped in cross validating these latter two lead times. The period of records for the 4-month and less hindcasts are much shorter since the primary predictors used for these lead times, CBRFC ESP ensemble mean and CESM-DPLE temperature, begin in 1983 and 1981, respectively (Table 1). Conversely, the longer lead hindcasts do not use those predictors and instead are temporally limited by the CESM-LE projections, which began in 1921.

Historical Observations

The predictand (dependent variable being predicted) at all lead times is the Lees Ferry naturalized spring flow (April 1–July 31) over the period 1983–2017 (Prairie et al. 2005; USBR 2020). Natural flow is constructed using a RiverWare (Zagona et al. 2001) model that removes anthropogenic influences like diversions, reservoir storage, and consumptive use.

The suite of predictors used in the model are hydroclimate variables, large-scale climate teleconnection indices, and climate model projections, which are described as follows. UCRB basin-average precipitation and maximum and minimum temperatures are derived from parameter-elevation regressions on independent slopes model (PRISM) data (Daly et al. 1994). Runoff efficiency is calculated as the ratio of annual natural flow at Lees Ferry to basin-average precipitation.

Large-scale climate teleconnection indices including the PDO; Mantua et al. 1997; Zhang et al. 1997) and the AMO; Enfield et al. 2001) are obtained from International Research Institute (IRI) data library.

Future Simulated Predictors

Future simulated predictors include temperature and precipitation projections from both the Community Earth System Model–Large Ensemble (CESM-LE; Kay et al. 2014) and the Community Earth System Model–Decadal Prediction Large Ensemble (CESM-DPLE; Yeager et al. 2018). The climate model projections are basin-average and basin-totals, respectively, for maximum and minimum temperatures and precipitation, calculated as seasonal averages dictated by forecast lead time. The CESM-LE was initialized on January 1, 1920, and is free running through 2080. Conversely, the CESM-DPLE projections are initialized on November 1 of a given year and then run for ∼10 years; this process is repeated for every year between 1981 and 2017 (e.g., projections span 1981–1991, 1982–1992, …, 2016–2026, 2017–2027). CESM projections have been shown to be skillful in projecting flows on the Colorado River decadal time scales (Woodson et al. 2021).

Additionally, we obtained ESP (Wood and Werner 2011) retrospective forecasts issued by NOAA's CBRFC and then used them as predictors in the RF forecast and as a stand-alone benchmark forecast for which to compare random forest forecasts. CBRFC ESP hindcasts are based on a physical hydrologic model initialized

Table 1. Selected predictors and resolution

Predictor type	Predictor	Variable name	Resolution	Used for lead time(s) (months)	Period of record for model training
Observed, past information	Atlantic Multidecadal Oscillation	amo	Seasonal mean	6, 7, 8, 18	1921–2017
	Pacific Decadal Oscillation	pdo	Seasonal mean	8, 12, 18	1921–2017
	Maximum Temperature	tmax	Seasonal mean	1, 7	1921–2017
	Minimum Temperature	tmin	Seasonal mean	12, 18	1921–2017
	November to September precipitation	win.pcp	Seasonal total	6	1921–2017
	Past year's flow	past.q.maf	Annual total	6, 7, 8	1921–2017
	Past year's runoff efficiency	past.re	Annual total	6, 7, 8	1921–2017
Simulated, future information	CBRFC ESP–Ensemble Mean	esp	Seasonal sum	0, 1, 2, 3, 4	1983–2017
	Decadal Prediction Large Ensemble–Maximum Temperature	dple.tmax	Seasonal mean	0, 1, 2 3, 4	1981–2017
	Decadal Prediction Large Ensemble–Minimum Temperature	dple.tmin	Seasonal mean	0, 1, 2	1981–2017
	Large Ensemble–Maximum Temperature (Summer, winter, or spring)	le.tmax.sum, le.tmax.win, or le.tmax.spg	Seasonal mean	7, 8, 12, 18	1921–2017
	Large Ensemble–Minimum Temperature (Summer, winter, or spring)	le.tmin.sum, le.tmin.win, or le.tmin.spg	Seasonal mean	12, 18	1921–2017
	Large Ensemble–Precipitation (Summer, Spring)	le.pcp, le.pcp.spg	Seasonal total	2, 12	1921–2017

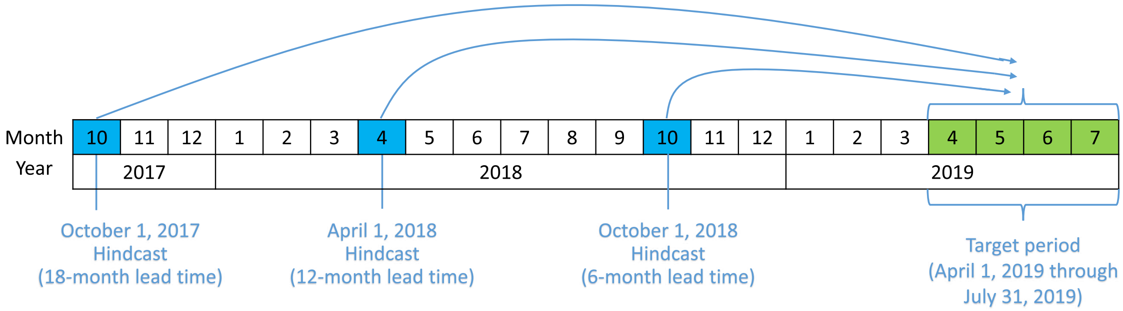

Fig. 1. Forecast framework for spring flows at three different lead times.

with basin snow, land surface conditions, and storage and forced with 5-year-long historical sequences of daily weather spanning 1981–2010. The ESP hindcasts are initialized and run on a monthly basis between 1983 and 2017, thus producing 30 ensemble members for each 5-year-long hindcast period in that time span. The ESP hindcasts for each lead time are initialized separately; for example, a 0-month lead time is initialized on April 1, a 3-month lead time is initialized on January 1, and a 12-month lead time is initialized on April 1 of the prior year. The CBRFC ESP forecasts are used in official water resources management decisions by Reclamation.

Ensemble means were calculated for CESM-LE projections and ESP. Further, they are aggregated to either seasonal or annual averages (or volumes in the case of precipitation and streamflow). For example, an RF forecast made on April 1 with a 12-month lead time might use a suite of predictors, including historical observations (e.g., the preceding winter average PDO and observed minimum temperature) as well as future simulations (e.g., CESM-LE projected future summer, winter, and spring mean temperature and precipitation, and ESP mean forecasted flows).

Modeling Framework and Predictor Justification

The research framework is schematically depicted in Fig. 1. It shows the lead times at which retrospective forecasts are made for the given spring runoff. As illustrated in this figure, the target is 2019 runoff season volume (April 1, 2019, through July 31, 2019), and this volumetric hindcast is made at three different lead times (18 months, 12 months, and 6 months): with hindcasts made on October 1, 2017, April 1, 2018, and October 1, 2018, respectively.

Any forecast framework requires the identification of predictors and a method for using them to generate hopefully skillful ensemble forecasts. Past research has shown that temperature plays an important role in modulating CRB flow, particularly during the ongoing Millennium Drought (Hoerling et al. 2019; Udall and Overpeck 2017), and that the Western United States climate is influenced by climate indices such as the ENSO, AMO and PDO (Kalra and Ahmad 2012; Lukas and Payton 2020; Nowak et al. 2012; Zhao et al. 2017, 2021, 2023; Zhao and Zhang 2022). From these studies, we hypothesize that the Pacific and Atlantic Oceans provide predictability at year-2 time scales, particularly with respect to land surface temperatures. Temperature predictability in turn impacts the basin runoff efficiency (the ratio of streamflow to precipitation), and consequently the streamflow, given the CRB's relatively high temperature sensitivity (higher temperatures lead to decreased runoff efficiency via increased evaporation and vice versa).

Preliminary analysis found moderate Spearman's rank correlation between winter sea surface temperatures (SSTs) and subsequent spring Lees Ferry flow as well as with flows one year ahead; correlation spatial patterns coincided with PDO and AMO regions (Fig. 2). SST correlation with Lees Ferry flow and UCRB precipitation persists into year-2 time scales (Zhao et al. 2017, 2021, 2023; Zhao and Zhang 2022). As such, the teleconnections captured by PDO and AMO indices along with antecedent conditions like temperature and precipitation are good candidate predictors for long-lead flow hindcasts.

As mentioned previously, the RF algorithm provides variable importance values for each variable in each fitted RF model. The variable importance values are provided as percent change in mean square error (MSE) if the given predictor is dropped from the model. Positive or negative values suggest that the variable is either beneficial or deleterious, respectively, to model performance. We used the variable importance metrics as a predictor selection tool for each lead time by starting with the 'full model' of all applicable predictors and then removing variables that yielded a 1983–2017 hindcast median percent change in MSE that was negative (i.e., the variable did not add value in over half of hindcast years). Nonlinear correlation of the tuned suite of predictors with spring flows at various lead times is shown in Fig. 3. Nonlinear correlation is calculated with the 'nlcor' R package (Ranjan 2020).

Based on our recent results (Woodson et al. 2021) in stochastic simulation of CRB flows at decadal time scales, we propose a machine learning approach based on RF as the forecasting model along with the suite of predictors. More background on the RF algorithm is provided in the following section.

Modeling Approach-RF

RFs are a commonly used machine learning approach originated by Ho (1995) and popularized by Breiman (2001). Random forests are a supervised approach based on the repeated bootstrapping of a training dataset to generate many different classification and regression trees (CART). Random forests are an attractive nonparametric approach due to their predictive ability, speed, robustness, stability, and capabilities in handling nonlinearity, noise, interactive effects, and small sample sizes (Tyralis et al. 2019). Particularly in hydrologic modeling, many predictors will be correlated (e.g., precipitation, SWE, and soil moisture), but the presence of correlated variables does not impact random forest prediction accuracy due to the random sampling of predictors involved in growing a single tree, although it may influence variable importance metrics (Boulesteix et al. 2012; Ziegler and König 2014). Additionally, random forests are less prone to overfitting than other techniques because of the bootstrap aggregation involved in growing a forest and again due to the random selection of a subset of predictors for each

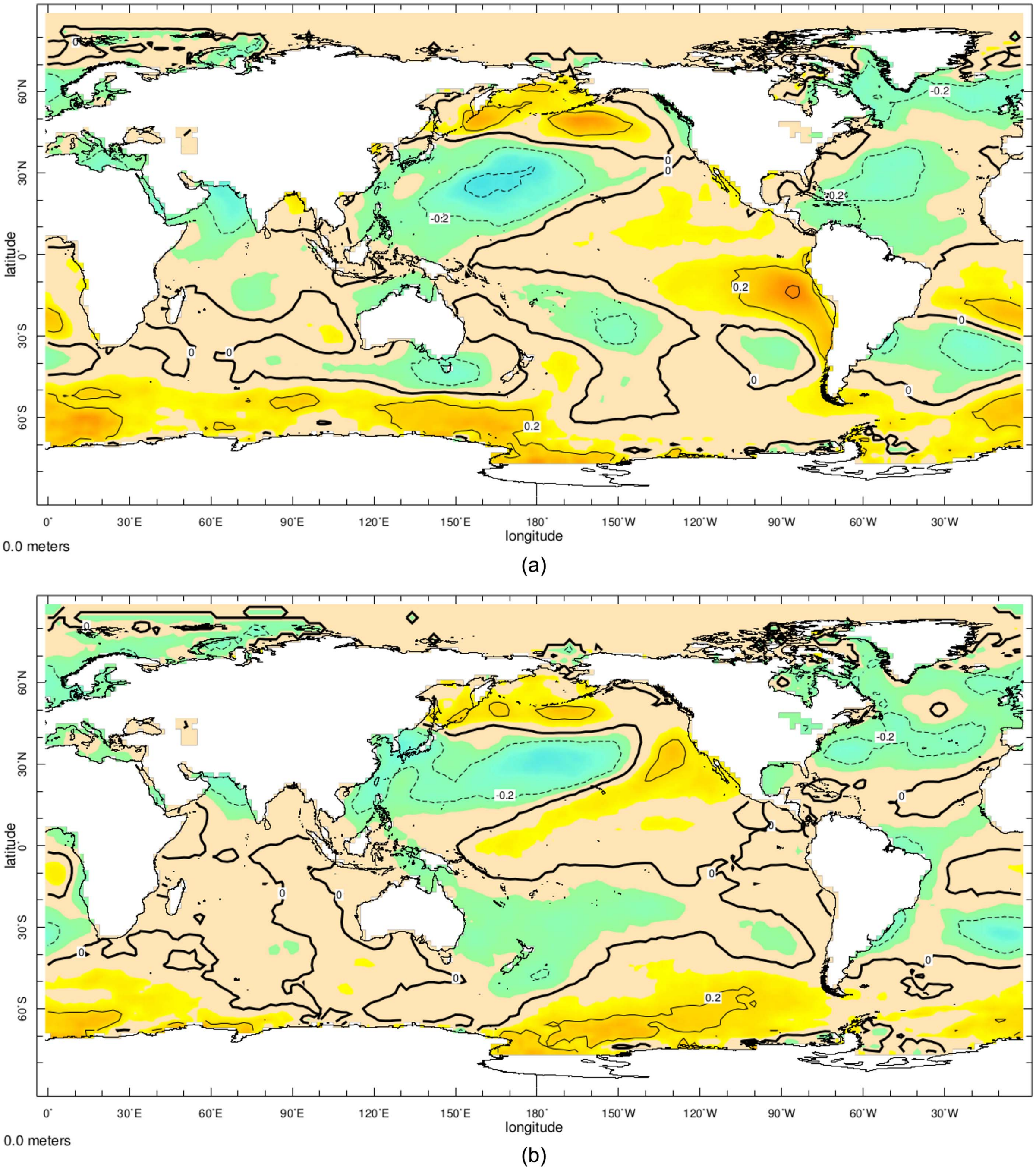

Fig. 2. (a) Rank correlation (1950–2021) between winter SST and concurrent, UCRB water year flow reminiscent of PDO, ENSO, and AMO teleconnections; and (b) rank correlation and spatial patterns persist when calculated between winter SSTs and *next year's* flow.

tree (Ziegler and König 2014). Others (Qi 2012) have reported that random forests can perform well even with small sample sizes; for example, Luan et al. (2020) found that a species distribution model's performance greatly improved after increasing the sample size from 10 to 30, but only marginal improvements occurred when the sample size increase to 50 to 80 samples.

Another valuable aspect of random forests is their ability to provide information about variable importance within the training dataset. Random forests have been used in many fields, with applications ranging from construction safety risk (Tixier et al. 2016) to water quality modeling (Suchetana et al. 2017). Researchers have also found utility in random forests for streamflow simulation, generally at monthly and daily time scales (Abbasi et al. 2020; Al-Juboori 2019; Ghorbani et al. 2020; Hussain and Khan 2020; Li et al. 2019; Liang et al. 2018; Muñoz et al. 2018; Papacharalampous and Tyralis 2018; Pham et al. 2020). Recently, Woodson et al. (2021) applied this for the projection of streamflows at decadal time scales in the CRB. The presented papers describe the methodology; however, for a good description of the method with implementation, we refer the readers to a book by Hastie et al. (2009).

In RF, the space of predictors (i.e., independent variables) is successively partitioned by randomly selecting one of the variables to partition at each step–hence, its name. The partitions continue until a stopping criterion is met that is often based on minimizing the mean squared error of the dependent variable, resulting in a 'tree.'

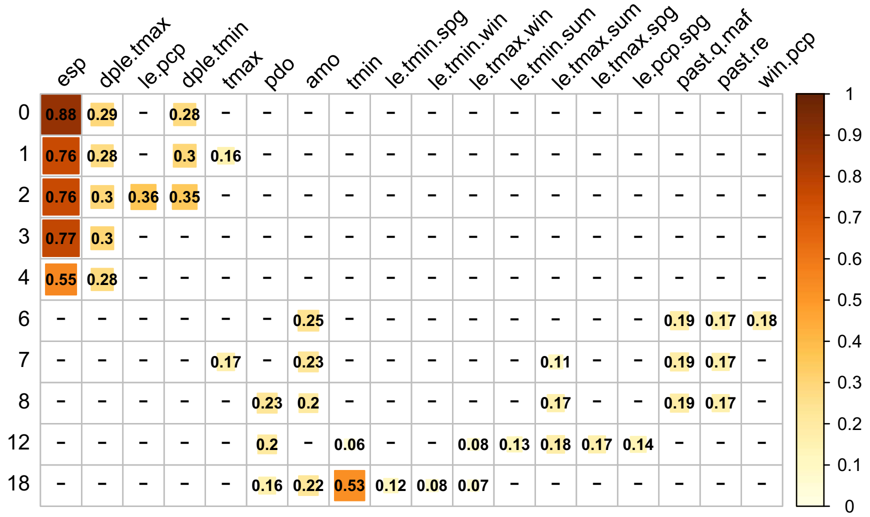

Fig. 3. Nonlinear correlation between predictors and predictand by lead time. Definitions for each variable are given in Table 1.

The process is repeated to generate a large number of trees. The estimate of the dependent variable is obtained from the average of the estimates from all the trees. However, to generate an ensemble, instead of computing the average across all the trees, we keep the estimates from each tree thus producing an 'ensemble.' This novel modification was proposed and outlined in Woodson et al. (2021) and is used here. We implemented this in R with the 'randomForest' package (Liaw and Wiener 2002; R Core Team 2019). Using the 'tuneRF' function within 'randomForest,' we optimize the 'mtry' parameter (number of predictors randomly sampled at each split in a tree) for each hindcast year and lead time based on the relevant training data. Further, we use a 'ntree' size of 2,000 trees for robust predictions.

While many other machine learning approaches exist and have been applied in hydrologic studies (e.g., support vector machines, Gaussian process regression, K-means clustering, long-and short-term memory networks, and other neural networks), we selected the RF algorithm due to its proven performance in streamflow forecasting, robustness, and computational efficiency.

Model Validation

For model validation, we tested both RF and CBRFC ESP using a Leave P Out Cross Validation (LPOCV) over the common period 1983–2017. In an LPOCV, P is equal to the number of data points dropped from the training set prior to model training. We set P = 5 and drop the year to be hindcast, as well as the two preceding years and the two following years. This provides for a 'blind' retrospective forecast, where no knowledge of the year in consideration is included in model training. For example, to hindcast water year 2001 spring flow, the following water years are dropped from the training set: 1999, 2000, 2001, 2002, and 2003. Similarly, to hindcast water year 2002 spring flow, data for water years 2000–2004 are dropped. This is applied to both the training dataset for the RF model and for the CBRFC ESP traces. There are 30 CBRFC ESP traces total for any given year, after dropping P = 5 years from this record, 25 traces remain to be used as an ESP hindcast of the given year. The LPOCV, while not truly blind as in real time forecast, attempts to mimic this and assess the hindcast models conservatively.

The LPOCV approach is applied to each year in the 1983–2017 period, resulting in 35 different hindcasts for each lead time from both the RF approach and ESP. For each hindcast, the RF generates a 2000-member ensemble, while the blind ESP hindcasts only have 25 traces. As mentioned in the description of the RF in an earlier section, traditionally, the mean of estimates of the predictand (in this case spring streamflow) from all the trees in the RF is used as the single output of the RF. However, as done in Woodson et al. (2021), we use the entirety of the forest as our ensemble.

Evaluation Metrics

Multiple metrics are used to evaluate the model performance of RF and ESP during the 1983–2017 validation period. The following paragraphs describe these metrics.

We use the ranked probability skill score (RPSS) to evaluate the RF and ESP forecast performance using historical flow terciles as the boundaries for the categories. RPSS is a categorical probabilistic skill metric (Epstein 1969; Murphy 1969, 1971; Weigel et al. 2007) that compares the forecasted categorical probabilities (e.g., a percent chance of being a high-, medium-, or low-flow year) with a climatological probabilistic hindcast based on the 1983–2017 natural flow record (~1/3 chance of being in any tercile). For a blind comparison, we drop the hindcast year from the flow record prior to use as climatology. RPSS values range from one to negative infinity. Scores greater or less than zero represent better or worse performance, respectively, of the hindcast relative to climatology. A score of one is considered a perfect hindcast. Another performance metric used is relative bias (RB), measuring accuracy via the deviation of the ensemble median from the observed value for each year [Eq. (1)]

$$\mathrm{RB}(\%) = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)}{\sum_{i=1}^{n} y_i} * 100 \tag{1}$$

where, $\hat{y}_i$ = Forecast ensemble median for year i; y_i = Observed value for year i; and n = number of years in the hindcast.

Additionally, we calculated the root MSE on the hindcast ensemble median for each lead time [Eq. (2)]

$$\mathrm{RMSE}_{\mathrm{median}} = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{2}$$

We also calculated an 'ensemble RMSE', which measures the performance of the entire ensemble for each year in the hindcast

[Eq. (3)] rather than the ensemble mean or median time series, as has been done in prior work (Baker et al. 2021b; Woodson et al. 2021). This is essentially the root of the average squared ensemble trace error

$$\text{RMSE}_{\text{ensemble,year}-i} = \sqrt{\sum_{j=1}^{m} \frac{(\hat{y}_{i,j} - y_i)^2}{m}} \tag{3}$$

where, $\hat{y}_{i,j}$ = Forecast ensemble member j for year i; and m = number of members in ensemble.

Finally, we evaluate the forecasts using reliability diagrams and rank histograms, which are general measures of reliability (Hashimoto et al. 1982) and show how well-observed frequencies (conditioned on the hindcast distribution) compare to hindcast probabilities (Baker et al. 2021b; Hartmann et al. 2002).

Overall, RPSS, ensemble RMSE, reliability diagrams, and rank histograms indicate overall ensemble performance, whereas RMSE and RB describe ensemble median performance.

Results

The predictor variables selected and their importance for the RF models at various lead times are first presented, followed by the hindcast skills.

Predictor Variable Importance

We fit a RF model for each year of the 35-year period in LPOCV mode; thus, for a given lead time we have 35 importance values for each variable that can be shown as a boxplot. The variables selected and their importance values for several lead times are shown as boxplots in Fig. 4 (see Table 1 for the full names of the abbreviations of the variables). We found that at shorter lead times (0–4 months), fewer (2–4) variables are important and used in the model. The CBRFC ESP ensemble mean hindcast was found to be the most important variable at these lead times along with seasonal projections of temperatures from the CESM-DPLE climate model (Fig. 4, top row). This is intuitive and expected, in that the CBRFC ESP's use of SAC-SMA and SNOW17 physical models captures all the hydrologic processes (e.g., snow, land surface conditions, and basin storage) quite robustly until the start of the forecast issuance, while the projections of temperatures during the spring season of interest provide additional complementary information not captured by the ESP model. Thus, together they help to capture all aspects of the land surface and projected atmospheric conditions.

The magnitude of variable importance was also generally greater at shorter lead times, indicating that the predictors add value to the model and are thus skillful. At shorter lead times (4 months or shorter), we initially used predictors like SWE and winter precipitation but later found similar performance by substituting these variables with CBRFC ESP ensemble mean and CESM-DPLE future temperature. The ESP hindcasts simulate SWE from winter precipitation inputs; thus, using ESP ensemble mean hindcasts as an input to our RF model was parsimonious. At longer lead times (6–18 months), the hydrologic information provided by the prior winter's snowpack and other land surface conditions diminishes. Consequently, the ESP forecasts tend to be closer to climatology and thus not very skillful. Therefore, the large-scale climate drivers and their teleconnection indices–PDO, AMO–provide new information about future climate conditions along with antecedent runoff efficiency and hindcasts of temperature and precipitation from the CESM-DPLE model. This is seen in the variables selected and their positive importance in Fig. 4 (bottom row). As lead time increases, there are no strong predictors; hence, combining information from several predictor variables is important. In addition, the variable importance magnitudes are lower in comparison to those at shorter lead times. For the 18-month hindcast, the most important variable was the preceding 3-month mean minimum temperature. Since the 18-month hindcast is issued on October 1st, this predictor would be the mean of the preceding July, August, and September

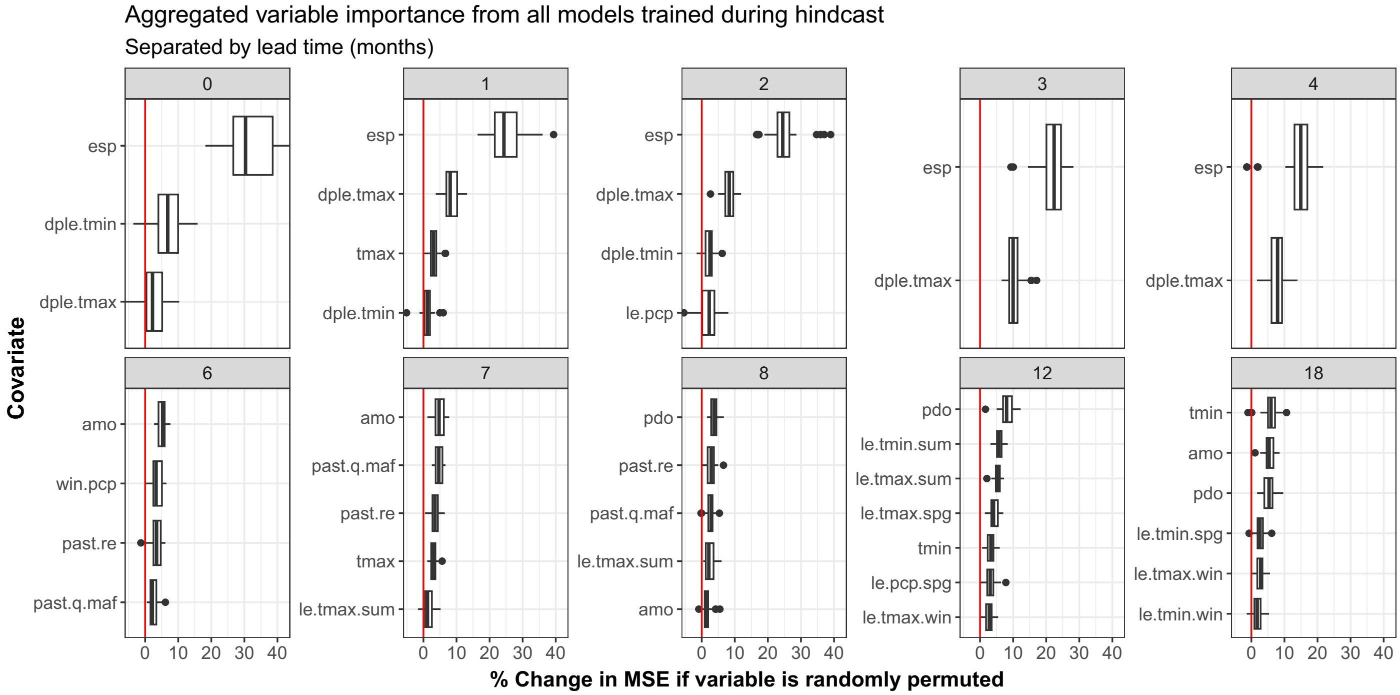

Fig. 4. Variable importance plots from RF flow hindcasts during 1983–2017 LPOCV (n = 35 years per boxplot).

monthly minimum temperatures. This suggests that the effects of antecedent basin conditions can extend far into the future.

Cross Validation Results and Skill

We generated LPOCV ensemble hindcasts for all 10 lead times and show boxplots of these at representative lead times of 8, 12, and 18 months in Fig. 5 to highlight that improving performance at longer lead times is a central goal of this work. Boxplots in Fig. 5 show the distribution of ensemble members for both the 25-member CBRFC ESP and 2,000-member RF ensembles as well as the observed spring flow value for each year. Even at an 8-month lead time, both ensembles are well dispersed, indicating comparable confidence and reliability. However, ESP's flow distribution generally does not change much over time, meaning that high- or low-flow years are not well captured by the ensemble median and interquartile range (IQR). The limited ensemble from ESP is also a likely factor in this. Conversely, the RF hindcast's median and IQR vary with time and better track with the observed flows, except in some years where it is off. For example, all three of the longer lead time hindcasts show, relative to CBRFC ESP, RF projections (as evidenced by the IQR) tend toward higher-than-average flows in the mid-1980s pluvial and lower-than-average flows during the early 2000s Millennium Drought. Conversely, CBRFC ESP looks more like a climatological ensemble during these extreme epochs.

Hindcast ensemble medians compared to observations are shown for each model type and lead time in Fig. S1. Performance deteriorates for all model types for lead times longer than 4 months. This shows the importance of probabilistic, rather than deterministic, predictions.

The first skill metric calculated was the RPSS on terciles for both CBRFC ESP and RF hindcasts at all lead times (Fig. 6). Individual points represent a single year in the retrospective forecast, while the boxplots represent aggregate performance during the entire 35-year period. At shorter lead times, the RPSS distribution

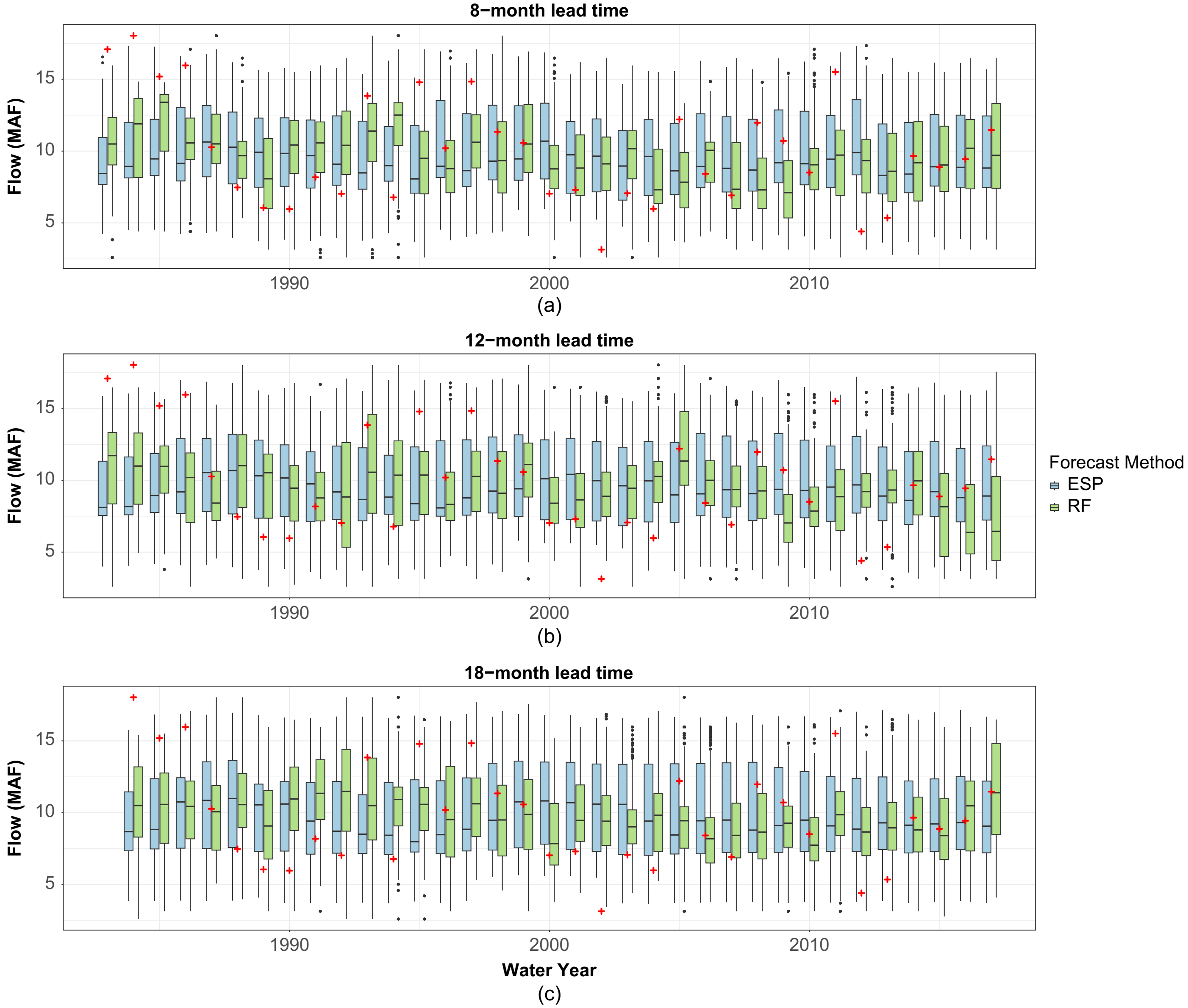

Fig. 5. LPOCV (P = 5) results for (a) 8-; (b) 12-; and (c) 18-month lead time hindcasts from ESP and 2000-member RF ensemble. Observed spring flows shown by cross marks. RF ensembles include values less than the minimum observed flow shown here because RF is trained on flows starting in 1921, some of which are below those in the 1983–2017 retrospective forecast period.

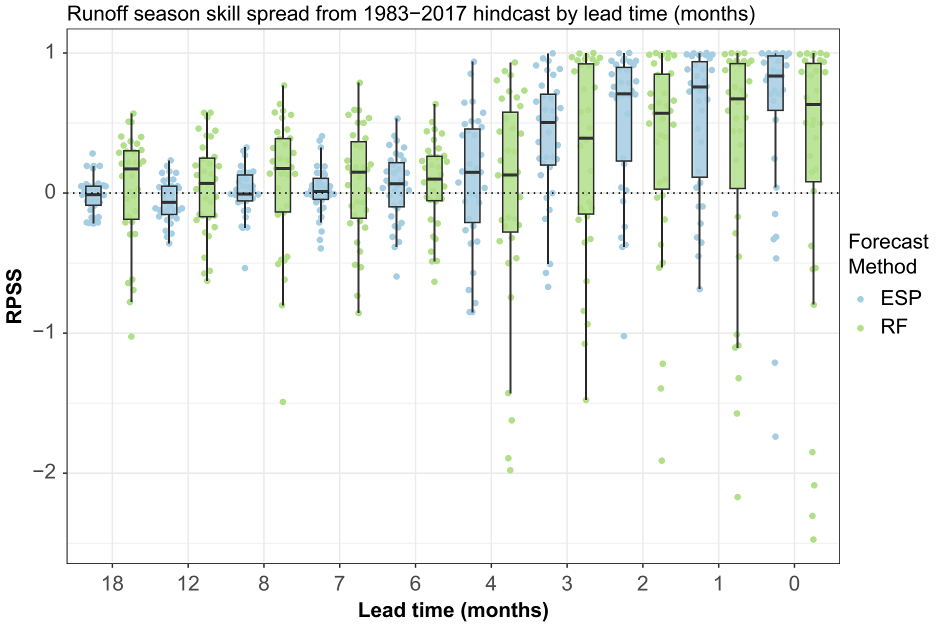

Fig. 6. RPSS results by lead time from LPOCV (P = 5) for 2,000-member RF ensemble and ESP. Climatology is 1983–2017 with the forecast year left out.

shows that ESP outperforms RF, although the latter outperforms climatology in ~75% of forecast years and has a median RPSS of around 0.5 or higher. As lead time increases, the performance of both methods decreases and ESP's median RPSS becomes zero or negative beyond 6 months. Conversely, the median RPSS of RF remains positive at all long lead times, with values spanning 0.07 to 0.18 for lead times of 7 to 18 months: a 0.14 to 0.18 improvement over the ESP median RPSS. RF's upper quartile and upper whisker also outperform ESP at longer leads. However, RF's lower quartile and lower whisker are generally lower than ESP's. RF's few but extreme failure cases may be a result of overconfident hindcasts in those poorly performing years (e.g., the RF IQR spread does not capture the observed value). This may be a result of predictor-predictand relationships not seen in the training set and motivates the need for longer datasets with more relevant predictors. Conversely, ESP does not experience such a great failure because at longer lead times it is underconfident and performs similarly to a climatological ensemble forecast (RPSS ~0).

Hindcast median RPSS values from each lead time and hindcast method are summarized in Table 2. RF shows the most improvement over ESP at 8- and 18-month lead times (Δ = 0.18). However, a 6-month lead time and longer RF shows increasing skill, while the skill of ESP declines below climatology. Continuous ranked probability skill scores (CRPSS) are also calculated for RF, ESP, and a linear model (Fig. S2).

Table 2. Median RPSS by lead time for RF and ESP hindcasts

Lead time (months)	Median RPSS		Δ Improvement (RF minus ESP)
	RF	ESP	
18	0.17	−0.01	0.18
12	0.07	−0.07	0.14
8	0.18	−0.01	0.18
7	0.15	0.01	0.14
6	0.10	0.07	0.03
4	0.13	0.15	−0.02
3	0.39	0.50	−0.11
2	0.57	0.71	−0.14
1	0.67	0.76	−0.08
0	0.63	0.84	−0.20

Similar to the results for RPSS, ESP generally has a lower RB and root MSE for lead times up to 3–5 months, but for leads of 6 months or longer, RF performs better than both ESP and climatology (Fig. 7). All three approaches appear to underpredict but are never lower than −10% in RB. ESP experiences the worst RB (−10%) during a 6-month lead time whereas RF is consistently between 0 and −6% in RB.

The data shown in Fig. 7 is numerically highlighted in Table 3. Percent change in the ensemble median RMSE was also calculated and shows how better or worse the RF ensemble median performs with respect to either ESP or climatology. For lead times of 6 months and greater, RF outperforms ESP with reductions in RMSE ranging from −10% to −15%. RF ensemble median outperforms climatology at all lead times, with RMSE reductions ranging from −9% to −35%. The ensemble RMSE (Fig. 8) shows a comparable pattern as the other metrics: RF begins to outperform ESP at a 6-month lead time and shows moderate improvement over ESP and climatology for lead times of 8, 12, and 18 months. At lead times of 8, 12, and 18 months, RF's median ensemble RMSE is 0.57-MAF (−13%), 0.23-MAF (−5%), and 0.56 (−12%) MAF lower, respectively, than ESP's median (Table 4).

Both methods appear to have high reliability, particularly at longer lead times, as shown by close agreement with the 1:1 line in the reliability diagram (Fig. 9). The high reliability at longer leads is likely due to the wide range both methods show at these leads, whereas at shorter lead times, the hindcasts may be overconfident and underdispersed.

Summary and Discussion

We developed a random forest-based machine learning modeling framework, for forecasting UCRB runoff season flow at several lead times spanning 0 to 18 months across years 1983–2017. The suite of predictors includes large-scale climate drivers–AMO,

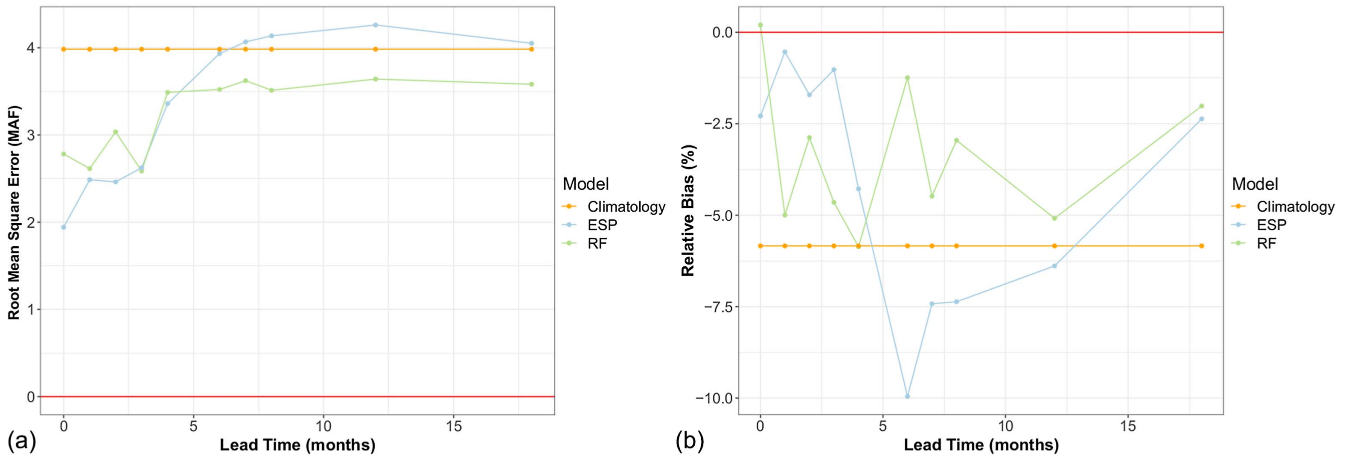

Fig. 7. (a) Root MSE; and (b) RB calculated on LPOCV ensemble median hindcast across lead time for three hindcast methods (RF, ESP, and climatology) over 1983–2017. Climatology is 1983–2017 with the hindcast year left out.

Table 3. RB and root MSE calculated on LPOCV ensemble medians by lead time for RF and ESP

	RB (%)			RMSE (MAF)				
Lead time (months)	RF	ESP	Climatology	RF	ESP	RF % change from ESP (%)	Climatology	RF % change from climatology (%)
18	−2.02	−2.37	−5.84	3.58	4.05	−12	3.98	−10
12	−5.09	−6.39	−5.84	3.64	4.26	−15	3.98	−9
8	−2.95	−7.37	−5.84	3.51	4.14	−15	3.98	−12
7	−4.48	−7.42	−5.84	3.62	4.07	−11	3.98	−9
6	−1.24	−9.95	−5.84	3.52	3.93	−10	3.98	−12
4	−5.86	−4.28	−5.84	3.49	3.36	4	3.98	−12
3	−4.65	−1.02	−5.84	2.59	2.62	−1	3.98	−35
2	−2.88	−1.71	−5.84	3.04	2.46	23	3.98	−24
1	−5.00	−0.54	−5.84	2.61	2.49	5	3.98	−34
0	0.20	−2.29	−5.84	2.78	1.94	43	3.98	−30

Note: RF % change indicates performance relative to ESP or climatology: negative values and positive values represent better and worse performance, respectively, from RF.

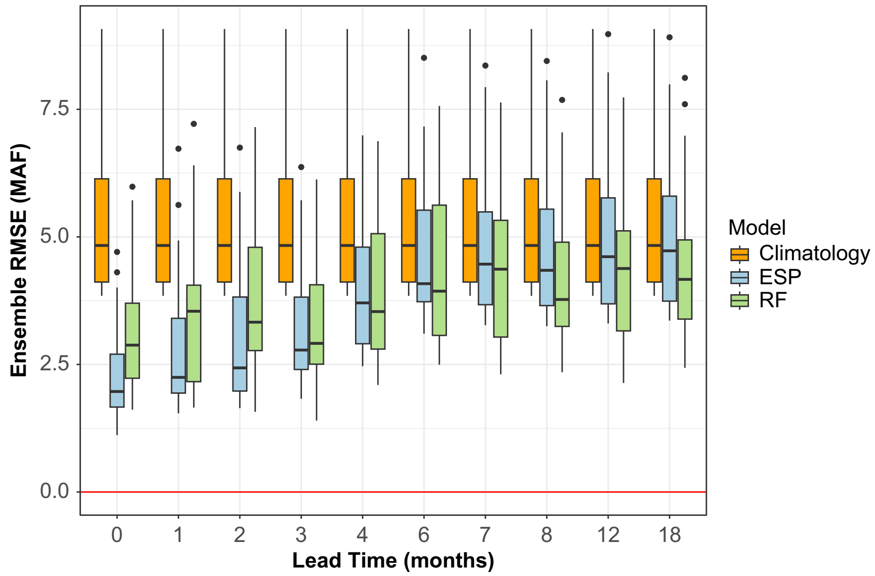

Fig. 8. Ensemble RMSE by the lead time for RF, ESP, and climatology during the 1983–2017 hindcast. Each point in a given boxplot is the RMSE calculated on the LPOCV forecast ensemble from a single hindcast year.

Table 4. Median ensemble RMSE by lead time for RF and ESP hindcasts

Lead time (months)	Median ensemble RMSE (MAF)		Δ Improvement (RF minus ESP, MAF)	% Change (RF over ESP) (%)
	RF	ESP		
18	4.17	4.73	−0.56	−12
12	4.38	4.61	−0.23	−5
8	3.77	4.34	−0.57	−13
7	4.37	4.47	−0.10	−2
6	3.94	4.08	−0.15	−4
4	3.54	3.71	−0.17	−5
3	2.91	2.78	0.13	5
2	3.33	2.43	0.90	37
1	3.54	2.25	1.30	58
0	2.88	1.97	0.91	46

PDO, hindcasts of temperature and precipitation from CESM-DPLE and CESM-LE climate models, basin runoff efficiency, and the ensemble mean of CBRFC ESP flow. At shorter lead times (0–4 months), when the basin's IHC have high persistence, CBRFC ESP retrospective forecasts forced with 1981–2010 precipitation and temperature were skillful in comparison to the RF. However, at longer lead times, CBRFC ESP was less skillful than RF, and in some cases, climatology. Based on multiple metrics, random forest-based hindcasts outperformed both CBRFC ESP and climatology for lead times of 8, 12, and 18 months while maintaining reliability over the course of a 35-year-long hindcast. At long leads, RF hindcasts also performed better than CBRFC ESP at capturing wet and dry extremes in the 1980s and early 21st century, respectively, as evidenced by the IQR spread of each model type.

At shorter lead times, the CBRFC ESP ensemble mean was the most important training variable for the RF hindcasts followed by CESM-DPLE temperature projections, indicating the usefulness of CBRFC ESP forecasts and associated IHC at these lead times. At lead times of 6 months or greater, the ESP ensemble mean had a negative hindcast median variable importance score, indicating that it added no value to RF hindcasts in over half of hindcast years. For these longer lead times, ocean teleconnections, antecedent minimum temperature, and to a lesser degree, climate projections, were the only variables that added value to the RF hindcasts.

Although only naturalized flows at Lees Ferry, AZ, were examined in this study, other studies have found slightly better performance when making UCRB subbasin forecasts then aggregating, compared to UCRB aggregate (i.e., Lees Ferry) forecasts (Baker et al. 2021a). Though outside of the scope of this research, model performance at a subbasin scale should be elucidated when using novel ML techniques trained on ocean teleconnections and climate model projections. Others have developed robust methods for disaggregating streamflows in space and time from a single gauge to subbasins (Nowak et al. 2010). For this study, the runoff season flow hindcasts could be disaggregated in space-time to all UCRB subbasins and be used as inputs to a decision model such as Reclamation's CRB Mid-term Modeling System that is used in 0–24 months operations and planning (Baker et al. 2021b; Reclamation 2019; Towler et al. 2021; Woodson et al. 2021).

Our proposed method showing improved retrospective forecasts over existing operational forecasts through the inclusion of additional basin and climate information serves as a proof of concept toward enhanced predictability of years 1 to 2 CRB flows. Mechanistic sources that contribute to the predictability of years 1 to 2 CRB flows need to be fully understood. However, machine learning methods can better exploit them and potentially the nonlinearities and interactions between land, ocean, and atmospheric variables, and augment traditional, physically based methods like SAC-SMA and SNOW-17 used in CBRFC's ESP forecasts. The relatively low computational burden of machine learning approaches is also attractive. Given the importance of Reclamation's forecasts to stakeholders for anticipating upcoming operations that affect water deliveries, hydropower production, recreation, and environmental flows under increasingly uncertain and critical conditions, even moderate improvements over climatological and operational forecasts are likely to be welcome tools.

Encountering streamflows that are outside of the historical envelope is a challenge for both RF and CBRFC's ESP technique and is a phenomenon that will likely increase with climate change's influence on variability and extremes. RF hindcasts performed better than CBRFC ESP in anticipating these unprecedented values but are susceptible to greater failure in cases where the relationship between predictors and predictand changes from what is learned in the training set, which is a potential reality, particularly for ocean

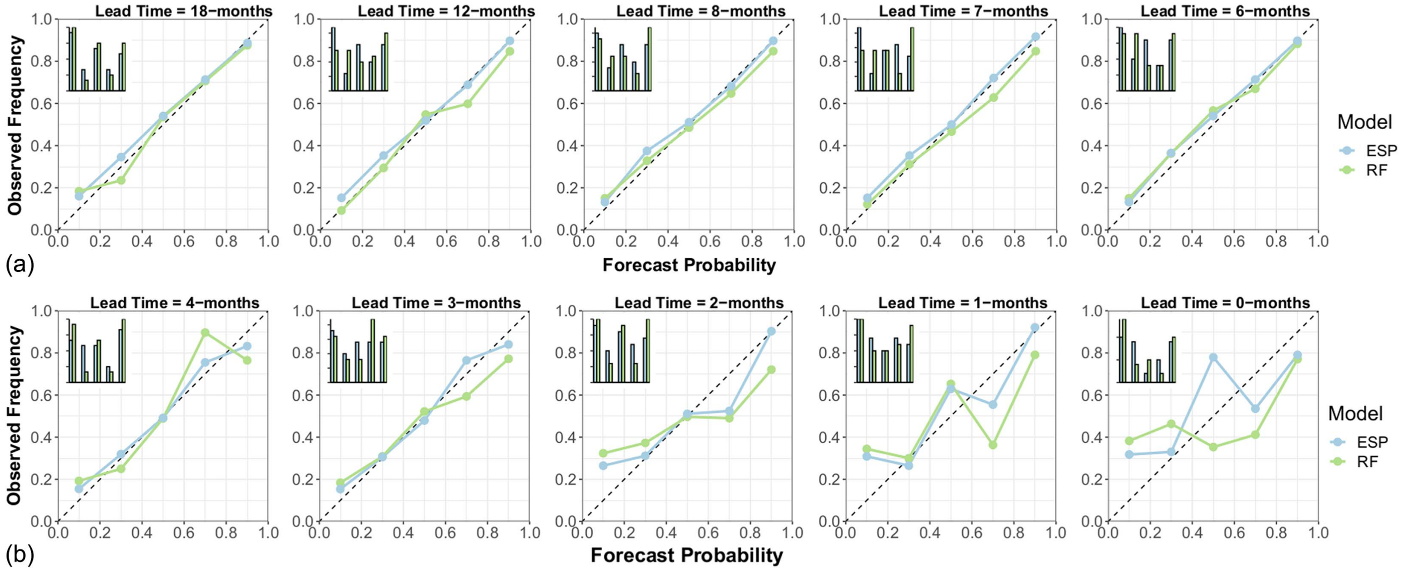

Fig. 9. Reliability diagram and rank histogram by lead time for RF and ESP hindcast ensembles: (a) shows long lead times (6 to 18 months); and (b) shows short lead times (0 to 4 months).

teleconnections as the oceans warm and salinity changes. As such, future work might examine the use of tailored SST predictors, potentially delineated using causal and convolution network methods or long- and short-term memory networks (e.g., Kratzert et al. 2019a, b; Sasou 2021; Song 2022; Zhang et al. 2021), to better exploit the high correlation zones observed in the Pacific and Atlantic Oceans, rather than simply using traditional ocean indices. Additionally, other machine learning approaches could be used instead of or in addition to the RF approach, with the latter having the potential to be a multimodel projection. Training ML models on the paleoreconstructed streamflow record (e.g., Gangopadhyay et al. 2022) could yield a rich predictive source, assuming quality predictors are available for such a time span. Finally, the assessment of translation of the skills in flow forecasts from this study to their potential skill in water resources decision variables is crucial for real time application, particularly with respect to how well such forecasts perform concerning robustness, reliability, and vulnerability of, for example, estimates of reservoir elevation from a decision model forced with streamflow forecasts.

Data Availability Statement

All data and code that support the findings of this study are available from the corresponding author upon reasonable request and are available in a repository online in accordance with funder data retention policies: https://www.hydroshare.org/resource/4358608e3fb343f3a62770306072e48b/.

Acknowledgments

We gratefully acknowledge funding from the US Bureau of Reclamation (Grant Number 19264) as well as the NOAA Weather Program Office (Project Number 04.FY19.2). We also thank and acknowledge NCAR scientists Dr. Erin Towler and Ming Ge for the provision of climate model projection data. Additionally, we would like to thank the two anonymous reviewers whose feedback and suggestions greatly strengthened this work.

Supplemental Materials

Figs. S1 and S2 are available online in the ASCE Library (www.ascelibrary.org).

References

Abbasi, M., A. Farokhnia, M. Bahreinimotlagh, and R. Roozbahani. 2020. "A hybrid of random forest and deep auto-encoder with support vector regression methods for accuracy improvement and uncertainty reduction of long-term streamflow prediction." *J. Hydrol.* 597 (Jun): 125717. https://doi.org/10.1016/j.jhydrol.2020.125717.

Al-Juboori, A. M. 2019. "Generating monthly stream flow using nearest river data: Assessing different trees models." *Water Resour. Manage.* 33 (9): 3257–3270. https://doi.org/10.1007/s11269-019-02299-4.

Baker, S. A., B. Rajagopalan, and A. W. Wood. 2021a. "Enhancing ensemble seasonal streamflow forecasts in the upper Colorado River basin using multi-model climate forecasts." *JAWRA 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, C. Jerla, E. Zagona, R. Butler, and R. Smith. 2021b. "The Colorado river basin operational prediction testbed: A framework for evaluating streamflow forecasts and reservoir operations." *JAWRA J. Am. Water Resour. Assoc.* 58 (5): 690–708. https://doi.org/10.1111/1752-1688.13038.

Boulesteix, A.-L., S. Janitza, J. Kruppa, and I. R. König. 2012. "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics." *WIREs Data Min. Knowl. Discovery* 2 (6): 493–507. https://doi.org/10.1002/widm.1072.

Bracken, C., B. Rajagopalan, and J. Prairie. 2010. "A multisite seasonal ensemble streamflow forecasting technique." *Water Resour. Res.* 46 (3): 2009WR007965. https://doi.org/10.1029/2009WR007965.

Bracken, C., B. Rajagopalan, and C. Woodhouse. 2016. "A Bayesian hierarchical nonhomogeneous hidden Markov model for multisite streamflow reconstructions." *Water Resour. Res.* 52 (10): 7837–7850. https://doi.org/10.1002/2016WR018887.

Breiman, L. 2001. "Random forests." *Mach. Learn.* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Chikamoto, Y., S.-Y. S. Wang, M. Yost, L. Yocom, and R. R. Gillies. 2020. "Colorado River water supply is predictable on multi-year timescales owing to long-term ocean memory." *Commun. Earth Environ.* 1 (1): 26. https://doi.org/10.1038/s43247-020-00027-0.

Christensen, N. S., A. W. Wood, N. Voisin, D. P. Lettenmaier, and R. N. Palmer. 2004. "The effects of climate change on the hydrology and water resources of the Colorado River Basin." *Clim. Change* 62 (1–3): 337–363. https://doi.org/10.1023/B:CLIM.0000013684.13621.1f.

Daly, C., R. Neilson, and D. Phillips. 1994. "A statistical-topographic model for mapping climatological precipitation over mountain terrain." *J. Appl. Meteorol.* 33 (2): 140–158. https://doi.org/10.1175/1520-0450(1994)033<0140:ASTMFM>2.0.CO;2.

Day, G. N. 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).

Enfield, D. B., A. M. Mestas-Nuñez, and P. J. Trimble. 2001. "The Atlantic multidecadal oscillation and its relation to rainfall and river flows in the continental U.S." *Geophys. Res. Lett.* 28 (10): 2077–2080. https://doi.org/10.1029/2000GL012745.

Epstein, E. S. 1969. "A scoring system for probability forecasts of ranked categories." *J. Appl. Meteorol. (1962-1982)* 8 (6): 985–987. https://doi.org/10.1175/1520-0450(1969)008<0985:ASSFPF>2.0.CO;2.

Erkyihun, S. T., B. Rajagopalan, E. Zagona, U. Lall, and K. Nowak. 2016. "Wavelet-based time series bootstrap model for multidecadal streamflow simulation using climate indicators: Wavelet-based time series bootstrap." *Water Resour. Res.* 52 (5): 4061–4077. https://doi.org/10.1002/2016WR018696.

Fleming, S. W., and A. G. Goodbody. 2019. "A machine learning metasystem for robust probabilistic nonlinear regression-based forecasting of seasonal water availability in the US West." *IEEE Access* 7 (Aug): 119943–119964. https://doi.org/10.1109/ACCESS.2019.2936989.

Gangopadhyay, S., C. A. Woodhouse, G. J. McCabe, C. C. Routson, and D. M. Meko. 2022. "Tree rings reveal unmatched 2nd century drought in the Colorado River Basin." *Geophys. Res. Lett.* 49 (11). https://doi.org/10.1029/2022GL098781.

Ghorbani, M. A., R. C. Deo, S. Kim, M. Hasanpour Kashani, V. Karimi, and M. Izadkhah. 2020. "Development and evaluation of the cascade correlation neural network and the random forest models for river stage and river flow prediction in Australia." *Soft Comput.* 24 (16): 12079–12090. https://doi.org/10.1007/s00500-019-04648-2.

Harpold, A. A., K. Sutcliffe, J. Clayton, A. Goodbody, and S. Vazquez. 2017. "Does including soil moisture observations improve operational streamflow forecasts in snow-dominated watersheds?" *JAWRA J. Am. Water Resour. Assoc.* 53 (1): 179–196. https://doi.org/10.1111/1752-1688.12490.

Hartmann, H. C., T. C. Pagano, S. Sorooshian, and R. Bales. 2002. "Confidence builders: Evaluating seasonal climate forecasts from user perspectives." *Bull. Am. Meteorol. Soc.* 83 (5): 683–698. https://doi.org/10.1175/1520-0477(2002)083<0683:CBESCF>2.3.CO;2.

Hashimoto, T., J. R. Stedinger, and D. P. Loucks. 1982. "Reliability, resiliency, and vulnerability criteria for water resource system performance evaluation." *Water Resour. Res.* 18 (1): 14–20. https://doi.org/10.1029/WR018i001p00014.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The elements of statistical learning: Data mining, inference, and prediction*. 2nd ed. New York: Springer.

Ho, T. K. 1995. “Random decision forests.” In Vol. 1 of *Proc., 3rd Int. Conf. on Document Analysis and Recognition*, 278–282. New York: IEEE. https://doi.org/10.1109/ICDAR.1995.598994.

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.

Hussain, D., and A. A. Khan. 2020. “Machine learning techniques for monthly river flow forecasting of Hunza River, Pakistan.” *Earth Sci. Inf.* 13 (3): 939–949. https://doi.org/10.1007/s12145-020-00450-z.

Kalra, A., and S. Ahmad. 2012. “Estimating annual precipitation for the Colorado River Basin using oceanic-atmospheric oscillations.” *Water Resour. Res.* 48 (6). https://doi.org/10.1029/2011WR010667.

Kay, J. E., et al. 2014. “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.

Kratzert, F., D. Klotz, M. Herrnegger, A. K. Sampson, S. Hochreiter, and G. S. Nearing. 2019a. “Toward improved predictions in ungauged basins: Exploiting the power of machine learning.” *Water Resour. Res.* 55 (12): 11344–11354. https://doi.org/10.1029/2019WR026065.

Kratzert, F., D. Klotz, G. Shalev, G. Klambauer, S. Hochreiter, and G. Nearing. 2019b. “Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets.” *Hydrol. Earth Syst. Sci.* 23 (12): 5089–5110. https://doi.org/10.5194/hess-23-5089-2019.

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 U.S. southwest.” *Geophys. Res. Lett.* 44 (24): 208–217. https://doi.org/10.1002/2017GL076043.

Li, X., J. Sha, and Z.-L. Wang. 2019. “Comparison of daily streamflow forecasts using extreme learning machines and the random forest method.” *Hydrol. Sci. J.* 64 (15): 1857–1866. https://doi.org/10.1080/02626667.2019.1680846.

Liang, Z., T. Tang, B. Li, T. Liu, J. Wang, and Y. Hu. 2018. “Long-term streamflow forecasting using SWAT through the integration of the random forests precipitation generator: Case study of Danjiangkou reservoir.” *Hydrol. Res.* 49 (5): 1513–1527. https://doi.org/10.2166/nh.2017.085.

Liaw, A., and M. Wiener. 2002. “Classification and regression by randomForest.” *R News* 2 (3): 18–22.

Luan, J., C. Zhang, B. Xu, Y. Xue, and Y. Ren. 2020. “The predictive performances of random forest models with limited sample size and different species traits.” *Fish. Res.* 227 (Jul): 105534. https://doi.org/10.1016/j.fishres.2020.105534.

Lukas, J., and E. Payton. 2020. *Colorado River Basin climate and hydrology: State of the science*. Boulder, CO: Western Water Assessment, Univ. of Colorado Boulder. https://doi.org/10.25810/3HCV-W477.

Mantua, N. J., S. R. Hare, Y. Zhang, J. M. Wallace, and R. C. Francis. 1997. “A Pacific interdecadal climate oscillation with impacts on salmon production.” *Bull. Am. Meteorol. Soc.* 78 (6): 1069. https://doi.org/10.1175/1520-0477(1997)078<1069:APICOW>2.0.CO;2.

McCabe, G. J., and D. M. Wolock. 2007. “Warming may create substantial water supply shortages in the Colorado River basin.” *Geophys. Res. Lett.* 34 (22): L22708. https://doi.org/10.1029/2007GL031764.

Meko, D. M., C. A. Woodhouse, C. A. Baisan, T. Knight, J. J. Lukas, M. K. Hughes, and M. W. Salzer. 2007. “Medieval drought in the upper Colorado River Basin.” *Geophys. Res. Lett.* 34 (10). https://doi.org/10.1029/2007GL029988.

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.

Muñoz, P., J. Orellana-Alvear, P. Willems, and R. Célleri. 2018. “Flash-flood forecasting in an Andean mountain catchment—Development of a step-wise methodology based on the random forest algorithm.” *Water* 10 (11): 1519. https://doi.org/10.3390/w10111519.

Murphy, A. H. 1969. “On the ‘ranked probability score.’” *J. Appl. Meteorol. Climatol.* 8 (6): 988–989. https://doi.org/10.1175/1520-0450(1969)008<0988:OTPS>2.0.CO;2.

Murphy, A. H. 1971. “A note on the ranked probability score.” *J. Appl. Meteorol. Climatol.* 10 (1): 155–156. https://doi.org/10.1175/1520-0450(1971)010<0155:ANOTRP>2.0.CO;2.

Nowak, K., M. Hoerling, B. Rajagopalan, and E. Zagona. 2012. “Colorado River Basin hydroclimatic variability.” *J. Clim.* 25 (12): 4389–4403. https://doi.org/10.1175/JCLI-D-11-00406.1.

Nowak, K., J. Prairie, B. Rajagopalan, and U. Lall. 2010. “A nonparametric stochastic approach for multisite disaggregation of annual to daily streamflow.” *Water Resou. Res.* 46 (8). https://doi.org/10.1029/2009WR008530.

Papacharalampous, G. A., and H. Tyralis. 2018. “Evaluation of random forests and Prophet for daily streamflow forecasting.” *Adv. Geosci.* 45 (Aug): 201–208. https://doi.org/10.5194/adgeo-45-201-2018.

Payton, E., R. Smith, C. Jerla, and J. Prairie. 2020. “Primary planning tools.” In *Chap. 3 in Colorado River Basin climate and hydrology: State of the science*, 82–111. Boulder, CO: Western Water Assessment, Univ. of Colorado Boulder.

Pham, L. T., L. Luo, and A. O. Finley. 2020. “Evaluation of random forest for short-term daily streamflow forecast in rainfall and snowmelt driven watersheds.” *Hydrol. Earth Syst. Sci. Discuss.* 2020 (Jun): 1–33. https://doi.org/10.5194/hess-2020-305.

Prairie, J. R., B. Rajagopalan, T. J. Fulp, and E. A. Zagona. 2005. “Statistical nonparametric model for natural salt estimation.” *J. Environ. Eng.* 131 (1): 130–138. https://doi.org/10.1061/(ASCE)0733-9372(2005)131:1(130).

Qi, Y. 2012. “Random Forest for Bioinformatics.” In *Ensemble machine learning: Methods and applications*, edited by C. Zhang, and Y. Ma, 307–323. New York: Springer. https://doi.org/10.1007/978-1-4419-9326-7_11.

Rajagopalan, B., S. T. Erkyihun, U. Lall, E. Zagona, and K. Nowak. 2019. “A nonlinear dynamical systems-based modeling approach for stochastic simulation of streamflow and understanding predictability.” *Water Resour. Res.* 55 (7): 6268–6284. https://doi.org/10.1029/2018WR023650.

Ranjan, C. 2020. “Package ‘nlcor’: Compute nonlinear correlations.” Research Gate. 10. https://doi.org/10.13140/RG.2.2.33716.68480.

R Core Team. 2019. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reclamation. 2015. *Colorado River Basin mid-term probabilistic operations model (MTOM) overview and description*. Denver: Bureau of Reclamation.

Reclamation. 2019. *Colorado River Basin mid-term probabilistic operations model (MTOM) technical user guide for stakeholders*. Denver: Bureau of Reclamation.

Reclamation. 2021. *Water reliability in the West—2021 SECURE Water Act report (prepared for the United States Congress)*. Denver: Bureau of Reclamation, Water Resources and Planning Office.

Reclamation. 2022. “Interior department announces actions to protect Colorado River system, sets 2023 Operating Conditions for Lake Powell and Lake Mead.” Accessed February 26, 2023. https://www.doi.gov/pressreleases/interior-department-announces-actions-protect-colorado-river-system-sets-2023.

Regonda, S. K., B. Rajagopalan, and M. Clark. 2006a. “A new method to produce categorical streamflow forecasts.” *Water Resour. Res.* 42 (9). https://doi.org/10.1029/2006WR004984.

Regonda, S. K., B. Rajagopalan, M. Clark, and E. Zagona. 2006b. “A multimodel ensemble forecast framework: Application to spring seasonal flows in the Gunnison River Basin.” *Water Resour. Res.* 42 (9). https://doi.org/10.1029/2005WR004653.

Rosenberg, E. A., A. W. Wood, and A. C. Steinemann. 2011. “Statistical applications of physically based hydrologic models to seasonal streamflow forecasts.” *Water Resour. Res.* 47 (3). https://doi.org/10.1029/2010WR010101.

Salehabadi, H., D. G. Tarboton, B. Udall, K. G. Wheeler, and J. C. Schmidt. 2022. “An assessment of potential severe droughts in the Colorado River Basin.” *JAWRA J. Am. Water Resour. Assoc.* 58 (6): 1053–1075. https://doi.org/10.1111/1752-1688.13061.

Sasou, A. 2021. “Deep residual learning with dilated causal convolution extreme learning machine.” *IEEE Access* 9 (Dec): 165708–165718. https://doi.org/10.1109/ACCESS.2021.3134700.

Smith, R., E. Zagona, J. Kasprzyk, N. Bonham, E. Alexander, A. Butler, J. Prairie, and C. Jerla. 2022. “Decision science can help address the challenges of long-term planning in the Colorado River Basin.” *JAWRA J. Am. Water Resour. Assoc.* 58 (5): 735–745. https://doi.org/10.1111/1752-1688.12985.

Song, C. M. 2022. “Data construction methodology for convolution neural network based daily runoff prediction and assessment of its applicability.” *J. Hydrol.* 605 (Feb): 127324. https://doi.org/10.1016/j.jhydrol.2021.127324.

Stern, C. 2023. *Responding to drought in the Colorado River Basin: Federal and state efforts*. Washington, DC: Congressional Research Service.

Suchetana, B., B. Rajagopalan, and J. Silverstein. 2017. “Assessment of wastewater treatment facility compliance with decreasing ammonia discharge limits using a regression tree model.” *Sci. Total Environ.* 598: 249–257. https://doi.org/10.1016/j.scitotenv.2017.03.236.

Switanek, M. B., and P. A. Troch. 2011. “Decadal prediction of Colorado River streamflow anomalies using ocean-atmosphere teleconnections.” *Geophys. Res. Lett.* 38 (23). https://doi.org/10.1029/2011GL049644.

Tixier, A. J.-P., M. R. Hallowell, B. Rajagopalan, and D. Bowman. 2016. “Application of machine learning to construction injury prediction.” *Autom. Constr.* 69: 102–114. https://doi.org/10.1016/j.autcon.2016.05.016.

Towler, E., D. PaiMazumder, and J. Done. 2018. “Toward 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., D. Woodson, S. Baker, M. Ge, J. Prairie, B. Rajagopalan, S. Shanahan, and R. Smith. 2021. “Incorporating mid-term temperature predictions into streamflow forecasts and operational reservoir projections in the Colorado River Basin.” *J. Water Resour. Plann. Manage.* 148 (4): 04022007. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001534.

Towler, E., and D. Yates. 2021. “Incorporating multiyear temperature predictions for water resources planning.” *J. Appl. Meteorol. Climatol.* 60 (2): 171–183. https://doi.org/10.1175/JAMC-D-20-0134.1.

Tyralis, H., G. Papacharalampous, and A. Langousis. 2019. “A brief review of random forests for water scientists and practitioners and their recent history in water resources.” *Water* 11 (5): 910. https://doi.org/10.3390/w11050910.

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

USBR. 2020. “Colorado River interim guidelines for lower basin shortages and coordinated operations for Lake Powell and Lake Mead.” Accessed April 14, 2020. https://www.usbr.gov/lc/region/programs/strategies.html#IGReview.

Vano, J. A., et al. 2014. “Understanding uncertainties in future Colorado River Streamflow.” *Bull. Am. Meteorol. Soc.* 95 (1): 59–78. https://doi.org/10.1175/BAMS-D-12-00228.1.

Vano, J. A., T. Das, and D. P. Lettenmaier. 2012. “Hydrologic sensitivities of Colorado River runoff to changes in precipitation and temperature.” *J. Hydrometeorol.* 13 (3): 932–949. https://doi.org/10.1175/JHM-D-11-069.1.

Wang, J., B. Udall, E. Kuhn, K. Wheeler, and Schmidt, J. C. (2022). *Evaluating the Accuracy of reclamation’s 24-month study Lake Powell projections (white paper no. 7)*, 20. Logan, UT: Center for Colorado River Studies.

Weigel, A. P., M. A. Liniger, and C. Appenzeller. 2007. “The discrete brier and ranked probability skill scores.” *Mon. Weather Rev.* 135 (1): 118–124. https://doi.org/10.1175/MWR3280.1.

Werner, K., and K. Yeager. 2013. “Challenges in forecasting the 2011 runoff season in the Colorado Basin.” *J. Hydrometeorol.* 14 (4): 1364–1371. https://doi.org/10.1175/JHM-D-12-055.1.

Williams, A. P., B. I. Cook, and J. E. Smerdon. 2022. “Rapid intensification of the emerging southwestern North American megadrought in 2020–2021.” *Nat. Clim. Change* 12 (3): 232–234. https://doi.org/10.1038/s41558-022-01290-z.

Williams, A. P., E. R. Cook, J. E. Smerdon, B. I. Cook, J. T. Abatzoglou, K. Bolles, S. H. Baek, A. M. Badger, and B. Livneh. 2020. “Large contribution from anthropogenic warming to an emerging North American megadrought.” *Science* 368 (6488): 314–318. https://doi.org/10.1126/science.aaz9600.

Wood, A., and K. Werner. 2011. “Development of a seasonal climate and streamflow forecasting testbed for the Colorado River Basin.” In *Science and technology infusion climate bulletin*. Silver Spring, MD: National Weather Service.

Woodhouse, C. A., K. E. Kunkel, D. R. Easterling, and E. R. Cook. 2005. “The twentieth-century pluvial in the western United States.” *Geophys. Res. Lett.* 32 (7). https://doi.org/10.1029/2005GL022413.

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: Temperature and Colorado 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.

Xiao, M., B. Udall, and D. P. Lettenmaier. 2018. “On the causes of declining Colorado River streamflows.” *Water Resour. Res.* 54 (9): 6739–6756. https://doi.org/10.1029/2018WR023153.

Yeager, S. G., et al. 2018. “Predicting near-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 modeling1.” *JAWRA J. Am. Water Resour. Assoc.* 37 (4): 913–929. https://doi.org/10.1111/j.1752-1688.2001.tb05522.x.

Zhang, G., W. Chen, C. Huo, H. Bai, J. Gao, J. He, S. Ma, and T. A. Liu. 2021. “A method of load forecasting based on temporal convolutional network.” In *Proc., 2021 4th Int. Conf. on Artificial Intelligence and Big Data (ICAIBD)*, 198–202. New York: IEEE. https://doi.org/10.1109/ICAIBD51990.2021.9458972.

Zhang, Y., J. M. Wallace, and D. S. Battisti. 1997. “ENSO-like Interdecadal Variability: 1900–93.” *J. Clim.* 10 (5): 1004–1020. https://doi.org/10.1175/1520-0442(1997)010<1004:ELIV>2.0.CO;2.

Zhao, S., Y. Deng, and R. X. Black. 2017. “Observed and simulated spring and summer dryness in the United States: The impact of the pacific sea surface temperature and beyond.” *J. Geophys. Res.: Atmos.* 122 (23): 12–713. https://doi.org/10.1002/2017JD027279.

Zhao, S., R. Fu, M. L. Anderson, S. Chakraborty, J. H. Jiang, H. Su, and Y. Gu. 2023. “Extended seasonal prediction of spring precipitation over the Upper Colorado River Basin.” *Clim. Dyn.* 60 (5): 1815–1829. https://doi.org/10.1007/s00382-022-06422-x.

Zhao, S., R. Fu, Y. Zhuang, and G. Wang. 2021. “Long-lead seasonal prediction of streamflow over the upper Colorado River basin: The role of the Pacific Sea surface temperature and beyond.” *J. Clim.* 34 (16): 6855–6873. https://doi.org/10.1175/JCLI-D-20-0824.1.

Zhao, S., and J. Zhang. 2022. “Causal effect of the tropical Pacific sea surface temperature on the Upper Colorado River Basin spring precipitation.” *Clim. Dyn.* 58 (3): 941–959. https://doi.org/10.1007/s00382-021-05944-0.

Ziegler, A., and I. R. König. 2014. “Mining data with random forests: Current options for real-world applications.” *WIREs Data Min. Knowl. Discovery* 4 (1): 55–63. https://doi.org/10.1002/widm.1114.