Stochastic Decadal Projections of Colorado River Streamflow and Reservoir Pool Elevations Conditioned on Temperature Projections

David Woodson[1], Balaji Rajagopalan[1,2], Sarah Baker[3], Rebecca Smith[3], James Prairie[3], Erin Towler[4], Ming Ge[4], and Edith Zagona[5]

[1]University of Colorado, Boulder.

[2]Cooperative Institute for Research in Environmental Sciences.

[3]United States Bureau of Reclamation.

[4]National Center for Atmospheric Research.

[5]Center for Advanced Decision Support for Water and Environmental Systems (CADSWES), University of Colorado, Boulder, CO

Corresponding author: David Woodson (david.woodson@colorado.edu)

Key Points:

- A random forest method for decadal (2 ~ 10 years) ensemble flow projections conditioned on temperature projections from climate models
- Midterm (1 ~ 5 years) ensemble flow projection of water resources management variables including reservoir pool elevations
- Encouraging prospects for translating skillful temperature projections to multi-year flow and water management at CRB and other basins

Abstract

Decadal (~10-years) scale flow projections in the Colorado River Basin (CRB) are increasingly important for water resources management and planning of its reservoir system. Physical models – Ensemble Streamflow Prediction (ESP) – do not have skill beyond interannual time scales. However, Global Climate Models have good skill in projecting decadal temperatures. This, combined with the sensitivity of CRB flows to temperature from recent studies, motivate the *research question* - can skill in decadal temperature projections be translated to operationally skillful flow projections and consequently, water resources management? To explore this, we used temperature projections from the Community Earth System Model – Decadal Prediction Large Ensemble (CESM-DPLE) along with past basin runoff efficiency as covariates in a Random Forest (RF) method to project ensembles of multi-year mean flow at the key aggregate gauge of Lees Ferry, Arizona. RF streamflow projections outperformed both ESP and climatology in a 1982-2017 hindcast, as measured by ranked probability skill score. The projections were disaggregated to monthly and sub-basin scales to drive the Colorado River Mid-term Modeling System (CRMMS) to generate ensembles of water management variables. The projections of pool elevations in Lakes Powell and Mead – the two largest U.S. reservoirs that are critical for water resources management in the basin – were found to reduce the hindcast median root mean square error by up to -20 and -30% at lead times of 48- and 60-months, respectively, relative to projections

generated from ESP. This suggests opportunities for enhancing water resources management in the CRB and potentially elsewhere.

Plain Language Summary

The Colorado River Basin (CRB) is a critical resource for tens of millions in the North American southwest but does not provide consistent streamflow year to year and is susceptible to multi-year droughts. While many good forecasts for Colorado River flow exist at seasonal time scales, projections beyond one year into the future perform little better than simply using a long-term average, despite the importance of these so-called "decadal" projections for water resource planning. As past studies have found that CRB streamflow is moderately influenced by air temperature, we seek to leverage relatively accurate climate model projections of temperature to improve CRB flow forecasts at decadal time scales using a statistical model. We find modest improvements in decadal projections of pool elevation at Lakes Powell and Mead compared to a standard method, supporting the potential for use of climate model projections in flow forecasting.

1 Introduction

The Colorado River Basin (CRB) is a heterogeneous, dynamic river basin that is a key source of water for tens of millions of people in the southwestern United States and Northwest Mexico. The CRB supports agriculture, hydropower, recreation, ecosystems, drinking water, etc. and boasts a storage capacity of ~60 million acre-feet (MAF), or approximately four years' worth of average annual flow (USBR, 2012). This storage is critical in light of the high variability in Colorado River annual flow (Figure 1) and its susceptibility to multi-year droughts (Nowak et al., 2012). Intricate legal frameworks, trans-basin diversions, growing populations, as well as over-allocation pose additional challenges for the CRB's water resource systems. There are many studies that try to provide clarity on future CRB flows using a variety of models and methods, but these approaches can exhibit significant variability and sometimes produce conflicting conclusions. Here we discuss the key issues complicating CRB flow analysis and prognostications, research gaps, and possible paths forwards to improved decadal flow projections.

Water resources management in this basin has become more challenging due to past and projected warming. Decreased precipitation is generally agreed upon as the primary driver of the ongoing drought since 2001, referred to as the "Millennium drought", mostly resulting from a shift to cooler Pacific sea surface temperatures (SSTs) related to the Pacific Decadal Oscillation (PDO) and Atlantic Multidecadal Oscillation (AMO) (Delworth et al., 2015; Martin Hoerling et al., 2010; Lehner et al., 2018; K. Nowak et al., 2012; Schubert et al., 2009; Seager et al., 2005; Seager & Ting, 2017; Zhao et al., 2017). Several recent studies suggest that this "hot drought", the worst on record with mean annual flow ~20% less than the long-term average, was due to both decreased precipitation and, in contrast to previous droughts, a mean temperature ~1°C above 20th century climatology (Udall & Overpeck, 2017; Woodhouse et al., 2016). Conversely, temperature increases in the CRB and elsewhere have been attributed to anthropogenic GHG forcings (Hoerling et al., 2019; Lukas et al., 2014), but estimates of temperature's relative influence on the recent drought vary widely. Hoerling et al. (2019) suggest that increased temperature was responsible for one-fifth of the recent flow decline, while Udall & Overpeck (2017) and Xiao et al. (2018) estimate that temperatures were responsible for approximately one-third and one-half of the flow decline, respectively. The difference in estimates of temperature influence is partially due to different

temperature sensitivities used in each study. Consider Hoerling et al.'s (2019) fully-coupled, land surface model (LSM)-derived sensitivity of -2.5% $°C^{-1}$ compared to Udall & Overpeck's (2017) sensitivity of -6.5% $°C^{-1}$, which itself was chosen because it was the mean of a range of -3% $°C^{-1}$ to -10% $°C^{-1}$ developed by Vano et al. (2012, 2014) from multi-model LSM simulations. Recently, Milly & Dunne (2020) reported a LSM model-derived sensitivity range of -7.8% $°C^{-1}$ to -12.2% $°C^{-1}$. Empirically derived sensitivities tend towards higher values, with estimates of -9% $°C^{-1}$ and -14% $°C^{-1}$ from McCabe & Wolock (2007) and Nowak et al. (2012), respectively. The true CRB temperature sensitivity, and possible changes in it, will play an important role in the degree of potential future flow declines induced by the warming trend.

Projections of mid- and end-of-century CRB flows, primarily predicated on climate model outputs, vary widely and are dependent on a variety of factors including both climate and land surface model selection, downscaling techniques, emissions scenarios, parameter estimation, resolution, etc. (Haddeland et al., 2002; Mendoza et al., 2016; Vano et al., 2014). While simulations from the Coupled Model Intercomparison Project (CMIP) all predict a warmer future in the CRB, approximately half of CMIP3 models project increases in CRB precipitation and streamflow by mid-century, while the other half project decreases, resulting in an ensemble median of approximately 0% change. Interestingly, approximately two-thirds of CMIP5 models project mid-century precipitation increases, but like CMIP3, have an ensemble median of zero for mid-century streamflow change (USBR, 2016). Many of the CMIP models exhibit a wet-bias, with approximately one-half and one-quarter of CMIP5 and CMIP3 models, respectively, unable to replicate the Millennium drought at any point (Udall & Overpeck, 2017). This supports earlier conclusions that many climate models cannot simulate decadal and multi-decadal drought cycles (Ault et al., 2012, 2013). However, some studies have projected an increase in the risk and severity of decadal and multi-decadal drought due to warming, despite any possible increases in precipitation (Ault et al., 2014, 2016; Cook et al., 2015). Further, among the CMIP3 and CMIP5 members that did replicate the Millennium drought in simulations, mean projected end-of-century flows were 85% and 91% of the 20^{th} century average observed flow at Lees Ferry, respectively. The CMIP3 and CMIP5 members that could not replicate the Millennium drought suggest mean end of century flows 109% and 113% higher, respectively, than 20^{th} century average observed flow (Udall & Overpeck, 2017). A separate analysis by Milly & Dunne (2020) forced CMIP members that had a high fidelity to observed Upper Colorado River Basin (UCRB) flows with projected mid-century changes in mean temperature and precipitation. This evaluation generated ensemble mean flow changes between -5 and -24% under Representative Concentration Pathway (RCP) 4.5 and a +3% to -40% range under RCP 8.5. This suggests that flows will either decrease or, if buffering is provided by increased precipitation, remain approximately unchanged. Flow decreases were largely attributed to decreased snow albedo resulting in higher potential evapotranspiration (PET).

The high variability of CRB flow, potential for multi-year droughts, and unknown, possibly deleterious future shifts in hydroclimate have prompted a large amount of research attention towards better understanding and predicting CRB flow behavior, with prognostic research efforts tending towards either seasonal forecasts or multidecadal projections (e.g. Baker et al., 2019, 2020, 2021; Baker, 2019; Bracken et al., 2010, 2014; Erkyihun et al., 2016, 2017; Prairie et al., 2006; Rajagopalan et al., 2019). Decadal flow projections, or forecasts with lead times greater than 1 year and less than 10 years, have received relatively less attention, despite their importance for multiyear planning, which is particularly relevant in the CRB due to a high storage capacity and susceptibility to multi-year droughts. Ensemble streamflow prediction (ESP) is a current,

commonly used method for generating seasonal and annual forecasts in the CRB and elsewhere. ESP involves initializing a hydrologic model with current conditions and forcing the model with sequences of observed past precipitation and temperature (Wood & Werner, 2011). While ESP has been shown to be significantly more skillful than climatology at seasonal time scales (Baker, 2021a; Franz et al., 2003; Harrigan et al., 2018), there is limited skill at a 12-month lead time and no skill relative to climatology in the CRB for annual forecasts with a lead time of 15-months or longer (Baker, 2017).

Unlike the uncertainty associated with climate model based projections of future precipitation and streamflow at decadal and multi-decadal scales, confidence in long-term temperature projections is higher due to hindcast skill in predicting the warming trend of the last several decades, primarily due to the strong signal imparted by GHG forcings (Hargreaves, 2010; Hausfather et al., 2020; Jiang et al., 2012; Kumar et al., 2016). Climate model-based projections have shown skill at high resolutions for seasonal time scales (Infanti & Kirtman, 2014; Ma et al., 2015; Mishra et al., 2019; Mo & Lyon, 2015; Slater et al., 2019), but generally only multi-year averages have been shown to be skillful past a 1-year lead time (Kim et al., 2012; Meehl et al., 2014; van Oldenborgh et al., 2012; Towler et al., 2018; Yeager et al., 2018). While earlier efforts at decadal prediction of streamflow have yielded modest skill (KC Nowak, 2011), decadal temperature predictions are becoming more common and are increasingly being used with non-parametric and other approaches to improve hydrologic projections at multi-year time scales (Esit et al., 2021; Kiem et al., 2021; Towler & Yates, 2021, Towler et. al, 2021 (in review)).

In watersheds with multi-year storage capacities, like the CRB, flow projections past a 12-month lead time are important for operational planning. However, climatological forecasts are often used past year-1 due to the lack of decadal skill from current forecast methods, potentially leading to sub-optimal decision making under multi-year drought scenarios. We investigate whether the skill found in climate model-based temperature projections can be transferred to decadal streamflow forecasts in the CRB. This critical need combined with the knowledge that (i) Global Climate Models have good skill in projecting decadal temperatures, and (ii) the documented sensitivity of CRB flows to temperature, motivates the present study with this *research question* - can skill in decadal temperature projections be translated to operationally skillful flow projections and consequently, water resources management? To this end, the paper is organized as follows. First, the data sets and the water resources management model are described briefly. Then, the proposed streamflow projection methodology, a Random Forest (RF) algorithm conditioned on temperature covariates is presented along with projections of basin water resources management variables. Results assessing the skill of these projections follows. The paper concludes with a summary and discussion.

2 Data and Water Resources Management Model

Naturalized flow at Lees Ferry, Arizona for 1906-2017 was obtained from the United States Bureau of Reclamation for this study (USBR, 2020). Naturalized flow is ideal when investigating hydroclimate teleconnections, as it has anthropogenic impacts like diversions and storage removed (Prairie et al., 2005). Streamflow at Lees Ferry, just downstream of Lake Powell, represents flow from the entire UCRB that on average produces ~90% of total CRB flow (Christensen et al., 2004). The UCRB yields a majority of Colorado River streamflow since its headwaters are situated in the Rocky Mountains and lies within the states of Colorado, New Mexico, Utah, and Wyoming. Conversely, the Lower Basin is in the much more arid states of Arizona, California, and Nevada

as well as portions of Mexico. The Lees Ferry streamflow gauge is used as the demarcation point between the Upper and Lower Colorado River Basins.

Observed hydroclimate data used in this study include precipitation and minimum and maximum temperatures from the Parameter-elevation Regressions on Independent Slopes (PRISM) model (Daly et al., 1994). Monthly 4-km data for the entire continental US was clipped to the UCRB boundary and spatially averaged, then either summed or averaged in time for precipitation and temperature, respectively, to generate water year values for the UCRB.

Climate model simulated data for 1981-2017 was obtained from the Community Earth System Model Decadal Prediction Large Ensemble (CESM-DPLE; Yeager et al., 2018). The CESM-DPLE is a 40-member ensemble at a monthly, 1-degree (nominally 111-km) resolution. The CESM-DPLE is a fully coupled model with atmosphere, ocean, land, and sea-ice components. The CESM-DPLE is initialized with "current" atmospheric conditions at the beginning of every 10-year projection in the 1981-2017 running hindcast. Since ensemble members are generated via perturbing initial conditions by round off errors, the ensemble spread is representative of natural variability and climate forcings only and does not include model biases (unlike other large ensembles). Calculating the anomaly correlation coefficient (ACC) indicates that the CESM-DPLE has good skill in projecting both maximum and minimum CRB temperature at multi-year time scales but has little to no skill in projecting precipitation (Figure 2). The ACC represents the correlation between forecast anomalies and observed anomalies on a spatial scale (Jolliffe & Stephenson, 2011; Miyakoda et al., 1972). An ACC of one means that the forecast and observed anomalies are perfectly correlated while a value of zero signals that there is no correlation. A negative ACC indicates that the forecast is predicting the opposite of what occurred, and an ACC of 0.6 or greater is recognized as having better skill than using a climatological forecast ("Anomaly Correlation Coefficient - Forecast User Guide - ECMWF Confluence Wiki," n.d.). The ACC for the DPLE minimum temperature is over 0.7 for all lead times. The DPLE maximum temperature projections have ACC values of 0.55, 0.46, and 0.35 for lead times of 1-5, 3-7, and 5-9 years, respectively, suggesting decreasing skill with increased lead time. Conversely, DPLE precipitation has an ACC of 0.10 for a lead time of 5-9 years and a value of zero for both other lead times, generally indicating no skill in projecting precipitation. Lead time indicates the time into the future for which projections are made, as well as the years being temporally averaged into the projection. For example, a lead time of 3-7 years represents a projection of the mean annual value between 3 and 7 years into the future.

The translation of hydrologic skill into operational skill of climate-conditioned RF flows was tested by comparing their performance to ESP when used to force the Colorado River Mid-term Modeling System (CRMMS), previously known as the Mid-term Operations Model (MTOM; USBR, 2015). CRMMS is a key modeling tool used by the United States Bureau of Reclamation ("Reclamation") for midterm planning in the CRB. It is driven by monthly streamflow at the twelve forecast locations in the UCRB (Figure 3), initial conditions of basin reservoirs, and a climatological approach for six sub-basins below Lees Ferry. CRMMS is a RiverWare model (Zagona et al., 2001) that simulates operational decisions such as releases and diversions according to the Law of the River and policies from the 2007 Interim Guidelines for shortages (USBR, 2015, 2020). Ensemble simulations are run via the Colorado River Basin Operational Prediction Testbed (Baker et al., 2021b)implemented within RiverSMART software ("RiverSMART," 2020), providing a platform for testing of many different hydrologic input scenarios and comparing their resulting operational variables (e.g. pool elevation). Reclamation provided CRMMS files and

forcing data (ESP, observed flows, and initial conditions). Several management actions related to delivery reductions (e.g., DCP and Interim Guidelines) are intimately tied to pool elevations at Lakes Mead and Powell, hence flow projections and subsequent simulations of reservoir levels are critically important for decision making.

3 Proposed Methodology

The proposed methodical framework consists of three modules: (1) Random Forest machine learning model, (2) disaggregation model, and (3) water resources decision model. In the CRB, future drought risk might be better mitigated if well-informed water resources management decisions could be reliably made on multi-year time horizons. However, operational flow forecast methods like ESP are generally only skillful for lead times up to one year, which is the same time frame that operational decisions in the CRB are limited to. Conversely, temperature projections, often from global climate model projections, have proved skillful at multi-year time scales. Figure 4 shows the scatterplots of PRISM-derived precipitation and temperatures with CRB flows for 5-year running means. Strong positive correlation with precipitation (0.8) and moderate to strong negative correlations with annual maximum (-0.74) and minimum (-0.39) temperatures can be seen, consistent with prior studies documenting temperature sensitivity to streamflow in the basin. Similar relationships and correlation strengths are seen between these hydroclimate variables with running mean lengths of 2- and 10-years. Motivated by these relationships in both the observed and model spaces, we propose our methodology to exploit these relationships and the skill in temperature projections to improve projections of streamflow and reservoir elevations.

3.1 Streamflow projection

We propose a RF machine learning approach to generate streamflow ensembles conditioned on a set of covariates.

3.1.1 Covariate Selection

Strong correlations between temperature and CRB flow were seen in the historical record in Figure 4 and Supplement Figures B1 through B4. To select the covariates, we assessed the relationship between the multi-year mean of UCRB flow and potential covariates: corresponding ensemble-mean minimum and maximum temperatures from CESM-DPLE and past multi-year basin runoff efficiencies. The temperatures are a result of large-scale ocean and atmospheric conditions, while the runoff efficiency (RE) captures the basin characteristics and antecedent conditions. We examined the relationships for "multi-year means" of 2- to 10-years. The past runoff efficiency for each multi-year period was selected based on maximum correlation within a 15-year window. For example, for 3-year mean future flow, the optimal past runoff efficiency was found to be an 8-year mean, while the optimal past runoff efficiency for 5- and 7-year mean future flow was 7- and 6-years, respectively. The scatterplots of flow and potential covariates for multi-year mean lengths of 3-, 5-, and 7-years are shown in Figure 5 as representative samples. The covariates for all other multi-year mean flow lengths are shown in Supplement Figures C1 through C6. At shorter mean lengths (e.g., 3-year means), the relationship between CESM-DPLE-simulated temperature and observed flow is inverse and relatively linear, consistent with the observed hydroclimate relationships (Figure 5). For 5-year means, simulated minimum temperature also exhibits a linear relationship with flow. While the simulated maximum temperature generally exhibits a negative correlation with flow for positive temperature anomalies, mimicking the observed relationship, flow unexpectedly decreases as negative temperature anomalies decrease. The relatively lower skill of DPLE maximum temperature projections compared to minimum temperature suggest that lower quality temperature projections might not

be able to capture well the observed trend between flow and temperature. From a physical perspective, we would expect that as temperature increases, evapotranspiration demand rises and consequently reduces flow as noted in several studies described earlier (M. Hoerling et al., 2019; Vano et al., 2012; Woodhouse et al., 2016; Xiao et al., 2018). In terms of impact, the cool temperature-low flow pairings occur in the late 1980s through early 1990s, a time when the warming signal was not as strong as in the last two decades, suggesting that this erroneous relationship should play a relatively minimal role when making forecasts forced with more recent temperature projection inputs. The relationship of the 7-year means is very similar to that of the 5-year means. We selected runoff efficiency as an effective indicator of the current basin hydrologic signature because streamflow is an interaction between temperature, precipitation, and land surface conditions that is well captured by runoff efficiency (Nowak et al., 2012).

Based on these diagnostic analyses, the covariates selected for various multi-year mean flow projections consists of three covariates – future multi-year annual mean of maximum and minimum temperatures from CESM-DPLE and past multi-year average RE. For example, for the 2-year mean flow projections from the year 1982, the feature vector consists of CESM-DPLE-projected mean maximum and minimum temperatures for 1982-1983 and 1973-1981 average runoff efficiency (past 9-year mean efficiency yielded the highest correlation with future 2-year mean flow). Conversely, for the 5-year mean flow projections in the same year, the feature vector consists of CESM-DPLE projected mean maximum and minimum temperatures for 1982-1986 and 1975-1981 average runoff efficiency (past 7-year mean efficiency yielded the highest correlation with future 5-year mean flow). Thus, for each historical year we compute this vector of covariates.

3.1.2 Random Forest

An RF machine learning approach was used to generate multi-year mean flow forecasts based on the covariates described above. Random forests, originally developed by Ho (1995) and later improved by Breiman (2001), are a Classification and Regression Tree (CART)-based supervised machine learning algorithm that can perform either classification or regression. Random forests are essentially generated through repeated bootstrapping of the training dataset, with a decision tree generated for each sample. Thus, unlike CART where a single tree is fitted to the data, here many trees (hence, forest) are fitted to bootstrapped samples of the historical data. For a covariate vector, estimates of the predictand (i.e., the dependent variable, here it is the multi-year CRB flow) are obtained from each tree in the forest and averaged. We use the estimates from all trees to make up an *ensemble*, which provides a distribution. The RF algorithm also provides *variable importance* details, or how much utility a given predictor in the training dataset imparts on prediction performance. A single value of mean square error (MSE) or node purity is produced with every forest. The MSE metric indicates how much the fitted MSE would increase if the given predictor is randomly permuted. For details of the RF algorithm, we refer the reader to (Hastie et al., 2009).

RFs are appealing for a multitude of reasons, including their predictive performance, robustness, speed, non-parametric nature, stability, diagnosis of variable importance, as well as their ability to handle non-linearity, interactions, noise, and small sample sizes in forcing data (Tyralis et al., 2019). RFs are being used in a variety of applications including water resources and water quality modeling (Suchetana et al., 2017), construction safety risk (Tixier et al., 2016), and used with success in recent years for flow forecasting, primarily at daily and monthly time scales

(Abbasi et al., 2020; Al-Juboori, 2019; Ghorbani et al., 2020; Hussain & Khan, 2020; Li et al., 2019; Liang et al., 2018; Muñoz et al., 2018; Papacharalampous & Tyralis, 2018; Pham et al., 2020).

3.1.3 RF Implementation

RF models are trained on historical, naturalized, water year flow at Lees Ferry for each multi-year block flow ranging from 2- to 10-year means ("multi-year means") along with the scaled anomalies of the covariates spanning the same time ranges. The water year flows begin in October while the climate projections begin in November; the one-month difference in start dates is likely negligible since multi-year averages are calculated and used for both datasets. All RF forecasts start on October 1, regardless of the mean length. Multi-year mean flows are generated from each of the trees in the RF, thus producing an ensemble. We implemented the RF model in a K-fold cross-validation mode, which is described below.

(i) To make a given forecast starting in year i with mean length N, a corresponding covariate vector $\boldsymbol{x_{i,N}}$ is fed into a RF model ensemble generated using the randomForest R package (Liaw & Wiener, 2002; R Core Team, 2019). Each RF projection is trained on a blind training dataset subset from the 1982-2017 historical period, $d_{i\text{-}N}$, generated through a K-fold cross validation approach, where K is the number of years dropped prior to model training. K is a function of both the selected mean length N and its past-runoff efficiency window length (re_p). So that no information from the N-year long period to be forecast is included in the training set (i.e., a "blind forecast"), only years less than the forecast year i minus N or greater than the forecast year i plus the maximum of N and re_p are included in the training set (Equation 1). This approach prevents any overlap of the training and forecast years given the nature of multi-year means. Using the maximum of N and re_p also removes any knowledge of the past annual runoff efficiency from the year to be forecast.

$$d_{i-N} = \left\{ \begin{array}{c|c} x_{1,N} & y_{1,N} \\ x_{2,N} & y_{2,N} \\ x_{3,N} & y_{3,N} \\ \vdots & \vdots \\ x_{j-N,N} & y_{j-N,N} \\ x_{j+\max(N,re_p),N} & y_{j+\max(N,re_p),N} \\ \vdots & \vdots \\ x_{l-1,N} & y_{l-1,N} \\ x_{l,N} & y_{l,N} \end{array} \right\} \quad (1)$$

Where,

$N = multi-year\ mean\ length\ (years)$
$d_{i-N} = blind\ training\ dataset\ for\ forecast\ year\ i\ and\ mean\ length\ N$
$x_{j,N} = predictors\ for\ year\ j\ of\ training\ dataset\ and\ mean\ length\ N$
$y_{j,N} = predictand\ for\ year\ j\ of\ training\ dataset\ and\ mean\ length\ N$
$re_p = past\ runoff\ efficiency\ mean\ window\ length\ (years)$
$l = number\ of\ years\ in\ the\ period\ of\ record$

For example, given a desired forecast of 2000-2004 mean flow (i = 2000 and N = 5, re_p = 7) the RF model is trained on running 5-year means spanning 1982-1995 and 2007-2013, where years indicate the beginning of the 5-year mean period (e.g., 2013 is the 2013-2017 5-year mean). For a desired forecast of 2010-2011 mean flow (i = 2010, N = 2, re_p = 9), the RF model is trained on running 2-year means spanning 1982-2008 only (e.g., 2008 is the 2008-2009 2-year mean) since $i + re_p = 2010 + 9 = 2019$ which is beyond the maximum year (2017) in the period of record used. The predictands include multi-year mean flows of a selected mean length (e.g., 5-year mean flow) and the predictors include CESM-DPLE projected mean minimum and maximum temperature of the same mean length, and past runoff efficiency of mean length determined through a correlation analysis that varies for each mean length.

(ii) The RF algorithm was set to generate 300 trees for each combination of forecast year and mean length. This number of trees was selected because while there were only marginal performance improvements past 100-200 trees, use of 300 trees allows for a robust and well-dispersed forecast with very reasonable computational expense. To generate a projection for a given forecast year and mean length, the relevant feature vector was fed into each tree in the forest resulting in an ensemble of 300 multi-year mean flow projections (Equation 2). Traditionally, the average of all trees' predictions is used as the random forest output in regression contexts, however our approach was to use individual trees' predictions to create an ensemble. Default values from the randomForest R package (Liaw & Wiener, 2002) for the remaining model parameters were used as tuning lead to only marginal differences in projection performance. The two most important model parameters include the number of trees, *nTree*, and the number of covariates randomly sampled for each split in a tree, *mTry*. *mTry* defaults to the number of covariates divided by three. In this study three covariates are used, resulting in a value of one for *mTry*.

$$y_{i-N} = \left\{ \begin{array}{c|c} x_{i,N} & tree_{1,N} \\ x_{i,N} & tree_{2,N} \\ x_{i,N} & tree_{3,N} \\ \vdots & \vdots \\ x_{i,N} & tree_{nTree-1,N} \\ x_{i,N} & tree_{nTree,N} \end{array} \right\} \quad (2)$$

Where,
$x_{i-N} = predictors\ for\ forecast\ year\ i\ and\ mean\ length\ N$
$y_{i-N} = predictands\ for\ forecast\ year\ i\ and\ mean\ length\ N$
$tree_{i,N} = i-th\ decision\ tree\ generated\ from\ blind\ training\ set$
$nTree = number\ of\ trees\ in\ forest$

(iii) The RF algorithm generates some tree outcomes that match exactly the input predictands (i.e., streamflow) and other tree outcomes that deviate slightly from the predictands due to introduced randomization. Consequently, to generate annual flow sequences from the RF-generated multi-year means, if a projected multi-year mean was not an exact match with an observed multi-year mean flow from the blind training dataset, then the closest observed value within the blind training dataset was used for

disaggregation. Disaggregation of multi-year means to annual flow sequences simply involved finding the sequence of annual flows that constituted a selected multi-year mean (Equations 3 and 4). One potential shortcoming of this approach is that it can only produce annual flow volumes sampled from the observed record, although use of paleo-reconstructed flows could lengthen the period of record.

$$\bar{q}_N = \frac{q_1+q_2+\cdots+q_{N-1}+q_N}{N} \quad (3)$$

$$\vec{q}_{N-annual} = \{q_1, q_2, \ldots, q_{N-1}, q_N\} \quad (4)$$

Where,
$\bar{q}_N = multi-year\ mean\ flow\ with\ mean\ length, N$
$q_i = annual\ flow\ from\ year\ i$
$\vec{q}_{N-annual} = sequence\ of\ annual\ flows\ constituting\ multi-year\ mean$

3.2 Space-time streamflow projection on the network via disaggregation

While there is value in skillful projections of annual and multi-year mean flow at midterm time scales, for it to be of impact to stakeholders, the skill in flow projections needs to be translated into reservoir pool elevations. As mentioned, to evaluate the utility of our novel climate conditioned decadal streamflow projections within an operational context, we use Reclamation's CRMMS, which runs on a monthly time scale. As such, we needed to disaggregate the RF-simulated annual flows, $\vec{q}_{N-annual}$, at Lees Ferry, the aggregate gauge, to a monthly time step at all 12 upstream sub-basins.

The annual flows from the RF were disaggregated to monthly flows at all the locations (i.e., sub-basins) of CRB that are input to CRMMS. For this space-time disaggregation we used the non-parametric disaggregation method developed by Nowak et al. (2010). In this, a K-nearest neighbor (KNN) method is applied to each annual flow in the multi-year block to select a historical year and its space-time proportion of the annual flows at Lees Ferry aggregate gauge. The selected proportion vector is multiplied with the annual flow to obtain flows for the 12 months and 12 sub-basins. For details of this simple and robust method and its application, we refer the readers to Nowak et al. (2010). This is repeated for all the annual flow values.

3.3 Projections of operationally critical reservoir elevations

Of the many variables CRMMS projects, we focus on simulated pool elevation at Lakes Mead and Powell, since various pool elevation thresholds in the two largest CRB reservoirs play a crucial role in determining water resources management throughout the basin. For example, when the August-projected end of year water level of Lake Mead drops below 1075 ft a Level 1 Shortage Condition is declared for the Lower Basin and deliveries are reduced. A Lake Powell related threshold exists at elevation 3700 ft, when the reservoir releases extra water to distribute high flows (i.e. the Equalization tier) (Baker, 2019; USBR, 2020). These policies then percolate to regional and local water management strategies across the basin.

CRMMS simulations were run for 32 hindcasts spanning 1982 to 2017 using the disaggregated RF projections from the 3-, 5-, and 7-year means as hydrologic flow forcings. The 3-, 5-, and 7-year means were used as CRMMS inputs to test the shortest, mid-length, and longest multi-year mean projections. The 10-year means only produced 27-hindcasts due to the longer mean length making 2008-2017 the last available full block of historical data. Each hindcast is a

3-, 5-, or 7-year-long block run at a monthly resolution. An ESP hindcast dataset trained on observed 1981-2010 precipitation and temperature sequences and originally produced by the National Weather Service Colorado River Basin Forecasting Center (CBRFC) was used as a "baseline" hydrologic forcing in CRMMS simulations covering the 1982-2017 hindcast period. For each CRMMS run of a retrospective forecast, the ESP trace that covered the same time period was dropped prior to CRMMS input (i.e., the ESP trace that started in the same month and year as the CRMMS simulation to be run was dropped). This made the number of ESP traces either 29 or 30 for forecasts made before or after 2010, respectively. ESP was used as a baseline model since it is used by Reclamation and other stakeholders in the CRB for seasonal- to annual- scale planning and risk assessment as well as for multi-year exploratory studies. Finally, a CRMMS simulation forced with historical flows in each sub-basin was run for each projection block to be used as a "truth" for calculating the projection skill of reservoir elevations. The historical simulation is not the true observed pool elevation since operational rules have changed over the course of the hindcast period but represents what would have been the observed pool elevation if the current Law of the River policies, including the Interim Guidelines, had been in place for the given historical unregulated flows. Since ESP and the historical simulation are at most 60 months long, the 7-year flow projection simulations were subset to a 5-year length for plotting and skill analysis. Similarly, the ESP and historical simulation were subset to a 36-month length when analyzed with the 3-year mean flow projections.

3.4 Validation of Models

We validated both the ensemble projections of flows and the monthly reservoir pool elevations, for the hindcast period of 1982 – 2017. Each framework was examined with separate performance metrics. For each forecast year, i, and mean length, N, the flow projections were made in a 'blind' K-fold cross validation model, where $K = N + maximum(N, re_p) - 1$, such that the feature vector of that year and any preceding or following running means with overlapping knowledge of the year to be forecast were not included in the RF training dataset. In other words, this avoids the prospect of selecting the flow of the N-years immediately prior to or following year i in training the projection model. The years selected in these blind training datasets were also used to disaggregate from multi-year mean projections to annual flows. A variety of performance metrics were calculated on both the multi-year mean flow projections and on the subsequent CRMMS simulation results. Applications of metrics such as the ranked probability skill score (RPSS), root mean square error (RMSE), continuous ranked probability skill score (CRPSS), Nash-Sutcliffe Efficiency (NSE), and reliability are further derived in the Supplement (Section A).

4 Results

Results from the flow and water resources decision variables are described below. For the multi-year mean projections of Lees Ferry naturalized streamflow, the 3-, 5-, and 7-year means are examined in Section 4.1 due to their skill, which generally outperforms both climatology and ESP, as well as their reliability, which is comparable with ESP's reliability. For the simulated reservoir pool elevations, disaggregated results from the 3-, 5-, and 7-year mean flows are presented in Section 4.2.

4.1 Streamflow projections

The blind K-fold cross validation ensemble projections of 3-, 5-, and 7-year mean flows for each year for the 1982-2017 period from RF are shown as boxplots along with the historically

observed multi-year mean flows (Figure 6). The flow projections can capture multi-year variability quite well for some mean lengths given the limited duration of training data. Hindcast projection results for the remaining multi-year means are shown in the Supplement (Figure D1).

Aggregating individual MSE values for each predictor variable from every year in the 1982-2017 hindcast shows that past runoff efficiency is the most important variable, followed by CESM-DPLE-simulated maximum and minimum temperature as can be seen for the 5-year projection in Figure 7. The runoff efficiency captures the combined impact of temperature with the basin hydrology and physical characteristics. This pattern is consistent for mean lengths of 2- through 9-years (Supplement Figures E1 through E7). The variable importance diagnosis was not performed on 10-year means due to a training dataset with too few years for adequate analysis after making the data 'blind' for 10-year means (i.e., dropping 19 out of 36 years in the 1982-2017 training record).

The RPSS was calculated by tercile thresholds for each year in the hindcast period. These tercile thresholds represent the 33rd and 66th percentiles but vary depending on the mean length selected. RPSS values greater or less than zero indicate better or worse performance, respectively, than a climatological ensemble forecast, with a score of one indicating a perfect forecast. For long-lead forecasts, any positive RPSS is considered satisfactory, while RPSS values of 0.5 are considered good and higher values considered excellent. The RF projections show a positive median skill score for all mean lengths except for the 8- and 10-year mean flows, the former of which is just below zero (Figure 8). The temporal pattern of the 10-year mean flow projections are fairly well captured, however the magnitudes were not. This we believe is due to the relatively long mean length necessitated dropping 19-years of data when training the RF in order to generate a blind forecast and thus yielded a constrained and under-dispersed sample space, hence the lower skill. Furthermore, the RF median RPSS mostly outperformed ESP median RPSS where applicable, with improvements in median RPSS ranging from 0.39 to 0.58 (Figure 8 and Table 1). ESP had a median RPSS of 0 for 2-year mean flow projections and a negative median RPSS for all other mean lengths, echoing prior studies that show little- to no- skill in ESP past a 1-year lead time.

Table 1. Median RPSS from 1982-2017 hindcast for RF and ESP multi-year mean flow projections

Mean Length (year)	Median RPSS		Δ Improvement (RF minus ESP)
	RF	ESP	
2	-0.05	0.00	-0.05
3	0.41	-0.15	0.56
4	0.24	-0.15	0.39
5	0.30	-0.28	0.58
6	0.33	-	-
7	0.21	-	-
8	-0.03	-	-
9	0.15	-	-
10	-0.70	-	-

Time series of RPSS for the 3-, 5-, and 7-year means show that RF outperforms ESP in most years of the hindcast, particularly when projecting 3-year mean flows during the beginning years of the Millennium drought where ESP has little to no skill (Figure 9). RPSS time series of the remaining multi-year means are shown in the Supplement (Figure D2). Reliability diagrams (Figure 10) show the relationship between the forecasted probability of an event and its observed frequency. A perfectly reliable forecast falls along the 1:1 line and in some cases (e.g., 5-year means) the RF is not only closer to the 1:1 line than ESP but also has a flatter rank histogram. A flat rank histogram indicates that, over the course of the hindcast, the observed flows being forecast have an equal probability of occurring in any rank of the ensemble projection. In other words, the forecasts and observations are random samples from the same probability distribution. Rank histograms that are skewed to high or low ranks suggest forecast quality issues such as bias or under-dispersion (Hamill, 2001). Reliability scores (Figure D3; Candille et al., 2007; Frenette, 2019; Hersbach, 2000) are calculated through a CRPS decomposition; a perfectly reliable forecast is represented by a reliability score of zero and higher values indicate lower reliability. Where comparable, ESP has consistently better reliability scores than RF; although RF reliability begins to improve for mean lengths of 6-, 7-, and 8-years potentially due to reduced variability induced by multi-year averaging. RF reliability worsens for the 9- and 10-year means, likely due to the smaller sample size available for training these forecasts.

4.2 Reservoir pool elevations

As mentioned, ensemble projections of future 3-, 5-, and 7-year annual, naturalized Lees Ferry flow from the RF method were disaggregated in space and time via the KNN approach described in the previous section. These space-time flows were then used to drive the CRMMS and compared with CRMMS simulations forced with ESP over the 1982-2017 hindcast period of running 3- or 5-year blocks. Since ESP's duration is 5-years, analysis of CRMMS results was stopped at the 60-month mark for the 7-year flow projections. Figures 12 and 13 show the hindcasts for Lakes Powell and Mead pool elevations, respectively, generated from the 5-year mean flow projections for the 1999 – 2003 period, which covers the worst years of the Millennium drought. At a 60-month lead time, the historical reservoir level at Lake Powell is not captured by either forecast method ensemble, although the RF median and quartiles skew more towards the historical pool elevation compared to ESP (Figure 11a). At Lake Mead, both forecast methods capture the historical reservoir levels at all lead times, but the RF distribution again tends towards lower elevations relative to ESP (Figure 11b). Both forecast methods have similar sharpness and spread. Similar results are seen in other hindcast blocks in the 1982-2017 period as well as from the other multi-year mean lengths (Sections G.1 through G.3 in Supplement).

For each of the 32-different hindcasts, an 'ensemble RMSE' was calculated on the end of water year (EOWY) simulated pool elevations at Lakes Powell and Mead and are shown as boxplots in Figure 12. For a detailed description of the ensemble RMSE calculation, see Supplement Section A. It can be seen from the boxplots that, while ESP performs similarly to, or better than, RF at shorter lead times ($\leq$ 24 months), RF generally outperforms ESP at lead times of 36 months or longer. This is also corroborated in the median RMSE at all lead times (Table 2a and 2b). RF forecasts at longer lead times generated reductions in hindcast median RMSE relative to ESP ranging from -7% to -30% at Mead and -3% to -20% at Powell. Generally, these results indicate translation of skill in flow projections informed by temperature to improvements in water resources management variables.

Table 2a. Median RMSE from CRMMS-simulated 1982-2017 hindcast at Lake Powell for both ESP and RF

Powell Median EOWY RMSE (ft)									
Lead Time (number of months)	**3-year mean (n = 32)**			**5-year mean (n = 32)**			**7-year mean (n = 30)**		
	ESP	**RF**	**% Change**	**ESP**	**RF**	**% Change**	**ESP**	**RF**	**% Change**
12	24.27	23.99	-1%	24.27	25.74	6%	23.75	25.45	7%
24	34.79	33.15	-4.7%	34.79	39.20	13%	34.40	36.52	6%
36	45.76	40.40	-11.7%	45.76	44.51	-3%	44.29	48.03	8%
48	-	-	-	50.38	40.54	-20%	47.85	46.76	-2%
60	-	-	-	49.62	43.71	-12%	49.33	44.09	-11%

Table 2b. As in Table 2a. but for Lake Mead

Mead Median EOWY RMSE (ft)									
Lead Time (number of months)	**3-year mean (n = 32)**			**5-year mean (n = 32)**			**7-year mean (n = 30)**		
	ESP	**RF**	**% Change**	**ESP**	**RF**	**% Change**	**ESP**	**RF**	**% Change**
12	7.88	9.53	21%	7.88	12.98	65%	7.72	13.96	81%
24	17.41	20.49	17.7%	17.41	23.93	37%	17.16	25.41	48%
36	33.06	26.51	-19.8%	33.06	30.76	-7%	30.11	36.69	22%
48	-	-	-	44.13	33.52	-24%	42.30	35.22	-17%
60	-	-	-	52.79	36.72	-30%	50.86	43.35	-15%

CRPSS and NSE were also applied to the CRMMS simulations and showed similar results as the ensemble RMSE. For more information, see Supplement Section H. Overall, these results suggest that ESP could remain the preferred forecast method during the first 12-months of a midterm forecast but improvements in skill during years 3-5 could be gained through RF or similar methods.

5 Summary and Discussion

A simple decadal flow projection method was developed based on RF machine learning trained on CESM-DPLE simulations of UCRB temperature, past RE, and observed UCRB flow. Motivation for the model came from moderate correlation and non-linearities between future multi-year mean flow and future multi-year mean CESM-DPLE simulated temperature, as well as between past runoff efficiency and future flow. RF projections of 3- through 7-year mean flow were able to broadly capture multi-year variability and outperformed climatology over half of the time. Based on hindcast median RPSS, RF outperformed ESP for multi-year mean projections of 3-years through 5-years. The greatest improvement over ESP occurred for 3-year means, where the RF and ESP hindcast median RPSS was 0.41 and -0.15, respectively, a 0.56 difference. In projecting 3-year means during the lowest flow years of the Millennium Drought (1999-2005 and 2012-2015), RF had RPSS values at or above ~0.5, whereas ESP RPSS values were at or below 0. Interestingly, in the only other low flow period in the hindcast (1987-1992) during which model temperatures projections were anomalously cool, RF performed worse than ESP in projecting 3-year means. This potentially indicates the importance of both the warming trend and quality of projected temperature covariates to model performance in predicting low flow years. Otherwise,

RF tends to outperform both ESP and climatology in low flow years for mean lengths of up to 5-years. Both methods perform similarly for high flow years and generally show improvements over climatology.

When used as a hydrologic flow forcing in an operational simulation model of the CRB, RF projections generally outperformed ESP in predicting pool elevations at Lakes Powell and Mead for lead times of 36 months or greater. The improvements of RF over ESP are noteworthy when using projections disaggregated from either 3- or 5-year means. The hindcast median ensemble RMSE of RF was, at best, between 12% and 30% lower than that of ESP for lead times between 36- and 60-months: equating improvements of approximately 5- to 16-feet. One caveat is that the RF algorithm cannot predict flows that occur outside of the historical envelope used to train the model. Although, any future projections trained with the RF method would include the extremes observed in the 1980s and early 2000s and, at least in theory, going forward there are not likely to be volumes far outside of the period of record. The training dataset is limited by the CESM-DPLE temporal domain, which only begins in 1982. Utilization of paleo-reconstructed flow data and overlapping climate model simulations like the CESM - Last Millennium Ensemble (Otto-Bliesner et al., 2015) would allow for a wider range of flow extrema but would also likely introduce greater projection uncertainty compared to usage of naturalized 20th and 21st century flows only.

We suggest that modest improvements upon current forecast methods can be gained in projecting reservoir elevations at midterm time scales using CESM-DPLE simulated temperature data along with RF machine learning models. Advancements in midterm flow projection techniques are critically important for multiyear planning of reservoir management. As shown during the Millennium drought, rapid drawdowns can occur with sustained low flows and a lack of midterm shortage planning. The warming trend, associated uncertainty with possible reductions in streamflow and increases in drought risk, as well as continued population growth all support the necessity of smart, efficient, and informed midterm planning that emphasizes conversation when facing multi-year droughts; temperature conditioned flow forecasts can help meet this need.

Future work might investigate methods to improve skill of the RF approach and the covariates used. Of particular importance is improving the skill for precipitation and pre-1980 temperature simulations, upon which the performance of any derivative flow projection is heavily dependent. Future studies should investigate any new iterations of CESM or other climate model-based projections to leverage potential advances in hydroclimate projection skill. While we used a correlation-informed heuristic analysis to determine model covariates, more objective techniques like Generalized Cross Validation (GCV) might be used during covariate selection, particularly if the parameter space is significantly enlarged with additional climate models. Additionally, the model training dataset might be further expanded through inclusion of paleo-reconstructed flow and overlapping simulated hydroclimate data. Though we used future 2- through 10-year block lengths, similar approaches could be applied to multi-decadal time scales for longer term planning. Finally, incorporation of other machine learning methods (e.g., neural networks) as well as wavelet decomposition of temperature, precipitation, and flow data may improve simulated flow projections by exploiting hydroclimate linkage patterns and regime-like flow behavior.

Acknowledgments, Samples, and Data

- David Woodson and Dr. Balaji Rajagopalan would like to thank the USBR (Science & Technology grant #19264) and CU-Boulder CEAE for research funding.
- Dr. Erin Towler would like to thank the Colorado River Climate and Hydrology Work Group and associated funding from the Central Arizona Water Conservation District, Denver Water, California's Six Agency Committee, and Southern Nevada Water Authority. Dr. Towler's work is 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.
- 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.
- The authors declare no financial or affiliation-related conflicts of interests.
- Code and data can be found on GitHub or in a Google Drive repository.

Figures

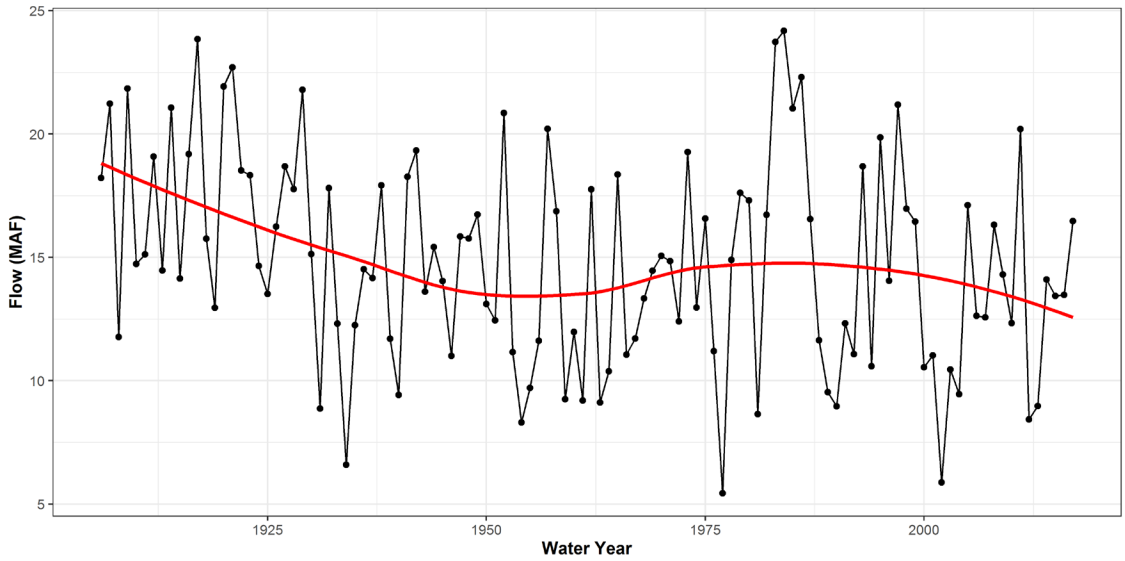

Figure 1. Annual naturalized Colorado River flow (million acre-feet) at Lees Ferry, Arizona. Red line shows a loess fit.

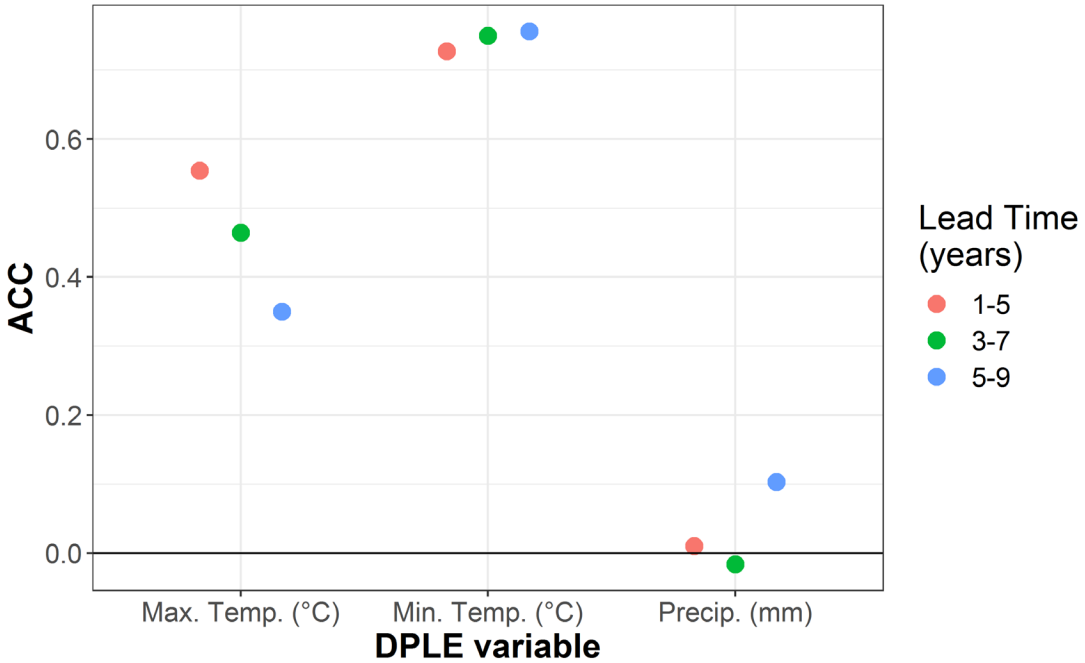

Figure 2. Anomaly correlation skill score (ACC) of CESM-DPLE projections of CRB multiyear mean temperature (maximum, minimum) and precipitation during a 30-year hindcast from 1981-2011. "Lead Time" indicates both the lead time and the specific years used in multi-year averaging.

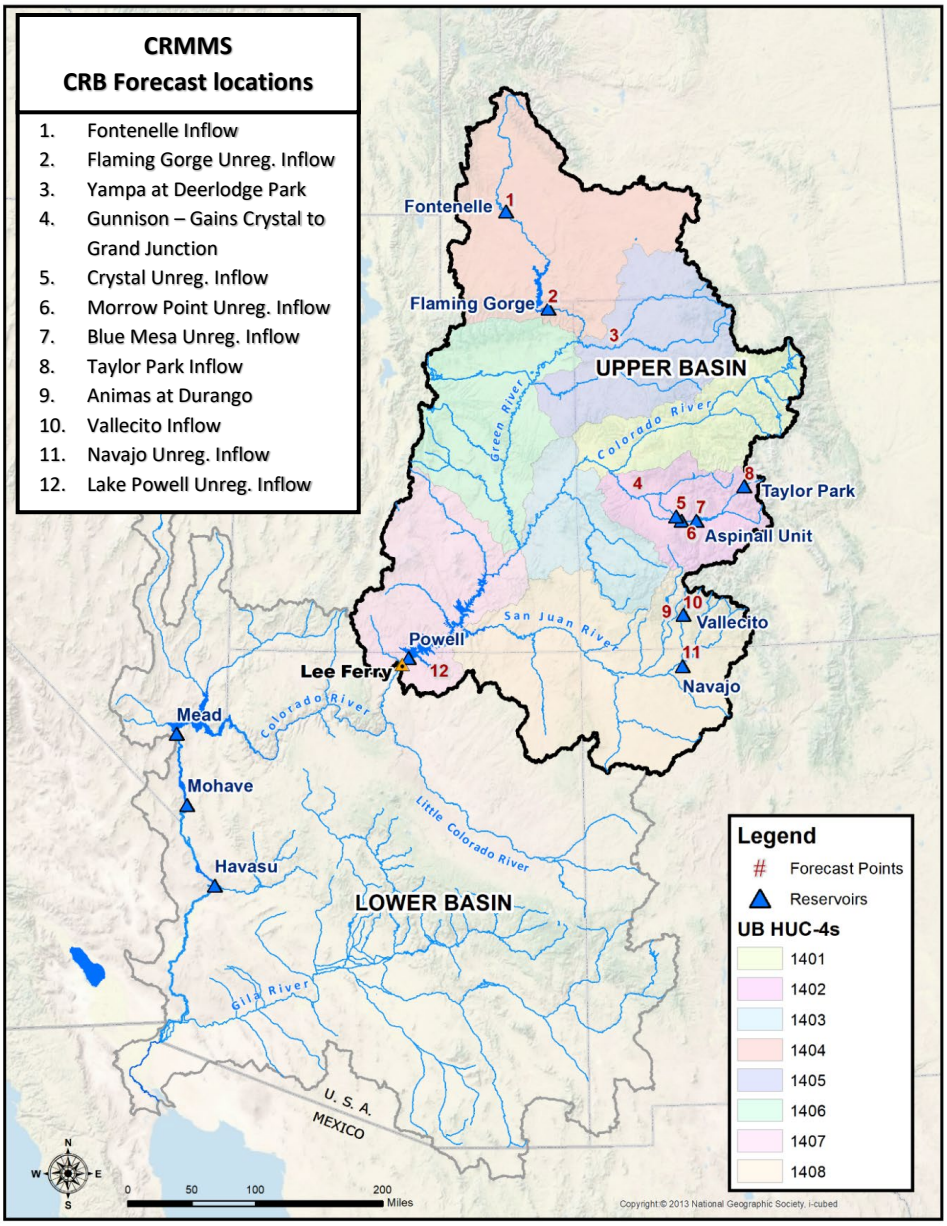

Figure 3. CRB subbasins, forecast locations, and reservoirs simulated by CRMMS. A prediction of monthly flow over the next 5-years is made for 12 UCRB forecast locations and used as CRMMS hydrologic input. Climatological approaches are used in the Lower Basin due to relatively lower flow contributions.

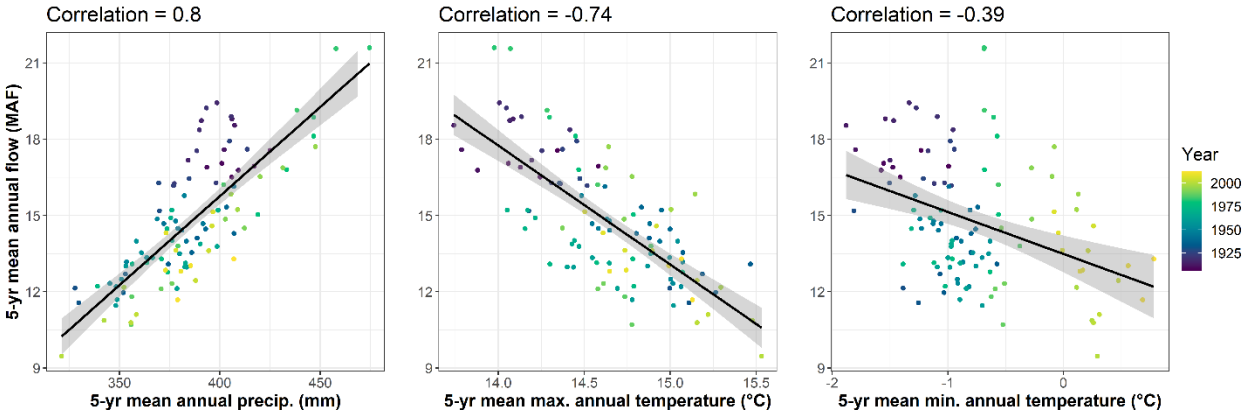

Figure 4. Relationship between naturalized Lees Ferry flow and PRISM derived precipitation, temperature in the UCRB for 5-year running means. Color indicates year in which observation occurred. Correlation was calculated for flow and each meteorological variable. Trend and 95% confidence from simple linear regression.

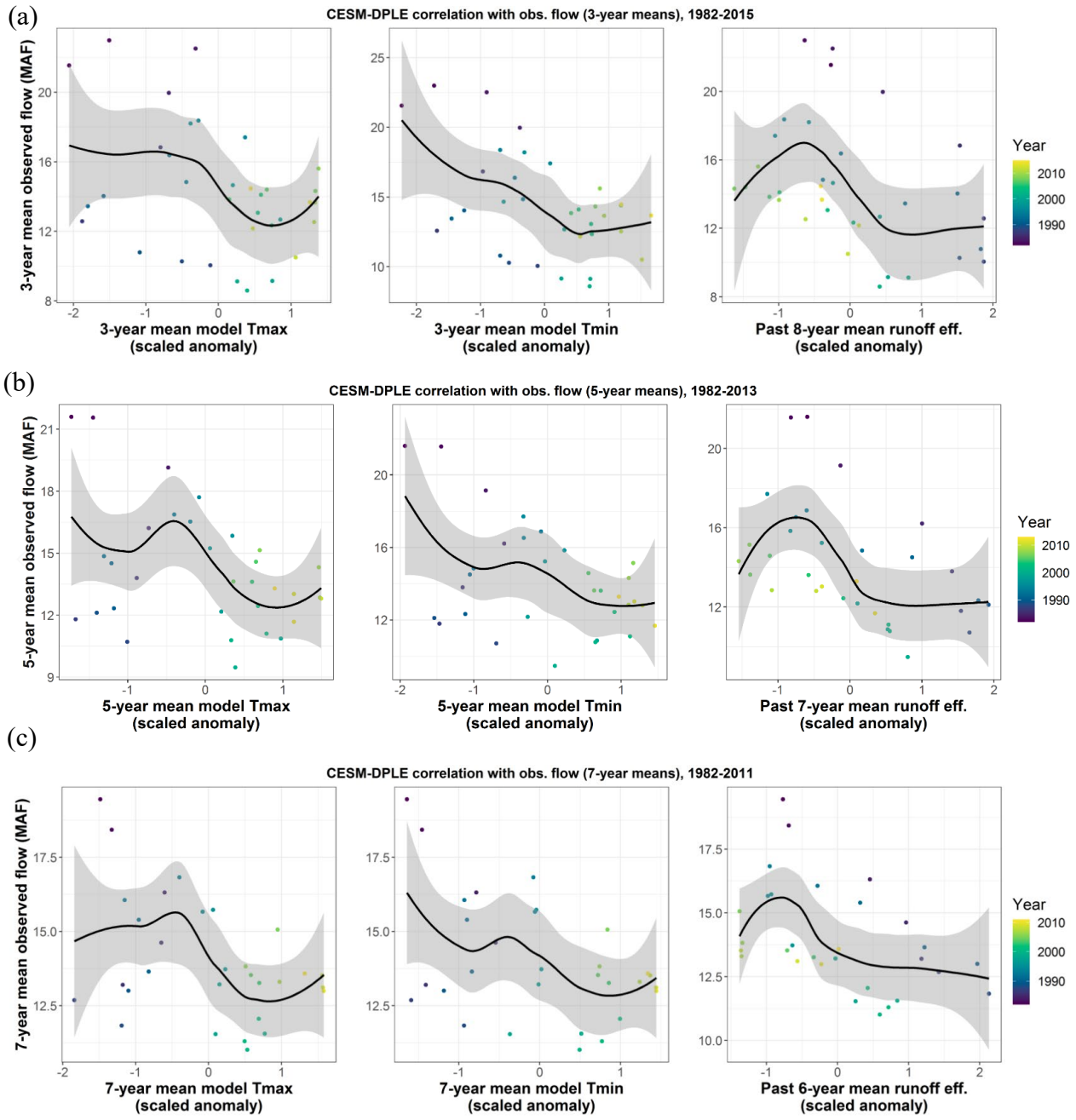

Figure 5. Relationship between future N-year mean flow and covariates. Upper (a), middle (b), and lower (c) plots show 3-, 5-, and 7-year means, respectively. Past runoff efficiency mean length was determined separately for each mean length based on a correlation analysis. Trend and 95% confidence from loess fit.

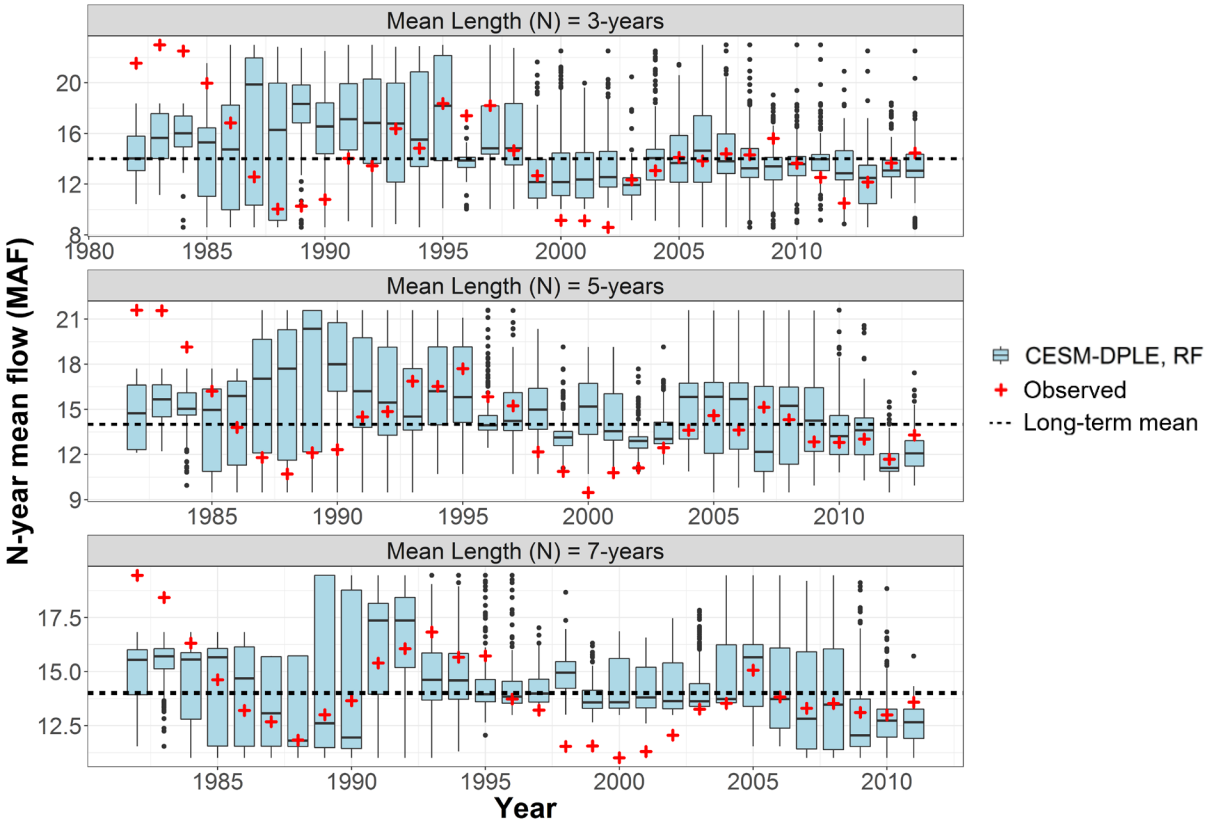

Figure 6. Running multi-year mean projections of naturalized Lees Ferry flow from the random forest method trained on blind, K-fold cross validation. Multi-year means include 3-, 5-, 7-years and are indicated above each graphic. Forecast ensembles are represented via blue boxplots, with the box representing the 25th-, 50th-, and 75th- percentiles, and the upper and lower whiskers representing 1.5 times the interquartile range (IQR) plus or minus the upper and lower quartiles, respectively. Black points above or below the boxplot whiskers are considered outliers based on the 1.5 times IQR threshold. Observed multi-year mean flows are indicated by red crosses. The years on the x-axis represent the starting year of each multi-year mean (e.g., 1982 for a mean length of 3-years represents 1982-1984 mean). Horizontal dashed line is the 1981-2017 average multi-year mean flow.

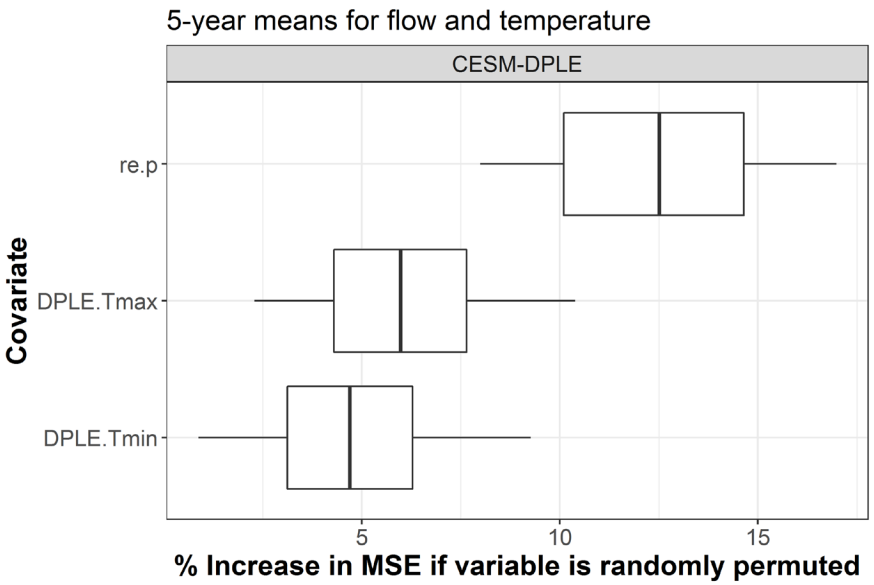

Figure 7. Variable importance plot for random forests generating 5-year mean flow projections for running 1982-2017 hindcasts. "re.p", "DPLE.Tmax", and "DPLE.Tmin", indicate past runoff efficiency, DPLE maximum temperature, and DPLE minimum temperature, respectively. Percent increase in MSE indicates how much the mean square error would increase if the select variable is randomly changed during the training period.

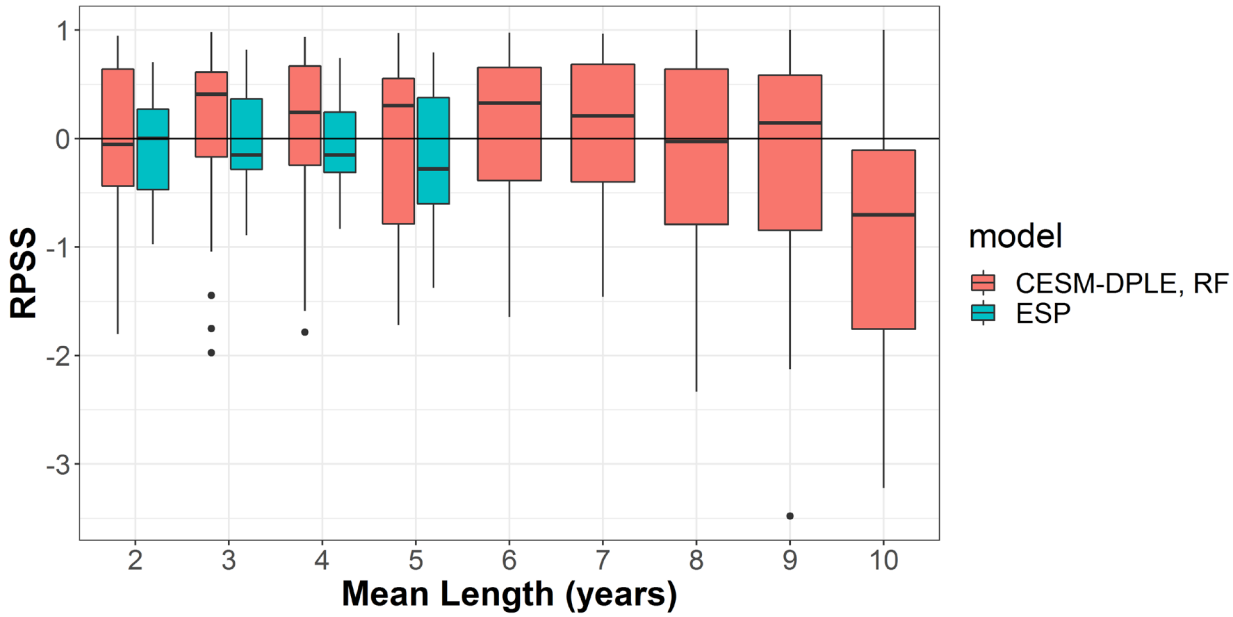

Figure 8. RPSS of RF- and ESP- projected multi-year mean flows for 1982-2017 from blind, K-fold cross validation. Skill shown for all years in hindcast with the number of RPSS values in hindcast varying from 35 to 27 depending on mean length. RPSS calculated against a 1981-2017 climatological ensemble of blind observed multi-year mean flows dependent on mean length selected. ESP's maximum length is 5-years.

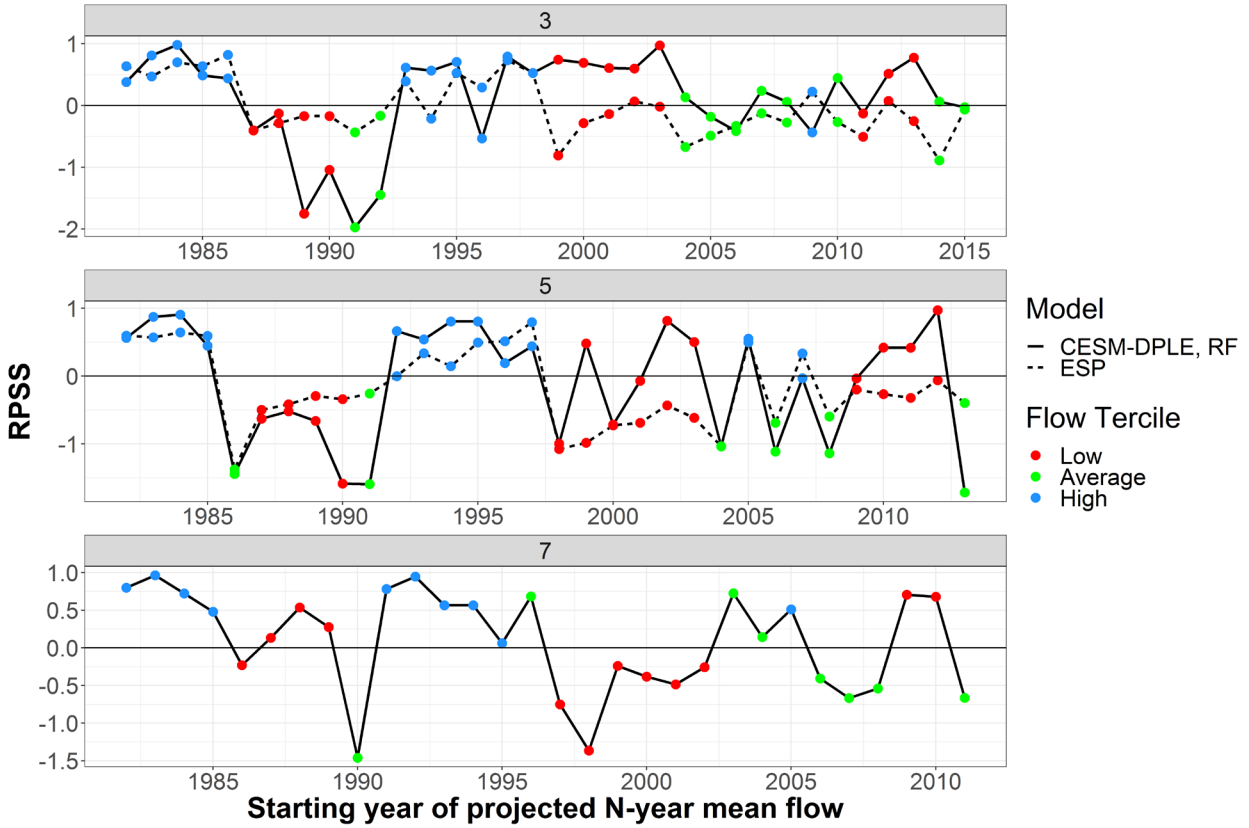

Figure 9. RPSS time series of RF- and ESP- projected multi-year mean flows for 1982-2017 from blind, K-fold cross validation. Points indicate RPSS for individual years in hindcast for the 3-, 5-, and 7-year mean flow projections. Point colors represent the observed flow tercile of each year (e.g., high-, low-, and average- flow years). The solid and dotted lines show skill time series for RF and ESP, respectively. RPSS is calculated against a 1981-2017 climatological ensemble of blind observed multi-year mean flows dependent on mean length selected. ESP's maximum length is 5-years.

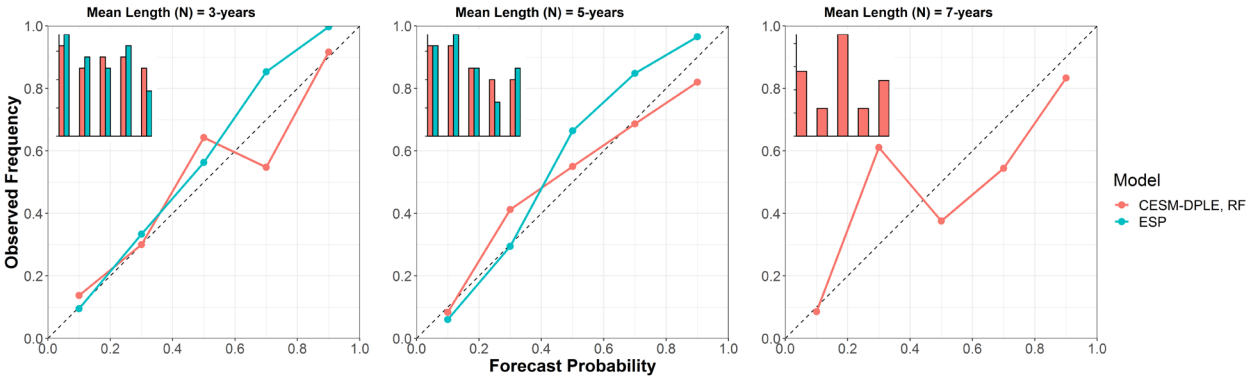

Figure 10. Reliability diagram with rank histogram for 3-, 5-, and 7-year mean flow projections during 1982-2017 hindcast for RF and ESP methods. Points that fall along the 1:1 line are considered perfectly reliable.

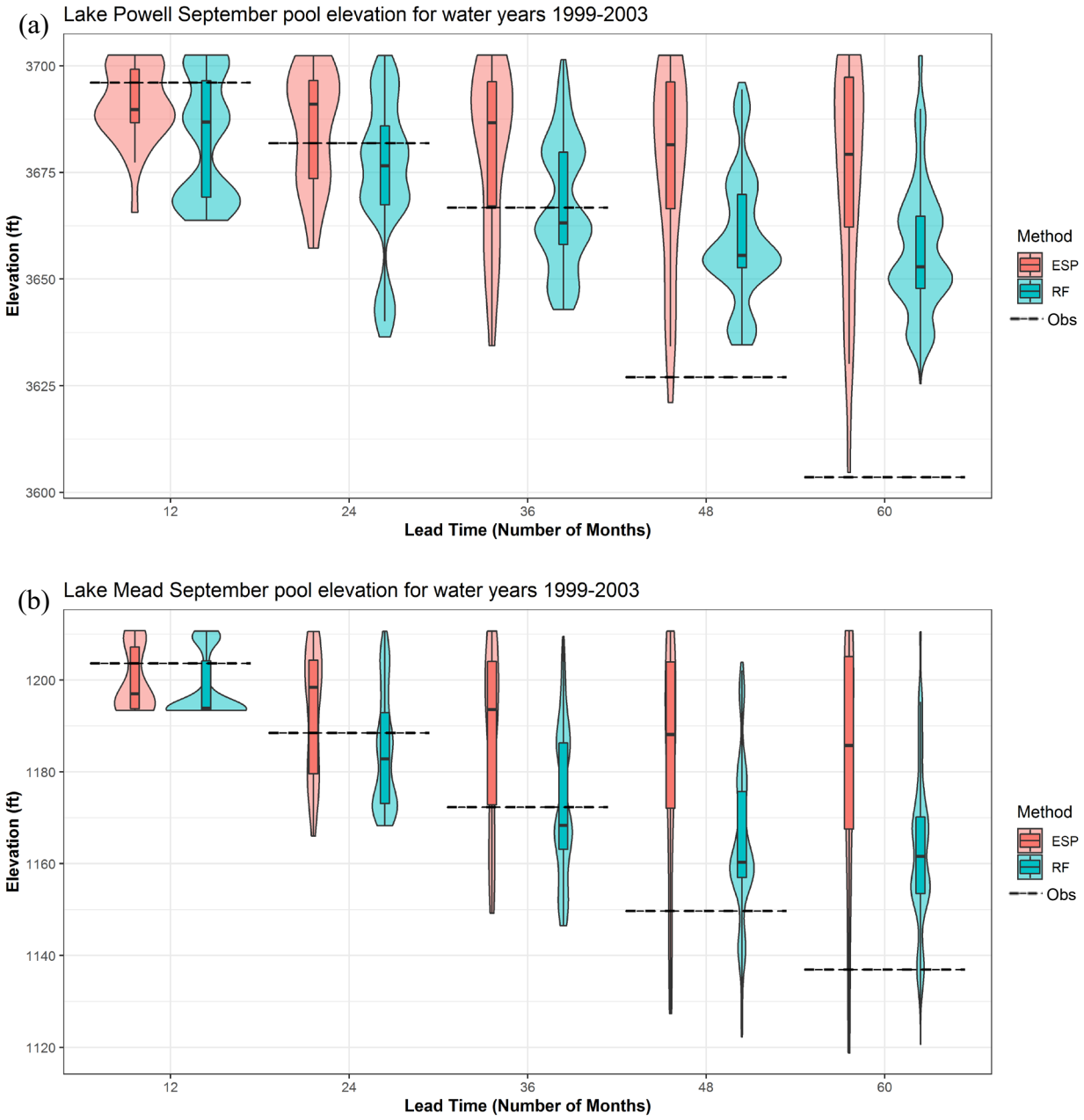

Figure 11. CRMMS-simulated EOWY pool elevation at Lake Powell (a) and Lake Mead (b) from ESP and 5-year mean RF flow forecasts as well as a historical simulation. Hindcast covers water years 2000-2004 (October 1999 – September 2004). X-axis values delineate lead times for each EOWY. Historical EOWY pool elevations indicated by horizontal black dashed lines. Violin plots show measures of central tendency as well as probability density.

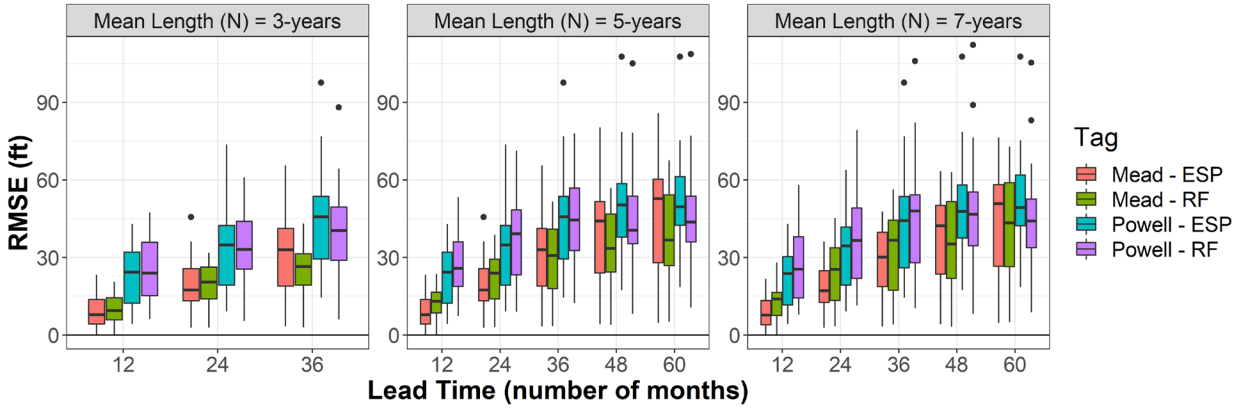

Figure 12. EOWY ensemble RMSE calculated on CRMMS-simulated pool elevation at Lakes Powell and Mead for running hindcasts (1982-2017) disaggregated from 3-, 5-, and 7-year mean flow projections as well as ESP. Historic simulations forced by historical unregulated flows were considered the observed value in error calculations.

References

Abbasi, M., Farokhnia, A., Bahreinimotlagh, M., & Roozbahani, R. (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. *Journal of Hydrology*, 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 Resources Management*, *33*(9), 3257–3270. https://doi.org/10.1007/s11269-019-02299-4

Anomaly Correlation Coefficient - Forecast User Guide - ECMWF Confluence Wiki. (n.d.). Retrieved November 7, 2021, from https://confluence.ecmwf.int/display/FUG/Anomaly+Correlation+Coefficient

Ault, T. R., Cole, J. E., & St. George, S. (2012). The amplitude of decadal to multidecadal variability in precipitation simulated by state-of-the-art climate models: CMIP5 DECADAL PRECIPITATION VARIABILITY. *Geophysical Research Letters*, *39*(21), n/a-n/a. https://doi.org/10.1029/2012GL053424

Ault, T. R., Cole, J. E., Overpeck, J. T., Pederson, G. T., St. George, S., Otto-Bliesner, B., et al. (2013). The Continuum of Hydroclimate Variability in Western North America during the Last Millennium. *Journal of Climate*, *26*(16), 5863–5878. https://doi.org/10.1175/JCLI-D-11-00732.1

Ault, T. R., Cole, J. E., Overpeck, J. T., Pederson, G. T., & Meko, D. M. (2014). Assessing the Risk of Persistent Drought Using Climate Model Simulations and Paleoclimate Data. *Journal of Climate*, *27*(20), 7529–7549. https://doi.org/10.1175/JCLI-D-12-00282.1

Ault, T. R., Mankin, J. S., Cook, B. I., & Smerdon, J. E. (2016). Relative impacts of mitigation, temperature, and precipitation on 21st-century megadrought risk in the American Southwest. *Science Advances*, *2*(10), e1600873. https://doi.org/10.1126/sciadv.1600873

Baker, S. A. (2017, May). *A New Framework to Evaluate the Skill of Different Hydrologic Forecasts used in the 24-MS and MTOM*. Presented at the Proceedings of the 2017 Colorado River Hydrology Research Symposium, Las Vegas, Nevada.

Baker, S. A., Wood, A. W., & Rajagopalan, B. (2019). Developing Subseasonal to Seasonal Climate Forecast Products for Hydrology and Water Management. *JAWRA Journal of the American Water Resources Association*, *55*(4), 1024–1037. https://doi.org/10.1111/1752-1688.12746

Baker, S. A., Wood, A. W., & Rajagopalan, B. (2020). Application of Postprocessing to Watershed-Scale Subseasonal Climate Forecasts over the Contiguous United States. *Journal of Hydrometeorology*, *21*(5), 971–987. https://doi.org/10.1175/JHM-D-19-0155.1

Baker, S. A., Wood, A. W., & Rajagopalan, B. (2021a, accepted). Enhancing Ensemble Seasonal Streamflow Forecasts in The Upper Colorado River Basin Using Multi-Model Climate Forecasts. *Journal of the American Water Resources Association*.

Baker, S. A., Wood, A. W., Rajagopalan, B., Jerla, C., Zagona, E., Butler, R., & Smith, R. (2021b). The Colorado River Basin Operational Prediction Testbed: A Framework for Evaluating Streamflow Forecasts and Reservoir Operations. *JAWRA Journal of the American Water Resources Association*.

Baker, S. A. (2019, January 1). *Development of Sub-seasonal to Seasonal Watershed-Scale Hydroclimate Forecast Techniques to Support Water Management*. ProQuest Dissertations Publishing.

Bracken, C., Rajagopalan, B., & Prairie, J. (2010). A multisite seasonal ensemble streamflow forecasting technique: MULTISITE SEASONAL ENSEMBLE STREAMFLOW. *Water Resources Research*, *46*(3). https://doi.org/10.1029/2009WR007965

Bracken, C., Rajagopalan, B., & Zagona, E. (2014). A hidden Markov model combined with climate indices for multidecadal streamflow simulation. *Water Resources Research*, *50*(10), 7836–7846. https://doi.org/10.1002/2014WR015567

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Candille, G., Côté, C., Houtekamer, P. L., & Pellerin, G. (2007). Verification of an Ensemble Prediction System against Observations. *Monthly Weather Review*, *135*(7), 2688–2699. https://doi.org/10.1175/MWR3414.1

Christensen, N. S., Wood, A. W., Voisin, N., Lettenmaier, D. P., & Palmer, R. N. (2004). The Effects of Climate Change on the Hydrology and Water Resources of the Colorado River Basin. *Climatic Change*, *62*(1–3), 337–363. https://doi.org/10.1023/B:CLIM.0000013684.13621.1f

Cook, B. I., Ault, T. R., & Smerdon, J. E. (2015). Unprecedented 21st century drought risk in the American Southwest and Central Plains. *Science Advances*, *1*(1), e1400082. https://doi.org/10.1126/sciadv.1400082

Daly, C., Neilson, R., & Phillips, D. (1994). A Statistical-Topographic Model for Mapping Climatological Precipitation Over Mountain Terrain. *J. Appl. Meteor.*, *33*(2), 140–158. https://doi.org/10.1175/1520-0450(1994)033<0140:ASTMFM>2.0.CO;2

Delworth, T. L., Zeng, F., Rosati, A., Vecchi, G. A., & Wittenberg, A. T. (2015). A Link between the Hiatus in Global Warming and North American Drought. *Journal of Climate*, *28*(9), 3834–3845. https://doi.org/10.1175/JCLI-D-14-00616.1

Erkyihun, S. T., Rajagopalan, B., Zagona, E., Lall, U., & Nowak, K. (2016). Wavelet-based time series bootstrap model for multidecadal streamflow simulation using climate indicators: WAVELET-BASED TIME SERIES BOOTSTRAP. *Water Resources Research*, *52*(5), 4061–4077. https://doi.org/10.1002/2016WR018696

Erkyihun, S. T., Zagona, E., & Rajagopalan, B. (2017). Wavelet and Hidden Markov-Based Stochastic Simulation Methods Comparison on Colorado River Streamflow. *Journal of Hydrologic Engineering*, *22*(9), 04017033. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001538

Esit, M., Kumar, S., Pandey, A., Lawrence, D. M., Rangwala, I., & Yeager, S. (2021). Seasonal to multi-year soil moisture drought forecasting. *Npj Climate and Atmospheric Science*, *4*(1), 16. https://doi.org/10.1038/s41612-021-00172-z

Franz, K., Hartmann, H., Sorooshian, S., & Bales, R. (2003). Verification of National Weather Service Ensemble Streamflow Predictions for Water Supply Forecasting in the Colorado River Basin. *Journal of Hydrometeorology*, *4*(6), 1105–1118.

Frenette, R. (2019). *crpsDecompostion: Decompostion of Continuous Ranked Probability Score in verification: Weather Forecast Verification Utilities*. Retrieved from https://rdrr.io/cran/verification/man/crpsDecompostion.html

Ghorbani, M. A., Deo, R. C., Kim, S., Hasanpour Kashani, M., Karimi, V., & Izadkhah, M. (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 Computing*, *24*(16), 12079–12090. https://doi.org/10.1007/s00500-019-04648-2

Haddeland, I., Matheussen, B. V., & Lettenmaier, D. P. (2002). Influence of spatial resolution on simulated streamflow in a macroscale hydrologic model: INFLUENCE OF SPATIAL RESOLUTION. *Water Resources Research*, *38*(7), 29-1-29–10. https://doi.org/10.1029/2001WR000854

Hamill, T. M. (2001). Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Monthly Weather Review*, *129*(3), 550–560. https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2

Hargreaves, J. C. (2010). Skill and uncertainty in climate models: Skill and uncertainty in climate models. *Wiley Interdisciplinary Reviews: Climate Change*, *1*(4), 556–564. https://doi.org/10.1002/wcc.58

Harrigan, S., Prudhomme, C., Parry, S., Smith, K., & Tanguy, M. (2018). Benchmarking ensemble streamflow prediction skill in the UK. *Hydrology and Earth System Sciences*, *22*(3), 2023–2039. https://doi.org/10.5194/hess-22-2023-2018

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Random Forests. In T. Hastie, R. Tibshirani, & J. Friedman, *The Elements of Statistical Learning* (pp. 587–604). New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-84858-7_15

Hausfather, Z., Drake, H. F., Abbott, T., & Schmidt, G. A. (2020). Evaluating the Performance of Past Climate Model Projections. *Geophysical Research Letters*, *47*(1). https://doi.org/10.1029/2019GL085378

Hersbach, H. (2000). Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather and Forecasting*, *15*(5), 559–570. https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition* (Vol. 1, pp. 278–282 vol.1). https://doi.org/10.1109/ICDAR.1995.598994

Hoerling, M., Barsugli, J., Livneh, B., Eischeid, J., Quan, X., & Badger, A. (2019). Causes for the Century-Long Decline in Colorado River Flow. *Journal of Climate*, *32*(23), 8181–8203. https://doi.org/10.1175/JCLI-D-19-0207.1

Hoerling, Martin, Eischeid, J., & Perlwitz, J. (2010). Regional Precipitation Trends: Distinguishing Natural Variability from Anthropogenic Forcing. *Journal of Climate*, *23*(8), 2131–2145. https://doi.org/10.1175/2009JCLI3420.1

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

Infanti, J. M., & Kirtman, B. P. (2014). Southeastern U.S. Rainfall Prediction in the North American Multi-Model Ensemble. *Journal of Hydrometeorology*, *15*(2), 529–550. https://doi.org/10.1175/JHM-D-13-072.1

Jiang, Z., Song, J., Li, L., Chen, W., Wang, Z., & Wang, J. (2012). Extreme climate events in China: IPCC-AR4 model evaluation and projection. *Climatic Change*, *110*(1–2), 385–401. https://doi.org/10.1007/s10584-011-0090-0

Jolliffe, I. T., & Stephenson, D. B. (2011). *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. New York, UNITED KINGDOM: John Wiley & Sons, Incorporated. Retrieved from http://ebookcentral.proquest.com/lib/ucb/detail.action?docID=849911

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., 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. *Bulletin of the American Meteorological Society*, *96*(8), 1333–1349. https://doi.org/10.1175/BAMS-D-13-00255.1

Kiem, A. S., Kuczera, G., Kozarovski, P., Zhang, L., & Willgoose, G. (2021). Stochastic Generation of Future Hydroclimate Using Temperature as a Climate Change Covariate. *Water Resources Research*, *57*(2). https://doi.org/10.1029/2020WR027331

Kim, H.-M., Webster, P. J., & Curry, J. A. (2012). Evaluation of short-term climate change prediction in multi-model CMIP5 decadal hindcasts: MULTI-MODEL CMIP5 DECADAL PREDICTIONS. *Geophysical Research Letters*, *39*(10), n/a-n/a. https://doi.org/10.1029/2012GL051644

Kumar, S., Kinter, J. L., Pan, Z., & Sheffield, J. (2016). Twentieth century temperature trends in CMIP3, CMIP5, and CESM-LE climate simulations: Spatial-temporal uncertainties, differences, and their potential sources:

THE 20TH CENTURY TEMPERATURE TRENDS. *Journal of Geophysical Research: Atmospheres*, *121*(16), 9561–9575. https://doi.org/10.1002/2015JD024382

Lehner, F., Deser, C., Simpson, I. R., & Terray, L. (2018). Attributing the U.S. Southwest's Recent Shift Into Drier Conditions. *Geophysical Research Letters*, *45*(12), 6251–6261. https://doi.org/10.1029/2018GL078312

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

Liang, Z., Tang, T., Li, B., Liu, T., Wang, J., & Hu, Y. (2018). Long-term streamflow forecasting using SWAT through the integration of the random forests precipitation generator: case study of Danjiangkou Reservoir. *Hydrology Research*, *49*(5), 1513–1527. https://doi.org/10.2166/nh.2017.085

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22.

Lukas, J., Barsugli, J., Rangwala, I., & Wolter, K. (2014). Climate Change in Colorado A Synthesis to Support Water Resources Management and Adaptation. Retrieved from https://wwa.colorado.edu/climate/co2014report/Climate_Change_CO_Report_2014_FINAL.pdf

Ma, F., Yuan, X., & Ye, A. (2015). Seasonal drought predictability and forecast skill over China. *Journal of Geophysical Research: Atmospheres*, *120*(16), 8264–8275. https://doi.org/10.1002/2015JD023185

McCabe, G. J., & Wolock, D. M. (2007). Warming may create substantial water supply shortages in the Colorado River basin. *Geophysical Research Letters*, *34*(22), L22708. https://doi.org/10.1029/2007GL031764

Meehl, G. A., Goddard, L., Boer, G., Burgman, R., Branstator, G., Cassou, C., et al. (2014). Decadal Climate Prediction: An Update from the Trenches. *Bulletin of the American Meteorological Society*, *95*(2), 243–267. https://doi.org/10.1175/BAMS-D-12-00241.1

Mendoza, P. A., Clark, M. P., Mizukami, N., Gutmann, E. D., Arnold, J. R., Brekke, L. D., & Rajagopalan, B. (2016). How do hydrologic modeling decisions affect the portrayal of climate change impacts?: Subjective Hydrologic Modelling Decisions in Climate Change Impacts. *Hydrological Processes*, *30*(7), 1071–1095. https://doi.org/10.1002/hyp.10684

Milly, P. C. D., & Dunne, K. A. (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

Mishra, N., Prodhomme, C., & Guemas, V. (2019). Multi-model skill assessment of seasonal temperature and precipitation forecasts over Europe. *Climate Dynamics*, *52*(7–8), 4207–4225. https://doi.org/10.1007/s00382-018-4404-z

Miyakoda, K., Hembree, G. D., Strickler, R. F., & Shulman, I. (1972). Cumulative Results of Extended Forecast Experiments I. Model Performance for Winter Cases. *Monthly Weather Review*, *100*(12), 836–855. https://doi.org/10.1175/1520-0493(1972)100<0836:CROEFE>2.3.CO;2

Mo, K. C., & Lyon, B. (2015). Global Meteorological Drought Prediction Using the North American Multi-Model Ensemble. *Journal of Hydrometeorology*, *16*(3), 1409–1424. https://doi.org/10.1175/JHM-D-14-0192.1

Muñoz, P., Orellana-Alvear, J., Willems, P., & Célleri, R. (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

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

Nowak, K., Hoerling, M., Rajagopalan, B., & Zagona, E. (2012). Colorado River Basin Hydroclimatic Variability. *Journal of Climate*, *25*(12), 4389–4403. https://doi.org/10.1175/JCLI-D-11-00406.1

Nowak, K. C. (2011). *Stochastic Streamflow Simulation at Interdecadal Times Scales and Implications for Water Resources Management in the Colorado River Basin* (Civil Engineering Graduate Theses & Dissertations). University of Colorado.

van Oldenborgh, G. J., Doblas-Reyes, F. J., Wouters, B., & Hazeleger, W. (2012). Decadal prediction skill in a multi-model ensemble. *Climate Dynamics*, *38*(7–8), 1263–1280. https://doi.org/10.1007/s00382-012-1313-4

Otto-Bliesner, B. L., Brady, E. C., Fasullo, J., Jahn, A., Landrum, L., Stevenson, S., et al. (2015). Climate Variability and Change since 850 CE: An Ensemble Approach with the Community Earth System Model. *Bulletin of the American Meteorological Society*, *97*(5), 735–754. https://doi.org/10.1175/BAMS-D-14-00233.1

Papacharalampous, G. A., & Tyralis, H. (2018). Evaluation of random forests and Prophet for daily streamflow forecasting. *Advances in Geosciences*, *45*, 201–208. https://doi.org/10.5194/adgeo-45-201-2018

Pham, L. T., Luo, L., & Finley, A. O. (2020). *Evaluation of Random Forest for short-term daily streamflow forecast in rainfall and snowmelt driven watersheds* (preprint). Catchment hydrology/Modelling approaches. https://doi.org/10.5194/hess-2020-305

Prairie, J. R., Rajagopalan, B., Fulp, T. J., & Zagona, E. A. (2005). Statistical Nonparametric Model for Natural Salt Estimation. *Journal of Environmental Engineering*, *131*(1), 130–138. https://doi.org/10.1061/(ASCE)0733-9372(2005)131:1(130)

Prairie, J. R., Rajagopalan, B., Fulp, T. J., & Zagona, E. A. (2006). Modified K-NN Model for Stochastic Streamflow Simulation. *Journal of Hydrologic Engineering*, *11*(4), 371–378. https://doi.org/10.1061/(ASCE)1084-0699(2006)11:4(371)

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rajagopalan, B., Erkyihun, S. T., Lall, U., Zagona, E., & Nowak, K. (2019). A Nonlinear Dynamical Systems-Based Modeling Approach for Stochastic Simulation of Streamflow and Understanding Predictability. *Water Resources Research*, *55*(7), 6268–6284. https://doi.org/10.1029/2018WR023650

RiverSMART. (2020). Retrieved from https://riverware.org/RiverSmart/RiverSmartSoftwareSuiteHelp.html#RiverSMART_Help

Schubert, S., Gutzler, D., Wang, H., Dai, A., Delworth, T., Deser, C., et al. (2009). A U.S. CLIVAR Project to Assess and Compare the Responses of Global Climate Models to Drought-Related SST Forcing Patterns: Overview and Results. *Journal of Climate*, *22*(19), 5251–5272. https://doi.org/10.1175/2009JCLI3060.1

Seager, R., & Ting, M. (2017). Decadal Drought Variability Over North America: Mechanisms and Predictability. *Current Climate Change Reports*, *3*(2), 141–149. https://doi.org/10.1007/s40641-017-0062-1

Seager, R., Kushnir, Y., Herweijer, C., Naik, N., & Velez, J. (2005). Modeling of Tropical Forcing of Persistent Droughts and Pluvials over Western North America: 1856–2000*. *Journal of Climate*, *18*(19), 4065–4088. https://doi.org/10.1175/JCLI3522.1

Slater, L. J., Villarini, G., & Bradley, A. A. (2019). Evaluation of the skill of North-American Multi-Model Ensemble (NMME) Global Climate Models in predicting average and extreme precipitation and temperature over the continental USA. *Climate Dynamics*, *53*(12), 7381–7396. https://doi.org/10.1007/s00382-016-3286-1

Towler, E., & Yates, D. (2021). Incorporating Multiyear Temperature Predictions for Water Resources Planning. *Journal of Applied Meteorology and Climatology*, *60*(2), 171–183. https://doi.org/10.1175/JAMC-D-20-0134.1

Towler, E., PaiMazumder, D., & Done, J. (2018). Toward the Application of Decadal Climate Predictions. *Journal of Applied Meteorology and Climatology*, *57*(3), 555–568. https://doi.org/10.1175/JAMC-D-17-0113.1

Tyralis, H., Papacharalampous, G., & Langousis, A. (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., & Overpeck, J. (2017). The twenty-first century Colorado River hot drought and implications for the future: COLORADO RIVER FLOW LOSS. *Water Resources Research*, *53*(3), 2404–2418. https://doi.org/10.1002/2016WR019638

USBR. (2012). Colorado River Basin Water Supply and Demand Study. United States Bureau of Reclamation. Retrieved from https://www.usbr.gov/watersmart/bsp/docs/finalreport/ColoradoRiver/CRBS_Executive_Summary_FINAL.pdf

USBR. (2015, May). Colorado River Basin Mid-Term Probabilistic Operations Model (MTOM) Overview and Description. U.S. Department of the Interior: Bureau of Reclamation.

USBR. (2016, March). SECURE Water Act Section 9503(c) – Reclamation Climate Change and Water. U.S. Department of the Interior Bureau of Reclamation. Retrieved from https://www.usbr.gov/climate/secure/

USBR. (2020, March 31). Colorado River Interim Guidelines for Lower Basin Shortages and Coordinated Operations for Lake Powell and Lake Mead. Retrieved April 14, 2020, from https://www.usbr.gov/lc/region/programs/strategies.html#IGReview

Vano, J. A., Das, T., & Lettenmaier, D. P. (2012). Hydrologic Sensitivities of Colorado River Runoff to Changes in Precipitation and Temperature*. *Journal of Hydrometeorology*, *13*(3), 932–949. https://doi.org/10.1175/JHM-D-11-069.1

Vano, J. A., Udall, B., Cayan, D. R., Overpeck, J. T., Brekke, L. D., Das, T., et al. (2014). Understanding Uncertainties in Future Colorado River Streamflow. *Bulletin of the American Meteorological Society*, *95*(1), 59–78. https://doi.org/10.1175/BAMS-D-12-00228.1

Wood, A., & Werner, K. (2011). Development of a seasonal climate and streamflow forecasting testbed for the Colorado River Basin. In *Science and Technology Infusion Climate Bulletin*. National Weather Service.

Woodhouse, C. A., Pederson, G. T., Morino, K., McAfee, S. A., & McCabe, G. J. (2016). Increasing influence of air temperature on upper Colorado River streamflow: TEMPERATURE AND COLORADO STREAMFLOW. *Geophysical Research Letters*, *43*(5), 2174–2181. https://doi.org/10.1002/2015GL067613

Xiao, M., Udall, B., & Lettenmaier, D. P. (2018). On the Causes of Declining Colorado River Streamflows. *Water Resources Research*, *54*(9), 6739–6756. https://doi.org/10.1029/2018WR023153

Yeager, S. G., Danabasoglu, G., Rosenbloom, N. A., Strand, W., Bates, S. C., Meehl, G. A., 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. *Bulletin of the American Meteorological Society*, *99*(9), 1867–1886. https://doi.org/10.1175/BAMS-D-17-0098.1

Zagona, E. A., Fulp, T. J., Shane, R., Magee, T., & Goranflo, H. M. (2001). Riverware: A Generalized Tool for Complex Reservoir System Modeling1. *JAWRA Journal of the American Water Resources Association*, *37*(4), 913–929. https://doi.org/10.1111/j.1752-1688.2001.tb05522.x

Zhao, S., Deng, Y., & Black, R. X. (2017). Observed and Simulated Spring and Summer Dryness in the United States: The Impact of the Pacific Sea Surface Temperature and Beyond. *Journal of Geophysical Research: Atmospheres*, *122*(23). https://doi.org/10.1002/2017JD027279