JOURNAL OF THE AMERICAN WATER RESOURCES ASSOCIATION

Vol. 57, No. 6 AMERICAN WATER RESOURCES ASSOCIATION December 2021

# Enhancing Ensemble Seasonal Streamflow Forecasts in the Upper Colorado River Basin Using Multi-Model Climate Forecasts

*Sarah A. Baker, Balaji Rajagopalan, and Andrew W. Wood*

**Research Impact Statement**: Conditioning ensemble streamflow forecasts with subseasonal-to-seasonal climate forecast information can improve the skill of seasonal streamflow predictions in the Upper Colorado River Basin.

ABSTRACT: In the Colorado River Basin (CRB), ensemble streamflow prediction (ESP) forecasts drive operational planning models that project future reservoir system conditions. CRB operational seasonal streamflow forecasts are produced using ESP, which represents climate using an ensemble of meteorological sequences of historical temperature and precipitation, but do not typically leverage additional real-time subseasonal-to-seasonal climate forecasts. Any improvements to streamflow forecasts would help stakeholders who depend on operational projections for decision making. We explore incorporating climate forecasts into ESP through variations on an ESP trace weighting approach, focusing on Colorado River unregulated inflows forecasts to Lake Powell. The k-nearest neighbors (kNN) technique is employed using North American Multi-Model Ensemble one- and three-month temperature and precipitation forecasts, and preceding three-month historical streamflow, as weighting factors. The benefit of disaggregated climate forecast information is assessed through the comparison of two kNN weighting strategies; a basin-wide kNN uses the same ESP weights over the entire basin, and a disaggregated-basin kNN applies ESP weights separately to four subbasins. We find in general that climate-informed forecasts add greater marginal skill in late winter and early spring, and that more spatially granular disaggregated-basin use of climate forecasts slightly improves skill over the basin-wide method at most lead times.

(KEYWORDS: streamflow forecasts; trace weighting; subseasonal-to-seasonal; climate forecasts; Colorado River; NMME; reservoir inflows.)

## INTRODUCTION

Many operational streamflow forecasts in the United States (U.S.) are provided by the National Weather Service (NWS) River Forecasting Centers and National Resource Conservation Service (NRCS) (Pagano et al. 2014; Lukas and Payton 2020). In the Colorado River Basin (CRB), the Colorado River Basin River Forecasting Center (CBRFC) produces streamflow forecasts using the ensemble streamflow prediction (ESP) method. ESP is a widely used operational forecasting method with streamflow forecasts produced using a land-surface or watershed model initialized with current basin conditions and forced into the forecast period with an ensemble of historical temperature and precipitation sequences, termed "traces" (Day 1985; Franz et al. 2003; Lukas and

Paper No. JAWR-20-0142-P of the *Journal of the American Water Resources Association* (JAWR). Received September 29, 2020; accepted October 13, 2021. © 2021 American Water Resources Association. **Discussions are open until six months from issue publication**.

Lower Colorado Region (Baker), U.S. Bureau of Reclamation Boulder, Colorado, USA; Department of Civil, Environmental, and Architectural Engineering (Rajagopalan), University of Colorado Boulder Boulder, Colorado, USA; and Climate and Global Dynamics (Wood), National Center for Atmospheric Research Boulder, Colorado, USA (Correspondence to Baker: sabaker@usbr.gov).

*Citation*: Baker, S.A., B. Rajagopalan, and A.W. Wood. 2021. "Enhancing Ensemble Seasonal Streamflow Forecasts in the Upper Colorado River Basin Using Multi-Model Climate Forecasts." *Journal of the American Water Resources Association* 57 (6): 906–922. https://doi.org/10.1111/1752-1688.12960.

Payton 2020). The initial days of the forecast period are typically informed by weather forecasts (e.g., five days of precipitation forecasts in the CRB, and out to 15 days in some other regions), but subseasonal-to-seasonal (S2S; Week 3 and longer) climate forecast information is not typically incorporated in NWS operations or in NRCS statistical forecasts.

The Bureau of Reclamation (Reclamation) uses ESP forecasts provided by the CBRFC in operational planning models to provide stakeholders with risk-based information regarding future basin conditions. The Colorado River Mid-term Modeling System (CRMMS; previously called the Mid-Term Operations Model) is one of these operational planning models which uses ESP to project monthly reservoir operations and basin conditions for storage and streamflow out to five-year lead times (Daugherty 2013). The results from CRMMS can be used for planning and risk assessment of potential shortage or surplus basin conditions that affect many communities and economies throughout the Southwestern U.S., which relies heavily on CRB water resources (Bureau of Reclamation 2015). Improving the skill of streamflow forecasts used in CRMMS would benefit stakeholders who use projections of future basin conditions in decision making and planning.

Over the last few decades, numerous variations to the traditional ESP approach have been proposed to improve its skill and reliability. The primary avenues for improving seasonal hydrologic prediction methods include improving the ability to describe initial watershed moisture conditions (in the soil and snowpack if present) and improving meteorological inputs during the forecast period (Wood and Lettenmaier 2008; Li et al. 2009; Arnal et al. 2017). In general, forecast skill is highly dependent on the quality of the initial conditions at shorter lead times, but at progressively longer leads becomes more dependent on climate forcings (Li et al. 2009; Wood et al. 2016). Improvements in initial conditions can be achieved through enhanced watershed-scale observations, data assimilation, and hydrology model upgrades, the latter of which also benefit the forecasts through a more physically realistic evolution of model states and fluxes (e.g., runoff) through the forecast period. Upgraded meteorological forecast inputs can come through incorporating weather and climate predictions, whether from statistical or dynamical (or hybrid) sources (Lukas and Payton 2020). This paper focuses on the climate-related opportunity for improving ESP.

Certain methods for enhancing meteorological inputs to ESP may be termed preadjustment (Werner et al. 2004) or pre-ESP, in contrast to strategies that impose climate or other information to the outputs of an ESP forecast (Mendoza et al. 2017). One pre-ESP approach is to generate weather inputs for ESP through a conditional resampling of historical weather sequences based on the state of climate indices — for example, the El Niño Southern Oscillation (ENSO) index, the Southern Oscillation Index, the North Atlantic Oscillation index, the Pacific Decadal Oscillation index, and others (Grantz et al. 2005; Bracken and Rajagopalan 2014; Beckers et al. 2016; Erkyihun and Zagona 2017). Increasingly, the advancing skill of dynamical climate models argues for using climate forecasts to drive pre-ESP methods. Several model generations ago, moderate benefits were achievable mainly in strong ENSO years (e.g., Wood and Kumar 2005; Wood and Lettenmaier 2006). More recently, Mo and Lettenmaier (2014) applied climate forecasts from the North American Multimodel Ensemble (NMME) over the contiguous U.S., finding that skill gains in one- to three-month lead runoff forecasts were seasonally and regionally dependent, and arose mainly after a one-month lead, before which initial conditions dominated the forecast skill. In the Upper CRB, for instance, NMME improved ESP runoff in April at a two- and three-month lead time. A recent special collection of papers by Wetterhall et al. (2016) contains a number of examples that broadly illustrate the potential improvements to streamflow forecasts through the use of S2S climate forecasts.

Other studies have explored improving streamflow forecasts through the postprocessing of ESP, which can be termed post-ESP methods. A common technique is to weigh ESP traces based on teleconnections or large-scale climate information (e.g., Hamlet and Lettenmaier 1999; Mendoza et al. 2017). Najafi and Moradkhani (2012) used climate indices to inform ESP weighting schemes in the East River Basin, Colorado, testing five different weighting methods and finding improvements with several. Bradley and Habib (2015) introduced another alternative, the Bayesian Climate Index Weighted method, finding that it performed similarly to traditional methods while suggesting that it did not require a calibration with hindcasts.

Werner et al. (2004) compared pre- and post-ESP weighting, obtaining favorable results in the CRB. The study assessed methods for weighting CBRFC ESP forecasts in three subbasins in the CRB using climate indices (e.g., Nino-3.4) either through a pre-ESP method of adjusting the precipitation and temperature ensembles input to a hydrology model or a post-ESP method by weighting streamflow ensemble members. Post-ESP methods were found to outperform preadjustment methods. Mendoza et al. (2017) assessed alternatives to traditional ESP methods in basins in the Pacific Northwest. They compared ESP to statistical and hybrid approaches including trace weighting schemes, which were informed by either climate indices or climate reanalysis components, depending on their skill. Climate predictors added seasonal forecast skill with the best results in watersheds that had stronger teleconnections, as might be expected. Wood and Schaake (2008) assessed the validity of postprocessing ESP forecasts through bias correction and calibration techniques to correct for errors in the mean and spread of the forecast in the Feather River Basin, California. Calibration was found to substantially improve forecast bias and spread relative to raw ESPs. These studies show that postprocessing ESP forecasts, for example, with the addition of climate information or statistical error correction, offers potential benefits.

Advancements in streamflow forecast skill would improve projections from operational planning models such as CRMMS (Raff et al. 2013), which in turn would benefit stakeholders in the CRB who depend on outlooks of potential water deliveries, and reservoir releases and levels to drive operational decisions. To investigate the extent to which the latest operational U.S. monthly to seasonal climate forecasts can benefit seasonal streamflow prediction, we investigate using an analog-based post-ESP trace weighting technique — k-nearest neighbors (kNN) — to improve streamflow forecasts in the Upper CRB, using S2S temperature and precipitation forecasts from the NMME. This paper is organized as follows. We first discuss CRMMS, ESP, and the watershed scale S2S climate forecasts in the Background and Data section. The Methods section discusses the different predictors, kNN weighting methods, and verification metrics. In the Results section, we compare kNN weighting methods on the runoff season scale, followed by the Discussion and Conclusion section, which summarizes results, limitations, and potential avenues of further method improvements.

## BACKGROUND AND DATA

### *CRMMS and ESP*

CRMMS is a Reclamation CRB mid-term operational projection model built-in RiverWare, a generalized river basin modeling software platform (Zagona et al. 2001). CRMMS runs unregulated inflow traces through a decision-making framework using rule-based logic to model system and reservoir operations. The model produces probabilistic operational projections of 12 major reservoirs (9 Upper Basin and 3 Lower Basin) in the CRB out five years at a monthly timestep. There are 12 Upper Basin forecast locations where ESP forecasts are input to CRMMS. An important forecast location is an unregulated inflow to Lake Powell, which combines other forecast locations and unaccounted for gains into one flow for the Upper CRB. A schematic of the CRB with the important reservoirs and forecast locations is shown in Figure 1.

The ESP forecasts used in CRMMS are for unregulated inflows — that is, streamflow that would have flowed through a location if there were no dams and or specific measured diversions upstream of the forecast point. The unregulated flows do not represent what would be observed, nor are they natural flows since natural flows would account for unmeasured depletions and return flows, including from irrigated acreage, evaporative losses from reservoirs, and losses due to groundwater pumping (Colorado Basin River Forecast Center, Accessed June 2021, https://www.cbrfc.noaa.gov/wsup/doc.php). The unregulated inflows used in this study are produced by the CBRFC specific for CRMMS and other Reclamation decision-making models, and only include three Upper Basin tunnel diversions. The CBRFC produced a hindcast dataset of ESP forecasts with calibrated NWS models (e.g., SacSMA and Snow-17), using 30 meteorological input traces driven by historical precipitation and temperature traces from the climatological record (1981–2010). The reforecast dataset provided by the CBRFC for 1981–2011 was extended using operational forecasts that were available from 2012 to 2016. For the 1981–2010 ESP reforecasts, a leave-one-year-out cross-validation method was employed in which the trace from the forecasted year's climate forcings was removed from the ensemble to avoid including a trace with perfect knowledge of the temperature and precipitation. Therefore, the ESP forecasts from 1981 to 2010 have 29 traces, while the forecasts from 2011 to 2016 have 30 traces. The ESP reforecasts do not include short-term forecasts of precipitation and temperature, which are included in operational ESP forecasts (temperature and precipitation forecast are included for up to seven days, with temperature forecasts blending back to climatology through day 10).

### *S2S Climate Forecasts*

NMME climate forecasts are used to inform the ESP trace weighting method for this study. NMME is a multi-model ensemble that combines global climate model forecasts, with forecasts available at a monthly time step for leads up to nine months (Kirtman et al. 2014). The version of NMME we are employing in this effort includes seven global climate models, which are summarized in Table 1. Individual model

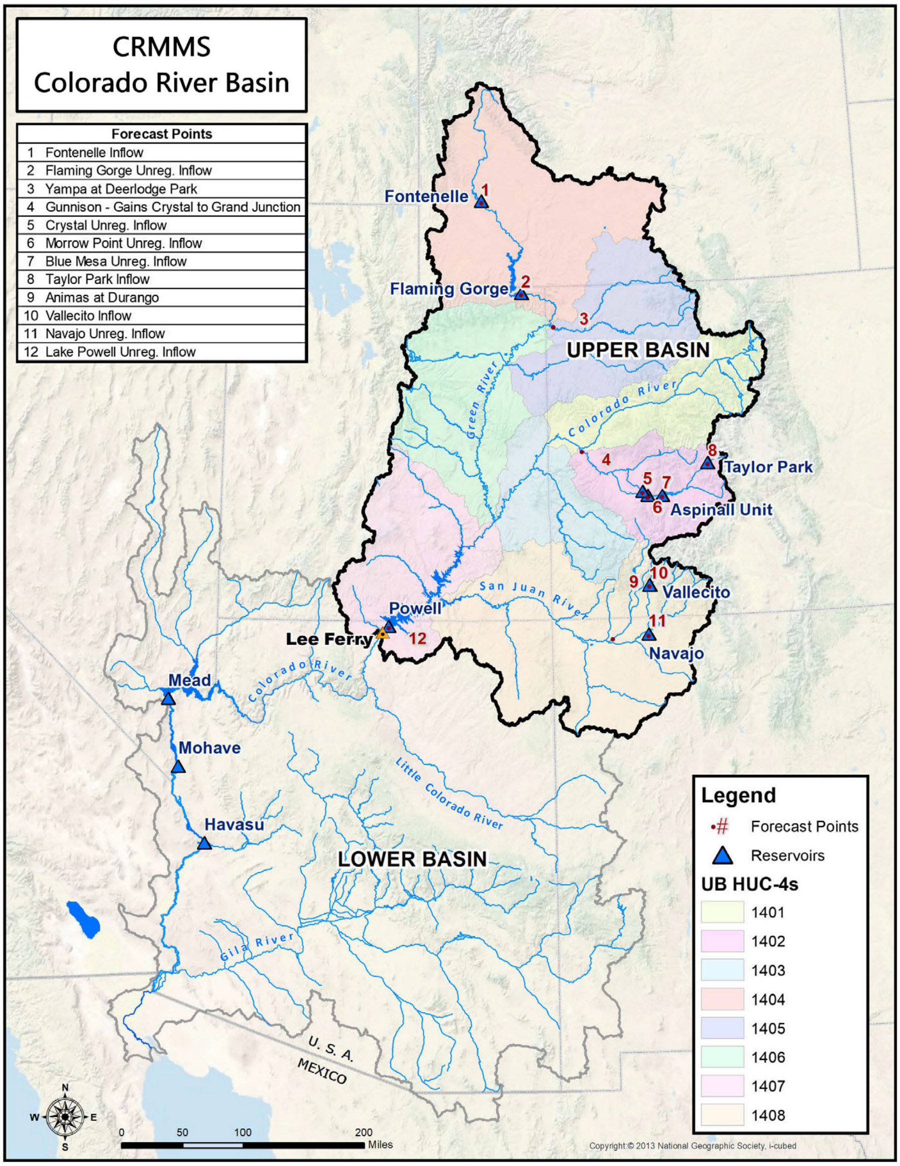

FIGURE 1. Schematic of the Colorado River Basin as setup in the Colorado River Mid-term Modeling System (CRMMS). The forecast locations are numbered from 1 to 12 with points representing the inflow locations, many of which overlap with reservoirs. Forecast names are located in the top left table in the figure. The eight Upper Basin hydrologic unit code 4 (HUC-4) watersheds are symbolized by different colors of shading. The names of the HUC-4s can be found in Table 2. The reservoirs are represented by blue triangles. The Aspinall Unit includes Blue Mesa, Morrow Point, and Crystal Reservoirs.

ensemble mean reforecasts and operational forecasts from 1982 to 2017 were used in this effort. Baker and Wood (2019) spatially averaged raw gridded NMME ensemble mean temperature and precipitation forecast to a watershed scale, using the U.S. Geological Survey (USGS) hydrologic unit code 4 (HUC-4) as the target spatial unit. NMME forecasts were verified at monthly and seasonal leads to calculate the forecast skill compared to the North American Land Data Assimilation System (NLDAS; Xia et al. 2012). Real-time watershed scale forecasts and benchmark assessments of hindcasts have been made available online since 2017 (http://hydro.rap.ucar.edu/s2s/). For a detailed description of raw NMME forecast processing to a watershed scale and a detailed skill assessment, see Baker et al. (2019).

## METHODS

The kNN method is described below with the proposed feature vector components. The method is applied to two different spatial scales, which are compared.

TABLE 1. North American Multi-Model Ensemble (NMME) model information.

| Model acronym | Model name | Reference |
|---|---|---|
| CFSv2 | NOAA NCEP Climate Forecast System version 2 | Saha et al. (2014) |
| NASA_GEOS5 | Goddard Earth Observing System version 5 | Vernieres et al. (2012); Molod (2012) |
| CCSM4 | NCAR/University of Miami Community Climate System Model version 4 | Lawrence et al. (2012) |
| GFDL-CM2.1 | Geophysical Fluid Dynamics Laboratory's (GFDL's) Climate Model version 2.1 | Zhang et al. (2007) |
| GFDL_FLOR-CM2.5 | GFDL's Climate Model version 2.5 [FLORa06 and FLORb01] | Vecchi et al. (2014) |
| CanCM3 | Third Generation Canadian Coupled Global Climate Model | Merryfield et al. (2013) |
| CanCM4 | Fourth Generation Canadian Coupled Global Climate Model | Merryfield et al. (2013) |

Notes: NCAR, National Center for Atmospheric Research; NCEP, National Centers for Environmental Prediction; NOAA, National Ocean and Atmospheric Administration.

*kNN Trace Weighting Scheme*

We adapted the post-ESP trace weighting scheme from Werner et al. (2004) to weight the ESP forecast ensemble members in the Upper CRB, creating weighted ensembles for the ESP forecasts initialized from January 1982 to September 2016. The kNN analog-based weighting method is applied to each equally weighted raw ESP forecast to create a conditionally weighted ESP forecast. kNN creates weights for each ESP ensemble member (trace) using a list of $n_v$ variables — a "feature vector," $\boldsymbol{x} = (x_1,\ x_2,\ \ldots x_{n_v})$ — that are chosen to quantify the similarity between the current forecast year and meteorological years represented in the ESP ensemble. The similarity of the feature vector elements between the forecast year and each of the ESP trace years is then used to assign a weight to each of the ESP traces. The technique is described as follows:

1. The feature vector $\boldsymbol{x}$ elements are selected and also given prescribed relative weights ($c_i$, $i = 1, n_v$) reflecting their importance in the vector, with the weights summing to 1.
2. Hindcast year ($t$) ESP ensemble traces are derived using the historical meteorological inputs from $n_{\text{esp}}$ years ($j$) in 1981–2010, so the variable data from these years ($j$) are used to assemble matching feature vectors to compare with the vector of the hindcast year (t). The hindcast year's data records are removed from this set of vectors so as to exclude the ESP trace that would include any knowledge of the future climate or antecedent conditions in a given hindcast. For example, the hindcast made in February 1991 could not assign a nonzero weighting for the ESP trace derived from meteorology observed in 1991 (i.e., enforcing $t \neq j$). The feature vectors are standardized for each start month.
3. For each hindcast made in year ($t$), a distance vector $\boldsymbol{d}$ containing one value ($d_j$) for each ESP trace meteorological year ($j \in$ 1981–2010, $j \neq t$), is computed. The distance $d_j$ for each trace is the weighted Euclidean distance of the $n_v$ elements in the feature vector for the forecast trace ($j$), as in Equations (1 and 2).

$$\boldsymbol{d} = (d_1,\ d_2,\ \ldots,\ d_j), \quad \text{where} \tag{1}$$

$$d_j = \sqrt{c_1\left(x_{t,1} - x_{j,1}\right)^2 + c_2\left(x_{t,2} - x_{j,2}\right)^2 + \cdots + c_{n_v}\left(x_{t,n_v} - x_{j,n_v}\right)^2}, \quad \text{for each}\, j \in (1981,\ \ 2010),\ \ j \neq t, \tag{2}$$

4. The corresponding hindcast trace weights, $\boldsymbol{w}$, are then calculated from the trace distances $\boldsymbol{d}$, as follows:

$$k = \text{NINT}\left(\frac{n_{\text{esp}}}{\alpha}\right), \tag{3}$$

$$w_j = \left[1 - \frac{d_j}{d_k}\right]^{\lambda}, \quad \text{where}\, d_j \leq d_k, \tag{4}$$

$$w_j = 0, \quad d_j > d_k, \tag{5}$$

$$\boldsymbol{w} = (w_1,\ w_2,\ \ldots,\ w_j), \tag{6}$$

where the $\alpha$ parameter determines the $k$ nearest neighbor traces that will have nonzero weights in the

ESP and $\lambda$ is the distance sensitive weighting parameter. The NINT is the nearest integer operator. In this study, we set $\lambda = 2.5$ and $\alpha = 1$ based on sensitivity experiments (not shown). This means that all ESP traces are included in the kNN forecast, albeit with each having a different weight.

5. The weights are then normalized so that their sum is equal to 1.

6 ESP traces are resampled based on the normalized ensemble weights to obtain a 100-member ensemble. We tested this method with larger ensembles (i.e., 500- and 1,000-member ensemble) but found little to no benefit to increasing the ensemble size.

### *Feature Vector*

The feature vector elements (or predictors) used in the kNN trace weighting scheme consist of the mean NMME temperature and precipitation watershed scale forecasts with associated NLDAS observations, and observed antecedent streamflows for different numbers of lagged months. These predictors were weighted based on prescribed weights ($\boldsymbol{w}$) that were assigned based on prior knowledge of their potential usability (Prairie et al. 2006; Baker et al. 2019) for influencing future runoff predictions and motivated by the desire to avoid including predictors with little theoretical relationship to runoff. We recognize that more sophisticated data-driven analysis in predictor screening and selection would also be appropriate for this objective. To the extent that the resulting weights are suboptimal for harnessing the full information content of the predictors, the study yields a conservative estimate of the potential benefit of the NMME dataset in this context.

The four NMME-based predictors were watershed-scale anomaly forecasts for the one-month mean temperature, one-month mean precipitation, three-month mean temperature, and three-month mean precipitation, all at a zero-month lead time. The difference between each NMME forecast and the NLDAS observed temperature and precipitation for the same time periods and locations were used to identify and weight the NMME forecasts' nearest neighbors. For example, an NMME forecast for a dry and warm Month 1 would give higher weights to ESP meteorological trace years in which the first month was relatively dry and warm, shifting the overall ESP ensemble toward dry and warm tendencies leading to reduced flows. An additional element — observed three-month antecedent streamflow near the outlet location of the Upper CRB for which the predictor feature is being applied — was also included in the predictor vector. The historical unregulated flow provides proxy information about antecedent moisture conditions in the basin, such as the amount of baseflow, which relates to basin soil moisture and modulates the impact of rainfall and snowmelt on streamflow during the spring melt season. Note, the modeled moisture conditions used to initialize ESP also contain such information, but due to potential errors in these unobserved modeled states, there is an opportunity for historical streamflow to augment the model contribution. Other studies have used alternative antecedent conditions metrics such as Palmer drought severity index (PDSI) when predicting streamflow (i.e., Regonda et al. 2006; Bracken and Rajagopalan 2010). The PDSI and other potential antecedent condition metrics were not explored here, as the information they would provide is similar to that of streamflow.

To include information about the skill of NMME forecasts, we distributed the weights for each climate forecast lead time (one month, three months) between temperature and precipitation based on their relative anomaly correlations with observations, averaged over the year. For example, a predictor with double the correlation of another predictor would get twice the weight in the feature vector. To a limited degree, the adjusted weights reflect differences in skill of the different NMME forecast elements, as opposed to using equal weights for the contribution of each element to the distance calculation, which is the default application of kNN for analog formation.

The skill of NMME forecasts differs depending on the variable and lead, as shown in Figure 2. Skill as measured by anomaly correlation is generally higher for the one-month lead compared to the three-month lead. There is more spatial consistency in skill for temperature than precipitation, with some basins such as the San Juan and Green exhibiting different patterns in skill compared to the basin-wide average. Figure 2 also shows a usability threshold of 0.3, which is used by forecast groups such as the NCEP Climate Prediction Center (O'Lenic et al. 2008). Many forecasts are close to this threshold, especially for the seasonal forecasts, and in some cases fall below it. This skill threshold could be used to remove certain forecast months use of NMME forecasts, though that was not pursued in this project. It could also be used to weight forecast elements differently in each initialization month, although here we used a single weighting based on mean skill for each predictor in the feature vector for all lead times. We acknowledge that an optimization approach toward predictor screening and selection, for which many techniques are available (e.g., Tibshirani 1996), would almost certainly be beneficial.

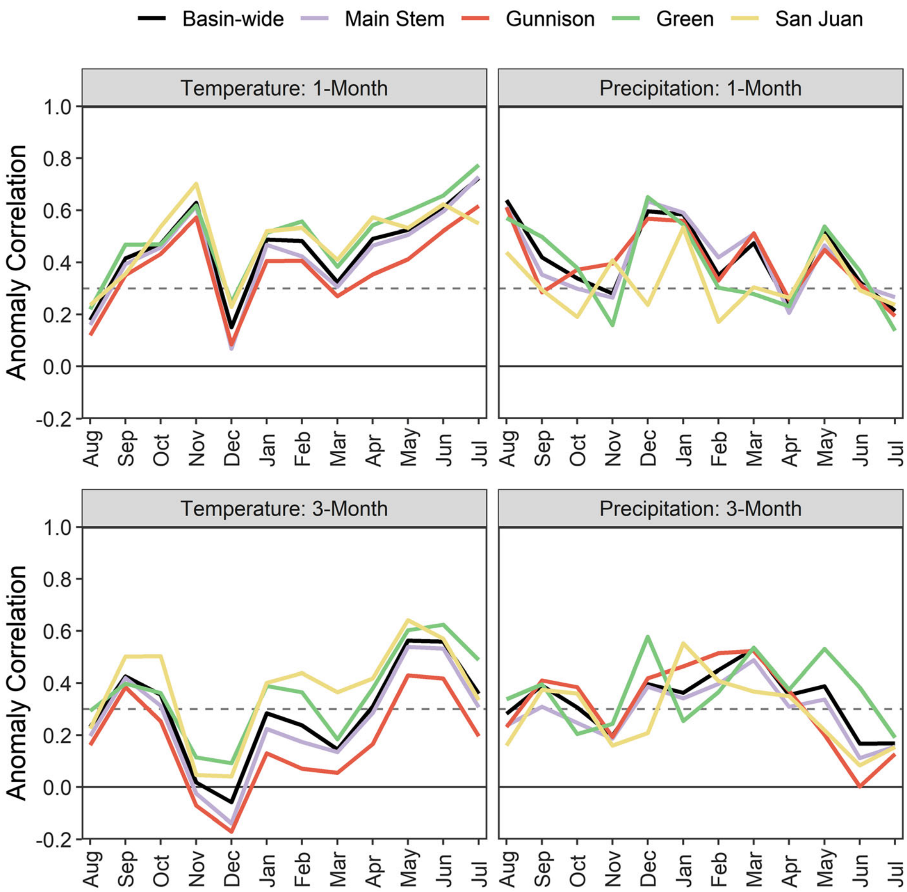

FIGURE 2. Anomaly correlation of NMME one- and three-month temperature and precipitation forecasts with respect to North American Land Data Assimilation System observed anomalies (1982–2016) for the entire basin and the four individual basins. The dashed line corresponds to an anomaly correlation of 0.3, which represents a usability threshold.

### *Spatial Evaluation Scenarios*

The kNN trace weighting method was applied for two different spatial scale strategies described in the following sections — Basin-wide method and the Disaggregated-basin method. Although the paper focuses on aggregate flow at the outlet of the entire Upper CRB (approx. 113,000 sq. mi.), we hypothesize that the NMME may offer useful information to discriminate subbasin scale climate variability (HUC-4s average 14,000 sq. mi.), such a wetter Green River basin in the north combined with a drier San Juan basin in the south, as occurs under La Niña ENSO conditions (Clark and Serreze 2001).

**Case 1: Basin-Wide Method.** The Basin-wide method of applying kNN weights the ESP forecast members based on NMME forecasted temperature and precipitation aggregated over the entire Upper CRB. The Upper CRB aggregated NMME forecasts were calculated from a flow-weighted average of individual HUC-4 NMME forecasts, as opposed to using area-weighted averages, because runoff's importance to Lake Powell unregulated inflow varies across the contributing HUC-4s. The weighted climate forecasts were based on the observed long-term mean ratios of contributing flows from each HUC-4 basin to the inflow to Lake Powell, as calculated based on USGS stream gages near HUC-4 outlets as unregulated flows were not available near all HUC-4 outlets. Though gaged flows were used in this part of the analysis, the relative subbasin contribution to basin runoff in the long-term mean is similar between the observed and unregulated data series. The ESP trace weights based on the Basin-wide method were applied uniformly to the ESPs in the Upper CRB.

**Case 2: Disaggregated-Basin Method.** The Disaggregated-basin (Disagg-basin) strategy splits the Upper CRB into four different subbasins based on the major tributaries in the Upper CRB and CRMMS forecast locations. The kNN weighting is estimated at each of the four basins separately, producing different ESP trace weights for each subbasin. This results in new ESP ensembles in the sense that the recombination of ESP members from the subbasins can create a wider range of forecasted unregulated inflows to Lake Powell than is available from the historical record combinations that are used in the Basin-wide ESP.

The four subbasins are the (1) Main Stem, (2) Green, (3) Gunnison, and (4) San Juan. The eight HUC-4s are categorized into the four basins as shown in Table 2, based on the forecast locations used in CRMMS. For instance, the Lower Green HUC-4 is along the Green River, but the contributing flows from this HUC-4 are included in the Lake Powell unregulated inflow forecast since the downstream most CRMMS forecast location on the Green River is above the Lower Green HUC-4.

The NMME temperature and precipitation forecasts for the four subbasins are weighted based on contributing flow from each subbasin's HUC-4 watersheds. The 12 CRMMS forecast locations are also split into the four subbasins as shown in Table 2. The historical preceding three-month averaged flows used in the feature vector are the total contributing or intervening flow for each of the four subbasins.

This requires calculating new total unregulated flow volumes for each subbasin. Since forecast locations are for total flow, this involves adding tributaries to the downstream most forecast location on each tributary. The Main Stem flow is calculated differently since the Green, Gunnison, and San Juan River flows must be subtracted from the Lake Powell unregulated inflow forecast location. The flows for each subbasin are described in the following equations:

$$Q_{\mathrm{Green}} = Q_{\mathrm{FG}} + Q_{\mathrm{Yampa}}, \tag{7}$$

$$Q_{\mathrm{Gunnison}} = Q_{\mathrm{Cry}} + Q_{\mathrm{CryGJ}}, \tag{8}$$

$$Q_{\mathrm{SanJuan}} = Q_{\mathrm{Nav}} + Q_{\mathrm{Animas}}, \tag{9}$$

$$Q_{\mathrm{MainStem}} = Q_{\mathrm{Powell}} - (Q_{\mathrm{Green}} + Q_{\mathrm{Gunnison}} + Q_{\mathrm{SanJuan}}), \tag{10}$$

with variables defined in Table 2.

Once kNN is performed on each of the four subbasins separately, the forecasts are recombined to calculate a new CRMMS input ensemble for the 12 forecast locations. For all forecast locations except the Lake Powell unregulated inflow location, the resampled 100-member ensembles are combined. The traces (ensembles) in each of the four subbasins traces were sorted by trace weights from highest to lowest then combined by matching traces across all subbasins with equal ranks or weights. This results in the traces with the highest weights or lowest weights being combined. The Lake Powell unregulated inflow forecast location was calculated from this new combination of traces from all four subbasins as follows:

$$Q_{\mathrm{Powell}} = Q_{\mathrm{MainStem}} + Q_{\mathrm{Green}} + Q_{\mathrm{Gunnison}} + Q_{\mathrm{SanJuan}}. \tag{11}$$

The Disagg-basin method results in a new ensemble with a different spread and distribution than the original ESP forecast and the Basin-wide method because it was not limited to combinations of subbasin results that had occurred in the historic record. This combination of flow sequences across subbasins without regard to their spatial covariation across basins would be problematic if the focus of the outcome was on higher frequency timeseries — that is, daily, weekly, or even monthly, in which case the synchronization of flow arising from weather- or climate-driven patterns is an important characteristic to preserve. Because our focus is on seasonal and longer volumetric flow input to Lake Powell, or in individual basins, ensuring realistic space-time subperiod sequencing of the flows is not a constraint to the choice of technique.

TABLE 2. HUC-4 and CRMMS forecast location assignment in Disaggregated-basin (Disagg-basin) method.

| Disagg-basin | HUC-4 | Forecast locations |
|---|---|---|
| Green | 1404 — Great Divide-Upper Green<br>1405 — White-Yampa | 1. Fontenelle Inflow ($Q_{\mathrm{Font}}$)<br>2. Flaming Gorge Unregulated Inflow ($Q_{\mathrm{FG}}$)<br>3. Yampa at Deerlodge Park ($Q_{\mathrm{Yampa}}$) |
| Gunnison | 1402 — Gunnison | 4. Gunnison — Gains Crystal to Grand Junction ($Q_{\mathrm{CryGJ}}$)<br>5. Crystal Unregulated Inflow ($Q_{\mathrm{Cry}}$)<br>6. Morrow Point Unregulated Inflow ($Q_{\mathrm{MP}}$)<br>7. Blue Mesa Unregulated Inflow ($Q_{\mathrm{BM}}$)<br>8. Taylor Park Inflow ($Q_{\mathrm{TP}}$) |
| San Juan | 1408 — San Juan | 9. Animas at Durango ($Q_{\mathrm{Animas}}$)<br>10. Vallecito Inflow ($Q_{\mathrm{Vall}}$)<br>11. Navajo Unregulated Inflow ($Q_{\mathrm{Nav}}$) |
| Main Stem | 1401 — Colorado Headwaters<br>1403 — Upper Colorado-Delores<br>1406 — Lower Green<br>1407 — Upper Colorado-Dirty Devil | 12. Lake Powell Unregulated Inflow ($Q_{\mathrm{Powell}}$) |

An alternative technique for combining traces from subbasins was also assessed, though the results are not shown here. Subbasin trace combinations were created by conditionally resampling each subbasin's traces separately according to the frequencies indicated by their weights. This strategy also allowed each trace to potentially have a different combination of historical years of temperature and precipitation across subbasins than exists in the historical record, also potentially allowing for different variability than is found in the historical record. We found that this strategy made a minimal difference in the resulting verification metrics in the CRB domain (which does not imply that this would be the case elsewhere).

### *Verification Metrics*

Streamflow forecasts from ESP and kNN methods are compared to historical unregulated flows at the inflow to Lake Powell and the four subbasins used in the Disagg-basin method. Verification metrics are calculated for the runoff season volume, April–July, for lead times of up to 12 months prior to the last forecast in July. Runoff season volume is measured in million acre-foot (MAF), a common volumetric measure used in U.S. water resources, which represents 1 foot of water covering an acre of land, or 0.81 billion cubic meters. Recognizing the likely audience of this paper, we present results in the units currently (and traditionally) used in operations and management, along with this conversion factor for international readers. The forecasts initialized in May, June, and July include historical unregulated flows for the part of the runoff seasonal that has already been observed (i.e., is before the initialization date). These forecast leads have higher skill and lower errors compared to other forecasts since there are fewer months in the runoff season to forecast, and the observed portion has zero error. Lead times in this paper are defined as months until the end of July or the end of the forecast period. For example, a January forecast would be defined as a seven-month lead since January is seven months before the end of July, even though metrics are validated on the April–July period. We recognize that it is common in many forecast contexts to define lead times differently, for example, between forecast issuance and the beginning of the forecast event. The definition used here aligns with the understanding of a widely used CRB management forecast product, the "24-Month Study," which provides a two-year outlook for the future conditions of the reservoir system.

The forecast verification measures used here include an accuracy metric, the root mean squared error (RMSE), which is the mean square root of the average of the squared differences between volume forecasts and observations, calculated for each ensemble member in a forecast and then averaged across a sample (sequence) of ensemble forecasts made on a single initialization date (e.g., March 1). Because the errors are squared, larger errors have a greater influence on RMSE than smaller errors. We also calculate a probabilistic error metric, the continuous rank probability score (CRPS; Hersbach 2000), which is the integrated squared difference between the cumulative distribution function of the forecast ensemble and the corresponding distribution of observations (a Heaviside function equaling 0 for values below each verifying observation and 1 for values equal to or greater than the observation). The CRPS can be decomposed into two elements, reliability and potential. The CRPS reliability can be interpreted similarly to a rank histogram, which summarizes the accuracy of the probabilistic spread of a forecast, while the CRPS potential represents the skill that could be achieved if the forecast were reliable, and is strongly influenced by systematic spread errors (i.e., too many or too few outliers) in the ensembles (Hersbach 2000). All these metrics are bounded at zero and lower values are desired.

Two skill score metrics are also used, allowing for comparison between the accuracy and skill of the forecasts relative to the performance of a reference forecast, and taking the form 1-forecast_metric/reference_metric (Wilks 2010). One skill score is the calculated using the forecast's CRPS error metric, and thus is called the continuous rank probability skill score (CRPSS). The second skill score is calculated using the mean squared error metric, and is called the mean squared skill score (MSSS) is the average squared difference between forecast ensembles and observations, compared to the reference forecast. Two reference forecasts are used for comparison to the experimental forecasts. The first is the observed runoff volume climatology from 1981 to 2010, meaning that each forecast is the same sample (ensemble) of historical values for the predictand (e.g., April–July flow) taken from this historical period. The second is the traditional ESP forecast (without climate weighting) that corresponds to the given forecast from the proposed methods, which contains the same yearly varying initialization information as the climate-weighted experimental forecasts. The CRPSS and MSSS skill scores range from 1 to $-\infty$, with a perfect skill score equal to 1. A skill score of 0 means the skill of the forecast is equal to that of the reference forecast, and a negative skill score means the forecast is less skillful than the reference forecast.

## RESULTS

The ESP, Basin-wide kNN, and Disagg-basin kNN streamflow forecasts were analyzed for their skill and accuracy when forecasting runoff season unregulated inflow to Lake Powell. The CRPSS at leads of 12-months to 1-month for the three forecasts compared to the reference forecast of climatology is shown in the left panel of Figure 3. Since we are computing a seasonal flow, the forecast includes observed unregulated flows once the lead is less than four months since the April flow would have been observed by the time of forecast. At longer leads of 12 and 11 months, the forecasts all perform relatively poorly with the median skill of the Basin-wide kNN forecast slightly outperforming other forecasts. The Disagg-basin kNN forecast has a wider range of CRPSS skill values than the other two forecasts with some forecasts having high skill, but overall, the median skill is lower than the other forecasts.

As leads decrease, the skill of all the forecasts increases. The forecast with the highest median skill varies by lead, but the two climate-informed ESPs have slightly improved CRPSS skill from December to March. After March, the median skill of all forecasts is relatively the same, with only small differences in the range of skill as shown by the boxes of each boxplot. Notably, the error (right panel of Figure 3) yields a different perspective: errors are larger (worse) at longer leads and decrease with lead, as expected for volume runoff forecasts, but the Disagg-basin kNN forecasts have a markedly lower error compared to the other forecasts at most lead times. The median error of Disagg-basin kNN is about 1 MAF lower than the other forecasts at longer leads and decreases as all forecasts start to perform better. The Basin-wide kNN forecast performs better than ESP at most leads, especially at shorter leads when the NMME forecasts have a greater impact on the analyzed flow. Based on these metrics, the NMME weighting helps in the Disagg-basin method for February and March, consistent with the determination of skill in the NMME precipitation or temperature during these months (see Figure 2).

To understand why the two metrics shown in Figure 3 suggest different outcomes (e.g., some leads showing lower error but also lower skill), we evaluate the CRPS decomposition elements, reliability, and potential. As seen in Figure 4, the CRPS reliability for the forecasts is relatively similar, with the Disagg-basin kNN method showing a slightly worse CRPS reliability on average and with a greater spread. For CRPS potential, the ESP and Basin-wide kNN forecasts perform worse than the Disagg-basin kNN forecasts, indicating that without such reliability errors this method would produce the best forecasts. This suggests that the benefits in the Disagg-basin kNN method arise at least in part from narrowing the forecast spread, as opposed to an increase in correlation skill for the median forecast. This is consistent with Figure 3, which shows that the Disagg-basin kNN method performs similarly to the other forecasts when considering CRPSS skill but performs better with RMSE.

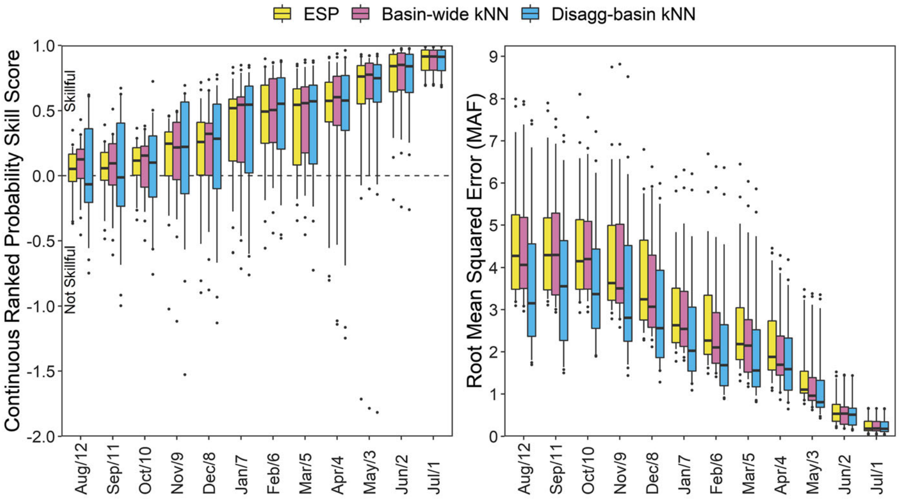

FIGURE 3. Continuous ranked probability skill score (CRPSS) and root mean squared error for runoff season streamflow forecasts of Lake Powell unregulated inflow. The streamflow forecasts ensemble streamflow prediction (ESP), Basin-wide k-nearest neighbors (kNN), and Disaggregated-basin (Disagg-basin) kNN are compared at leads of 12-months to 1-month. The reference forecast for the CRPSS calculations is climatology. The boxplot boxes represent the 25th–75th percentiles, with the median as the midline. The whiskers extend to the 5th and 95th percentiles, and the outliers are any data outside this range. MAF, million acre-foot.

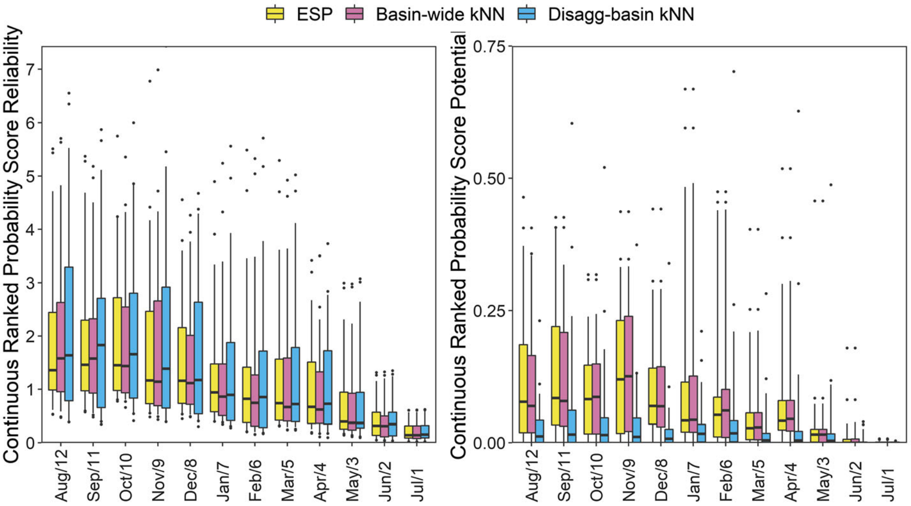

FIGURE 4. Continuous rank probability score reliability and potential for runoff season streamflow forecasts of Lake Powell unregulated inflow. The streamflow forecasts ESP, Basin-wide kNN, and Disagg-basin kNN are compared at leads of 12-months to 1-month. See Figure 3 for the description of the boxplots.

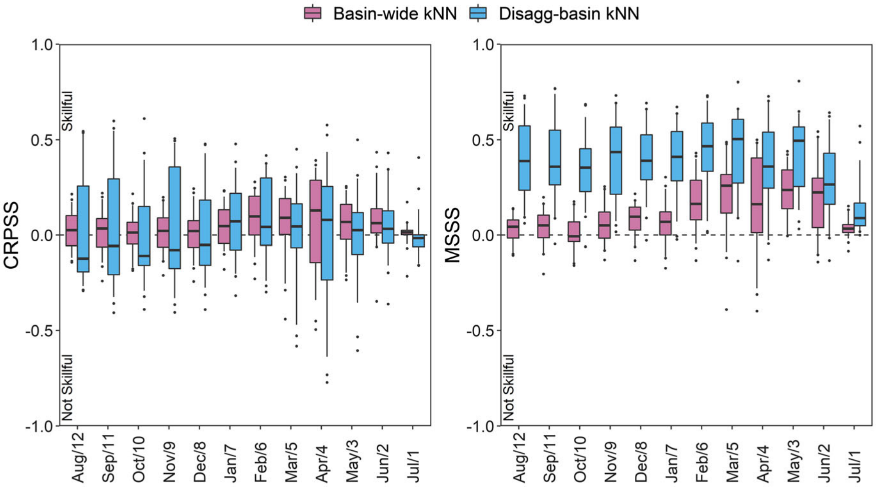

FIGURE 5. CRPSS and mean squared error skill score (MSSS) with reference forecast of ESP for runoff season streamflow forecasts of Lake Powell unregulated inflow. The streamflow forecasts Basin-wide kNN and Disagg-basin kNN are compared at leads of 12-months to 1-month. See Figure 3 for the description of the boxplots.

To explicitly compare the two kNN forecasting methods to ESP, we can use skill scores with a reference forecast of ESP, as opposed to climatology (Figure 5). The CRPSS results for Disagg-basin kNN show no improvement over the ESP forecast except at leads during February and March, with some minor improvements in other months. The Basin-wide kNN forecasts outperform the Disagg-basin kNN forecasts in most months, with the Disagg-basin kNN method exhibiting larger ranges in CRPSS. For MSSS skill, both forecasts show improvements over ESP for most leads. The Disagg-basin kNN forecasts have much higher MSSS skill compared to the Basin-wide method. During February through June, the MSSS skill of the Basin-wide kNN method improves compared to ESP, illustrating that weighting the raw ESP traces with NMME forecast can improve the mean squared error of the runoff season streamflow ensemble. The February through June period is also a period when Basin-wide NMME temperature and precipitation forecasts have generally good performance, though the three-month temperature forecasts exhibit poor skill during February and March (see Figure 2). The striking improvement in the MSSS for the Disagg-basin KNN approach can be traced to the reduced appearance of high outlier flows in the resulting ensembles.

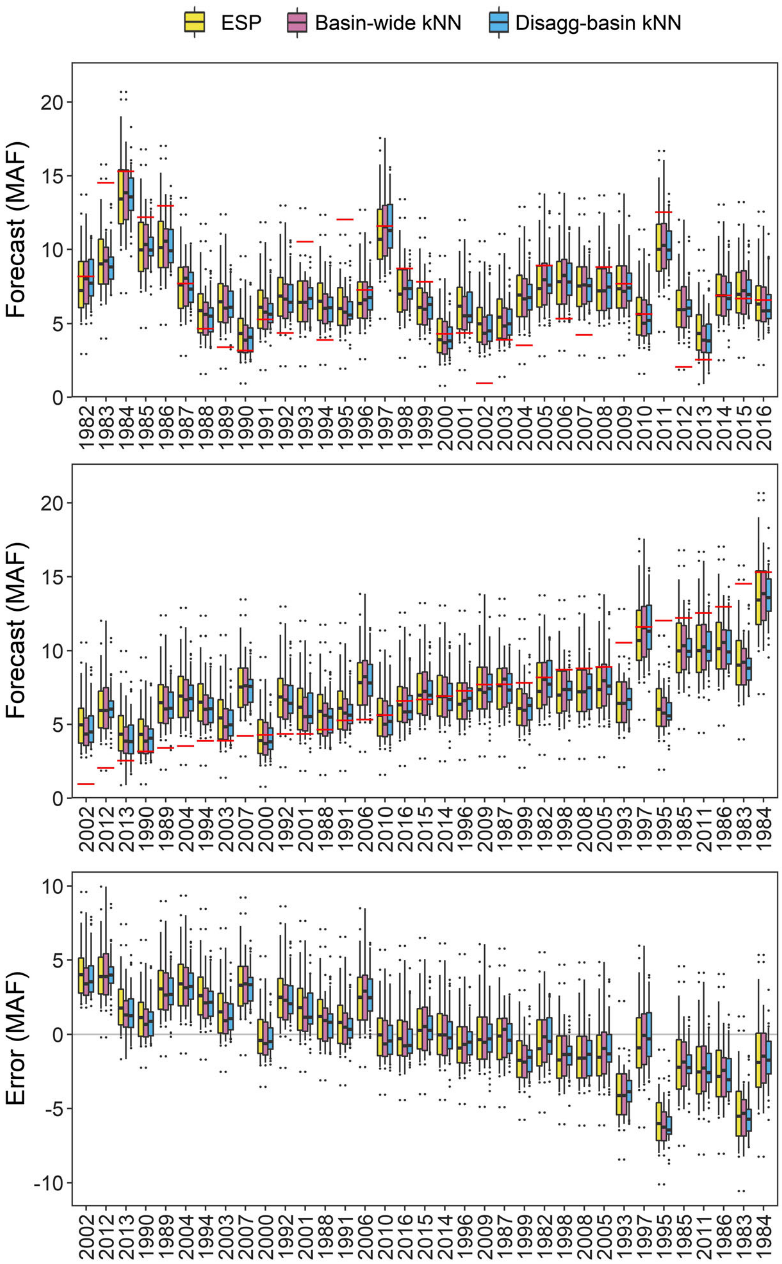

FIGURE 6. Runoff season ensemble forecasts for 1982–2016 compared to observations ranked by year (top) and observed value (middle), and the forecast trace error ranked by observed value (bottom). ESP, Basin-wide kNN, and Disagg-basin kNN forecasts of Lake Powell April–July unregulated inflow are compared for each year at a seven-month lead in January. Observed unregulated inflow is represented by a red horizontal line. See Figure 3 for the description of the boxplots.

To illustrate (qualitatively) the outcome of the trace-weighting for forecasts, Figure 6 shows an example of the streamflow forecasts for each year compared to observations, using seven-month lead forecasts (e.g., initialized January 1) for the period 1982–2016. The figure shows forecasts ranked by observed value (top) and chronologically (bottom), and the ensemble trace errors (bottom). Note, at a seven-month lead, the runoff forecasts for the Upper CRB have limited discrimination because the runoff variability is largely driven by snow accumulation, which is still nascent at this time of year. Higher flow years are detected in part because such years often have an earlier than normal snow accumulation. The ESP forecasts generally have a wider spread than the Disagg-basin kNN forecasts. The Basin-wide or Disagg-basin kNN forecasted median flows improve on ESP in most years (25 of 35), based on whether the median is closer to the observation or if similar, the forecast has a smaller spread. The kNN-based

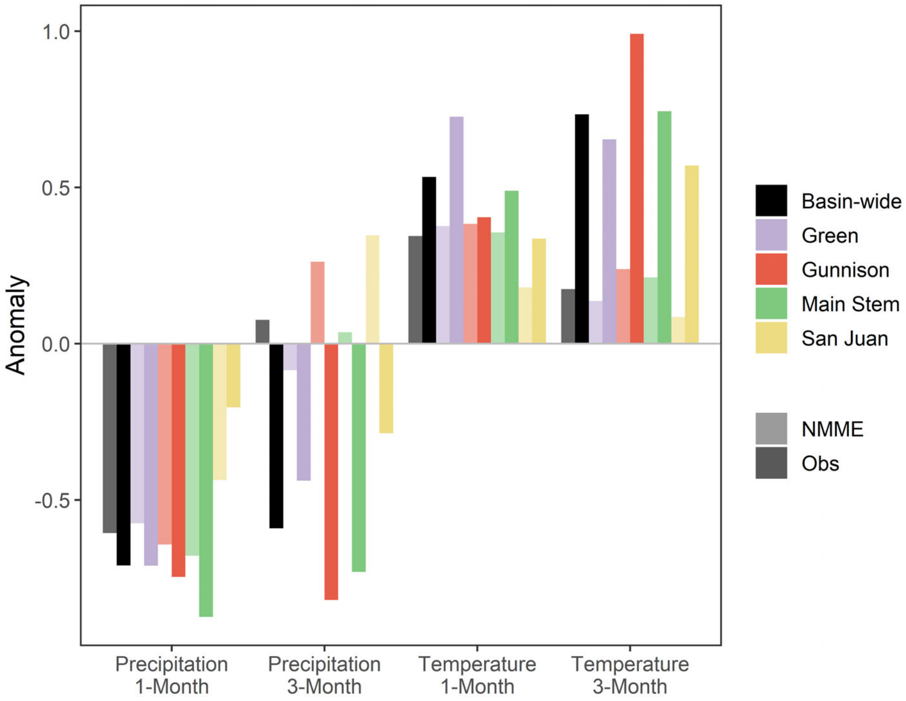

FIGURE 7. Analysis of January 1990 climate forecasts from NMME compared to observations for various basins. The lighter shades of the colors are the NMME forecast and the darker shaded colors are the observations.

trace-weighting tends to perform better during above or below average flow years but does not add much potential value during average flow years. This finding is consistent with previous research that examines why near-normal forecasts are difficult to skillfully predict (van den Dool and Toth 1991).

To understand whether these improvements are related to the use of NMME-based predictors, we analyzed specific forecasts to determine whether NMME forecast anomalies (e.g., wet/dry, warm/cold) were consistent with observed climate anomalies, particularly for kNN-based forecasts that performed better than the associated baseline ESP forecasts. As an example, Figure 7 shows the January 1990 forecasted NMME temperature and precipitation anomaly forecasts, corresponding to a runoff season in which the flow was slightly below average. The negative runoff anomaly was driven in part by temperatures that were above average coupled with below-average precipitation (the observation bars in Figure 7) in the forecast timeframe (January for one month and January–March for three months). The NMME forecast captures the below-average one-month precipitation well, and forecasts above-average temperatures for both leads, though the forecasted magnitudes are not as large. The three-month NMME precipitation forecast shows slightly above average precipitation forecasted for most basins, though the forecasts are relatively close to zero. In aggregate, the analysis suggests that the NMME forecasts do play a role in benefitting the runoff forecast — in the case illustrated, NMME forecasts conditioned the volumetric runoff forecast toward below average.

## DISCUSSION AND CONCLUSIONS

The post-ESP method explored in this work illustrates the opportunity for modern climate forecasts from the NMME to inform and improve seasonal spring runoff streamflow forecasts through a postprocessing framework in the Upper CRB. NMME one- and three-month temperature and precipitation forecasts, along with three-month preceding averaged flow were shown to be useful predictors for weighting ESP streamflow traces. Two approaches to apply the climate forecast information from NMME were explored, testing whether the use of forecasts on a level of greater spatial / watershed resolution would provide more skill than using them at more spatially aggregated resolution, applying different climate signals in different subbasins of the Upper CRB. The established weighted analog method (kNN) was applied to translate the climate forecast signals into alternative weightings for ESP. Both trace-weighting strategies led to forecasts that were generally (but not always) more accurate than ESP for all lead times. The greater spatial granularity of the Disagg-basin kNN showed substantially reduced error compared to other forecasts at all leads through May. The CRPSS-based skill results were more equivocal than RMSE-based results, likely because the spread of the Disagg-basin kNN forecasts can be too narrow, especially at times of low skill, compared to reference climatology forecasts. The narrower spread of the Disagg-basin kNN method could arise from the ability of NMME to successfully weight traces in the ESP toward the observed runoff, or it could be an artifact of the method coupled with NMME forecasts erroneously and systematically down weighting some traces while upweighting others. For instance, if NMME circulation misrepresents the spatial variability and co-variability of precipitation and temperature, ESP traces based on observations with such patterns would never receive appropriate weight. In general, a reduced spread is desirable in a forecast, to the extent that the forecast spread still gives an unbiased expectation of the frequency of the observed outcome (i.e., it is commensurate with forecast skill improvements). Here, the narrow spread slightly degrades reliability for the Disagg-basin strategy (e.g., Figure 4), thus it is making some of the forecasts overconfident. The ability to ascertain such forecast qualities through efforts such as hindcasting is important so that operators can calibrate their use, if deficiencies cannot be eliminated.

A simplification made during this analysis was to treat temperature and precipitation independently, though these variables are correlated in nature. Due to their cross correlation, it is likely that the inclusion of both variables adds less information to the analog selection than it would if they were fully independent, although there is cause to include both. Cross correlations between precipitation and temperature tend to be in the range of −0.4 to −0.6, which means they only explain 25%–36% of the variability in their occurrence. This means that all combinations of precipitation and temperature anomalies do occur in nature (e.g., cold-wet, cold-dry, warm-wet, warm-dry). In this basin, due to variability driven from a variety of mechanisms (such as easternly frontal systems and northerly tropical connections, among others), including both predictors allows for discriminating between such patterns.

When comparing the two trace-weighted methods to traditional ESP, the Basin-wide kNN method performs better than ESP when considering skill for February through July, though by variable amounts. The Disagg-basin kNN approach shows much higher MSSS-based skill compared to ESP and the Basin-wide kNN method for all leads, though the Basin-wide method also outperforms ESP especially in the late winter to early summer. These differing results illustrate that the choice of metrics can have a large impact on the interpretation of the benefits of using the NMME-informed trace-weighting, thus it is important to focus on metrics that respond to characteristics of forecasts that are important for particular decision uses (e.g., the accuracy of the median forecast vs. the forecast spread and statistical reliability). The results of the study support the potential use of NMME forecasts at a HUC-4 level of disaggregation, and also outline a strategy for combining weighted traces from different subbasins to generate a broader range of ESP traces than is typically available from resampling basin-wide hydrology.

This work is not an exhaustive study of post-ESP methods but it does apply pragmatic choices toward implementing a new, state of the science climate forecast resource (the NMME) through post-ESP methods for potential water management in the Upper CRB. Further research could explore more extensive data-driven approaches to predictor screening and selections, weighting schemes, distance calculations, predictors, or postprocessing calibration techniques. In particular, it is likely that greater climate skill could be harnessed from shorter range forecasts that are more skillful, including weather scale forecasts (which are used at the CBRFC), or S2S forecasts — for example, forecasts for 2–3 weeks lead times, which are not currently used in operations but tend to have higher skill than the NMME one- and three-month forecasts. In separate work by the authors, S2S forecasts from an individual NMME model, Climate Forecast System version 2, were postprocessed and assessed on a HUC-4 watershed scale, but were not incorporated into this study (Baker and Wood 2020) due to their shorter length of hindcast record.

The method could also be improved through a broader investigation of potential predictors included in the kNN analog feature vector. In this case, the selection of antecedent streamflow and forecasted precipitation and temperature was based on decades of experience in the communities of research and practice documenting their predictive relationship to future runoff. Other watershed variables known to influence streamflow, such as soil moisture or snow water equivalent were not included here. Both are inherently reflected in the modeled watershed states used to initialize the ESP forecasts, although we did include the common predictor antecedent streamflow, which is in part a proxy for soil moisture variability. Nonetheless, a data-driven approach to predictor selection, including large-scale climate system fields such as sea surface temperatures (as in Baker et al. 2020) may lead to improved performance compared to the narrowly focused approach used here. It is also likely that varying the predictors by forecast month — for example, to remove NMME forecasts or other predictors when they do not benefit skill, while keeping them when they do, would provide a benefit. Despite various limitations in certain aspects of the approach, the overall results of the work recommend the adoption of climate forecasts to inform streamflow forecasting at certain leads, such as in the winter when ESP forecasts are less skillful due to limited knowledge of potential anomalous winter precipitation that could lead to higher or lower streamflow. The proposed post-ESP methodology can be applied to other basins that depend on ESP streamflow forecasts throughout the U.S., for which S2S watershed scale climate forecasts are available. This strategy will likely be most useful in watersheds where NMME climate forecast skill is high, such as southern California in winter and spring or the southeastern U.S. during fall and winter.

The streamflow forecasts analyzed in this work are on the scale used in Reclamation's mid-term operations and planning model, CRMMS. The improved streamflow forecasts with the kNN weighting schemes could be applied to CRMMS to assess how improved streamflow forecasts translate into operation projections. Improvements to streamflow forecasts, even if modest, could improve projections of future streamflow and reservoir system conditions in the CRB, which would benefit stakeholders who depend on operational projections of future basin conditions for their decision making.

## ACKNOWLEDGMENTS

This study was supported by NOAA's Climate Program Office's Modeling, Analysis, Predictions, and Projections (MAPP) Program under Awards NA16OAR4310138 and NA14OAR4310238, and the Bureau of Reclamation under Cooperative Agreement R16AC00039. The project also participated in the MAPP Subseasonal to Seasonal Climate Testbed. The project acknowledges high-performance computing support from Cheyenne (https://doi.org/10.5065/d6rx99hx) provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation. We are grateful for the Colorado Basin River Forecast Center's contribution in creating the streamflow hindcast record that was used in this project. We also appreciate the scientific and pragmatic insights from Jim Prairie and Edith Zagona.

## AUTHOR CONTRIBUTIONS

**Sarah A. Baker:** Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Software; Validation; Visualization; Writing-original draft; Writing-review & editing. **Balaji Rajagopalan:** Conceptualization; Funding acquisition; Methodology; Project administration; Resources; Supervision; Visualization; Writing-review & editing. **Andrew W. Wood:** Conceptualization; Funding acquisition; Methodology; Project administration; Resources; Visualization; Writing-review & editing.

## LITERATURE CITED

Arnal, L., A.W. Wood, E. Stephens, H.L. Cloke, and F. Pappenberger. 2017. "An Efficient Approach for Estimating Streamflow Forecast Skill Elasticity." *Journal of Hydrometeorology* 18 (6): 1715–29. https://doi.org/10.1175/JHM-D-16-0259.1.

Baker, S.A., A.W. Wood, and B. Rajagopalan. 2019. "Developing Subseasonal to Seasonal Climate Forecast Products for Hydrology and Water Management." *Journal of the American Water Resources Association* 55 (4): 1024–37. https://doi.org/10.1111/1752-1688.12746.

Baker, S.A., A.W. Wood, and B. Rajagopalan. 2020. "Application of Postprocessing to Watershed-Scale Subseasonal Climate Forecasts over the Contiguous United States." *Journal of Hydrometeorology* 21 (5): 971–87. https://doi.org/10.1175/JHM-D-19-0155.1.

Beckers, J.V.L., A.H. Weerts, E. Tijdeman, and E. Welles. 2016. "ENSO-Conditioned Weather Resampling Method for Seasonal Ensemble Streamflow Prediction." *Hydrology and Earth System Sciences* 20 (8): 3277–87. https://doi.org/10.5194/hess-20-3277-2016.

Bracken, C., B. Rajagopalan, and J. Prairie. 2010. "A multisite Seasonal Ensemble Streamflow Forecasting Technique." *Water Resources Research* 46 (3): W03532. https://doi.org/10.1029/2009WR007965.

Bracken, C., B. Rajagopalan, and E. Zagona. 2014. "A Hidden Markov Model Combined with Climate Indices for Multidecadal Streamflow Simulation." *Water Resources Research* 50 (10): 7836–46. https://doi.org/10.1002/2014WR015567.

Bradley, A.A., M. Habib, and S.S. Schwartz. 2015. "Climate Index Weighting of Ensemble Streamflow Forecasts Using a Simple Bayesian Approach." *Water Resources Research* 51 (9): 7382–400. https://doi.org/10.1002/2014WR016811.

Bureau of Reclamation. 2015. *Colorado River Basin Mid-Term Probabilistic Operations Model (MTOM)*.

Clark, M.P., M.C. Serreze, and G.J. McCabe. 2001. "Historical Effects of El Nino and La Nina Events on the Seasonal Evolution of the Montane Snowpack in the Columbia and Colorado River Basins." *Water Resources Research* 37 (3): 741–57. https://doi.org/10.1029/2000WR900305.

Colorado Basin River Forecast Center. 2021. "Water Supply Documentation." https://www.cbrfc.noaa.gov/wsup/doc.php.

Daugherty, L. 2013. "An End-to-End Framework for Seasonal Forecasting in Water Resources Management in the San Juan River Basin Using Stochastic Weather Generator Based Ensemble Streamflow Predictions." MS thesis, University of Colorado Boulder.

Day, G. 1985. "Extended Streamflow Forecasting Using NWSRFS." *Journal of Water Resources Planning and Management* 111 (2): 157–70. https://doi.org/10.1061/(ASCE)0733-9496(1985)111:2(157).

van den Dool, H.M., and Z. Toth. 1991. "Why Do Forecasts for "Near Normal" Often Fail?" *Weather and Forecasting* 6 (1): 76–85. https://doi.org/10.5194/hess-15-3529-2011.

Erkyihun, S.T., E.A. Zagona, and B. Rajagopalan 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.

Franz, K.J., H.C. Hartmann, S. Sorooshian, and R. Bales. 2003. "Verification of National Weather Service Ensemble Streamflow Predictions for Water Supply Forecasting in the Colorado River Basin." *Journal of Hydrometeorology* 4 (6): 1105–18. https://doi.org/10.1175/1525-7541(2003)004<1105:VONWSE>2.0.CO;2.

Grantz, K., B. Rajagopalan, M. Clark, and E. Zagona. 2005. "A Technique for Incorporating Large-Scale Climate Information in Basin-Scale Ensemble Streamflow Forecasts." *Water Resources Research* 41 (10): https://doi.org/10.1029/2004WR003467.

Hamlet, A.F., and D. Lettenmaier. 1999. "Columbia River Streamflow Forecasting Based on ENSO and PDO Climate Signals." *Journal of Water Resources Planning and Management* 125 (6): 333–41. https://doi.org/10.1061/(ASCE)0733-9496(1999)125:6(333).

Hersbach, H. 2000. "Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems." *Weather and Forecasting* 15 (5): 559–70. https://doi.org/10.1175/1520-0434(2000)015%3C0559:DOTCRP%3E2.0.CO;2.

Kirtman, B.P., D. Min, J.M. Infanti, J.L. Kinter, D.A. Paolino, Q. Zhang, H. van den Dool *et al*. 2014. "The North American Multimodel Ensemble: Phase-1 Seasonal-to-Interannual Prediction; Phase-2 toward Developing Intraseasonal Prediction." *Bulletin of the American Meteorological Society* 95 (4): 585–601. https://doi.org/10.1175/BAMS-D-12-00050.1.

Lawrence, D.M., K.W. Oleson, M.G. Flanner, C.G. Fletcher, P.J. Lawrence, S. Levis, S.C. Swenson, and G.B. Bonan. 2012. "The CCSM4 Land Simulation, 1850–2005: Assessment of Surface Climate and New Capabilities." *Journal of Climate* 25: 2240–60.

Li, H., L. Luo, E.F. Wood, and J. Schaake. 2009. "The Role of Initial Conditions and Forcing Uncertainties in Seasonal Hydrologic Forecasting." *Journal of Geophysical Research: Atmospheres* 114 (D4): D04114. https://doi.org/10.1029/2008JD010969.

Lukas, J., and E. Payton, eds. 2020. "Colorado River Basin Climate and Hydrology: State of the Science." https://doi.org/10.25810/3HCV-W477.

Mendoza, P.A., A.W. Wood, E. Clark, E. Rothwell, M.P. Clark, B. Nijssen, L.D. Brekke, and J.R. Arnold. 2017. "An Intercomparison of Approaches for Improving Operational Seasonal Streamflow Forecasts." *Hydrology and Earth System Sciences* 21 (7): 3915–35. https://doi.org/10.5194/hess-21-3915-2017.

Merryfield, W.J., W.-S. Lee, G.J. Boer, V.V. Kharin, J.F. Scinocca, G.M. Flato, R.S. Ajayamohan, J.C. Fyfe, Y. Tang, and S. Polavarapu. 2013. "The Canadian Seasonal to Interannual Prediction System. Part I: Models and Initialization." *Monthly Weather Review* 141 (8): 2910–45. https://doi.org/10.1175/MWR-D-12-00216.1.

Mo, K.C., and D.P. Lettenmaier. 2014. "Hydrologic Prediction over the Conterminous United States Using the National Multi-Model Ensemble." *Journal of Hydrometeorology* 15 (4): 1457–72. https://doi.org/10.1175/JHM-D-13-0197.1.

Molod, A. 2012. "The GEOS-5 Atmospheric General Circulation Model: Mean Climate and Development from MERRA to Fortuna." https://ntrs.nasa.gov/search.jsp?R=20120011790.

Najafi, M.R., H. Moradkhani, and T.C. Piechota. 2012. "Ensemble Streamflow Prediction: Climate Signal Weighting Methods vs. Climate Forecast System Reanalysis." *Journal of Hydrology* 442–443: 105–16. https://doi.org/10.1016/j.jhydrol.2012.04.003.

O'Lenic, E.A., D.A. Unger, M.S. Halpert, and K.S. Pelman. 2008. "Developments in Operational Long-Range Climate Prediction at CPC." *Weather and Forecasting* 23 (3): 496–515. https://doi.org/10.1175/2007WAF2007042.1.

Pagano, T.C., A.W. Robertson, K. Werner, and R. Tama-Sweet. 2014. "Western U.S. Water Supply Forecasting: A Tradition Evolves." *Eos, Transactions American Geophysical Union* 95 (3): 28–29. https://doi.org/10.1002/2014EO030007.

Prairie, J.R., B. Rajagopalan, T.J. Fulp, and E.A. Zagona. 2006. "Modified K-NN Model for Stochastic Streamflow Simulation." *Journal of Hydrologic Engineering* 11 (4): 371–78. https://doi.org/10.1061/(ASCE)1084-0699(2006)11:4(371).

Raff, D.A., L. Brekke, K. Werner, A. Wood, and K. White. 2013. "Short-Term Water Management Decisions -User Needs for Improved Climate, Weather, and Hydrologic Information". U.S. Army Corps of Engineers; Bureau of Reclamation; National Oceanic and Atmospheric Administration." Technical Report CWTS 2013-1. http://www.ccawwg.us/docs/Short-Term_Water_Management_Decisions_Final_3_Jan_2013.pdf.

Regonda, S.K., B. Rajagopalan, M. Clark, and E. Zagona. 2006. "A multimodel Ensemble Forecast Framework: Application to Spring Seasonal Flows in the Gunnison River." *Water Resources Research* 42 (9). https://doi.org/10.1029/2005WR004653.

Saha, S., S. Moorthi, X. Wu, J. Wang, S. Nadiga, P. Tripp, D. Behringer *et al*. 2014. "The NCEP Climate Forecast System Version 2." *Journal of Climate* 27 (6): 2185–208. https://doi.org/10.1175/JCLI-D-12-00823.1.

Tibshirani, R. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1): 267–88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Vecchi, G.A., T. Delworth, R. Gudgel, S. Kapnick, A. Rosati, A.T. Wittenberg, F. Zeng *et al*. 2014. "On the Seasonal Forecasting of Regional Tropical Cyclone Activity." *Journal of Climate* 27 (21): 7994–8016. https://doi.org/10.1175/JCLI-D-14-00158.1.

Vernieres, G., M. Rienecker, R. Kovach, and C. Keppenne. 2012. "The GEOS-IODAS: Description and Evaluation." Technical Report Series on Global Modeling and Data Assimilation Volume 30. Goddard Space Flight Center. http://gmao.gsfc.nasa.gov/pubs/docs/Vernieres589.pdf.

Werner, K., D. Brandon, M. Clark, and S. Gangopadhyay. 2004. "Climate Index Weighting Schemes for NWS ESP-Based Seasonal Volume Forecasts." *Journal of Hydrometeorology* 5 (6): 1076–90. https://doi.org/10.1175/JHM-381.1.

Wetterhall, F., I.G. Pechlivanidis, M.-H. Ramos, A.W. Wood, Q.J. Wang, E. Zehe, and U. Ehret. 2016. "Sub-Seasonal to Seasonal Hydrological Forecasting." *Hydrology & Earth System Sciences* 22. https://hess.copernicus.org/articles/special_issue824.html.

Wilks, D.S. 2010. "Sampling Distributions of the Brier Score and Brier Skill Score under Serial Dependence: Brier Score and Brier Skill Score." *Quarterly Journal of the Royal Meteorological Society* 136 (653): 2109–18. https://doi.org/10.1002/qj.709.

Wood, A.W., T. Hopson, A. Newman, L. Brekke, J. Arnold, and M. Clark 2016. "Quantifying Streamflow Forecast Skill Elasticity to Initial Condition and Climate Prediction Skill." *Journal of Hydrometeorology* 17 (2): 651–68.

Wood, A.W., A. Kumar, and D.P. Lettenmaier. 2005. "A Retrospective Assessment of National Centers for Environmental Prediction Climate Model–Based Ensemble Hydrologic Forecasting in the Western United States." *Journal of Geophysical Research: Atmospheres* 110 (D4): D04105. https://doi.org/10.1029/2004JD004508.

Wood, A.W., and D. Lettenmaier. 2006. "A Test Bed for New Seasonal Hydrologic Forecasting Approaches in the Western US." *Bulletin of the American Meteorological Society* 87 (12): 1699–712. https://doi.org/10.1175/BAMS-87-12-1699.

Wood, A.W., and D.P. Lettenmaier. 2008. "An Ensemble Approach for Attribution of Hydrologic Prediction Uncertainty." *Geophysical Research Letters* 35 (14): L14401. https://doi.org/10.1029/2008GL034648.

Wood, A.W., and J.C. Schaake. 2008. "Correcting Errors in Streamflow Forecast Ensemble Mean and Spread." *Journal of Hydrometeorology* 9 (1): 132–48. https://doi.org/10.1175/2007JHM862.1.

Xia, Y., K. Mitchell, M. Ek, J. Sheffield, B. Cosgrove, E. Wood, L. Luo *et al*. 2012. "Continental-Scale Water and Energy Flux

Analysis and Validation for the North American Land Data Assimilation System Project Phase 2 (NLDAS-2): 1. Intercomparison and Application of Model Products." *Journal of Geophysical Research: Atmospheres* 117 (D3): D03109. https://doi.org/10.1029/2011JD016048.

Zagona, E.A., T.J. Fulp, R. Shane, T. Magee, and H.M. Goranflo. 2001. "Riverware: A Generalized Tool for Complex Reservoir System Modeling." *Journal of the American Water Resources Association* 37 (4): 913–29. https://doi.org/10.1111/j.1752-1688.2001.tb05522.x.

Zhang, S., M.J. Harrison, A. Rosati, and A. Wittenberg. 2007. "System Design and Evaluation of Coupled Ensemble Data Assimilation for Global Oceanic Climate Studies." *Monthly Weather Review* 135 (10): 3541–64. https://doi.org/10.1175/MWR3466.1.