JOURNAL OF THE AMERICAN WATER RESOURCES ASSOCIATION

Vol. 58, No. 5 AMERICAN WATER RESOURCES ASSOCIATION October 2022

The Colorado River Basin Operational Prediction Testbed: A Framework for Evaluating Streamflow Forecasts and Reservoir Operations

Sarah A. Baker, Andy W. Wood, Balaji Rajagopalan, James Prairie, Carly Jerla, Edith Zagona, Robert A. Butler, and Rebecca Smith

Research Impact Statement: A new tool to assess water management in the Colorado River Basin will enable a systematic benchmarking of improvements in streamflow forecasts and reservoir system projections.

ABSTRACT: The Bureau of Reclamation (Reclamation) plays a central management role in the Colorado River Basin (CRB), with an increasing focus on meeting the needs of stakeholders during the current drought. One aspect of this role involves generating five-year projections of reservoir operating conditions in the federal multi-reservoir system. These projections are the basis for estimating the probability of shortage conditions, which are relied on by stakeholders, and are particularly important during drought. Currently, Ensemble Streamflow Prediction (ESP) forecasts drive Reclamation's Colorado River Mid-term Modeling System to produce probabilistic reservoir projections to be used in risk-based analysis and decision support for the first two years of the outlook period. The lack of significant forecast skill beyond the first year motivates interest in alternative forecasting approaches. The CRB Operational Prediction Testbed was created to provide a quantitative and consistent framework for assessing the skill of streamflow forecasts and their impact on associated reservoir system projections. Reservoir system projections are evaluated by analyzing Lakes Powell and Mead operations, including projected pool elevation and operating tiers. In an initial application of this testbed, ESP forecasts were compared to experimental streamflow forecasts to assess their skill impact on two-year reservoir projections, which are critical information for managing drought.

(KEYWORDS: benchmarking, streamflow forecasts; Colorado River; drought; reservoir operations; testbed; water resources management.)

INTRODUCTION

The Colorado River is a critical resource in the Western United States (U.S.) that is used to meet water supply, power, environmental, recreational, and cultural needs across seven states, 30 federally recognized tribes, and northern Mexico. In 1995, a special edition of the Journal of the American Water Resources Association contemplated how the Colorado River System could be managed under a hypothetical severe, sustained drought (JAWRA 1995). Since its publication, a multi-decadal drought has descended upon the Colorado River Basin (CRB), greatly diminishing the vast storage designed to weather periods of low flows. Recognizing this hydroclimate reality, a

Paper No. JAWR-21-0044-P of the *Journal of the American Water Resources Association* (JAWR). Received March 15, 2021; accepted May 13, 2022. **Discussions are open until six months from issue publication**.

Lower Colorado Basin Region (Baker, Butler, Smith), U.S. Bureau of Reclamation Boulder, Colorado, USA; Climate and Global Dynamics (Wood), National Center for Atmospheric Research Boulder, Colorado, USA; Department of Civil, Environmental, and Architectural Engineering (Rajagopalan, Zagona), University of Colorado Boulder Boulder, Colorado, USA; Upper Colorado Basin Region (Prairie), U.S. Bureau of Reclamation Boulder, Colorado, USA; U.S. Bureau of Reclamation (Jerla), Boulder, Colorado, USA; and Center for Advanced Decision Support for Water and Environmental Systems (Zagona), University of Colorado Boulder Boulder, Colorado, USA (Correspondence to Baker: sabaker@usbr.gov).

Citation: Baker, S.A., A.W. Wood, B. Rajagopalan, J. Prairie, C. Jerla, E. Zagona, R.A. Butler, and R. Smith. 2022. "The Colorado River Basin Operational Prediction Testbed: A Framework for Evaluating Streamflow Forecasts and Reservoir Operations." *JAWRA Journal of the American Water Resources Association* 58 (5): 690–708. https://doi.org/10.1111/1752-1688.13038.

new special edition of papers commemorating the 25th anniversary of the earlier collection has been assembled, to which this paper contributes. Since 1995, Lakes Powell and Mead, two reservoirs that make up over 80% of the storage in the CRB, declined from being over 90% full in 2000 to hovering around 30% full in early 2022, and are now nearing critical, drought-relevant operational thresholds (U.S. Bureau of Reclamation 2021). Beginning with the adoption of the Colorado River Interim Guidelines for Lower Basin Shortages and Coordinated Operations of Lake Powell and Lake Mead (Interim Guidelines) in 2007, multiple policies designed to manage the system through drought and low reservoir conditions have been implemented (U.S. Bureau of Reclamation 2020). Major components of these policies hinge on Bureau of Reclamation (Reclamation) projections of future reservoir elevations, and many stakeholders rely on mid-term projections (1- to 5-year lead times) of reservoir conditions to plan for their own future water use and management activities.

Reclamation uses operational planning models to produce outlooks of reservoir elevations and operating conditions five years to multiple decades into the future. Reclamation's Colorado River Mid-term Modeling System (CRMMS) (formerly referred to as the Mid-Term Operations Model, or MTOM) is one of the models used to produce probabilistic reservoir operational projections that provide risk-based analysis and decision support for five-year outlooks, though in the past, simulations for only the first two years were produced operationally. CRMMS is driven by streamflow forecasts produced by the National Weather Service (NWS) Colorado Basin River Forecast Center (CBRFC) using Ensemble Streamflow Prediction (ESP; Day 1985). ESP forecasts are produced using a hydrologic or land surface model initialized with recent weather information to represent current basin conditions, and then forced into the future forecast period with temperature and precipitation sequences. The future meteorological inputs typically incorporate weather forecasts for the first 5–15 days, followed by an ensemble of historical temperature and precipitation sequences that extend to the end of the forecast period. ESP forecasts can be highly skillful when initial watershed conditions such as observed snow water equivalent (SWE) and soil moisture strongly influence the forecast, but lack skill at longer leads (e.g., six months and longer) when climate forcings drive forecast skill (Wood and Lettenmaier 2008; Wood et al. 2016). The resulting ensemble of forecasted streamflow sequences, also called "traces" (Franz et al. 2003), outline the combination of estimated watershed conditions with uncertainty about future watershed climate inputs.

Skillful projections can give advance warning of the need for actions to mitigate future stresses on the system. Despite the central importance of these projections in preparing Reclamation and stakeholders to take action in the midst of severe, sustained drought, few substantive changes have been made to the forecasting methods that inform these operating policies since the models and methods were introduced decades ago. The current drought has underscored the need to improve the accuracy of future basin hydroclimate projections and storage conditions to meet the needs of CRB stakeholders. Although various CRB-focused studies have explored improvements to streamflow forecasts in the context of drought and a changing climate (e.g., Regonda et al. 2011; Lehner et al. 2017; Baker et al. 2021; Woodson et al. 2021; Towler et al. 2022), it has proven difficult to incorporate new approaches that might improve projections of future CRB system conditions without an effective and quantitative way to assess the new methods' potential value to reservoir operations. A structured approach is needed to ingest and evaluate the latest science, tools, and datasets to inform decisions regarding upgrades to improve mid-term projections. With this motivation, we have created the first objective testbed system and framework that can be used to evaluate the impacts of different streamflow forecasts on CRB reservoir operations. The Colorado River Basin Operational Prediction Testbed (CRBOPT) adopts a standard set of metrics that are applied to intercompare different inflow forecast approaches — the focus of this paper's demonstration.

This paper is organized as follows: the next section provides background perspective on streamflow forecasting and mid-term operations modeling and reservoir operating policies in the CRB. This is followed by a description of the CRBOPT framework, performance metrics, and the streamflow forecast approaches that are evaluated in this initial CRBOPT study. The results of CRBOPT are then presented, followed by discussion and conclusions.

BACKGROUND

Streamflow Forecasting

The water management community has long been interested in improved inflow forecasts, as well as refining the understanding of the skill of streamflow forecasts and their proper interpretation and value for resulting reservoir operations, with the ultimate goal of developing operating policies that optimally utilize hydrologic forecasts. Since the 1940s, U.S. federal agencies have provided reservoir inflow forecasts that are used in various ways by the reservoir management community (Pagano, Robertson, et al. 2014;

Pagano, Wood, et al. 2014; Wood et al. 2016; Lukas and Payton 2020). Reservoir management approaches vary by agency and institution and with the complexity of operational objectives; in some settings, operations policies have inherent flexibility to harness forecast information, while in others, operating policies are more constrained (i.e., involving fixed reservoir storage or elevation rule curves) and may minimize the risk over average hydrologic conditions but potentially sacrifice efficiency during unusual hydrologic regimes (Raff et al. 2013). Inflow forecasts have been used for multiple decades in varying degrees around the world to inform reservoir operations, with relatively more advancement and sophistication in private sector settings that have a financial incentive for operational efficiency (such as for hydropower) (Duan et al. 2018). In recent years, the term "Forecast-informed Reservoir Operations" (FIRO) has been popularized in the U.S. as a result of a California-based pilot initiative to improve reservoir management through balancing flood prevention releases with water storage retention for water supply with the aid of forecasts (Jasperse et al. 2017; Delaney et al. 2020). This expansion of reservoir management strategies in the FIRO project was facilitated, for the particular reservoirs of the pilot study, by a 2016 policy change of the U.S. Army Corps of Engineers to allow reservoirs to be operated in response to forecasts, vs. solely knowledge of "water on the ground" — that is, current observations (but not forecasts) of streamflow and reservoir levels.

Improving operations by utilizing streamflow forecasts ideally requires a detailed understanding of the skill characteristics of the forecasts as well as their potential impact on release decisions. Recent studies have demonstrated that the precise connection between streamflow forecasts and reservoir operations (projections of releases) can be complex and difficult to characterize comprehensively (Denaro et al. 2017; Turner et al. 2017; Giuliani et al. 2020). Anghileri et al. (2016) created a framework to explore the response of reservoir operations to streamflow forecasts at seasonal and inter-annual time horizons in snow-dominated river catchments. They found optimal reservoir management needed reliable streamflow forecasts at inter-annual leads, especially during periods of drought, while seasonal scale forecasts were less useful. Turner et al. (2017) assessed streamflow forecast impact based on the particular operating objective of a reservoir. They found that accurate forecasts substantially improved reservoir operations in reservoirs that operate to meet a target water elevation, while forecast accuracy did not necessarily translate into improved reservoir operations to meet a supply objective. These studies show that streamflow forecast skill does not translate linearly into improved reservoir operations and may need to be investigated based on the specific needs of each particular basin of interest. On the whole, these findings are not surprising: in more single-objective locations, there may be a direct relationship in which the release is a function of the storage and projected inflow, while in multi-objective locations or multi-reservoir systems, myriad factors in addition to projected inflows typically determine or constrain releases, and the constraints are system state dependent (i.e., becoming active near operating thresholds).

The general finding that skillful inflow forecasts can increase the efficiency of reservoir management has nonetheless been demonstrated in many settings. In a CRB-relevant example, Regonda et al. (2011) evaluated how increased streamflow forecast skill translates to improvements in operations and decision variables in the Gunnison River Basin, a tributary within the CRB. A nonlinear regression was used to create a multimodel ensemble streamflow forecast informed by large-scale climate information. Streamflow forecasts were then run through an operations model, which projected outflow, storage, and power production at Blue Mesa Reservoir. The study found that streamflow forecast skill transferred to operational skill at long lead times, though nonlinearly. Another study by Sankarasubramanian et al. (2009) investigated the role of streamflow forecasts produced by principal component regression and informed by monthly updated precipitation forecasts in a reservoir simulation model in the Philippines. Using forecasts was found to reduce spill, increase allocation for hydropower during above-normal years, and help meet end of season storage targets for below-normal years. Due to the readily achievable management benefits of inflow forecasts, they are widely used in reservoir operations, including the CRB.

CRB Reservoir Operations and Management

Unlike many other river systems, the CRB can store substantial volumes of water — estimated at about four times the historical annual flow (Christensen et al. 2004). Thus, the CRB has a large storage buffer to supply water during prolonged periods of drought, though this system reliability depends on periodic wet years to refill the system. This rare storage to inflow ratio has led to a focus on multi-year projections of inflow as operational decision inputs in the CRB. In contrast, common long-lead forecasts in the Eastern U.S., where hydrology does not have a pronounced seasonal summer dry period, have a three-month lead time, and most systems in the western U.S. that have only a storage for a year's supply rely on only seasonal to one-year long forecasts. In the CRB, the two-year projection of reservoir system conditions (the "24-Month Study") has long been an official information product for management, and more recently operational inflow forecasts issued from CBRFC and reservoir system projections issued by Reclamation extend to five years. The ability for the large storage and associated releases to have state-dependent predictability for five years (vs. systems that empty and refill every year), and the complex planning and inter-party negotiation (involving multiple state and treaty obligations) required to plan for and manage critical conditions has driven the demand for the CRB's unique multi-year predictions.

The importance of advancing capabilities for skillful inflow and system projections, particularly in the face of the current multi-decade drought and depleted storage, motivates the need for an objective framework that can be used to evaluate the impacts of different streamflow forecasts on CRB reservoir operations, to allow for benchmarking alternative strategies for improving such forecasts. This capability did not exist in prior years, and the adoption of new forecasting approaches has been relatively slow. The same watershed models and fundamental forecasting approach (ESP) have been used operationally for almost two decades for inflow prediction, albeit with software changes and improvements to meteorological inputs. Changes to such a critical information input for a complex and high-stakes, multi-party water resource are inherently challenging, but the inability to effectively quantify strengths and weaknesses of new forecasting models, data, and methods surely presents one hurdle to assessing the impact of new advances and making informed decisions about their adoption.

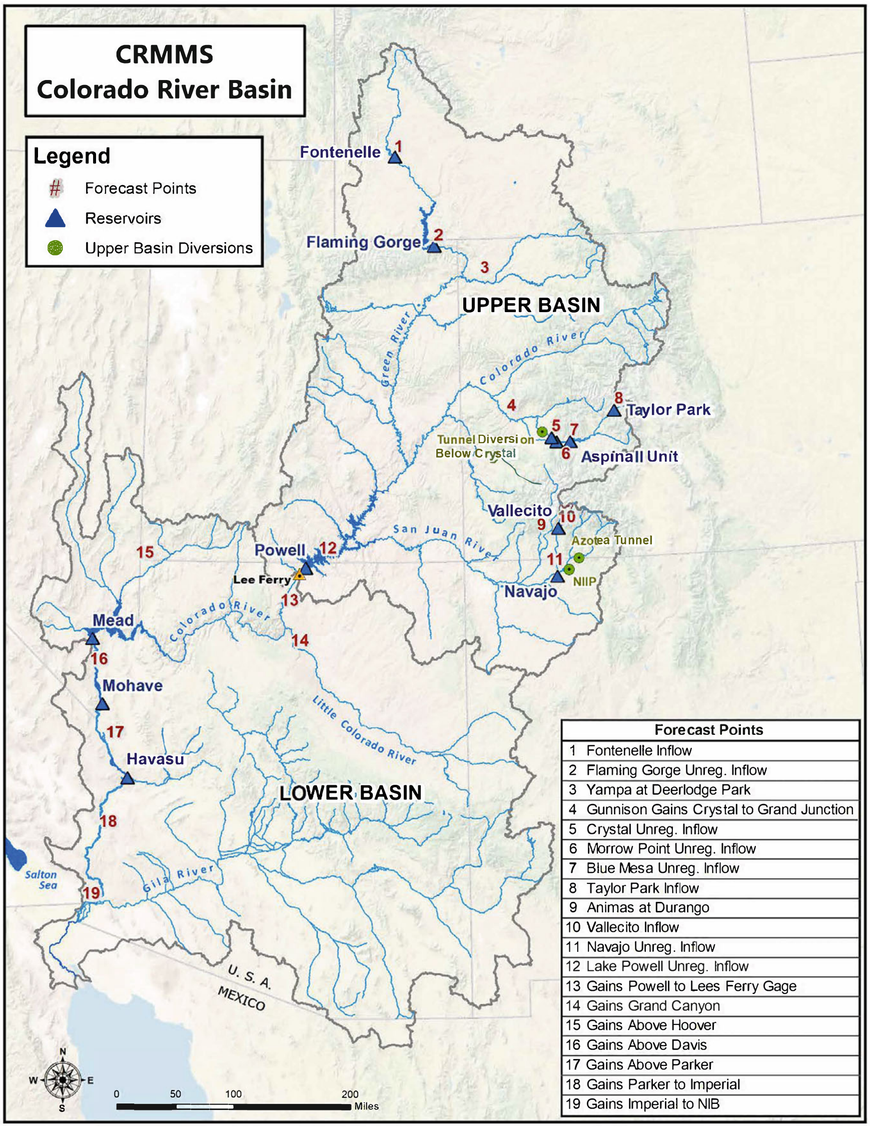
To address this situation, we have developed the CRBOPT for assessing the skill of both streamflow forecast inputs and the associated accuracy of CRMMS operational projections. This paper demonstrates the aspects of the CRBOPT capability through the analysis of a streamflow hindcast dataset (produced in a previous study from Baker et al. 2021) and associated operational projections. In particular, the water year (October 1 through September 30) accumulated flow is analyzed at lead times from 1 to 24 months ahead of the end of the accumulation period. The associated CRMMS projections of Lake Powell and Lake Mead operations, including pool elevation and outflow, and accuracy in predicting operating tiers and shortage and surplus conditions are analyzed with CRBOPT. CRBOPT is described in more detail in the following section.

CRB Water System Modeling

Colorado River Mid-term Modeling System is built in RiverWare, a generalized river basin modeling software platform (Zagona et al. 2001), and produce monthly reservoir projections and system conditions out five years, though we will focus on the two-year planning horizon for this analysis due to the type of forecasts that are analyzed. CRMMS can be run in two different modeling modes, Ensemble Mode and 24-Month Study Mode. The 24-Month Study Mode is a deterministic simulation used to produce the 24-Month Study, which is used for official operating decisions for operations of Lake Powell and Lake Mead. Reservoir operations for the 24-Month Study are manually set by reservoir operators. In contrast, CRMMS in Ensemble Mode, which we will refer to as CRMMS in this paper, is primarily used for risk-based analysis and planning with a focus on producing mid-term probabilistic projections of future CRB conditions. CRMMS uses rule-based "if-then" logic scripted using the RiverWare software to automate simulation of reservoir operations. Since CRMMS uses rule-based operations, it can be readily used to study how changes to model inputs, such as streamflow forecasts, affect reservoir operations.

The CRB is split into two subbasins with distinct geographic and climatic differences. The Upper Basin is the watershed above Lee Ferry, Arizona, where 80%–90% of the Colorado River flow originates from snowmelt in the Rocky Mountains (Christensen et al. 2004; Vano et al. 2012; Lukas and Payton 2020). The Lower Basin, located below Lees Ferry, is more arid and tends to experience convective storm systems that contribute flashy runoff and flow to the Colorado River. The CBRFC provides ESP forecasts at 12 locations in the Upper Basin, as shown in Figure 1. The streamflow forecasts are unregulated, meaning the forecasted flows are the streamflows that would have occurred if there were no regulation due to dams upstream of the forecast point. These forecasted unregulated flows include Upper Basin water use, which is incorporated into streamflow through calibration of the CBRFC's model, except for three tunnel diversions that are projected by their operators. The Lower Basin water uses are input to CRMMS with projected water use schedules. Lower Basin intervening flows at seven locations are set to sequences of historical flows in the operational CRMMS modeling, though in this study intervening flows will be set to historical values.

Colorado River Mid-term Modeling System simulates operations according to the "Law of the River," a collection of documents specifying how the Colorado River is managed and operated (Lukas and Payton 2020). The documents include interstate compacts, court decrees, the 1944 U.S.-Mexico Water Treaty, the 2007 Interim Guidelines, the 2019 Drought Contingency Plans, and other agreements relating to the use of the water of the Colorado River

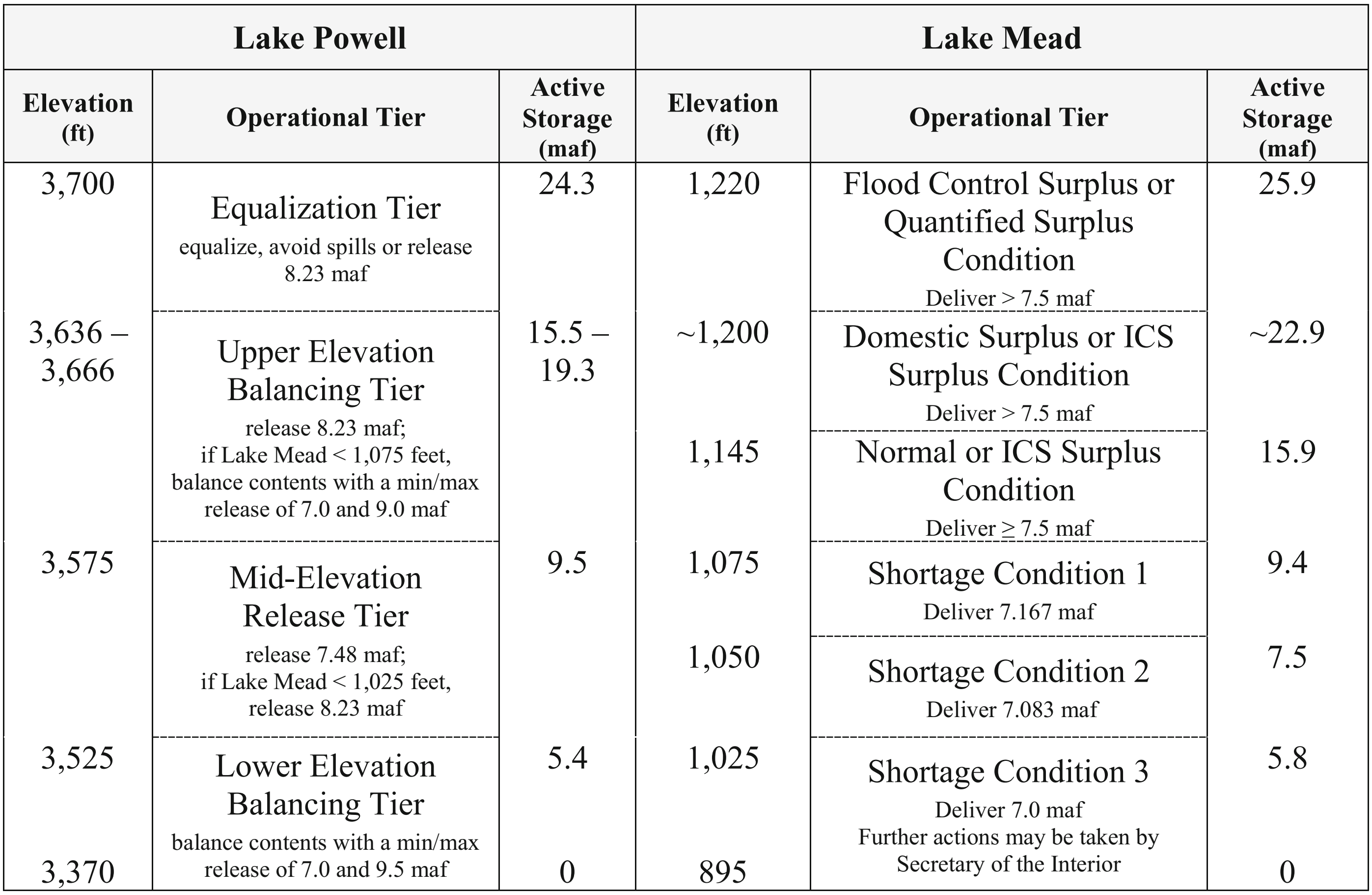


Forecast Points
1 Fontenelle Inflow
2 Flaming Gorge Unreg. Inflow
3 Yampa at Deerlodge Park
4 Gunnison Gains Crystal to Grand Junction
5 Crystal Unreg. Inflow
6 Morrow Point Unreg. Inflow
7 Blue Mesa Unreg. Inflow
8 Taylor Park Inflow
9 Animas at Durango
10 Vallecito Inflow
11 Navajo Unreg. Inflow
12 Lake Powell Unreg. Inflow
13 Gains Powell to Lees Ferry Gage
14 Gains Grand Canyon
15 Gains Above Hoover
16 Gains Above Davis
17 Gains Above Parker
18 Gains Parker to Imperial
19 Gains Imperial to NIB

FIGURE 1. Map of the Colorado River Basin with Colorado River Mid-term Modeling System (CRMMS) reservoirs, forecast points, and explicitly modeled Upper Basin diversions. A table in the bottom right describes the names of each numbered forecast location; forecast points 1–12 are in the Upper Basin and 13–19 are in the Lower Basin. The Aspinall Unit is a series of three reservoirs: Blue Mesa, Morrow Point, and Crystal.

(U.S. Bureau of Reclamation 2021). The Interim Guidelines provide operating criteria for Lakes Powell and Mead, including provisions designating potential delivery reductions, and allowing for greater flexibility to conserve and store water in the system (U.S. Department of Interior 2007). The coordinated annual operations of Lakes Powell and Mead, specified by operating tiers, are shown in Figure 2. Pool elevations are important criteria for the tier determinations at Lakes Powell and Mead. The end-of-year projections from the August 24-Month Study are used to set the Annual Operating Plan for the following year, which sets operations for Lakes Powell and Mead. The Annual Operating Plan can be changed mid-year due to an April adjustment, which is based on April 24-Month Study's projected pool elevations. An April adjustment can cause the annual operating tier to switch to Equalization or to balance the contents of Lakes Powell and Mead in the Upper Elevation Balancing Tier.

Lake Powell has four operating tiers that prescribe water year release volumes from Lake Powell. Release volumes are measured in million acre-ft (maf), a common volumetric measure used in U.S. water resources management, which represents 1 foot of water covering an acre of land, or 0.81 billion cubic meters. Lake Mead has three main operating tiers that specify deliveries to the Lower Basin states:



Lake Powell			Lake Mead		
Elevation (ft)	Operational Tier	Active Storage (maf)	Elevation (ft)	Operational Tier	Active Storage (maf)
3,700	Equalization Tier equalize, avoid spills or release 8.23 maf	24.3	1,220	Flood Control Surplus or Quantified Surplus Condition Deliver > 7.5 maf	25.9
3,636 – 3,666	Upper Elevation Balancing Tier release 8.23 maf; if Lake Mead < 1,075 feet, balance contents with a min/max release of 7.0 and 9.0 maf	15.5 – 19.3	~1,200	Domestic Surplus or ICS Surplus Condition Deliver > 7.5 maf	~22.9
			1,145	Normal or ICS Surplus Condition Deliver ≥ 7.5 maf	15.9

3,575	Mid-Elevation Release Tier
	release 7.48 maf;
	if Lake Mead < 1,025 feet, release 8.23 maf
		9.5	1,075	Shortage Condition 1
				Deliver 7.167 maf
					9.4

			1,050	Shortage Condition 2 Deliver 7.083 maf	7.5
3,525	Lower Elevation Balancing Tier balance contents with a min/max release of 7.0 and 9.5 maf	5.4	1,025	Shortage Condition 3 Deliver 7.0 maf Further actions may be taken by Secretary of the Interior	5.8
3,370		0	895		0

FIGURE 2. Schematic of the 2007 Interim Guidelines for the operating tiers of Lake Powell and Lake Mead with reservoir elevations, storage, and description of releases measured in million acre-ft (maf). Operating tiers are based on reservoir elevations at the end of the year. Lake Powell operating tiers set releases from Lake Powell. In operating tiers where balancing or equalization are specified, the releases from Lake Powell are set to the amount that would result in equal storage in Lakes Powell and Mead with release constraints based on tier. The elevation between the Equalization and Upper Elevation Balancing Tiers in Lake Powell increases each year between 2007 and 2026. Lake Mead operating conditions specify the available water for delivery or Intentionally Created Surplus (ICS) conditions in the Lower Basin.

shortage, surplus, and normal conditions. Under normal conditions, the 7.5 maf apportioned to the Lower Basin states is available for consumptive use; in surplus conditions, water in excess of 7.5 maf is available; and in shortage conditions, water less than 7.5 maf is available.

DATA AND METHODS

CRBOPT Framework

The CRBOPT framework, summarized in Figure 3, comprises multiple components. The first component ingests streamflow forecasts and then runs the flow forecasts through CRMMS to simulate future monthly reservoir operations (also termed reservoir system "projections"). This component of the CRBOPT is implemented using the RiverWare Study Manager and Research Tool (RiverSMART), a software that was created to facilitate large and complex planning studies and allows for simulating multiple alternative hydrology or demand scenarios (i.e., different inflow datasets and demand estimates), run start dates and run lengths, and different operating policies. CRBOPT uses the capabilities of RiverSMART to run several streamflow forecast ensemble datasets with varying numbers of traces through CRMMS to produce operational reservoir projections for the major reservoirs, for each of the inflow forecast datasets. The CRBOPT "metrics" components then run outputs from RiverSMART simulations through a series of scripts. These scripts calculate performance metrics for the streamflow forecasts and the reservoir projections. The linkage of these components forms the testbed, which in this case involves the configuration of RiverSMART with the CRMMS model.

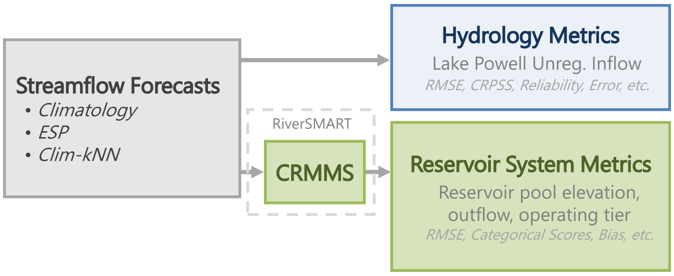

Streamflow Forecasts
• Climatology
• ESP
• Clim-kNN
RiverSMART
CRMMS
Hydrology Metrics
Lake Powell Unreg. Inflow
RMSE, CRPSS, Reliability, Error, etc.
Reservoir System Metrics
Reservoir pool elevation, outflow, operating tier
RMSE, Categorical Scores, Bias, etc.

FIGURE 3. Diagram of the Colorado River Basin Operational Prediction Testbed layout. Streamflow forecasts input to CRMMS. This study specifically assesses the three forecasts named in the left box: Climatology, ensemble streamflow prediction (ESP), and Climate Informed k-nearest neighbors (Clim-kNN). The streamflow forecasts are run through the water management model, CRMMS, using the RiverWare Study Manager and Research Tool (RiverSMART), which allows for many streamflow types and forecast dates to be efficiently run through RiverWare. The streamflow forecasts and reservoir operations output from CRMMS are then analyzed with scripts through two different sets of metrics that focus on different aspects of performance. The text in italics are example entries that are tested in this study and could be adjusted for future studies. CRPSS, Continuous Ranked Probability Skill Score.

The study described in this paper illustrates the CRBOPT use through comparing different Upper Basin streamflow forecasts and associated reservoir system outcomes, as an estimated 80%–90% of the flow in the CRB originates in the Upper Basin. The forecasts assessed in this study include ESP (the current forecasting method provided by CBRFC and used by Reclamation), Climatology, and an experimental forecast termed "Climate Informed k-nearest neighbors" (Clim-kNN), which are described in the following section. Upper Basin streamflow forecasts are input to CRMMS, which simulates reservoir operations according to the current implementation of the "Law of the River." A hindcast version of CRMMS that represents reservoir operations logic back to 1981 was used in CRBOPT. To make the forecast comparison relevant to the present day, we applied current-day operations for the entire hindcast period. CRBOPT hindcast simulations are initialized on the eighth day of each month from water year 1982 through 2016 and run for a forecast period of two years. Since the climate data used to create the Clim-kNN forecasts are not available until the eighth day of each month, forecasts would only available after this, though ESP is typically available earlier in the month.

The Interim Guidelines were implemented in 2008, though the hindcast period dates back to water year 1982. To explore operational skill for the longer hindcast period, we extended the set of historical "observed" reservoir operations. A series of reservoir operations approximating what would have occurred if the Interim Guidelines had been in place since water year 1982 was created by running CRMMS with historical (observed) inflow values in the Upper Basin. That is, the model was given "perfect" (to the extent possible) streamflow information and the resulting operations were used in place of observed operations for the purpose of comparisons in this study. These simulations start on each date of the hindcast range and are used to evaluate reservoir operational projections throughout the hindcast period. We note that CRBOPT greatly benefits from the earlier development of CRMMS, a significant effort to codify reservoir operations policies and rules into an automatable system model, as well as the ability to automate and generate streamflow hindcasting. A testbed-like approach such as CRBOPT could include systems models and forecasts of varying complexity (including, e.g., reservoir optimization procedures), provided they can be automated and run in a hindcast model to generate sufficient numbers of trials for a robust assessment.

Streamflow Forecast Datasets

The CRBOPT was used to compare several alternatives in streamflow forecasting methods described below. The forecasts for water year flow are made for lead times that both precede the water year and also are within the predicted water year. In the latter case, the observed flows for the water year up to the time of the forecast are combined with the forecast flows for the portion of the water year remaining in the future. All streamflow forecast datasets were available on the eighth of the month for each month of water years 1982–2016 (i.e., October 8, 1981 through September 8, 2016). These datasets are termed hindcasts to denote forecasts that were made for past dates.

Climatology. The observed historical streamflow ("Climatology") was used as a forecasting baseline to determine whether forecasts perform better than an estimate of forecast period flow derived from the historical record. Climatology ensembles are assembled from historical flows (in this case, water years 1981–2010), with each ensemble member or "trace" being a historical flow sequence from a given year starting on the date of the forecast and extending through the forecast lead time. These ensemble hindcasts thus contain 29 or 30 traces, due to ensembles that contain a verification year dropping the observed year's trace to avoid including a trace with perfect knowledge of the streamflow. When used as a forecast here, Climatology begins to show skill for the water year prediction for lead times shorter than 12 months, when prior observed flows are combined with the forecast.

Climatology-Reference. A version of Climatology where observed flows are not included is used as a reference forecast against which the skill of the three forecast alternatives are calculated. It is a forecast of a distribution of inflows equal to historical inflows, leaving out the hindcast year. This baseline is "naive" in that it cannot distinguish the hydroclimate of one year from another, for example, the ensemble spread mean are similar for every year's forecast, allowing for the removed trace.

ESP. Ensemble Streamflow Prediction forms the current official streamflow forecasting method used when running CRMMS. ESP forecasts used in this study were produced by the CBRFC using the NWS Snow-17 and Sacramento Soil Moisture Accounting (Sac-SMA) models (Burnash et al. 1973), which were calibrated by CBRFC to reproduce historical conditions from 1981 to 2010. Streamflows were produced by running the watershed models with historical temperature and precipitation sequences from the climatological period of record (1981–2010) out 60 months, though only the first two years of the forecast are assessed in this study. Archives of published ESP operational forecasts were only available from 2011 to present, which necessitated the generation of a longer period of hindcasts (from 1981 to 2010); these were contributed by CBRFC for the CRBOPT development effort. The ESP hindcasts differ from operational ESP forecasts in that they do not include short-term temperature and precipitation forecasts or real-time forecaster modifications to model states. Nonetheless, the hindcasts viewed by the NWS as providing a useful though approximate indicator of operational ESP forecast skill (Wells et al. 2011). The ESP hindcast ensembles used in this study have 29 or 30 traces corresponding to the 30 years of historical precipitation and temperature traces, with the trace depending on the forecasted year's climate inputs (i.e., temperature and precipitation sequences) removed from the ensemble to avoid including a trace with perfect knowledge of the climate.

Clim-kNN. Baker et al. (2021) created and evaluated the disaggregated basin k-nearest neighbors streamflow forecasting method, referred to as Clim-kNN in this study. The Clim-kNN trace-weighting approach weights ESP traces using North American Multi-Model Ensemble (NMME) 1- and 3-month watershed-scale temperature and precipitation forecasts, along with the preceding three-month average observed streamflow. Using the information from NMME, which is available by the eighth day of each month, ESP traces are conditionally resampled, such that ESP traces with historical climate that more closely match NMME forecasts are weighted higher during the resampling, to create new ensembles containing 100 traces of inflows. The method is performed on four Upper CRB sub-basins: the Main Stem, Green, Gunnison, and San Juan. See Baker et al. (2021) for a detailed description of the method.

Performance Metrics

The CRBOPT calculates a number of statistical measures (metrics) to analyze the performance of the input hydrology (i.e., the streamflow forecast itself) and of the CRMMS reservoir system variables (e.g., releases, storages) arising from running the streamflow forecasts through CRMMS. Performance metrics have been used for forecast and model evaluation in many fields for decades and can be found in many sources from handbooks and textbooks (e.g., Wilks 2011; Duan et al. 2018) to agency guidelines as well as academic literature. Specific fields such as reservoir system design may have unique metrics, such as the reliability, resilience and vulnerability metrics introduced by Hashimoto et al. (1982). In a typical forecasting paper, authors select a small number (e.g., 1–5) of performance metrics that illustrate behavior with respect to different characteristics such as forecast bias, correlation, reliability or spread errors, and different sets of metrics are appropriate for different types of forecasts (e.g., deterministic, probabilistic, categorical). CRBOPT is expected to expand as it is used for future testing of different types of inputs; thus, the metrics shown in this study represent an initial set of potential metrics that can be calculated by CRBOPT. For the demonstration of CRBOPT presented in this paper, the metrics were evaluated for a 24- to 1-month lead time from the projection target date, which is the end of the second water year (September 30).

Hydrology Metrics. Forecast performance is measured in CRBOPT for different forecast attributes. Due to the ubiquity of published verification information, here we briefly summarize several common forecast attributes and concepts (Wilks 2011; Duan et al. 2018) that we select metrics to measure. Note, many of these forecast performance attributes can be described by more than one statistical metric.

Accuracy is a concept reflecting the strength of agreement between a deterministic forecast (or ensemble forecast central tendency) and observations. Accuracy is a measure of overall quality and may be measured by a number of statistical metrics, including correlation or error (for which various forms exist, e.g., mean absolute error, root mean squared error [RMSE], and their relative forms).

Bias, or unconditional bias, is a measure of the average error of the forecasts as calculated by the difference between the mean forecast and mean observations. Bias and accuracy differ; bias describes the overall difference between average forecasts and average observations, while accuracy is the average difference between individual forecasts and observations.

Reliability, or conditional bias, is a measure of the agreement between the forecast probabilities and the observed frequency of an event. Reliability characterizes the conditional distribution of the observations given a set of forecasts.

Resolution is a measure of the ability of forecasts to resolve the set of sample events into a subset of different outcomes. Forecasts that are nearly the same but have two different outcomes are said to have poor resolution, while forecasts that are different and exhibit different observed outcomes have good resolution. Discrimination is another measure that relates to resolution, which is a measure of how a forecast system can discriminate between two different events. In probabilistic forecasts, forecasts with no resolution have no discrimination and vice versa (Bröcker 2015).

Sharpness refers to the relative spread of a forecast and is a measure of the forecast alone (vs. its correspondence with observations). Forecasts that have a similar spread to climatology are said to have low sharpness, whereas forecasts with spread much narrower than climatology are sharp. Forecasts that are too sharp may fail to include the observed event at forecasted frequencies, in which case their reliability attribute would be poor. Such forecasts are often described as being overconfident or under-dispersed, respectively.

Skill refers to the performance of a forecast relative to a reference forecast and is calculated through a skill score formulation that compares the two and translates the result into a format that can be interpreted as a degree of improvement (e.g., percent improvement). The actual metric used in the skill calculation for both the forecast and the reference can vary but is often an accuracy metric.

In this study, we demonstrate the assessment of streamflow forecast inputs to CRBOPT by evaluating the annual water year Lake Powell unregulated inflow. The inflow to Lake Powell is an aggregate of all Upper Basin forecast locations and is therefore useful in evaluating the overall quality of each forecast. We report results for RMSE, a continuous rank probability skill score, and an ensemble spread visualization. These hydrology metrics are briefly described below. Other skill metrics, including rank histograms for reliability, were evaluated while exploring the forecasts, though results are not shown here. Not all attributes discussed above are reported through metrics in the demonstration selected for this paper.

Root mean squared error is the square root of the squared differences between individual ensemble traces (y_i) and observations ($\hat{y}$) divided by the number of traces in the ensemble (n); hence, it is the ensemble mean RMSE for each forecast.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y})^2}{n}}. \tag{1}$$

Since the errors are squared, larger errors have a greater influence on RMSE than smaller errors. RMSE is a measure of accuracy, and another common calculation of RMSE is for the ensemble mean vs. observation.

Continuous Ranked Probability Skill Score (CRPSS) measures the accuracy of a probabilistic or ensemble forecast relative to the accuracy of a reference forecast such as climatology (Hersbach 2000). The CRPSS contains the continuous ranked probability score (CRPS) metric, which is the integrated squared difference between the cumulative distribution function of the forecasts and the corresponding distribution of the observations. CRPS is similar to the common ranked probability score (RPS) except it uses a continuous distribution instead of categories, and the units of the CRPS are the units of the forecast variable. Like many skill scores, the CRPSS ranges from 1 (perfect) to $-\infty$, where a score of 0 means the skill of the forecast is equal to that of the reference forecast, and a negative score means the forecast is less skillful than the reference. The climatology-reference is used as the reference forecast in CRPSS calculation.

In addition to summary forecast verification metrics, graphical verification is also a valuable element of forecast or model performance evaluation. CBROPT generates an *Ensemble Spread Visualization* — a scatter plot view of the forecast vs. observations in which the forecast ensemble is represented using box and whisker symbols. Each symbol denotes one ensemble forecast at a specified lead time. The visualizations illustrate forecast spread, sharpness, bias, and discrimination at multiple lead times. A forecast with minimal bias and acceptable reliability would typically straddle the 1:1 line, indicating that the ensemble forecast range contains the observed flow. An "overconfident" forecast can be seen to be under-dispersed, with overly narrow spread, while the reverse is also possible. Ideally, an ensemble forecast is systematically unbiased and as sharp (narrow) as possible (so as to be able to discriminate between high and low flow events) while maintaining a spread that verifies with correct frequency against observations. For example, to have reliable spread, the 10th to 90th percentile range of the forecast should enclose the observations 80% of the time, and neither more nor less.

Reservoir System Metrics. Operational reservoir system metrics are designed to assess and intercompare operational projections resulting from variations in inputs (e.g., flow forecasts) and reservoir operations policies. We demonstrate this CRBOPT evaluation capability to compare different streamflow forecasting strategies, focusing on the ability to project Lakes Powell and Mead end-of-water year pool elevation, using the RMSE as the performance metric. In addition, because the pool elevations are operationally significant from the standpoint of different pool elevation thresholds, categorical forecast metrics are particularly appropriate — that is, the ability to predict the elevation category that the pool elevation will reach in the future accurately. Categorical scores, including the Percent Correct and Heidke Skill Score (HSS), were calculated on projected operating tiers at Lakes Powell and Mead per the Interim Guidelines.

Percent Correct is a categorical score ranging from 0% to 100% that measures accuracy as the percent of ensemble members (traces) for which the model projects the correct operating tier for each reservoir.

Heidke Skill Score is a categorical score that assesses the accuracy of the forecast in predicting the correct operating tier relative to that of random chance. The HSS is a measure of skill for categorical events (Heidke 1926). The score ranges from 1 to −∞, where 1 is a perfect skill score, 0 indicates skill equal to random chance, and negative values indicate skill worse than random chance.

We note that performance of the reservoir system projections is determined both by the accuracy of the inflow forecasts and the agreement between the CRMMS representation of operational policies compared to actual operations. In this study, we do not attempt to separate the two contributions to potential forecast skill or error, but rather illustrate how CRBOPT can be used to quantify the impacts of choosing between several different inflow forecasting strategies.

RESULTS

Hydrologic Forecast Comparison

The streamflow forecasts were run through the hydrology metrics scripts to analyze the annual water year unregulated inflow to Lake Powell. Inflows on a water year basis were used in this analysis because a water year is a widely used time scale to assess river flow in water management, and it is operationally important in the Upper Basin since Lake Powell operates to meet a water year release volume. The RMSE of Climatology, ESP, and Clim-kNN are shown in Figure 4 for a 24- to 1-month lead time. In this paper, we define the lead time to be the time from the initialization of the forecasts to the end of the forecast volume accumulation period so that all lead times shown in the analyses below are positive. In other forecast literature, the convention of defining lead time from the forecast date to the start of the forecast period is also common. Since we are computing an annual flow volume, the forecast includes observed flows once the lead time is less than 12 months — that is, the forecast is initialized within the predicted annual flow period. Therefore, as the lead decreases beyond the 12-month initialization, more months of observed flows are included in the forecast causing the skill and accuracy to increase.

In the out-year or second year, defined as leads of 24 to 13 months, forecasted RMSE values remain relatively constant. A forecast's errors start to decrease into the forecasted water year at leads less than 13 months. This is partially due to observed flows being incorporated into the water year volume but is also largely due to knowledge of antecedent basin conditions such as soil moisture. Early season soil moisture often contributes moderately to streamflow forecast accuracy and skill (Wood and Lettenmaier 2008; Koster et al. 2010; Wood et al. 2016), with fall soil moisture impacts thought to contribute a minor influence (up to approximately 10%, depending on location) of the spring runoff volume variability (Harpold et al. 2017). The seasonal (April–July) runoff or flow predictability associated with both initial soil and snow conditions can be as high as 90% in terms of forecast variance explained or reduction in error (Franz et al. 2003; Pagano et al. 2009) with the obtainable skill depending on variations in watershed hydroclimate and the date of forecast. The effect of antecedent basin conditions can be seen with ESP forecast improvements over Climatology in the fall of the forecasted year. CRB streamflow forecast accuracy is highly dependent on snow in the high mountain regions in the spring. This relationship is highlighted for ESP and Clim-kNN forecasts in late winter from January at a nine-month lead to April at a six-month lead, as errors decrease substantially compared to Climatology.

Throughout the 24-month leads, the Clim-kNN forecast is more accurate than other forecasts. At shorter leads (i.e., April six-month lead to the end of the water year), the Clim-kNN forecast has smaller errors than ESP, but the relative improvement is

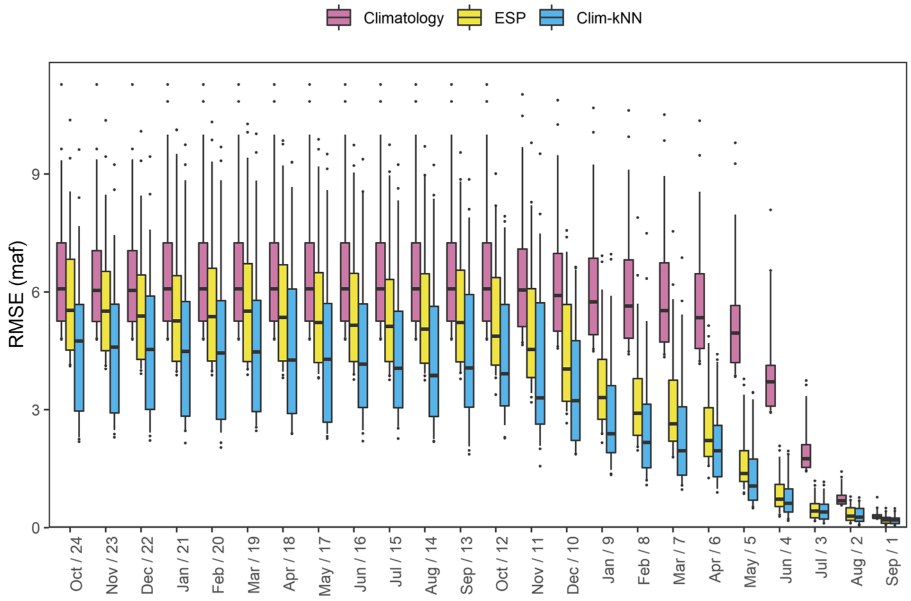


FIGURE 4. Root mean squared error (RMSE) of water year Lake Powell unregulated inflow. The RMSE at a 24- to 1-month lead is compared for Climatology, ESP, and Clim-kNN. The forecasts are available from 1982 to 2016. Each boxplot includes one data point for each year, totaling 35 data points. The x-axis shows the "month/number of lead months" to the end of the water year.

smaller than at longer leads. ESP and Clim-kNN both have improved accuracy over Climatology for all leads as they are informed by initial conditions or, in the case of Clim-kNN, by NMME climate forecasts. Clim-kNN forecasts are found to be better than the ESP forecasts for the RMSE metric at a statistical significance level of p = 1e-7 depending on lead time, using a two-sided t-test.

At long leads from 24 to 13 months, the CRPSS of most forecasts are close to climatology (zero line), as seen in Figure 5. The Clim-kNN forecast has a larger range of skill compared to ESP, with ESP and Clim-kNN forecasts having median skill close to zero. Climatology is being compared to itself, though the observed trace is dropped from the Climatology forecast, introducing a small amount of noise to the skill. When only considering forecast skill, there is no benefit to using the ESP or Clim-kNN forecast over Climatology during this out-year period.

The skill of both the ESP and Clim-kNN forecasts increases above Climatology in the fall of the out-year, starting in November at an 11-month lead. This is consistent with RMSE results and show that soil moisture, SWE, and other initial basin conditions have a positive impact on the forecast. As the season progresses through winter and early spring (leads of 10 to 6 months), the CRPSS of ESP and Clim-kNN increases as more snow accumulation is observed, which represents a storage of water that will melt during the runoff season. The skill continues to increase through the runoff season as more months in the annual inflow are observed.

From the CRPSS perspective, the median skill of the Clim-kNN method is slightly higher than the ESP forecast during certain months in the current year (e.g., December, January, and June at the 10-, 9-, and 4-month leads), but during most of this period their skill is roughly equal. The range of skill with the Clim-kNN method is larger than ESP at longer leads until May at a five-month lead. The skill of Climatology performs poorly since it has no knowledge of initial conditions in the basin or what the future climate may look like. During the last two months of the water year, Climatology performs slightly better than ESP and Clim-kNN forecasts, which can too tightly constrain forecasted streamflow for August and September.

A visualization of annual water year Lake Powell unregulated inflow spread for Climatology, ESP, and Clim-kNN is shown in Figure 6. At longer leads of 24 to 12 months, the ESP and Climatology forecasts are very similar and lack discrimination with all forecasts projecting similar flows. The Clim-kNN forecasts have a smaller spread than the ESP forecasts (i.e., smaller boxplot range), showing that the forecast may be too sharp and overconfident for such long leads. The smaller spread effects the accuracy of the forecast, resulting in higher accuracy for Clim-kNN as shown by lower RMSE values in Figure 4. The ESP forecasts start to discriminate from Climatology in the fall (October at a 12-month lead), with the ESP forecasts range and median forecasting different ensemble medians and range compared to the Climatology ensemble. By a 10-month lead in December,

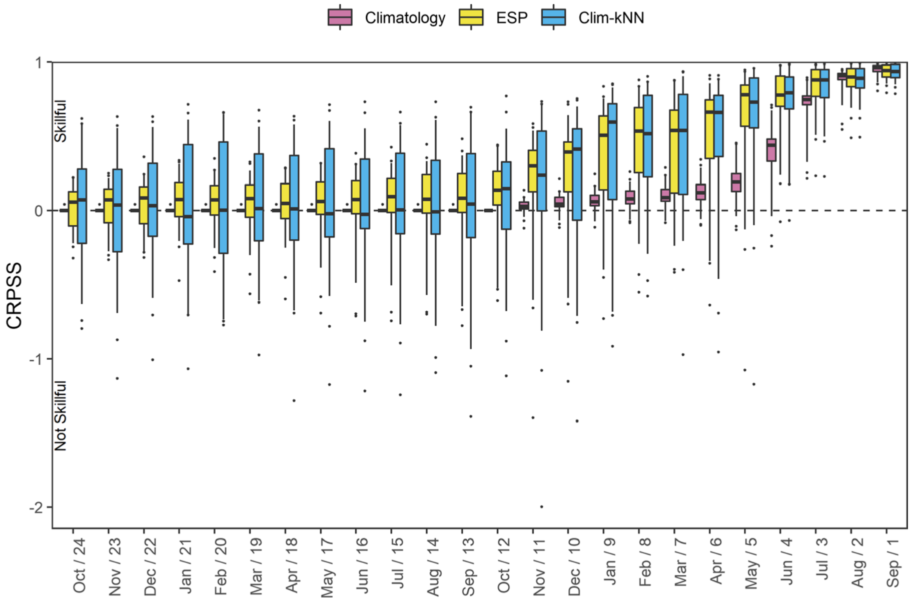


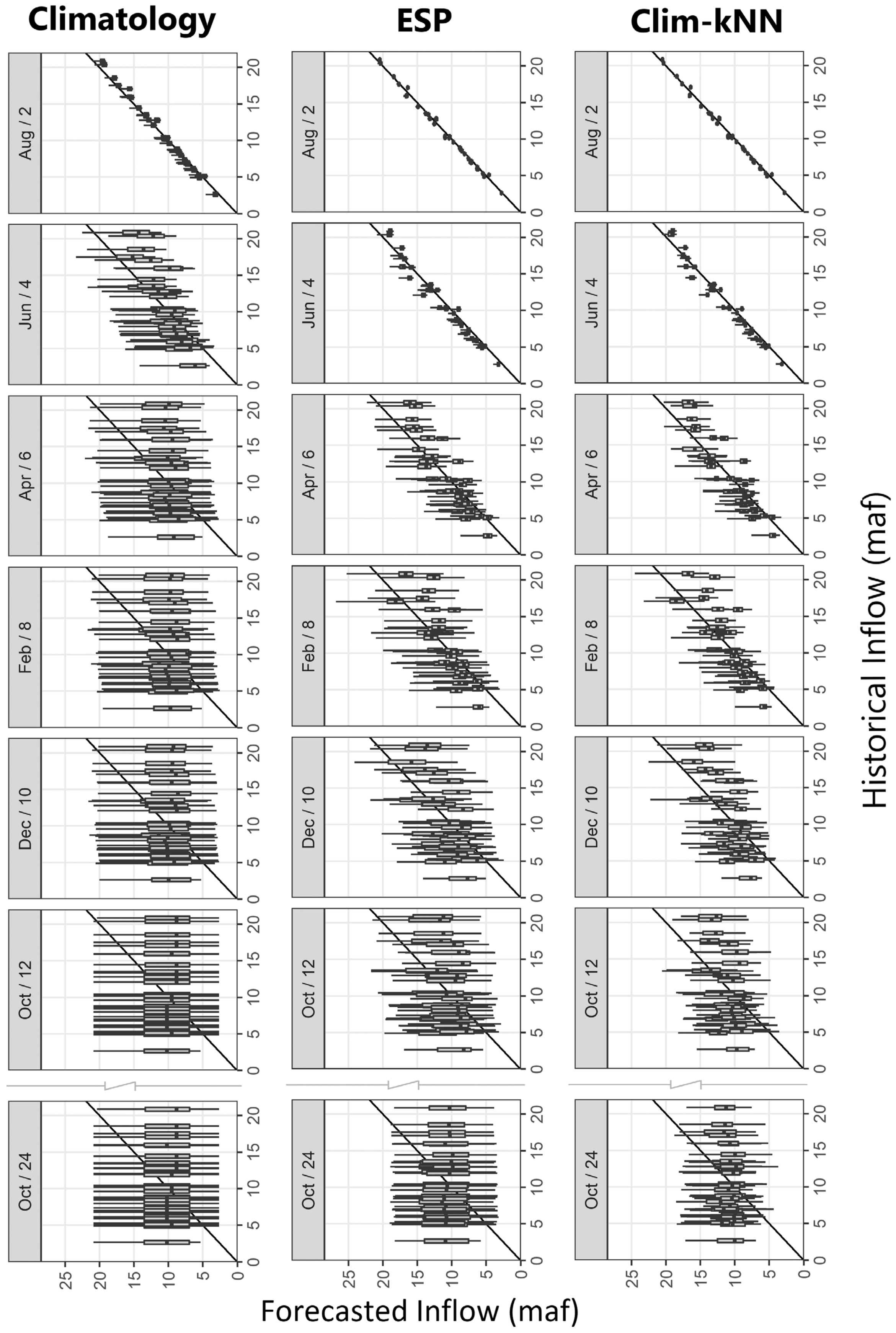
FIGURE 5. CRPSS of water year Lake Powell unregulated inflow. CRPSS at a 24- to 1-month lead is compared for Climatology, ESP, and Clim-kNN. The forecasts are available from 1982 to 2016. Each boxplot includes one data point for each year, totaling 35 data points. The *x*-axis shows the "month/number of lead months" to the end of the water year.

the ESP and Clim-kNN forecasts tighten; forecasts in the highest observed years starting to discriminate from lower and average flow years as shown by boxplots move closer to the observed flows. The spread of the Climatology forecast remains large, as the forecast only narrows when large flows are observed (e.g., starting in the runoff season). The smaller spread of the ESP and Clim-kNN forecasts compared to Climatology is the attribute that causes these forecasts to have higher accuracy, as shown as a lower RMSE in Figure 4. This is also apparent when comparing Clim-kNN and ESP, as the range of Clim-kNN is narrower at all leads causing higher accuracy compared to ESP. The reduction in spread does not necessarily improve the forecast performance, as shown by CRPSS (see Figure 5). The Clim-kNN skill is more variable than ESP, especially at longer leads as shown by the wide box and whisker range in the out-year, which is not a desired forecast quality.

By February (eight-month lead), the ESP and Clim-kNN forecasts' spreads narrow, especially the 25th–75th quartiles of the ensembles. The Clim-kNN forecast spread remains narrower than the ESP forecast through the end of the water year. Both forecasts capture the high inflows well as seen by the boxplots of higher observed inflows (greater than ~12 maf) moving upwards toward forecasting an ensemble of above average inflows. In April at a 6-month lead, both forecasts' spreads have narrowed significantly, in some instances to exclude the observed streamflow. The Clim-kNN forecasts have a narrower spread, with median flows closer to the 1:1 line, though the ESP forecast's spreads capture observed flows better. By June (four-month lead), forecasts show overconfidence; many ensembles are too narrow and do not capture the observed streamflow. At shorter leads than June, it is hard to discern the spread of the forecasts since most forecasts have converged on the 1:1 line. Because it is based on ESP, Clim-kNN relies on a strategy that does not represent uncertainty or bias in modeled initial watershed conditions, as discussed in Wood and Schaake (2008). All forecasts made via ESP or ESP-based methods therefore suffer from overconfidence at times in the year when the predictability associated with initial conditions is high, such as the spring snowmelt season or early summer in the Western U.S.

The most extreme years in this hindcast period are difficult to capture in the ensemble spread at extended lead times since the trace containing the historical weather that produced the extreme streamflow was removed from the forecast. For instance,

FIGURE 6. Ensemble spread visualization of water year Lake Powell unregulated inflow for Climatology, ESP, and Clim-kNN with boxplots of the forecast vs. the observed Lake Powell water year annual unregulated inflow (1982–2016) for leads of 24, 12, 10, 8, 6, 4, and 2 months. The boxplot's whiskers represent the full range of the forecast, the box is the 25th–and 75th quantiles, and the midline represents the forecast median. The shortest lead times appear in the scatter plots toward the top of the figure. The shortest lead times appear in the scatter plots toward the top of the figure.


Climatology
ESP
Clim-kNN
Aug / 2
Jun / 4
Apr / 6
Feb / 8
Dec / 10
Oct / 12
Oct / 24
0
5
10
15
20
25
Forecasted Inflow (maf)
Historical Inflow (maf)

2002 has the lowest streamflow in the analyzed period. The 2002 ESP ensemble (farthest left in boxplot) does not capture the observed streamflow (1:1 line) because no weather traces from ESP climatology (1981–2010) were as dry or close to as dry as the 2002 trace. Therefore, the forecast cannot produce such low streamflow until very dry initial basin conditions drive the forecast, as opposed to most of the signal coming from precipitation and temperature. This limitation is also a common concern with ESP, in that the meteorological drivers of a future streamflow prediction are taken only from historically observed sequences, which can fail to include sufficient extreme members to characterize their risk of occurrence.

Reservoir System Projection Evaluation

Reservoir system projections resulting from running streamflow forecasts through CRMMS using RiverSMART are evaluated using reservoir system metrics scripts, which have the capability of processing many different reservoir variables. This analysis will compare projected end-of-water year pool elevation and annual operating tiers as these variables provide a summary of other reservoir related variables including reservoir outflow.

The RMSE of projected Lakes Powell and Mead end-of-water year pool elevations for Climatology, ESP, and Clim-kNN forecasts are compared in Figure 7. Lake Powell has larger RMSEs at all leads compared to Lake Mead because the inflows to Lake Mead are controlled by Lake Powell, which releases a smaller range of flows compared to the potential variability of Lake Powell inflow. At all leads, Clim-kNN outperforms ESP, and Climatology performs the worst. This result is consistent with the RMSE of Lake Powell unregulated inflow (Figure 4), as the inflow is the main factor in projected pool elevation. For Lake Powell at a 24-month lead, the RMSEs for the forecasts are large with a median RMSE of 44.0, 35.6, and 24.6 ft for pool elevation and 4,920, 3,990, and 2,770 kaf for storage for Climatology, ESP, and Clim-kNN, respectively. The RMSEs for ESP and Clim-kNN are much smaller than Climatology from January through July of the current year, as streamflow forecasts gain more skill and sharpness. By April at a six-month lead, the median RMSE for forecasts have decreased to 27.3, 12.2, and 7.8 ft (or 3,220, 1,340, and 880 kaf) for Climatology, ESP, and Clim-kNN, respectively. The RMSE continues to decrease through the end of the water year, with the errors from ESP and Clim-kNN converging to similar values.

For Lake Mead, RMSE decreases relatively linearly with median RMSE values at a 24-month lead of 25.3, 19.5, and 14.0–11.6, 3.5, and 1.7 ft at a six-month lead for Climatology, ESP, and Clim-kNN, respectively. For median RMSEs in storage, this is equivalent to a decrease from 2,950, 2,300, and 1,580 kaf to 1,360, 433, and 211 kaf for Climatology, ESP, and Clim-kNN, respectively. Similar to the Lake Powell results, the Clim-kNN forecast has the lowest errors, followed by the ESP forecasts. Climatology has a slightly different trend in reduction of RMSE compared to the other forecasts, with minimal decrease in RMSE until the runoff seasons during the out-year and the current year. This is because the Climatology forecast does not have accuracy until part of the runoff volume has been observed. By the end of the runoff season, ESP, and Clim-kNN have very small errors. One of the main reasons for this is that Lake Powell's releases for the remainder of the water year are mostly known after April.

Another important performance metric to assess is correctness of the projected operating tiers for Lakes Powell and Mead, as compared to the observed tiers, given reservoir operations. The operating tiers, which are defined in the Interim Guidelines and summarized in Table 1, determine the releases from Lake Powell. The operating condition at Lake Mead sets deliveries in the Lower Basin. We evaluate the forecasted operating tiers using categorical scores based on the tier alone as well as the combined tier and Lake Powell release or Lake Mead condition. The Lake Powell releases in each release category can be within a given range as exact release volume can vary within a given tier.

The forecast metrics Percent Correct and HSS are categorical metrics used to analyze projections of operating tiers and releases or conditions for Lakes Powell and Mead. The categorical metrics are calculated from a multi-category contingency table that represents the frequency of forecasts and observations in each category for the 35 forecasts. The metrics are evaluated on two different contingency tables, each representing a different scale and represented by the columns in Table 1. The "Tier" results are based on broad categories and therefore have higher scores than the "Tier & Release/Condition" categories, which require the simulation to correctly determine the release for Lake Powell or condition for Lake Mead as well as the operating tier. The categorical scores for Climatology, ESP, and Clim-kNN streamflow forecasts are compared to historical streamflow projected operations for leads in January, April, and August of the out-year for 1982–2016.

The Percent Correct results in Table 2 show that Lake Mead projections perform better than Lake Powell projections when predicting both the tier and release or condition. This is expected as inflows to Lake Mead are mostly determined by Lake Powell's

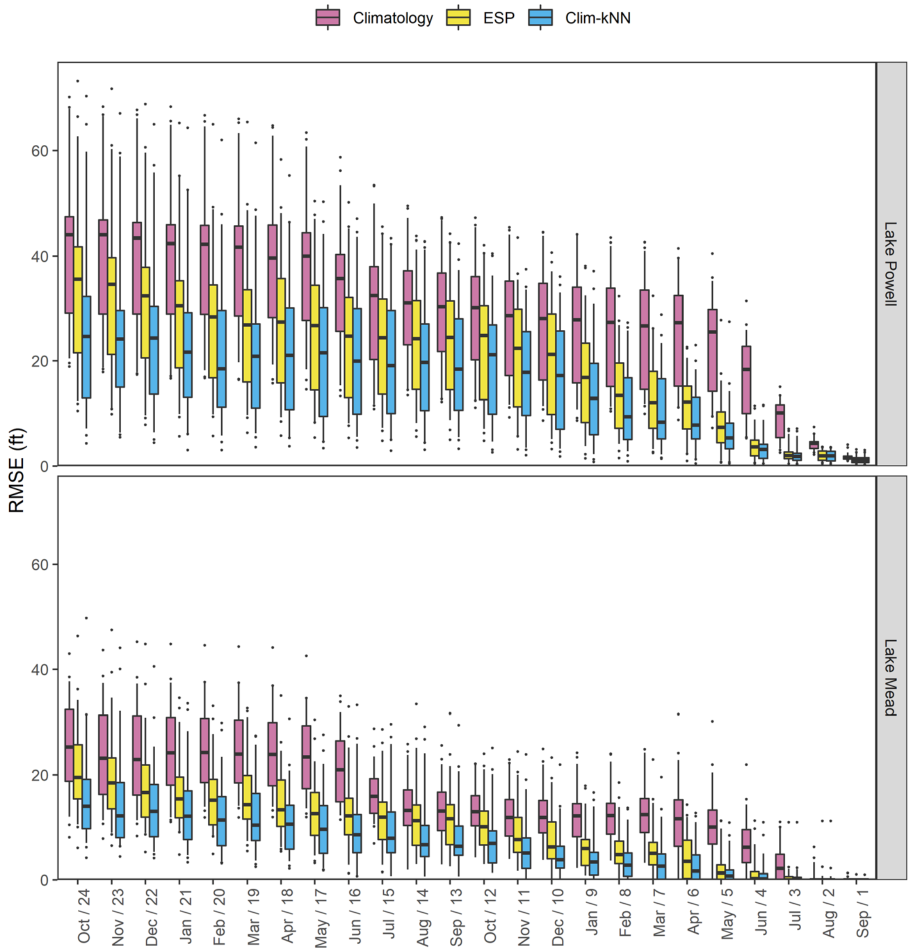


FIGURE 7. RMSE of end-of-water year pool elevation of Lakes Powell and Mead. Climatology, ESP, and Clim-kNN are compared to historical streamflow projected pool elevations (1982–2016).

TABLE 1. Operating tiers and releases or conditions used in categorical scores based on the Interim Guidelines.

Reservoir	Tier	Release/condition
Lake Powell	Equalization	Annual release > 8.23 maf Annual release = 8.23 maf
	Upper Elevation Balancing	Annual release >8.23 maf Annual release = 8.23 maf Annual release <8.23 maf
	Mid-Elevation Release	Annual release = 8.23 maf Annual release = 7.48 maf
	Lower Elevation Balancing	Lower Elevation Balancing Tier
Lake Mead	Shortage	First level (Mead ≤1,075 and ≥1,050) Second level (Mead < 1,050 and ≥1,025) Third level (Mead < 1,025)
	Surplus	Any except flood control Flood control
	Normal	Normal or ICS surplus condition

releases, which have less variability due to regulation and thus are easier to predict than Lake Powell's inflows. Lake Mead's pool elevation also has smaller errors at longer leads compared to Lake Powell (see Figure 7) for the same reason. All streamflow forecast operational projections for Lake Mead perform well at the longest lead in January, including Climatology. For Lake Powell, the predicted "Tier" determinations perform relatively well, especially by August; however, the "Tier & Release/Condition" projections are not as accurate. It is more difficult to get both these categories correct when there is a wide variety of different forecasted inflows at longer leads. The HSS in Table 3 shows similar results to the Percent Correct, except with lower values since we are comparing the forecasts to random chance.

When comparing the forecasts, Clim-kNN and ESP always perform better than Climatology, since Climatology is an uninformed forecast, while Clim-kNN performs slightly better than ESP. The largest

TABLE 2. Percent correct for climatology, ESP, and Clim-kNN vs. historical streamflow projected operating tiers from the out-year in January, April, and August at a 12-, 8-, and 5-month lead to the end of the calendar year when operational decisions are determined.

		Tier			Tier & release/condition		
Reservoir	**Streamflow forecast**	**January, %**	**April, %**	**August, %**	**January, %**	**April, %**	**August, %**
Lake Powell	Climatology	68	69	83	54	55	67
	ESP	71	75	86	57	62	69
	Clim-kNN	71	76	86	60	63	71
Lake Mead	Climatology	94	95	100	94	95	100
	ESP	99	99	100	99	99	100
	Clim-kNN	99	100	100	99	100	100

TABLE 3. Heidke Skill Score for climatology, ESP, and Clim-kNN vs. historical streamflow projected operating tiers from the out-year in January, April, and c at a 12-, 8-, and 5-month lead to the end of the calendar year when operational decisions are determined.

		Tier			Tier & release/condition		
Reservoir	**Streamflow forecast**	**January**	**April**	**April**	**January**	**April**	**April**
Lake Powell	Climatology	0.37	0.41	0.67	0.31	0.35	0.52
	ESP	0.39	0.50	0.68	0.34	0.43	0.55
	Clim-kNN	0.41	0.52	0.72	0.37	0.45	0.58
Lake Mead	Climatology	0.87	0.90	1.00	0.87	0.90	1.00
	ESP	0.97	0.98	1.00	0.97	0.98	1.00
	Clim-kNN	0.98	1.00	1.00	0.98	1.00	1.00

performance differences are in August, likely due to climate forecasts improving Clim-kNN performance slightly over ESP. Specifically, these improvements are due to Lake Powell's projections of Equalization and Upper Elevation Balancing, especially when the observed tier is an Equalization release equal to 8.23 maf or an Upper Elevation Balancing release above 8.23 maf. The Climatology forecasts for the "Tier & Release/Condition" perform fairly well, especially at Lake Mead, because the starting reservoir system storage levels provide a dominant part of the forecast signal for the end-of-period levels. This feature is likely to be truer of Lakes Powell and Mead, which combined can hold up to four years of the long-term average natural flow, than for smaller reservoir storage systems.

DISCUSSION AND CONCLUSIONS

The CRBOPT provides a foundation for the systematic evaluation of inflow forecasts that represent the Upper Basin hydrologic and basin-wide operational projections. While this study used a two-year forecast, the framework is extensible to longer periods. CRBOPT is a framework for analyzing streamflow forecasts through metrics assessing the error, skill, spread, and discrimination of the Lake Powell annual water year unregulated inflow. Streamflow forecasts are run through CRMMS to simulate operational projections and are evaluated using metrics including CRMMS projected pool elevations and operating tiers and conditions at Lakes Powell and Mead. The testbed is built to process any streamflow forecasts with a specific protocol that allows for an objective comparison of operational and experimental streamflow forecasts, though its scope could be expanded for other projects. The value of CRBOPT is that it will identify, quantitively and consistently, for the first time, the relative merits (strengths and weaknesses, skill attributes) of streamflow inputs to CRMMS, and how those propagate into reservoir variable prediction performance; it can also allow for intercomparison of different constraints or strategies in CRMMS that might be warranted given extreme conditions during drought or even as presented with hypothetical future climate or system change-related factors.

This study used CRBOPT to compare three ensemble streamflow forecasts: Climatology, ESP, and Clim-kNN. The baseline forecast for this analysis was Climatology, though a baseline forecast could be an existing forecast. Both hydrology and reservoir system projection metrics were processed, yielding a variety of metric-dependent results. It is important to assess both sets of metrics, as the relationship between streamflow forecasts and water resources management is inherently nonlinear. For instance, a marginal streamflow forecast skill increase may not make any difference in reservoir projections due to the operating policy, or it may cause a large improvement in operational projections. When comparing metric results, it is important consider tradeoffs between forecast attributes such as accuracy, overconfidence or under-confidence, and discrimination, as well as the timing of improvements of streamflow forecasts over the baseline.

The results from the hydrology metrics in this study showed that at long leads (greater than one year), all forecasts have good resolution, sharpness, and reliability, but lack discrimination and correlation skill. Clim-kNN forecasts are narrower with smaller errors than ESP and Climatology, and likely exhibiting minor overconfidence at long leads. ESP and Clim-kNN outperform Climatology starting in the fall months before the forecasted water year when antecedent basin conditions such as soil moisture begin to have influence over the coming spring runoff season, which can impact the runoff efficiency. By April of the forecasted year, the skill of these forecasts is much better than Climatology since there is better information about basin conditions such as snowpack. ESP has better skill than the Clim-kNN at most leads, except a few months in the winter of the forecasted year since Clim-kNN has knowledge of climate forecasts for 1- and 3-month leads. At shorter leads, ESP and Clim-kNN are overconfident and have reduced statistical reliability. These forecasts are too sharp, with narrow spread that excludes the observed annual streamflows more often than is predicted by the ensemble percentiles.

For the reservoir system projection metrics, all streamflow forecasts showed larger errors in projected end-of-water year pool elevation at longer leads. These errors decreased during the current year, especially by April for both reservoirs. The error (RMSE) of Clim-kNN was lower than ESP at all leads, likely due to the smaller spread in the Clim-kNN forecast. The categorical scores showed that Clim-kNN performed slightly better than ESP in April and August of the out-year with more traces projecting into the correct operating tier.

Each of these metrics highlights a different aspect of the forecast performance. When considering which forecast is "best," it is important to look at a range of metrics, both for hydrology and reservoir projections. With differing performance between metrics, it can be difficult to assess if a forecast is truly better than another, which is where knowledge of the application context in a specific basin can be useful. For instance, users may find projections of reservoir elevations or releases more relevant depending on their resource of interest (e.g., recreation, hydropower, ecology, etc.). The particular configuration and state of a reservoir system clearly also influence the impact of differences in forecast qualities (e.g., spread, mean skill) on reservoir operations and performance. As this study does not focus on the full cost of forecast errors to other resource types such as hydropower or flood damages, a potential direction of future expansion for CRBOPT would be to include more tailored stakeholder-focused metrics, translating the impacts of inflow or policy alternatives beyond pool levels to sectoral impacts. This may be a challenging step, however, because such calculations require input and cooperation from the stakeholder groups to develop and share critical data and policy insights; thus, there are added institutional considerations to contend with should the CRBOPT evolve in that direction.

Considering the forecasts analyzed here, the differences in the Clim-kNN relative to ESP from the CRPSS and RMSE perspective were moderate, showing 5%–25% improvement at most lead times. This improvement was reduced for reservoir system variables, with Clim-kNN offering slightly higher accuracy for both inflow to Lake Powell and projected pool elevation, showing that improved streamflow forecasts can result in improvements to reservoir system projections. For projected operating tiers, the Clim-kNN method performs better than ESP, and further refinement of the method is likely warranted to explore opportunities for adding additional predictors that could add forecast skill (such as sub-seasonal climate forecasts, year 2 climate forecasts, improved watershed modeling or data assimilation) and to balance augmentations with the potential for overconfidence, especially at longer leads.

Traditional ESP can only reflect how past climate combines with initial conditions and produces streamflow, and therefore has difficulty capturing our current prolonged drought, especially at leads beyond the current winter. Other experimental forecasts could be processed with CRBOPT and would be valuable for future use if they are able to capture a warmer and potentially drier climates' impacts on streamflow. It is critically important to improve the prediction of extreme years both on the low and high end, and CRBOPT will be useful in gaging the success of new approaches that may pursue such advances. These may come through improved estimates of initial conditions (through better watershed modeling or data assimilation of new earth observations) or advances in weather and climate forecasting coupled with improved usage of such forecasts in water supply prediction. Reclamation is already using CRBOPT for various other research projects aimed at improving reservoir system projections in the CRB (e.g., Woodson et al. 2021; Towler et al. 2022). These include exploring alternative forecasting methods for Lower Basin inflows and further exploration of Upper Basin forecasting methods for the 2- to 5-year range, with the hope that new methods can capture drought and hydrologic variability more effectively, seeking to improve projections of reservoir operations into the future. The CRBOPT enables these studies to become more automated, streamlining the process of exploring streamflow forecasts and reservoir system projections with a common set of robust scientific metrics to meet the needs of the CRB community.

Mid-term projections support an important planning timeframe during which Reclamation and stakeholders prepare to act in accordance with long-term policies. The ability to evaluate and ingest the latest science related to mid-term forecasts will become increasingly important as our climate warms, and the CRB experiences more frequent periods of severe, sustained drought not seen in recent history. This study introduces the CRBOPT framework and illustrates how it can be used to explore different inflow forecasts as they become available.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study include NMME and ESP hindcast datasets and observed hydrology datasets that are available from various sources. The NMME data are openly available in IRI/LDEO Climate Data Library at https://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME/. The ESP data are available from the corresponding author upon reasonable request. The observed hydrology datasets are available at the Bureau of Reclamation's Hydrologic Database at https://www.usbr.gov/lc/region/g4000/riverops/_HdbWebQuery.html.

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 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 would also like to thank the Bureau of Reclamation's Upper and Lower Colorado Regions for their continued support, development, and maintenance of the Colorado River Mid-term Modeling System and the Colorado River Basin Operational Prediction Testbed.

AUTHOR CONTRIBUTIONS

Sarah A. Baker: Conceptualization; data curation; formal analysis; investigation; methodology; project administration; software; validation; visualization; writing – original draft; writing – review and editing. Andy W. Wood: Conceptualization; data curation; funding acquisition; investigation; methodology; project administration; resources; supervision; validation; visualization; writing – review and editing. Balaji Rajagopalan: Conceptualization; funding acquisition; methodology; project administration; resources; software; supervision; validation; visualization; writing – review and editing. James Prairie: Conceptualization; funding acquisition; methodology; project administration; resources; software; supervision; validation; visualization; writing – review and editing. Carly Jerla: Conceptualization; funding acquisition; methodology; project administration; resources; software; validation; visualization; writing – review and editing. Robert A. Butler: Conceptualization; methodology; validation; visualization; writing – review and editing. Rebecca Smith: Conceptualization; methodology; validation; visualization; writing – review and editing. Edith Zagona: Conceptualization; funding acquisition; methodology; project administration; resources; software; supervision; validation; visualization; writing – review and editing.

LITERATURE CITED

Anghileri, D., N. Voisin, A. Castelletti, F. Pianosi, B. Nijssen, and D.P. Lettenmaier. 2016. "Value of Long-Term Streamflow Forecasts to Reservoir Operations for Water Supply in Snow-Dominated River Catchments." *Water Resources Research* 52 (6): 4209–25.

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): 906–22. https://doi.org/10.1111/1752-1688.12960.

Bröcker, J. 2015. "Resolution and Discrimination–Two Sides of the Same Coin." *Quarterly Journal of the Royal Meteorological Society* 141 (689): 1277–82. https://doi.org/10.1002/qj.2434.

Burnash, R.J.C., R.L. Ferral, and R.A. McGuire. 1973. *A Generalized Streamflow Simulation System, Conceptual Modeling for Digital Computers*. Sacramento, CA: Joint Federal State River Forecasts Center.

Christensen, N.S., A.W. Wood, N. Voisin, D.P. Lettenmaier, and R.N. Palmer. 2004. "The Effects of Climate Change on the Hydrology and Water Resources of the Colorado River Basin." *Climatic Change* 62: 337–63.

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).

Delaney, C.J., R.K. Hartman, J. Mendoza, M. Dettinger, L.D. Monache, J. Jasperse, F. Martin Ralph, et al. 2020. "Forecast Informed Reservoir Operations Using Ensemble Streamflow Predictions for a Multipurpose Reservoir in Northern

California." *Water Resources Research* 56 (9). https://doi.org/10.1029/2019WR026604.

Denaro, S., D. Anghileri, M. Giuliani, and A. Castelletti. 2017. "Informing the Operations of Water Reservoirs over Multiple Temporal Scales by Direct Use of Hydro-Meteorological Data." *Advances in Water Resources* 103: 51–63.

Duan, Q., F. Pappenberger, A.W. Wood, H. Cloke, and J.C. Schaake. 2018. *Handbook of Hydrometeorological Ensemble Forecasting*. Berlin Heidelberg: Springer-Verlag GmbH.eds

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.

Giuliani, M., L. Crochemore, I. Pechlivanidis, and A. Castelletti. 2020. "From Skill to Value: Isolating the Influence of End-User Behaviour on Seasonal Forecast Assessment." *Hydrology Earth System Science Discussion* 2020: 1–20.

Harpold, A.A., K. Sutcliffe, J. Clayton, A. Goodbody, and S. Vazquez. 2017. "Does Including Soil Moisture Observations Improve Operational Streamflow Forecasts in Snow-Dominated Watersheds?" *Journal of the American Water Resources Association* 53 (1): 179–96. https://doi.org/10.1111/1752-1688.12490.

Hashimoto, T., D.P. Loucks, and J.R. Stedinger. 1982. "Robustness of Water Resources Systems." *Water Resources Research* 18 (1): 21–26. https://doi.org/10.1029/WR018i001p00021.

Heidke, P. 1926. "Berechnung des Erfolges Und Der Güte Der Windstärkevorhersagen Im Sturmwarnungsdienst (Measures of Success and Goodness of Wind Force Forecasts by the Gale-Warning Service)." *Geografiska Annaler* 8 (4): 301–49. https://doi.org/10.1080/20014422.1926.11881138.

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<0559:DOTCRP>2.0.CO;2.

Jasperse, J., F.M. Ralph, M. Anderson, L.D. Brekke, M. Dillabough, M. Dettinger, A. Haynes, et al. 2017. *Preliminary Viability Assessment of Lake Mendocino Forecast Informed Reservoir Operations*. La Jolla, CA: Center for Western Weather and Water Extremes.

JAWRA (Journal of the American Water Resources Association). 1995. "Severe Sustained Drought: Managing the Colorado River System in Times of Water Shortage." *Journal of the American Water Resources Association* 31 (5): 779–944.

Koster, R.D., S.P. Mahanama, B. Livneh, D.P. Lettenmaier, and R.H. Reichle. 2010. "Skill in Streamflow Forecasts Derived from Large-Scale Estimates of Soil Moisture and Snow." *Nature Geoscience* 3 (9): 613–16. https://doi.org/10.1038/ngeo944.

Lehner, F., A.W. Wood, D. Llewellyn, D.B. Blatchford, A.G. Goodbody, and Pappenberger, F. 2017. "Mitigating the Impacts of Climate Nonstationarity on Seasonal Streamflow Predictability In the U.S. Southwest." *Geophysical Research Letters*, 44 (24). https://doi.org/10.1002/2017gl076043.

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

Pagano, T.C., David C. Garen, Tom R. Perkins, and Phillip A. Pasteris. 2009. "Daily Updating of Operational Statistical Seasonal Water Supply Forecasts for the Western U.S." *Journal of the American Water Resources Association* 45 (3): 767–78. https://doi.org/10.1111/j.1752-1688.2009.00321.x.

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: 28–29.

Pagano, T.C., A.W. Wood, M.-H. Ramos, H.L. Cloke, F. Pappenberger, M.P. Clark, M. Cranston, et al. 2014. "Challenges of Operational River Forecasting." *Journal of Hydrometeorology* 15: 1692–707.

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." Technical Report, U.S. Army Corps of Engineers; Bureau of Reclamation; National Oceanic and Atmospheric Administration.

Regonda, S., E.A. Zagona, and B. Rajagopalan. 2011. "Prototype Decision Support System for Operations on the Gunnison Basin with Improved Forecasts." *Journal of Water Resources Planning and Management* 137 (5): 428–38. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000133.

Sankarasubramanian, A., U. Lall, N. Devineni, and S. Espinueva. 2009. "The Role of Monthly Updated Climate Forecasts in Improving Intraseasonal Water Allocation." *Journal of Applied Meteorology and Climatology* 48 (7): 1464–82. https://doi.org/10.1175/2009JAMC2122.1.

Towler, E., D. Woodson, S. Baker, M. Ge, J. Prairie, B. Rajagopalan, S. Shanahan, and R. Smith. 2022. "Incorporating Mid-Term Temperature Predictions into Streamflow Forecasts and Operational Reservoir Projections in the Colorado River Basin." *Journal of Water Resources Planning and Management* 148 (4): 04022007. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001534.

Turner, S.W.D., J.C. Bennett, D.E. Robertson, and S. Galelli. 2017. "Complex Relationship between Seasonal Streamflow Forecast Skill and Value in Reservoir Operations." *Hydrology Earth System Science* 21 (9): 4841–59.

U.S. Bureau of Reclamation. 2020. *Review of the Colorado River Interim Guidelines for Lower Basin Shortages and Coordinated Operations for Lake Powell and Lake Mead*. Washington, DC: United States Bureau of Reclamation.

U.S. Bureau of Reclamation. 2021. "Annual Operating Plan for Colorado River Reservoirs." https://www.usbr.gov/uc/water/rsvrs/ops/aop/.

U.S. Department of Interior. 2007. *Record of Decision — Colorado River Interim Guidelines for Lower Basin Shortages and the Coordinated Operations for Lake Powell and Lake Mead*. Washington, DC: United States Bureau of Reclamation.

Vano, J.A., T. Das, and D.P. Lettenmaier. 2012. "Hydrologic Sensitivities of Colorado River Runoff to Changes in Precipitation and Temperature." *Journal of Hydrometeorology* 13: 932–49.

Wells, E., A.W. Wood, E. Jones, J. Ostrowski, and K. He. 2011. "Hydrologic Ensemble Forecast Service: Requirements." Report to U.S. Department of Commerce, National Oceanic and Atmospheric Administration, National Weather Service Office of Hydrologic Development, Silver Spring, Maryland.

Wilks, D.S. 2011. *Statistical Methods in the Atmospheric Sciences* (Third Edition). Oxford: Academic Press.

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. https://doi.org/10.1175/JHM-D-14-0213.1.

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.

Woodson, D., B. Rajagopalan, S. Baker, R. Smith, J. Prairie, E. Towler, et al. 2021. "Stochastic Decadal Projections of Colorado River Streamflow and Reservoir Pool Elevations Conditioned on Temperature Projections." *Water Resources Research* 57: e2021WR030936. https://doi.org/10.1029/2021WR030936.

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.