Development of Sub-seasonal to Seasonal Watershed-scale Hydroclimate Forecast Techniques to Support Water Management

by

Sarah Ann Baker

B.S., Oregon State University, 2014

M.S., University of Colorado at Boulder, 2016

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirement for the degree of

Doctor of Philosophy

Department of Civil, Architectural, and Environmental Engineering

2019

This thesis entitled:
Development of Sub-seasonal to Seasonal Watershed-scale Hydroclimate Forecast Techniques to Support Water Management
written by Sarah Ann Baker
has been approved for the Department of Civil, Architectural,
and Environmental Engineering

Dr. Balaji Rajagopalan

Dr. Andrew Wood

Dr. Edith Zagona

Dr. Jim Prairie

Dr. Ben Livneh

Date_____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above-mentioned discipline

Abstract

Baker, Sarah Ann (Ph.D., Department of Civil, Architectural, and Environmental Engineering)

Development of Sub-seasonal to Seasonal Watershed-scale Hydroclimate Forecast Techniques to Support Water Management

Thesis directed by Professor Balaji Rajagopalan

Operational sub-seasonal to seasonal (S2S) climate predictions have advanced in skill in recent years but are not yet broadly utilized by stakeholders in the water management sector. While some of the challenges that relate to fundamental predictability are difficult or impossible to surmount, other hurdles related to forecast product formulation, translation, and accessibility can be directly addressed. An example of S2S climate forecast use in water management is through streamflow forecasting. Streamflow forecasts inform many water management decisions such as reservoir operations, water allocation, flood control, and instream supported releases. More skillful streamflow forecasts would benefit water managers through improved projections of future basin conditions for planning and decision making purposes.

This dissertation is motivated by the need to reduce hurdles in water manager adoption of S2S climate forecasts. To this end, this dissertation makes four

contributions. (1) Two S2S climate forecast products, Climate Forecast System version 2 (CFSv2) and North American Multi-model Ensemble (NMME), are processed to develop real-time watershed-based climate forecast products. A prototype S2S climate data products website was built to disseminate real-time forecasts of CFSv2-based bi-weekly climate forecasts (weeks 1-2, 2-3, and 3-4) and NMME-based monthly and seasonal prediction products on a watershed scale. (2) Bi-weekly S2S climate forecast of temperature and precipitation were post-processed to enhance the skill and reliability of raw CFSv2 climate forecasts using partial least squares regression (PLSR). (3) An experimental streamflow forecasting method was developed with a simple stochastic trace weighting technique that ingests watershed-based climate forecasts in the Colorado River Basin. The experimental forecasting technique was compared to the traditional streamflow forecasting method, Ensemble Streamflow Prediction (ESP). (4) The experimental and operational streamflow forecasts were compared and analyzed through a testbed framework that was developed to assess how streamflow forecast performance affects operational projections in the Colorado River Basin at a lead time of two years using the Bureau of Reclamation's Mid-term Probabilistic Operations Model (MTOM).

Acknowledgements

First and foremost, I would like to thank my co-advisors Balaji Rajagopalan and Andy Wood. Their guidance and enthusiasm throughout my PhD pushed me to improve and expand my scientific knowledge. I feel fortunate for the opportunities they have provided me and for the many hours they spent providing feedback and answering my questions. I am also grateful for my committee members, Edie Zagona, Jim Prairie, and Ben Livneh, for their thoughtful comments and suggestions throughout this project.

Thanks to CADSWES for their continued support throughout all my phases at the University of Colorado and for software developments that enabled this research to progress. I would like to give a special thanks to Edie Zagona for pushing me to progress my engineering education and for the opportunity to work on my first research project at the University of Colorado.

I would like to thank Reclamation for their support throughout this process, which allowed me to continue my education beyond a Master's degree. This work would not have been possible without the support from members of the Boulder Reclamation team, including Carly Jerla, Jim Prairie, Alan Butler and many others. I am grateful for their modeling expertise and knowledge of water management in the Colorado River Basin. I would also like to thank members of the NCAR RAL team for their supported and answering of Linux and data processing questions.

I am grateful for the funding and support provided by NOAA's Climate Program Office's Modeling, Analysis, Predictions, and Projections (MAPP) Program, (awards #NA16OAR4310138 & #NA14OAR4310238) and the Bureau of Reclamation.

Finally, I would like to thank my family and friends. Thank you Mom and Dad for always being supportive, teaching me to always aim high, and explore the world. Thanks to my sister, Christine, for your adventurous attitude and being a role model to me. And thank you to my partner, Tom, for your love, humor, and positive support throughout this journey.

Table of Contents

List of Tables

List of Figures

1 CHAPTER I: Introduction

1.1 Background

Water managers make many operational decisions on a sub-seasonal to seasonal (S2S) timescale, but do not yet widely use climate forecasts to inform decision making. Surveys indicate that water managers are reluctant to use available climate forecasts due to perceived poor reliability of forecasts, institutional reasons such as traditional reliance on built infrastructure, organization or regulatory restraints, risk aversion, and mismatched temporal or spatial scale (Callahan, Miles, & Fluharty, 1999; Kirchhoff, Lemos, & Engle, 2013; Raff, Brekke, Werner, Wood, & White, 2013; Rayner, Lach, & Ingram, 2005; White et al., 2017). Furthermore, water managers may be unaware of sources of seasonal climate forecasts or lack the skill set and resources to ingest forecasts in a usable format, especially managers at smaller utilities (Bolson, Martinez, Breuer, Srivastava, & Knox, 2013). Issues presented through these academic surveys can be addressed through a closer relationship between the forecast producer and user, increased institutional flexibility, and demonstration of effective forecast skill and use (Dilling & Lemos, 2011; Feldman & Ingram, 2009; Pagano, Hartmann, & Sorooshian, 2001).

Climate forecasts produced by global climate models (GCMs) have recently shown improved skill at the S2S timescale, which extends from two weeks to months in the sub-seasonal timeframe, out multiple seasons in the seasonal timeframe. One such model is the dynamical, fully coupled atmosphere–ocean–land model Climate Forecast System version 2 (CFSv2), which demonstrates skill in projecting climate

and land surface variables at various leads and seasons over the US and improves upon its predecessor CFSv1 (Saha et al., 2014; Tian, Wood, & Yuan, 2017; Yuan, Wood, Luo, & Pan, 2011). Multi-model ensembles have also shown improved S2S skill over single models (Becker, den Dool, & Zhang, 2014; Doblas-Reyes, Hagedorn, & Palmer, 2005; Hagedorn, Doblas-Reyes, & Palmer, 2005). The North American Multi-model Ensemble (NMME) is an operational seasonal climate forecast system that includes ensemble forecasts (for climate and land surface variables) from seven GCMs, leading to more skillful seasonal climate forecasts than from any individual GCM (Becker & van den Dool, 2016; Kirtman et al., 2014; Slater, Villarini, & Bradley, 2016). This work focuses on making CFSv2 and NMME forecasts more useable to water managers by applying them to a watershed scale in real-time and displaying useful verification metrics.

GCMs can also be used to assess the skill of seasonal extreme climate events such as heat waves or droughts. Slater et al. (2016) assessed the skill of extreme events using NMME and found that seasonal prediction of drought events are better forecasted than floods, and high temperature and low precipitation events are predicted equally well. Tian et al. (2017) found that CFSv2 exhibited skill when predicting consecutive rainy and dry days in the US, especially over the west coast. Other studies have analyzed the skill of extremes prediction in other GCMs (Barnston & Mason, 2011; Hamilton et al., 2012; Mo & Lyon, 2015).

Raw GCM forecasts, such as CFSv2, have shown predictable and skillful forecasts of temperature and precipitation for the 3-4 week period. DelSole et al.

(2017) found that raw CFSv2 reforecasts demonstrated predictability in 3-4 week forecasts of temperature and precipitation over parts of the US during January and July by decomposing anomalies in terms of an orthogonal set of patterns for each grid point. The analysis found that predictability of temperature and precipitation was related to El Nino-Southern Oscillation (ENSO) and Madden-Julian oscillation (MJO) events. Though the study did illustrate predictability over the contiguous US (CONUS), the results were no equally promising for all regions and especially weak for summer precipitation. This suggests the need for additional information when post-processing precipitation and temperature forecasts from GCMs through the use of large-scale climate features (e.g. ENSO or MJO), or through climate fields such as geopotential height or precipitable water.

In addition to assessing the raw skill of CFSv2 and NMME, researchers have used various methods to improve the process of translating the raw, large scale outputs to the regional scale that are useful to water managers. Downscaling and bias-correction are methods of improving temperature and precipitation forecasts and allow users to move forecasts to a finer grid. Tian et al. (2014) compared downscaling techniques for NMME precipitation and temperature forecasts for Alabama, Georgia, and Florida. They found that the locally weighted polynomial regression downscaling method showed higher skill than direct spatial disaggregation and bias-correction for this region. Many other studies of downscaling techniques have shown improvements to GCM outputs for other regions including the CONUS-wide domain and the Colorado River basin (Gutmann et al., 2014; Pablo A. Mendoza, Rajagopalan, Clark,

Cortés, & McPhee, 2014; Andrew W. Wood, Leung, Sridhar, & Lettenmaier, 2004). Downscaling and bias-correction techniques could be applied to raw CFSv2 temperature and precipitation forecasts for the CONUS domain to improve forecast skill.

Model skill can also be improved through ensemble weighting methods. Slater et al. (2017) explored five different multi-model weighting approaches for NMME temperature and precipitation forecasts to enhance skill of four climatic regions in Europe. The weighting approaches tested were equal weighting, Bayesian updating (BU), and Bayesian updating of principal components (BU-PCA) of both the eight single model means and all NMME ensemble members. The study found that BU and BU-PCA reduced unconditional bias and negative skill, but sometimes diminished positive skill in the raw forecasts. Other work has concluded that multi-model weighting methods can improve ensemble prediction (Krakauer, 2017; Wanders & Wood, 2016). Despite the improvements to raw temperature and precipitation forecasts through post-processing, the water management sector has not widely incorporated S2S climate forecasts into their decision making framework.

Water management decisions made at the S2S timescale, such as reservoir operations, water allocation, flood control, and instream supported releases, depend largely on streamflow forecasts. Many of the streamflow forecasts in the US are provided by the National Weather Service River Forecasting Centers (RFCs) and National Resource Conservation Service (NRCS) (Pagano, Robertson, Werner, & Tama-Sweet, 2014). A significant example of the use of these streamflow forecasts is

in the Colorado River Basin. In their management of the Colorado River, the Bureau of Reclamation uses streamflow forecasts produced by the Colorado Basin RFC as inputs to their operations and planning models, which are used for decision making and risk assessment of potential Lower Basin shortage or surplus conditions that affect many communities and economies (Bureau of Reclamation, 2015).

Most operational streamflow forecasts are not informed by S2S projections, but instead rely on land-surface models initialized with current basin conditions and forced with historical temperature and precipitation traces (Raff et al., 2013). This method, Ensemble Streamflow Prediction (ESP), is widely used throughout the water management community (Day, 1985; Franz, Hartmann, Sorooshian, & Bales, 2003). The skill of ESP streamflow forecasts are initially highly dependent on the initial conditions, but at leads longer than one month, the skill is more dependent on climate forcings (Li, Luo, Wood, & Schaake, 2009; Shukla & Lettenmaier, 2011). This leaves room for potential improvement to climate forcings through use of GCM outputs.

Statistical methods are also used to produce seasonal water supply forecasts (Garen, 1992; Pagano, Wood, et al., 2014). These monthly forecasts traditionally use principal component regression models trained on historical data such as seasonal precipitation and snow water equivalent (SWE). As with ESP, statistical water supply forecasting methods do not utilize the skill of S2S climate forecasts.

The potential value of S2S climate forecasts for use in streamflow prediction has been explored through different academic studies. Studies have shown that using climate forecasts in land-surface models can improve streamflow forecasts. Wood and

Lettenmaier (2006) showed improvement to ESP forecasts in the western US through the use of a land-surface model driven with climate forecasts ensembles from NASA's Seasonal-to-Interannual Prediction Project and other seasonal climate forecasts. Mo and Lettenmaier (2014) completed a similar study over the CONUS domain using NMME forecasts. They found that skill is seasonally and regionally dependent and that NMME forecasts contributed to skill after month 1 during which initial conditions were dominant. Many other studies illustrated the added streamflow forecasts skill through the use of climate forecasts in land-surface models (Li et al., 2009; Luo & Wood, 2008; Sankarasubramanian, Lall, Devineni, & Espinueva, 2009; Werner, Brandon, Clark, & Gangopadhyay, 2005; Yuan & Wood, 2012; Yuan, Wood, & Liang, 2014; Yuan, Wood, Roundy, & Pan, 2013). Despite the improvements to streamflow skill shown in these studies, most operational streamflow forecasts do not use climate forecasts.

Statistical water supply methods have also shown improvement when climate information is used as a predictor. Studies have used regression-based methods informed by large scale climate predictors such as ENSO to improve streamflow projections (Clark, Serreze, & McCabe, 2001; Gobena & Gan, 2010; Grantz, Rajagopalan, Clark, & Zagona, 2005; Moradkhani & Meier, 2010; S. K. Regonda, Rajagopalan, Clark, & Zagona, 2006; van Dijk, Peña-Arancibia, Wood, Sheffield, & Beck, 2013). Lehner et al. (2017) illustrated reduced error in seasonal streamflow forecasts when using monthly temperature forecasts from NMME and System 4 from European Center for Medium-Range Weather Forecast (ECMWF) to drive statistical

models in the Colorado and Rio Grande River Basins. Slater et al. (2017) presented a statistical-dynamical approach showing skillful seasonal streamflow forecast in Iowa using agricultural land cover, antecedent precipitation, and NMME monthly precipitation ensemble. These studies further illustrate the potential improvements to both land-surface models and statistical streamflow forecasting methods through the use of S2S climate forecasts.

As described above, many studies have explored the potential for S2S climate forecasts to improve the skill of streamflow forecasting. However, less work has been devoted to overcoming the spatial and temporal barriers noted by Rayner et al. (2005) that prevent water managers from incorporating the forecasts into their decision making processes. Though Hartmann et al. (2002) displayed digestible skill metrics for forecasts at the NOAA's Climate Prediction Center (CPC) climate division scale (as opposed to the typical gridded scale) and Bolinger et al. (2017) created a web-based tool for monthly water-level forecasts in the Great Lakes region, neither project provided the real-time CONUS-wide, watershed-scale forecasts and metrics that would be most useful to water managers.

This dissertation is designed to enhance climate prediction quality, specificity, and accessibility by applying S2S climate forecasts to a watershed scale and by improving climate forecast skill through stochastic post-processing methods. Skillful streamflow forecasts are important to water managers who plan reservoir operations and water allocation (Raff et al., 2013). S2S climate forecasts can be used to inform streamflow forecasts. Lehner et al. (2017) and Slater et al. (2017) applied statistical

methods to S2S climate forecasts to improve seasonal streamflow forecasts. This project will go further by using S2S climate forecasts with improved skill to project streamflow, and compare to current methods to analyze the effects of improved streamflow forecasting on reservoir operations in the Colorado River Basin.

1.2 Scope of Work and Thesis Structure

The dissertation seeks to reduce hurdles to S2S climate forecasts use by water managers through translating forecasts to a usable watershed-based spatial scale and demonstrating operational use in the Colorado River Basin through applications in streamflow forecasting. The chapters in this dissertation are as follows. Chapter 2 describes the formulation of S2S climate forecast products on a watershed scale. Bi-weekly CFSv2 and monthly and seasonal NMME watershed scale forecasts are evaluated and displayed on a real-time web based product. In Chapter 3, PLSR is explored as a potential post-processing technique for improving the skill of bi-weekly CFSv2 temperature and precipitation forecast using concurrent CFSv2 climate and land surface fields as predictors. Chapter 4 describes a stochastic trace weighting scheme for streamflow forecasting which ingests S2S climate forecasts. This experimental forecast is compared to traditional forecasting method, ESP, in the Colorado River Basin. Chapter 5 presents a testbed framework for assessing the performance of streamflow forecasts and operational projections in the Colorado River Basin. Chapter 6 provides a brief summary of conclusions presented in the previous four chapters and a description of future work.

2 CHAPTER II: Developing Sub-seasonal to Seasonal Climate Forecast Products for Hydrology and Water Management

2.1 Abstract

We describe a new effort to enhance climate forecast relevance and usability through the development of a system for evaluating and displaying real-time sub-seasonal to seasonal (S2S) climate forecasts on a watershed scale. Water managers may not use climate forecasts to their full potential due to perceived low skill, mismatched spatial and temporal resolutions, or lack of knowledge or tools to ingest data. Most forecasts are disseminated as large-domain maps or gridded datasets and may be systematically biased relative to watershed climatologies. Forecasts presented on a watershed scale allow water managers to view forecasts for their specific basins, thereby increasing the usability and relevance of climate forecasts. This paper describes the formulation of S2S climate forecast products based on the Climate Forecast System version 2 (CFSv2) and the North American Multi-model Ensemble (NMME). Forecast products include bi-weekly CFSv2 forecasts, and monthly and seasonal NMME forecasts. Precipitation and temperature forecasts are aggregated spatially to a USGS HUC-4 watershed scale. Forecast verification reveals appreciable skill in the first two bi-weekly periods (weeks 1-2 and 2-3) from CFSv2, and usable skill in NMME month 1 forecast with varying skills at longer lead times dependent on the season. Application of a bias-correction technique (quantile mapping) eliminates forecast bias in the CFSv2 reforecasts, without adding significantly to

correlation skill.[1]

[1] ***A version of this chapter has been published:***
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 1–14. https://doi.org/10.1111/1752-1688.12746.

2.2 Introduction & Background

Hydrologists and water managers make many operational decisions on a sub-seasonal to seasonal (S2S) time scale, but under-utilize climate prediction to inform decision making from a quantitative standpoint. Surveys indicate that water managers are reluctant to use climate forecast due to perceived poor reliability of forecasts, mismatched temporal or spatial scale, institutional reasons such as traditional reliance on built infrastructure, organization or regulatory restraints, and risk aversion (Callahan et al., 1999; Kirchhoff et al., 2013; Rayner et al., 2005; White et al., 2017). Water managers may be unaware of sources of seasonal climate forecasts or lack the skill set and resources to ingest forecasts in a usable format, especially managers at smaller utilities (Bolson et al., 2013). Issues presented in these academic surveys can be addressed through a closer relationship between forecast producer and user, increased institutional flexibility, and demonstration of effective climate forecast skill and use (Dilling & Lemos, 2011; Feldman & Ingram, 2009; Pagano et al., 2001).

Numerous water management short-term and mid-term decisions are made on the S2S time scale including reservoir operations, water allocation, flood control, hydropower generation, water treatment, and in-stream supported releases (Bolson et al., 2013). Decisions depend largely on streamflow forecasts, many of which are provided by the National Weather Service River Forecasting Centers (RFCs) and National Resource Conservation Service (NRCS) in the United States (US) (T. Pagano et al., 2014). In the Colorado River Basin, a river managed by the Bureau of

Reclamation, streamflow forecasts produced by the Colorado Basin RFC (CBRFC) are used as inputs to operations and planning models that are used for decision making and risk assessment of potential shortage or surplus basin conditions (Bracken, 2011). These streamflow forecasts are not informed by climate forecasts, even though recent work shows benefits (Lehner et al. 2017). Raff et al. (2013) identified enhancements to climate forecasts to meet the needs of water resource managers in the Bureau of Reclamation and US Army Corps of Engineers in a report documenting short-term water management decisions. Water managers interviewed in the report emphasized the need for better understanding of the skill and reliability of climate forecast products, easily accessible products on different time scales, and products presented in a format easily accessible by operators.

Recently, dynamical climate forecasts generated using initialized global climate models (GCMs) have shown skill improvements at the S2S time scale. One of these dynamical models, the fully coupled atmosphere–ocean–land model Climate Forecast System version 2 (CFSv2), which is run at the National Centers for Environmental Prediction (NCEP), demonstrates skill in projecting climate variables at various leads and seasons over the US and improves upon its predecessor CFSv1 (Saha et al., 2014; Tian et al., 2017; Yuan et al., 2011). Multi-model climate forecast ensembles have also demonstrated improved skill over single models (Becker et al., 2014; Doblas-Reyes et al., 2005; Hagedorn et al., 2005). The North American Multi-model Ensemble (NMME) is an operational seasonal climate forecast system that includes ensemble forecasts (for climate and land surface variables) from seven

GCMs, leading to more skillful seasonal climate predictions than from any individual GCM (Becker & van den Dool, 2016; Kirtman et al., 2014; Slater et al., 2016).

These climate model forecasts and verifications are normally presented at a system grid resolution or on North American-wide maps, or for all forecast initializations and lead times, rather than particular seasons that are easily related to local watershed scales. From a water management perspective climate forecast utility is highly specific to location, time of year, and predictand (Wood et al., 2016). There is a gap between the type of verification, data, and product diagnostics provided by forecast production centers and the skill information most readily interpretable and usable by the water community (Wood & Werner, 2011).

A number of the studies referenced above have explored S2S climate forecast skill, but more can be done to support water managers in incorporating climate forecasts into decision making. Some studies have attempted to address these issues by presenting seasonal climate forecasts on a different spatial scale than the typical gridded scale and by displaying skill metrics that are useful to water managers. Hartmann et al. (2002) explored a framework for evaluating seasonal temperature and precipitation projection performance with metrics more easily digestible by users. The metrics were displayed on the 344 Climate Divisions specified by NOAA's Climate Prediction Center (CPC). Although this spatial scale can be useful, the Climate Divisions are not, by design, aligned with hydrologic boundaries that may be relevant for areas of water manager responsibility. More recently, Bolinger et al. (2017) explored the use of a web-based tool to provide monthly updated water-level

projections informed by NMME forecasts in the Great Lakes region. The tool allows users to look at individual NMME model results and probabilities of hydrologic variables for specific regions. It represents an example of a regional water group processing climate outlooks onto spatial scales of interest, which underscores the need to develop a centralized, nationwide system to achieve similar goals.

With this motivation in mind, we present work to address some of the hurdles confronting the widespread use of S2S climate predictions in water management applications, and to bridge the gap for potential stakeholders by enhancing the quality, specificity, and accessibility of S2S predictions. To make S2S prediction more usable, this project aligns climate forecasts with users space-time needs, presents data in real-time in user friendly formats (such as CSV files by watershed area), removes systematic climatology biases in forecast products, and produces verification information that is relevant to water sector users.

This paper describes a new real-time experimental effort to develop and demonstrate climate forecasts tailored to water managers by presenting real-time forecasts and verification on a watershed scale over the conterminous United States (CONUS) domain. The effort contributes to a sequence of milestones required to transition research toward implementation in an agency operational center such as CPC. For prototyping and demonstration purposes, this effort adopts the United States Geological Survey (USGS) hydrologic unit code 4 (HUC-4) delineation, which includes of 202 watersheds, which is a suitable spatial scale to show meaningful variability in climate forecasts, given the de-correlation length scales of common

climate variables. In Section 2-3, we describe reforecast and real-time CFSv2 and NMME forecasts, and forcing datasets. Data processing, verification, and basic bias-correction methodologies for precipitation and temperature reforecasts at bi-weekly, monthly, and seasonal time steps are presented in Section 2-4. Results from watershed-scale verification are evaluated in Section 2-5, followed by a discussion of water sector responses to the new products and possible improvements to the S2S watershed climate forecasting system in Section 2-6.

2.3 Data Sources & Processing

2.3.1 CFSv2 Climate Forecasts

The leading operational S2S climate forecast dataset in the US is generated by CFSv2, a fully coupled atmosphere-ocean-land operational model (Saha et al., 2014). CFSv2 forecasts of temperature and precipitation rate are supported by a separate S2S-scale reforecast dataset, which has a 100 km (0.93 degree) grid resolution at a 6-hour time step from 1999 through 2010. The reforecasts were initialized each day at four synoptic times: 0000 UTC, 0006 UTC, 0012 UTC, and 0018 UTC. The 0000 UTC forecast extends to the end of a full season (end of the fourth month), while the 0006, 0012, and 0018 UTC forecasts extend for 45 days. Less frequent CFSv2 reforecasts, not used in this work, extend to 9 months lead time.

For this work, the raw CFSv2 temperature and precipitation reforecasts were re-projected from a native Gaussian grid to a 1/2-degree grid, temporally averaged to a daily time step, and areally averaged to USGS HUC-4 spatial units through spatially conservative remapping.

Figure 2-1 displays the 202 HUC-4 watersheds in the CONUS domain. Daily ensemble means were calculated for CFSv2 reforecasts and were temporally averaged to bi-weekly time periods (e.g. 1-2 week, 2-3 week, 3-4 week) to support a skill analysis on the sub-seasonal scale. Climatologies for each watershed, lead, and day of year (DOY) are based on a 15-day window (+/- 7 days from forecasted date). CFSv2 data were obtained online from the NOAA National Center for Environmental Information.

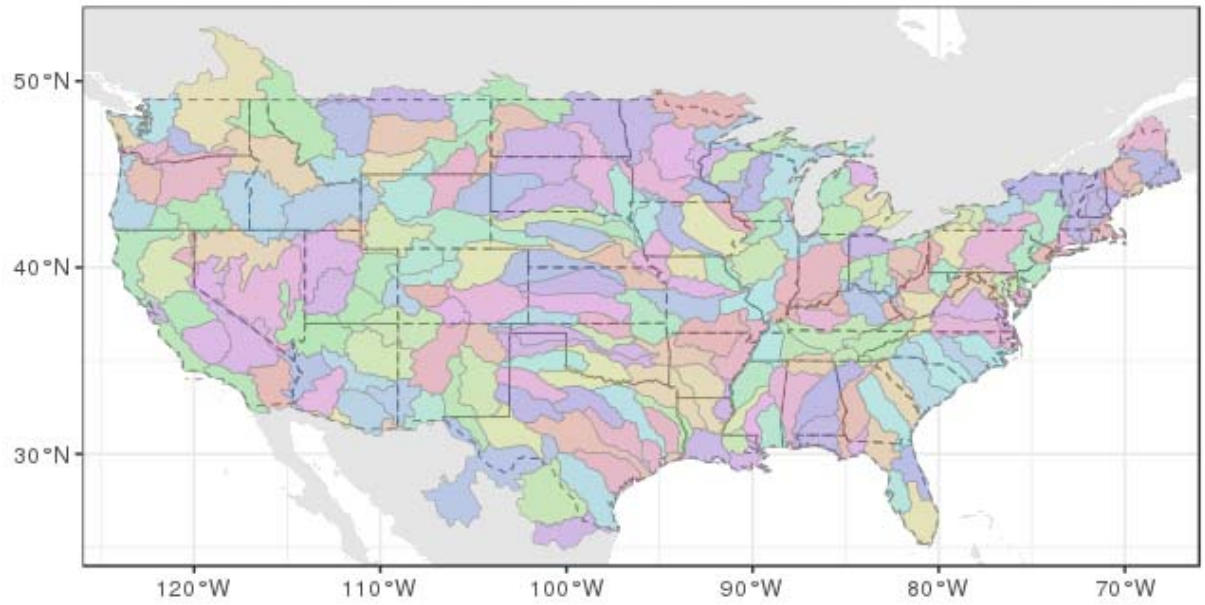

Figure 2-1: USGS HUC-4 watersheds. USGS hydrologic unit code 4 (HUC-4) watersheds over the conterminous United States (CONUS) domain overlaid by state outlines.

Real-time CFSv2 forecast are initialized each day at the four synoptic times, but in contrast to the retrospective runs, each initialization produces four ensemble members for a total of 16 forecasts each day of various lengths: four extend out to 9 months, three to 1 season, and nine to 45 days. The CFSv2 operational 16 member ensemble is downloaded each day and processed similarly to the reforecasts.

2.3.2 NMME Climate Forecasts

The NMME Phase 2 is a combination of seven global climate models that predict precipitation and temperature (among other variables) at a monthly time step for leads up to 7 months (Kirtman et al., 2014). Reforecasts are available for 1982 to 2010 and real-time model forecasts are available for 2011 to present. The models included in NMME are summarized in Table 2-1. For more information about each model in NMME, see Kirtman et al. (2014) or Slater et al. (2016), but note that the models included in the NMME have changed over time.

Table 2-1: NMME models

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 et al., 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

Raw temperature and precipitation reforecasts are re-projected from a 1-degree grid onto a 1/2-degree grid and spatially averaged to HUC-4 spatial units using the same method as CFSv2. The NMME forecast ensemble mean, which is

used in calculating several of the evaluation metrics, is calculated by equally weighting each model's ensemble average. A seasonal forecast is calculated by temporally averaging the first three months for the forecast for each model. Climatologies are then established for each NMME model, watershed, and forecasted month or season. Real-time NMME forecasts are updated monthly by the 8th day of each month. The ensembles for each of the 7 models are downloaded and processed to watershed scale monthly. Reforecasts were downloaded from the Climate Prediction Center's website and real-time forecasts are downloaded from the IRI Data Library.

2.3.3 NLDAS Climate Observations

The observational data for this study are derived from Phase 2 of the North American Land Data Assimilation System (NLDAS; Xia et al., 2012). NLDAS data are available at 1/8th-degree grid spacing from 1979 to the present at an hourly temporal resolution. Similar to the CFSv2 reforecasts, NLDAS precipitation and temperature data were spatially and temporally aggregated to a daily time step on a 1/2-degree grid (common to both datasets) before further aggregation to the sub-seasonal HUC-4 space-time resolution to match CFSv2 and NMME time scales. The choice to move to common grid spacing was for ease of analysis and to reduce disk space used during data processing. NLDAS data were obtained from NASA's Earth Science Data Systems Program websites.

2.4 Methods

2.4.1 Post-processing of Climate Forecasts to Reduce Systematic Biases

Raw GCM forecasts require post-processing due to systematic biases, unreliable ensemble spread, and forecasts not being skillful. Post-processing can take the form of statistical or downscaling to improve the raw output of GCMs. In this project, raw CFSv2 forecasts were bias-corrected using the Quantile Mapping (QM) method. QM removes systematic bias between the forecasted and observed climatologies, but does not further calibrate the forecasts to improve their skill. QM is a general method that has long been applied to weather forecasts (Panofsky & Brier, 1968) and later to climate forecasts (Wood et al., 2002). QM is an effective approach to removing bias, but does not address forecast deficiencies in attributes such as reliability and correlation skill (in some cases QM reduces the skill of the forecast). The distinction between bias-correction and probabilistic forecast calibration is further described in Wood and Schaake (2008) and Zhao et al. (2017), and the effectiveness of QM for post-processing climate model outputs for various applications, including extremes projection, is discussed in Ning et al. (2015), Cannon et al. (2015) and Maraun (2013).

When applied to CFSv2, QM replaces the forecast value with a value from the observed climatology (NLDAS) that has the same quantile. This is done by estimating a pair of cumulative distribution functions (CDFs) for the CFSv2 reforecasts and NLDAS data for each variable, lead, watershed, and time scale (climatologies based on 15-day window of +/- 7 days from forecast date). When forecasted values lie outside the quantile range, the two closest quantiles are used to

linearly extrapolate the new value. While QM corrects systematic biases in the first and second moments of the climate forecast distribution, concerns have been raised in various studies about its ability to preserve the extreme indices in the observed distribution and to preserve future trends, and it also does not guarantee that biases will be eliminated for durations not explicitly addressed in the CDF mapping. Nonetheless, it serves well as a first step to addressing major bias-related deficiencies in climate model forecast outputs.

2.4.2 Production of Real-time Web-based S2S Climate Outlooks

After spatial remapping and bias correction, prototype S2S climate data products – forecasts and associated skill analyses – are operationally disseminated by the National Center for Atmospheric Research (NCAR) on a public website to facilitate further product development through interactions with water managers. The website (http://hydro.rap.ucar.edu/s2s/) was built in R using the R package Shiny, which supports the staging of websites that link data and geospatial mapping. Climate products on the website include CFSv2-based bi-weekly climate forecasts for HUC-4 watersheds, and NMME-based monthly and seasonal prediction products. The workflow for product generation is summarized in **Error! Reference source not found.** and described in the previous sections. Raw and bias-corrected CFSv2 products are updated daily on the site and NMME products are updated once per month, when NMME forecast outputs are updated.

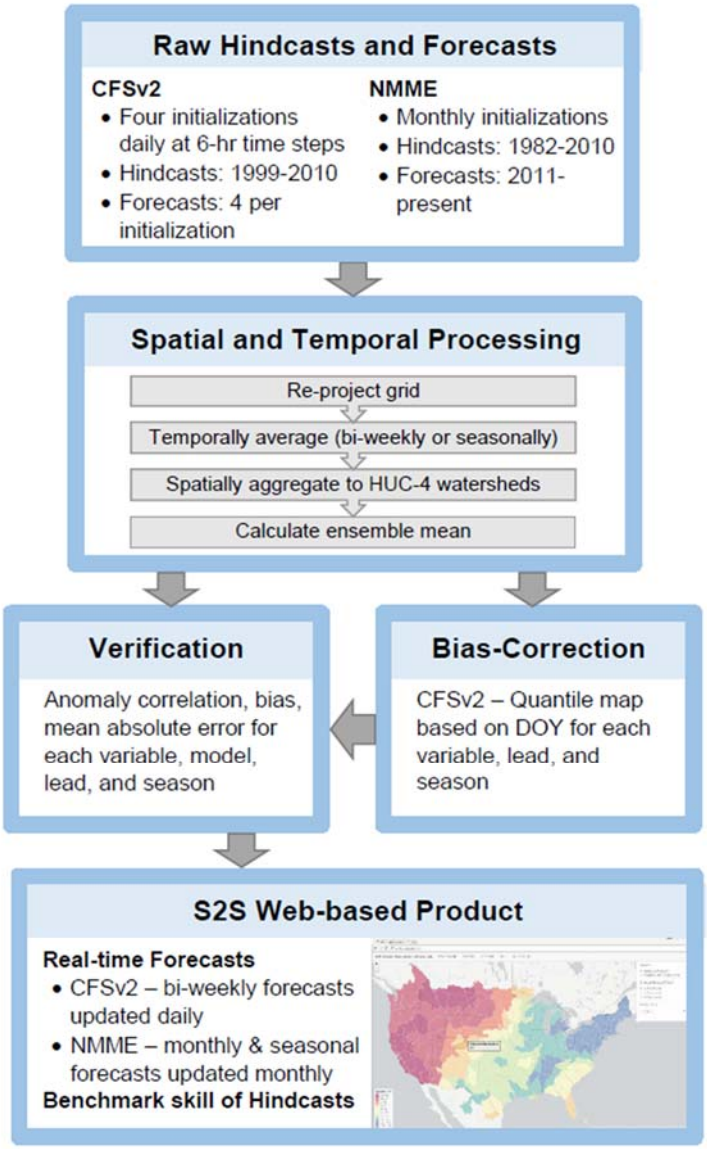

Figure 2-2: Methods of processing S2S forecasts. Summary of methods for processing of Climate Forecast version 2 (CFSv2) and North American Multi-model Ensemble (NMME) data, and delivering them to an online dissemination platform.

2.4.3 Forecast Verification

The anomaly correlation coefficient (ACC) metric is widely used in the climate prediction community to measure the degree of association between the forecast mean and the observations. The square of the ACC represents the fraction of climatological variance (uncertainty) explained by the forecast, where a score of 1

indicates that it provides perfect information and a score of zero means the forecast contains no information. For the purposes of prototyping, the ACC was used here to quantify the skill of the forecasts by calculating the correlation between the reforecasts and observations (or forcing data), as follows (Murphy & Epstein, 1989):

$$ACC = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2} - \sqrt{n\sum y^2 - (\sum y)^2}} \quad \text{Eq. 2-1}$$

where x is CFSv2 or NMME reforecast anomalies for each watershed and lead of temperature or precipitation and y is NLDAS anomalies for the same variable, watershed and lead, n is the number of forecasts, and ACC is the anomaly correlation coefficient for the reforecasts and forcing data. Anomalies for CFSv2, NMME, and NLDAS for S2S time scales were calculated using the climatologies described in the previous section.

We also calculate other standard deterministic forecast quality metrics that are familiar to water managers, including forecast bias (i.e. the mean error as a percent of observations for precipitation and as a difference from observations for temperature), and mean absolute error, and we plan to assess probabilistic metrics in the future. Forecast 'skill' is a multi-faceted concept, generally reflecting the quality of the forecast as described by various dimensions of forecast performance, such as reliability, discrimination, resolution, error, accuracy, correlation and bias. For the demonstration purposes of this paper, we discuss only the ACC and bias.

2.5 Results

The raw and bias-corrected real-time climate forecast products being staged on the website are complemented by maps showing skill metrics for different products,

seasons and lead times, which we summarize here. The anomaly correlation coefficient for CFSv2 bi-weekly forecasts (**Error! Reference source not found.**), shows that temperature has high skill for the first two bi-weekly periods, especially for weeks 1-2. The skill tends to be lower in the mid to south western half of the CONUS domain. By weeks 3-4, there are areas with skill exceeding a 'usability' threshold used by the CPC of ACC = 0.3 along the Atlantic and Gulf coasts but the rest of the domain has very low to no skill (O'Lenic et al., 2008). Precipitation forecasts have high skill (reaching values of 0.72) in the first bi-weekly period, especially on the west coast. Skill drops off significantly for weeks 2-3, especially in the central and eastern CONUS domain, and by weeks 3-4, the forecast has negligible skill.

The climate forecast skill varies considerably depending on the season. **Error! Reference source not found.** depicts CFSv2 weeks 2-3 anomaly correlation of precipitation forecast for four seasons. The west coast, especially watersheds in southern Arizona, and the Midwestern US have the highest skill in the December-February (DJF) season. In March-May (MAM) season, the skill is not as high, but the spatial pattern doesn't vary significantly compared to DJF. During the June-August (JJA) period, the pattern shifts and the watersheds in Nevada and Idaho have the highest skill while the remainder of the CONUS domain has low skill. In the September-November (SON) period, the region of highest skill shifts to the southeastern US. The maps in **Error! Reference source not found.** display different patterns of forecast skill compared to the corresponding map in **Error!**

Reference source not found.. This seasonal dependence on skill over the CONUS domain is apparent for all other leads and variables (not shown here).

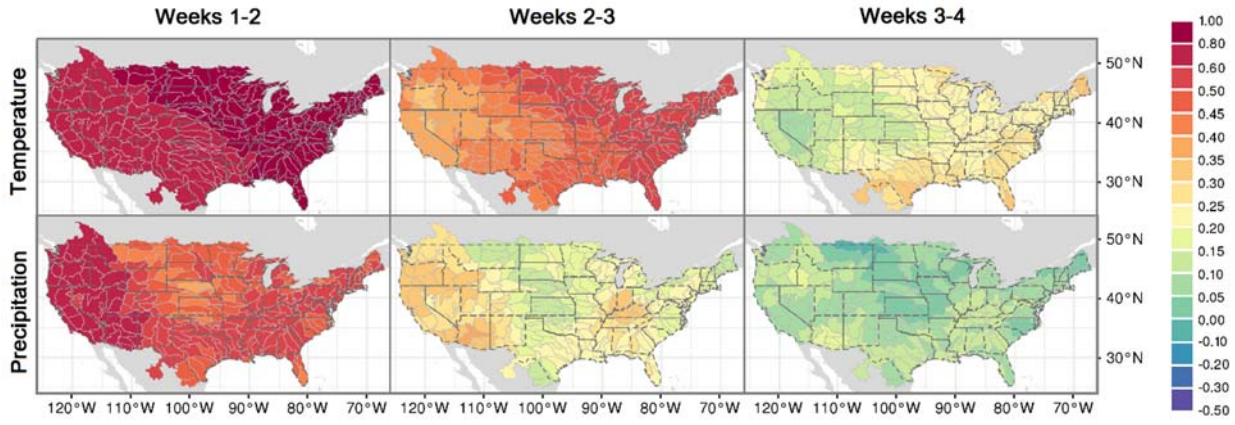

Figure 2-3: CFSv2 bi-weekly anomaly correlation. CFSv2 anomaly correlation at bi-weekly time step for temperature and precipitation at a hydrologic unit code 4 (HUC-4) watershed scale.

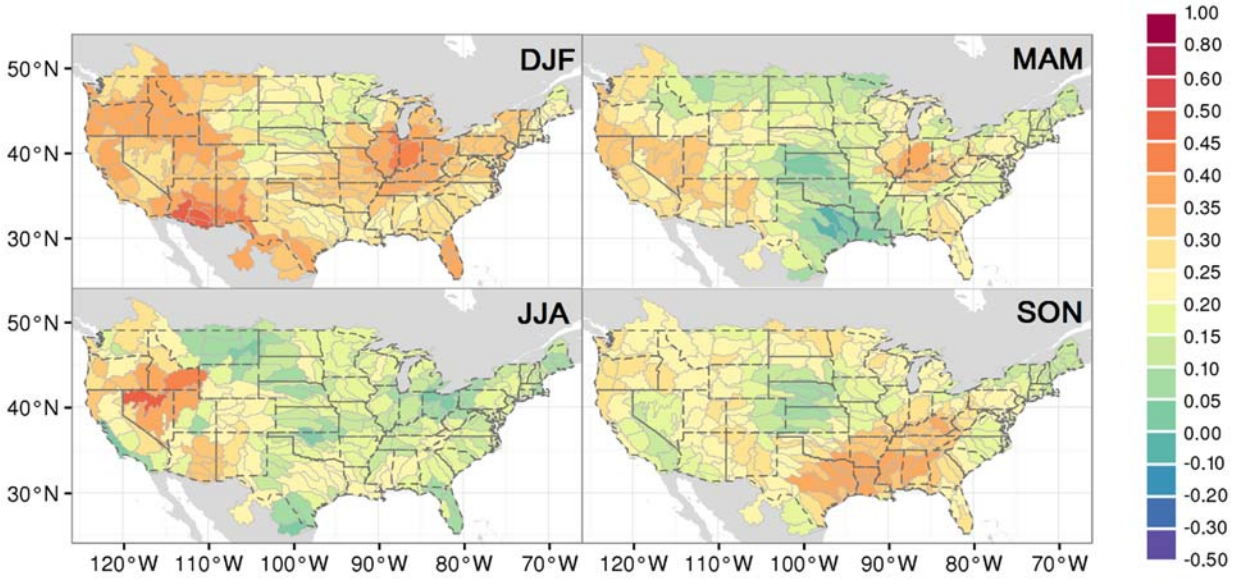

Figure 2-4: Seasonal CFSv2 anomaly correlation. Anomaly correlation of CFSv2 2-3 week precipitation reforecast for four seasons. Season acronyms contain the first letter of each month included in the season.

NMME monthly anomaly correlation for mean temperature and precipitation are shown in **Error! Reference source not found.**. There are three leads shown in the figure, which are labeled as months. Month 1 refers to the forecast initialization

month, or a lead 0, e.g. for a January NMME forecast, Month 1 would refer to January, Month 2 would be February, and Month 3 would be March. As has been found by other authors (Becker & van den Dool, 2016; Slater et al., 2016), temperature forecasts exhibit skill in month 1, especially in the north central US, but this skill drops off significantly in months 2 and 3. Precipitation has some skill in watersheds within California and the south east, but other areas of CONUS display low skill. The anomaly correlation of precipitation forecasts in months 2 and 3 are much lower. These trends in skill are highly seasonally dependent; therefore, there may be skill in months 2 and 3 for specific seasons not observed in the annual figures.

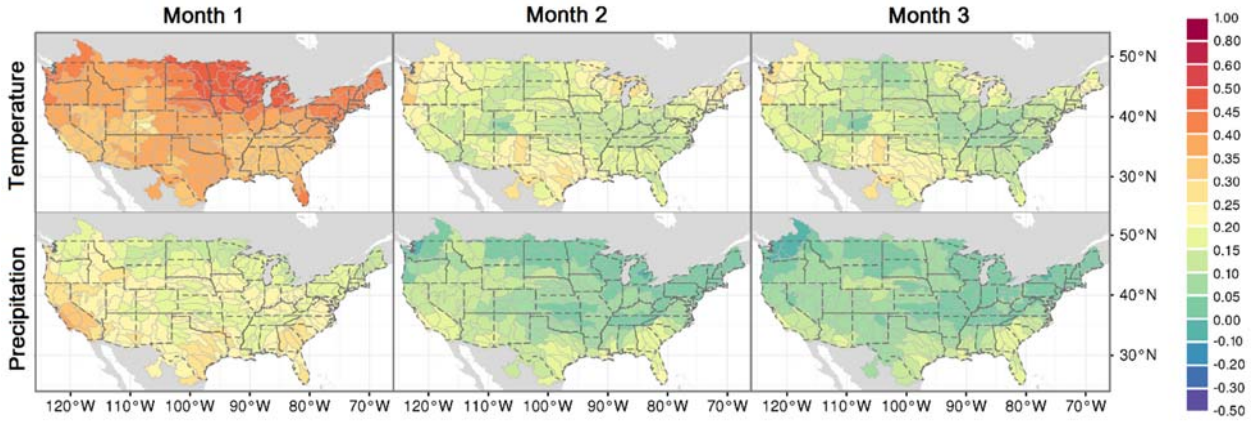

Figure 2-5: Monthly NMME anomaly correlation. NMME anomaly correlation of monthly time periods for temperature and precipitation at a HUC-4 watershed scale.

A basic skill assessment is presented, but additional analysis into the sources of predictability were not a component of this work. Many other studies have focused of predictability with CFSv2 and NMME. Sources of predictability in the S2S timescale is dependent on the season and lead. Infanti and Kirkman (2016) explored the relationship between ENSO and NMME forecasts of North American precipitation and temperature forecasts. Dirmeyer and Halder (2016) evaluated the

sensitivity, variability, and memory of land surface states in CFSv2 and found that soil moisture memory was important in improving forecast skill during spring and summer.

Quantile mapping was used to remove bias from CFSv2 forecasts. The bias prior to quantile mapping is shown in **Error! Reference source not found.**. Temperature bias is positive, meaning CFSv2 is over-forecasting temperature compared to NLDAS. The warm bias in temperature appears to grow with lead time. Climate model forecasts are known to drift (i.e. climatologies changing with lead time). To address any drift in bias, the quantile mapping adjustment is performed as a function of lead time. Precipitation exhibits the opposite trend and is mainly under-forecasted, except in a couple watersheds on the west coast and Texas. The spatial patterns in bias do not vary greatly between time periods. Figure 2-7 illustrates the result of bias-correction and shows that quantile mapping successfully removed bias from the CFSv2 reforecasts.

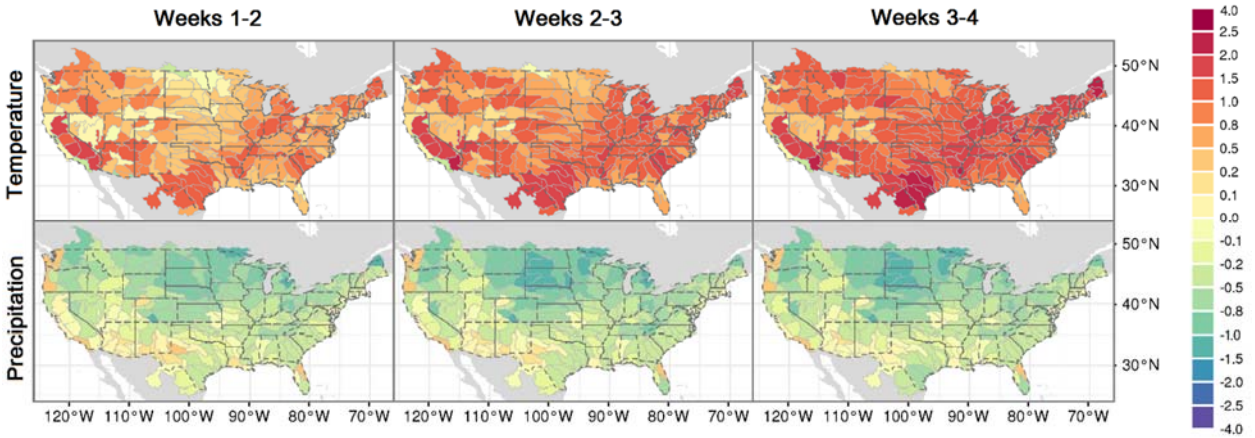

Figure 2-6: CFSv2 bi-weekly bias. Bias of raw CFSv2 temperature (degrees Centigrade) and precipitation rate (mm/d) for each bi-weekly period.

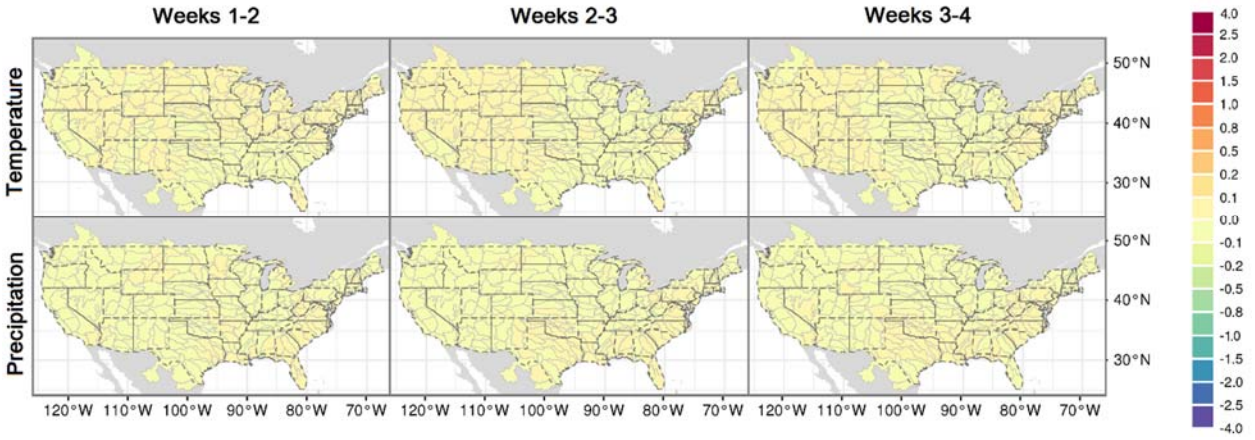

Figure 2-7: Quantile mapped CFSv2 bi-weekly bias. Bias of quantile mapped (QM) CFSv2 bi-weekly forecasts of temperature (degrees Centigrade) and precipitation rate (mm/d) over the 14-day period.

Bias-correction removes the average bias but does not necessarily improve the forecast skill. While some studies have shown that bias-correction can slightly degrade correlation skill (e.g. Mendoza et al., 2017), here the sample of forecasts used in training the bias-correction does not have this impact. The CFSv2 anomaly correlation skill shown in Figure 2-8 shows a similar skill when compared to the raw anomaly correlation shown in Figure 2-3. A watershed specific example of the QM results is shown in Figure 2-9 for the week 2-3 temperature forecast from raw CFSv2 and the QM approach for the Rio Grande-Amistad watershed in southern Texas. The top pair of 1:1 plots show the modeled versus observed forecasts for the raw and QM methods. The raw CFSv2 forecast shows systematic bias as it slightly under-forecasts temperature. The QM approach illustrates the removal of bias as the forecast shift higher and overlaps the 1:1 line. This can also be seen in the time series plot of temperature forecasts and observations for 2000. The QM forecast shifts the forecast up towards the observed temperature throughout the entire year. Other watersheds show similar results of removal of systematic bias where present.

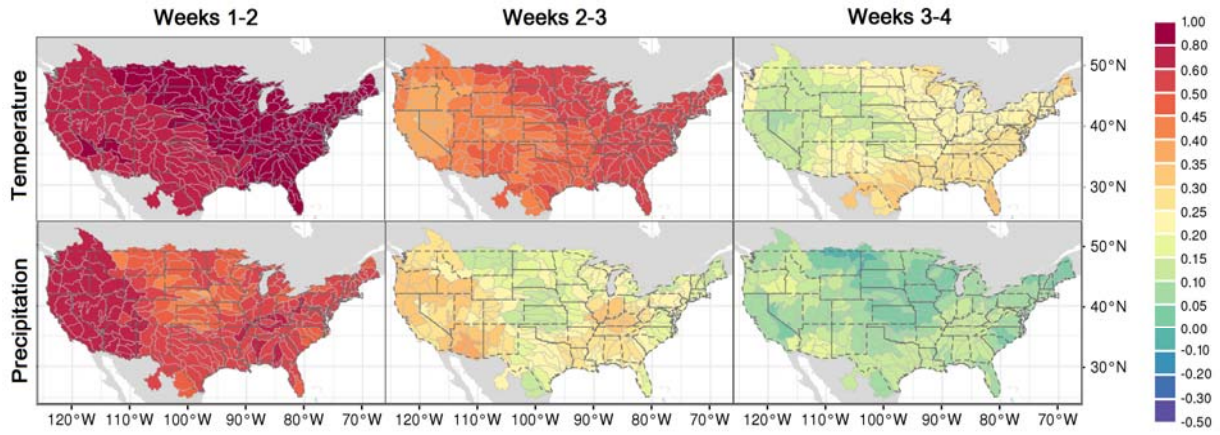

Figure 2-8: Quantile mappedCFSv2 bi-weekly anomaly correlation. CFSv2 anomaly correlation at bi-weekly time step for temperature and precipitation at a HUC-4 watershed scale.

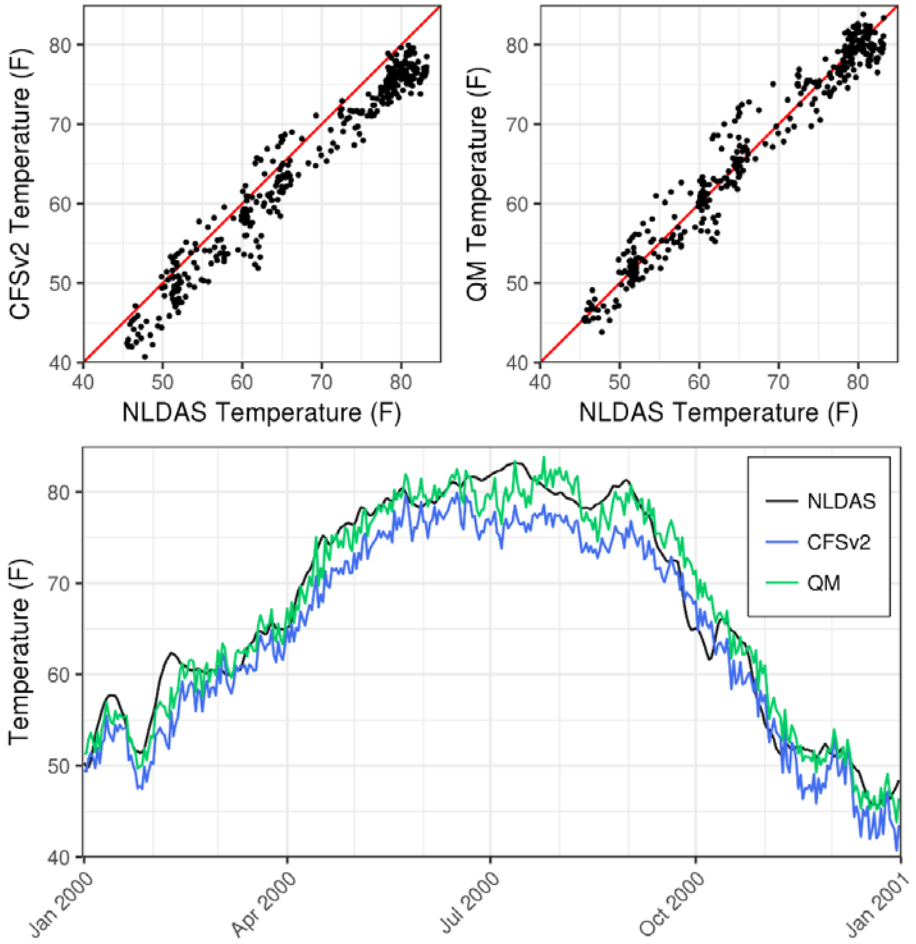

Figure 2-9: Example of raw and quantile mapped CFSv2 forecasts. Comparison of Rio Grande-Amistad watershed 2-3 week temperature forecast from CFSv2 and QM. The pair of figures top display the modeled vs. observed forecast for the full time period (1999-2010). The time series plot at the bottom displays the forecast for CFSv2 and QM in comparison to North American Land Data Assimilation System (NLDAS) for 2000.

In addition to the issues with capturing extreme events, QM can alter the modeled covariance of temperature and precipitation by QM treating them independently. In downscaling of daily weather data, it is common (and important) to preserve interrelationships between precipitation, temperature, and other fields because there are strong observable relationships linked by synoptic atmospheric dynamics. For instance, wet/precipitating days typically have a compressed temperature range versus clear days. At the sub-seasonal timescale, this covariance is typically weaker. The impact of QM on cross-correlations between precipitation and temperature for sub-seasonal bi-weekly CFSv2 predictands is discussed further in Appendix 8.1.

All results shown above are displayed on the S2S Climate Outlooks for Watersheds web-based tool. The results from the verification assessment on an annual and seasonal basis are displayed in tabs for each climate model. Real-time climate forecasts are available as shown in Figure 2-10. The tool allows the user to choose the lead, variable, and forecast displayed. They can hover over watersheds to view the forecasted anomaly and choose to view the raw or the bias-corrected output. This allows users to view their specific watersheds forecast as well as verification.

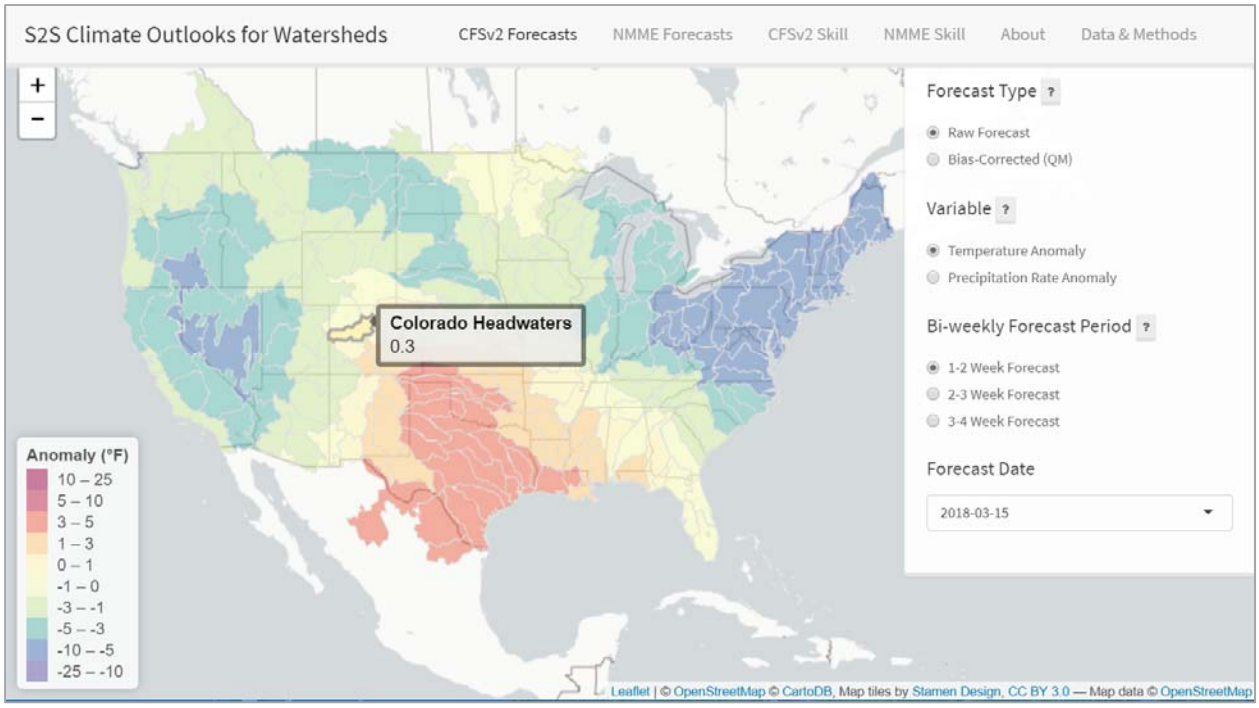

Figure 2-10: S2S Climate Outlooks for Watersheds web-based tool. Sub-seasonal to seasonal (S2S) Climate Outlooks for Watersheds web-based tool allows users to look at specific watersheds forecasts and verification metrics.

2.6 Discussion & Conclusion

The new watershed-scale S2S Climate Outlooks for Watersheds web-based tool offers a new medium for water managers to use climate products. Many academic studies and reports from within the industry indicate that S2S forecasts are obtained and assessed qualitatively by water managers, adding to situational awareness, but are less widely used quantitatively, as data streams input into water management decision support tools and models. Water managers cite perceived poor forecast reliability and skill, mismatched temporal or spatial scales, and lack of resources to ingest forecasts (Bolson et al., 2013; Rayner et al., 2005).

Through this new watershed-scale climate product, we aim to overcome several of these hurdles. The skill and accuracy of climate model forecasts for individual

watersheds has been explored using an initial, small set of forecast metrics, and these results are presented in an accessible format. Water managers can use the web-based tool to view real-time forecasts of precipitation and temperature at a bi-weekly, monthly, or seasonal outlook, and work is underway to provide data access to the watershed forecasts (and hindcasts) in accessible formats (both text and NetCDF). These products aim to bridge the gap of accessibility, spatial and temporal scale, and perception of unusable skill by allowing water managers to look at the climate forecast for their region as well as the skill of the forecast itself based on an analysis of the model reforecast.

We presented the new S2S Climate Outlooks web-tool to water managers at agencies across the US. The site was presented to Southern Nevada Water Authority (SNWA) in early 2017. SNWA is a wholesale water provider in the Las Vegas region who use reservoir levels at Lake Powell and Lake Mead to plan for their future water management needs. They were interested in how this product could be used to better inform streamflow forecasts in the Colorado River Basin.

Reservoir operators in Reclamation found this tool informative and useful. Operators use forecasts of streamflow quantity and timing to project operations of their reservoirs, and operators in the western US stated that this tool could be useful for timing reservoir releases based on projected temperatures during the snowmelt runoff season. Examples of this operation include the Upper Colorado River basin, where releases from Flaming Gorge Reservoir are timed to meet the natural peak in runoff from the Yampa River. Temperature and precipitation S2S forecasts could

also be useful for determining reservoir releases when attempting to meet storage targets in early summer when reservoirs are being filled. Reservoir system operators, however, expressed an interest in a wider variety of forecast products, including time series of past forecasts showing their evolution and agreement with observations, and forecasts of full precipitation and temperature fields rather than forecast anomalies. A California-based watershed manager requested the addition of finer scale watershed breakdowns for the climate forecasts. This work would benefit from a structured analysis of user utility but that focus was not part of this work, which was to demonstrate the concept of watershed-based climate forecast products. Formal surveys could be conducted with a diverse audience of public and private stakeholders to provide feedback and inform future tool development.

The S2S Climate Outlooks web-tool presents a skill assessment of raw CFSv2 forecasts. We show that CFSv2 temperature reforecasts exhibit significant correlation skill in the first two bi-weekly periods, especially in weeks 1-2, while moderate in weeks 2-3, followed by limited skill in weeks 3-4. CFSv2 precipitation forecast show skill over the CONUS domain in weeks 1-2 and regionally in weeks 2-3. In addition to being lead dependent, skill varies seasonally as exemplified in the analysis of CFSv2 precipitation reforecast of weeks 2-3. NMME reforecasts displays skill in Month 1, especially when predicting temperature. Months 2 and 3 show lower skill, especially for precipitation.

In general, one expects that as skill in a forecast improves, the forecast has greater utility, but the actual utility for decision making varies greatly among users

due to a range of factors. These include the resilience of their system to forecast busts, the characteristics of their penalty functions for forecast errors, their sensitivity to forecast accuracy in different parts of the forecast distribution (e.g. high or low flows), among others. Thus, we can only suggest utility based on climate forecast skill. For example, a water manager on the west coast can have reasonable confidence in the spring (MAM) 2-3 week CFSv2 precipitation forecast, but a water manager in Upper Colorado should not have high confidence in the spring 2-3 week precipitation forecast. Similarly, winter (DJF) NMME forecasts of Month 2 temperature for the Pacific Northwest are skillful but show limited skill during the summer.

The raw model forecasts also contain substantial biases, and we find that the application of quantile mapping to post-process the CFSv2 successfully removed bias from CFSv2 bi-weekly reforecasts for precipitation and temperature. Quantile mapping removes systematic bias between the forecasts and observations but does not improve skill or alter forecast reliability directly as a forecast calibration method might. To improve the skill of the climate forecasts, further work is underway to develop statistical post-processing procedures on a watershed by watershed scale that harness larger scale circulation patterns, variability and potential predictability.

At present, this paper describes the first steps toward addressing hurdles to widespread use of S2S prediction in water management applications. The S2S Climate Outlooks for Watersheds tool presented here enhances the quality,

specificity, and accessibility of S2S climate prediction. With wider use of the web-based tool, we intend to improve the product based on user feedback.

3 CHAPTER III: Enhancing Sub-seasonal Climate Forecast Skill through Post-processing at the Scales of Water Management

3.1 Abstract

Sub-seasonal to seasonal (S2S) climate forecasting has become a central component of climate services aimed at improving water management. In some cases, operational S2S climate predictions are translated into inputs for follow-on analyses or models, whereas the S2S predictions on their own may provide for qualitative situational awareness. At the scales of water management, however, S2S climate forecasts often suffer from systematic biases, and low skill and reliability. We assess the potential to improve S2S forecast skill and salience for watershed applications through the use of post-processing to harness additional information from the climate model forecast outputs. To this end, we use a components-based technique – Partial Least Squares Regression (PLSR) – to improve the skill of bi-weekly temperature and precipitation forecasts from the Climate Forecast System version 2 (CFSv2). The PLSR method forms predictor components based on a cross-validated analysis of hindcasts from CFSv2 climate and land surface fields, and the results are benchmarked against raw CFSv2 forecasts, remapped to intermediate-scale watershed areas. We find that post-processing affords marginal to moderate gains in skill in many watersheds, raising climate forecast skill above a usability threshold over the four seasons analyzed. In other locations, however, post-processing fails to improve skill, particularly for extreme events, and can lead to unreliably narrow forecast ranges. This work presents evidence that statistical post-

processing climate forecast system outputs has potential to improve forecast skill, suggesting that a more comprehensive study of approaches for post-processing climate forecasts may be fruitful.

3.2 Introduction & Background

Sub-seasonal to seasonal (S2S) climate forecast skill has received greater attention in recent years due to the potential applications of climate forecasts. Many sectors including public health, disaster preparedness, energy, agriculture, and water management would benefit by applying S2S climate forecasts to their specific needs (White et al., 2017). In the public health sector, S2S forecasts could help predict the probability of floods and droughts at longer leads, which in turn could inform disaster responses and warnings for extreme events. Skillful forecasts would help the energy sector anticipate energy demands and could inform the production of renewable energy sources, such as wind or solar power. Seasonal outlooks are already being used in the agricultural sector to make operational decisions on crop management, planting, irrigation scheduling, fertilizer application, and commodity pricing.

In the water management sector, more skillful forecasts of precipitation and temperature could improve streamflow forecasts informing projections of runoff volume, water levels in rivers and reservoirs, and water supply availability (Raff et al., 2013). Academic studies have indicated that water managers are reluctant to use climate forecasts due to perceived poor forecast skill, inadequate or misaligned temporal or spatial scale, institutional hurdles such as mandated decision

workflows, organizational restraints, and risk aversion (Callahan et al., 1999; Kirchhoff et al., 2013; Rayner et al., 2005; White et al., 2017). Baker et al. (2019) sought to address some of these hurdles by translating and bias-correcting S2S climate forecasts to a watershed spatial unit -- United States Geological Survey (USGS) hydrologic unit code 4 (HUC-4) watersheds -- for bi-weekly, monthly, and seasonal prediction periods. This aggregated forecast product was made available in real-time on the S2S Climate Outlooks for Watersheds web-based tool (http://hydro.rap.ucar.edu/s2s/). Baker et al. (2019) bias-correction to watershed climatologies improved forecast relevance through tailoring forecast outputs, and reduced bias, but did not improve S2S forecast performance for skill metrics other than bias (e.g., correlation).

The increased demand from applications sectors for S2S climate forecast information motivates an exploration of the potential for multi-variate post-processing methods to increase the skill of forecasts. The S2S timescale (2 weeks to 2 months) is a challenging period for climate forecast skill because it falls between shorter and longer, more aggregated timescales when weather forecasts and seasonal climate projections, respectively, exhibit skill (F. Vitart et al., 2016). In weather forecasting, skill comes from initial atmospheric and land surface conditions that tend to have less influence with increasing lead time. Seasonal prediction is influenced by land and ocean conditions such as sea surface temperature (SST) and to a lesser extent soil moisture, and their influence via large scale ocean-climate teleconnection patterns such as El Nino Southern Oscillation

(ENSO) and North Atlantic Oscillation (NAO). The S2S timescale falls in the gap between when initial conditions dominate forecast skill and when coupled climate system dynamics provide sources of atmospheric predictability.

Many studies have investigated the predictability of this time frame, with an increasing recent emphasis on the 3-4 week period. DelSole et al. (2017) explored the predictability of raw CFSv2 precipitation and temperature forecasts during January and July, and found that winter exhibited more predictability than summer and that predictability was linked to large scale climate features such as ENSO and the Madden Julian Oscillation (MJO). This analysis suggests that precipitation and temperature alone exhibit some predictability, but other climate and land surface fields (e.g. SST) could be used to improve week 3-4 forecasts.

There are several strategies to improve S2S climate prediction skill. One approach to improving climate forecast skill is through enhancements to the coupled dynamical climate or earth system models used to make the climate forecasts. This effort is strongly and steadily pursued by the centers that maintain and develop these large-scale dynamical models. For instance, NOAA's operational dynamical model, Climate Forecast System version 2 (CFSv2) improved upon its predecessor, CFSv1, through upgrades to nearly all aspects of the prediction system, including to data assimilation systems, the models' physics and parameterizations, dynamical core, resolution and coupling strategies, which resulted in major improvements to forecast skill (Saha et al., 2014).

A second strategy is to improve climate forecast skill through statistical post-processing of dynamical forecast model outputs. Post-processing is applied through statistically translating raw, large scale dynamical model outputs to a regional scale that is useful for local water management applications (D. Maraun et al., 2010). Raw dynamical model output typically requires post-processing or downscaling (a form of post-processing) to be used in follow-on applications due to systematic biases, unreliable ensemble spread, and/or forecasts' lack of skill. Common statistical post-processing methods include bias-correction, different forms of regression, and circulation pattern based approaches that harness information from large-scale climate predictors. In weather prediction, techniques such as model output statistics (Glahn & Lowry, 1972)that regress atmospheric predictors from numerical weather prediction (NWP) onto surface meteorological variables have been common for decades. More recently, Hamill and Whitaker (2006) popularized hindcast or reforecast datasets by showing analog techniques applied to precipitation could significantly raise the skill of NWP predictions.

Downscaling methods are a class of dynamical or statistical post-processing techniques that translate model-based forecasts to a finer spatial resolution and reduce bias, and are often applied in the context of climate change projection or seasonal forecasting. Tian et al. (2014) compared downscaling techniques for North American Multi-Model Ensemble (NMME) precipitation and temperature forecasts for Alabama, Georgia, and Florida. They found that the locally weighted polynomial regression downscaling method showed higher skill than direct spatial

disaggregation and bias-correction for this region. Many other studies of downscaling techniques have shown improvements to dynamical model outputs for other regions including the entire conterminous United States (CONUS) domain (Gutmann et al., 2014; Pablo A. Mendoza et al., 2014; Andrew W. Wood et al., 2004; Yoon, Mo, & Wood, 2011). Zhao et al. (2017) clarifies the distinction between bias correction methods such as quantile-mapping (Andrew W. Wood et al., 2004), which does not consider forecast skill and merely applies a climatological correction, and forecast calibration, which accounts for forecast skill by adjusting not only forecast mean but also forecast spread, in the case of an ensemble (Zhao et al., 2017). This study will go beyond the removal of bias and attempt to improve the skill of S2S forecasts, which is one of the main hurdles to widespread use by the water management community.

Other statistical post-processing techniques employ additional information from large-scale climate fields to improve dynamical model forecasts. Many studies have focused on improving seasonal precipitation and temperature forecasts (DelSole & Banerjee, 2016; Madadgar et al., 2016; Schepen, Wang, & Robertson, 2014; Ward & Folland, 1991; Xing, Wang, & Yim, 2016). Methods include analog-year models, regression methods, and empirical orthogonal function (EOF) mode techniques. Madadgar et al. (2016) explored forecasting seasonal precipitation over the southwestern United States (US) using a hybrid statistical-dynamical approach. The statistical approach used an analog-year technique based on copula functions informed by teleconnections such as Pacific Decadal Oscillation (PDO), Multivariate

ENSO Index, and Atlantic Multi-decadal Oscillation, and generated weighted NMME model combinations that showed improvements over the raw NMME ensemble mean seasonal precipitation.

Other studies have found value in using model-predicted SSTs instead of empirical climate indices or atmospheric fields. Xing et al. (2016) used Partial Least- Squares Regression (PLSR; Wold, 1966) to predict the principal component (PC) of EOF modes to forecast China summer rainfall using winter SSTs and temperature over land. They found that the summer rainfall prediction skill of the PLSR-EOF method at 4-month lead was significantly higher compared to 1-month lead dynamical model prediction. Another study by McIntosh et al. (2005), explored using PLSR to predict plant growth days using global SSTs. Plant growth days were predicted because they produced higher skill than rainfall predictions.

In this study, we use PLSR to assess the potential to enhance model-based sub-seasonal forecasts of week 2-3 and week 3-4 precipitation and temperature at watershed scales. PLSR has been used in a wide variety of fields, from the first applications in economics (Wold, 1966), to more recently being applied in the physical sciences to predict streamflow (Abudu, King J. Phillip, & Pagano Thomas C., 2010; P. A. Mendoza et al., 2017; Tootle, Singh, Piechota, & Farnham, 2007), teleconnections (Black et al., 2017), precipitation (Xing et al., 2016), and climate variability (Smoliak, Wallace, Lin, & Fu, 2015). Many of these studies have used climate fields to develop empirical forecasts of a predictand of interest. Black et al. (2017) employed PLSR with predictor fields of outgoing longwave radiation (OLR),

300 hPa geopotential height, and 50 hPa geopotential height to predict Northern Hemisphere teleconnection patterns at leads of 3-4 weeks. Tootle et al. (2007) showed improvements to long lead streamflow forecasts at gauges in the US using PLSR with previous spring and summer's SSTs. We apply PLSR with climate forecast model output fields from CFSv2 to predict surface precipitation and temperature.

This paper is organized as follows. We first outline the data and preliminary data processing, summarizing the precipitation and temperature watershed scale processing and CFSv2 predictor field processing. We then describe the PLSR method, verification metrics, and predictor selection. The Results section summarizes results on a seasonal scale and explore individual watershed results, followed by a discussion of the potential use and hurdles associated with the PLSR post-processing method.

3.3 Data

3.3.1 Precipitation and Temperature Analysis at Watershed Scales

The observational dataset used in this study is Phase 2 of the near real-time North American Land Data Assimilation System (NLDAS-2; Xia et al, 2012). NLDAS-2 data are available from 1979 to present at an hour temporal resolution at a 1/8th degree grid spacing. Precipitation and temperature fields from NLDAS-2 are spatially and temporally aggregated bi-weekly forecasts at a USGS HUC-4 watershed scale over the CONUS domain. The process presented in Baker et al. (2019) is summarized here. NLDAS-2 fields are translated to a 1/2-degree grid and

temporally averaged to a daily time step. The fields are then areally aggregated to 202 USGS HUC-4 watersheds through spatially conservative remapping, and then temporally averaged to bi-weekly periods (1-2 week, 2-3 week, and 3-4 week).

3.3.2 CFSv2 Climate and Surface Variable Forecasts

The dynamical climate forecasts used in this study are from the operational fully coupled atmosphere-ocean-land model CFSv2 (Saha et al., 2014). CFSv2 forecasts a variety of climate and land surface variables, including temperature and precipitation rate (here after referred to as precipitation), on a 6-hour time step with a ~100 km grid resolution. Reforecasts are available from 1999 – 2010 with 4 initializations each day at synoptic times 0000 UTC, 0006 UTC, 0012 UTC, and 0018 UTC. Reforecasts lead times extend from 45 days to 9 months depending on the forecast initialization time. The CFSv2 reforecasts are processed in the same fashion as the NLDAS-2 fields, yielding a CFSv2-based HUC-4 forecast dataset that can be directly compared to the NLDAS-2 analysis. We pooled forecasts over a 2-day period (creating 8-member ensembles) to smooth variability in forecast ensemble means from one day to the next.

The CFSv2 climate fields identified as potential predictors are listed in Table 3-1, which also outlines their spatial extent. These candidate predictors were chosen because they linked to North American atmospheric circulation and surface climate and many have been used in prior post-processing studies (e.g., Koster *et al.*, 2017).

The spatial extent for each predictor was prescribed based on each field's region of influence on the CONUS domain, as informed by prior literature (Doblas-

Reyes, García-Serrano, Lienert, Biescas, & Rodrigues, 2013). For instance, Quan et al. (2006) identified tropical and subtropical west SSTs as a source of seasonal temperature and precipitation forecast skill for the CONUS domain. Scaife et al. (2014) found sources of predictability for North American winters in large scale climate circulation patterns such as NAO, jet stream winds, and sea level pressures.

The chosen fields in Table 3-1 were spatially aggregated to a 2-degree grid resolution to reduce computational processing time. The climate and land surface fields were then aggregated to bi-weekly periods and pooled into 8-member lagged ensembles as in the processing of precipitation and temperature forecasts on a watershed scale.

Table 3-1. CFSv2 climate and surface predictor fields

Predictor Name	***Variable Name***	***Spatial Extent***
500 mb Geopotential Height	*hgt*	25 N – 80 N x 100 E – 340 E
Specific Humidity (2m)	*q2m*	20 S – 70 N x 100 E – 340 E
Surface Pressure	*prs*	20 S – 30 N x 100 E – 340 E
Sea Level Pressure	*slp*	20 S – 30 N x 100 E – 340 E
Precipitable Water	*pwt*	20 S – 70 N x 100 E – 340 E
Zonal Winds (850 mb)	*uwnd*	0 N – 80 N x 100 E – 340 E
Meridional Winds (850 mb)	*vwnd*	0 N – 80 N x 100 E – 340 E
Sea Surface Temperature	*sst*	20 S – 80 N x 100 E – 360 E
Outgoing Longwave Radiation	*olr*	20 S – 20 N x 100 E – 340 E
Surface Temperature	*tmp*	5 N – 75 N x 5 E – 125 E
Surface Precipitation Rate	*prt*	5 N – 75 N x 5 E – 125 E

3.4 Methods

3.4.1 Partial Least Squares Regression (PLSR)

PLSR is a components-based regression method similar to principal component regression (PCR) and canonical correlation analysis (CCA) that combines features of principal component analysis (PCA) and multiple linear regression (Abdi, 2010). PLSR differs from PCR since it forms predictor components that are ordered to explain the maximum covariance of the predictors with a single-valued predictand, while PCR first uses principal component analysis (PCA) to form components that are ordered to maximize only the explained variance of the predictors, and then regresses the components against the predictand. CCA is similar to PLSR in maximizing covariance between predictors and predictand, but allows for multi-variate predictands. PLSR provides for dimension reduction and avoids multi-collinearity in analyses with large sets of predictors, such as gridded model fields.

The PLSR method is detailed in papers such as Abdi et al. (2010) and Smoliak et al. (2010), and is summarized here. The predictors **X** (independent variables) can be decomposed through the following relationship:

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{P}^T \quad with \quad \boldsymbol{Z}^T\boldsymbol{Z} = \boldsymbol{I} \qquad \text{Eq. 3-1}$$

where **Z** is the latent vectors or scores (sometimes referred to as PLS (partial least squares) predictors), **P** is the loadings, and **I** is the identity vector. This equation is also used in PCA to decompose **X**. Similarly, the predictand **Y** (dependent variable) can be estimated through the relationship:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{Z}\boldsymbol{B}\boldsymbol{C}^T \qquad \text{Eq. 3-2}$$

where $\hat{\boldsymbol{Y}}$ is the estimate of **Y**, **B** is the regression weights, and **C** is the weights of the predictand. This system of equations does not alone have enough information to be solved; additional conditions are required to solve for the latent vectors **Z**. To find the latent vectors, two sets of weights, **w** and **c**, are found that form linear combinations of **X** and **Y** that maximize the covariance**:**

$$\boldsymbol{z} = \boldsymbol{Xw} \quad and \quad \boldsymbol{u} = \boldsymbol{Yc} \qquad \text{Eq. 3-3}$$

with the following constraints

$$\boldsymbol{w}^T\boldsymbol{W} = 1, \quad \boldsymbol{z}^T\boldsymbol{z} = 1, \quad and \quad \boldsymbol{z}^T\boldsymbol{u} = maximal \qquad \text{Eq. 3-4}$$

This process is done iteratively. Once the first latent vector **Z** is solved such that $\mathbf{z}^{\mathrm{T}}\mathbf{u}$ is maximized, it is subtracted from **X** and **Y** through an ordinary least squares regression to form residual matrices. The processes are then re-iterated to solve for the predictor from these partially deflated residual matrices. This process can be done using algorithms such as the SIMPLS (de Jong, 1993) and ensures that the latent vectors are mutually orthogonal components with respect to the predictors and predictand.

Smoliak et al. (2015) investigated PLSR performance related to Northern Hemisphere air temperature variability. The predictand types tested were: (1) point-wise where the predictand is a single grid point or an area average, (2) PC-wise where the target is a PC, and (3) field-wise where the predictand is an entire field. They found that point-wise and PC-wise PLSR methods explained more variance in the predictand with a lower number of field predictors and that all performed slightly better than PCR. In this analysis, we focus on the point-wise

predictand approach where we predict watershed aggregate NLDAS-2 precipitation and temperature at bi-weekly periods of 2-3 and 3-4 weeks. Both the predictors and predictands are normalized with a mean of 0 and a standard deviation of 1 to remove emphasis on predictor regions with relatively large amplitudes of variation.

A separate PLSR analysis is performed for each watershed with one model variable for each month. PLSR models are trained using data from the adjacent months meaning each year of data has 3 months of data available to train the model. For example, the PLSR model for a forecast of January 1st for the week 2-3 predictand would be trained using CFSv2 predictors and NLDAS-2 analyses from all forecast-analysis pairs in December, January, and February. The PLSR models are cross-validated by separating the 12-year reforecast period into training and verification periods – in this case, by dropping the year in which forecasts are verified from the training period. The nominal training and test sample sizes are approximately 1001 (11*91) and 91, respectively, although the use of lagged ensembles reduces the effective sample sizes due to a lack of serial independence. The analysis utilizes the R statistical software package *pls* to perform PLSR, and separates the training and test periods outside the *pls* function so that test period data cannot influence the component training.

3.4.2 Verification Metrics

Verification metrics are applied to compare the performance of ensemble-mean precipitation and temperature forecasts from PLSR-based post-processing with raw watershed-scale forecasts from CFSv2. The main verification metric presented in

this paper is the anomaly correlation (ACC), which is commonly used in the climate prediction community to measure the association of forecasts and observations. A score of 1 indicates a perfect forecast and a score of 0 or below represents a forecast that is not skillful. Other deterministic forecast verification metrics calculated for this study include mean absolute error (MAE) and bias (not shown), metrics that are familiar to water managers. The metrics are calculated separately for all forecasts in each 3-month seasonal basis to show seasonal variability in forecast performance. To translate forecasts and observations into anomalies, climatologies for each watershed, lead, and day of year were estimated based on averaging across a 15-day window (+/- 7 days from forecast date).

3.4.3 CFSv2 Predictor Selection

In operational S2S empirical prediction, there is discomfort with models that are entirely data-driven – i.e., in which predictors are free to vary in space and time – due to the risk that predictor selection is spuriously driven by training sample noise. There is also the practical difficulty of linking changes over time or across space in prediction outcomes based on changes in predictors, if predictors change from initialization date to initialization date, or from location to location. On the other hand, it is likely that dynamics do vary in space and by season, such that an optimal predictor set will also vary, and that data driven predictor selection can exploit varying sources of predictability if allowed to vary within a prediction approach. Another choice that must be made along the purely prescriptive to data-driven spectrum is the number of components or predictors to include. Though

there exist quantitative metrics for predictor adoption(e.g., the Bayesian Information Criterion, BIC; Schwarz, 1978) or regularization approaches to reduce the risk of overfitting (e.g., least absolute shrinkage and selection operator, LASSO; Santosa and Symes, 1986), they are also vulnerable to sampling uncertainty. In this study, we explore the variation in optimal predictor selection and the optimal number of predictor variables and components (based on cross-validated results), but we present conservative findings based on PLSR models that use only a limited number of generally strong predictors that are limited to two components. The goal of the study is not to exhaustively optimize empirical post-processing models but to present a more general outlook for the potential enhancement of raw climate model forecast outputs at watershed scales through the addition of circulation-scale predictors in a post-processing framework.

To investigate the relative importance of the eleven potential CFSv2 predictors (see Table 3-1) for predicting precipitation and temperature, we test each predictor individually, performing cross-validated PLSR for each predictand, watershed, lead time, and forecast month. An example of predicting July weeks 2-3 precipitation illustrates this predictor evaluation in Figure 3-1. We calculate the ACC of the PLSR forecast and the raw CFSv2 forecast for each watershed, and identify forecasts for which a predictor could improve skill relative to the raw CFSv2. The raw CFSv2 ACC is shown if Figure 3-1a, and the increase in ACC using the best performing PLSR variable is shown by colored watersheds in Figure 1b. The maximum ACC (either raw CFSv2 or PLSR) is displayed in Figure 3-1c with

the predictor that resulted in the highest ACC in Figure 3-1d. If the raw watershed scale CFSv2 forecast was not outperformed by a PLSR model, the watershed was not colored in Figure 3-1b and Figure 3-1d.

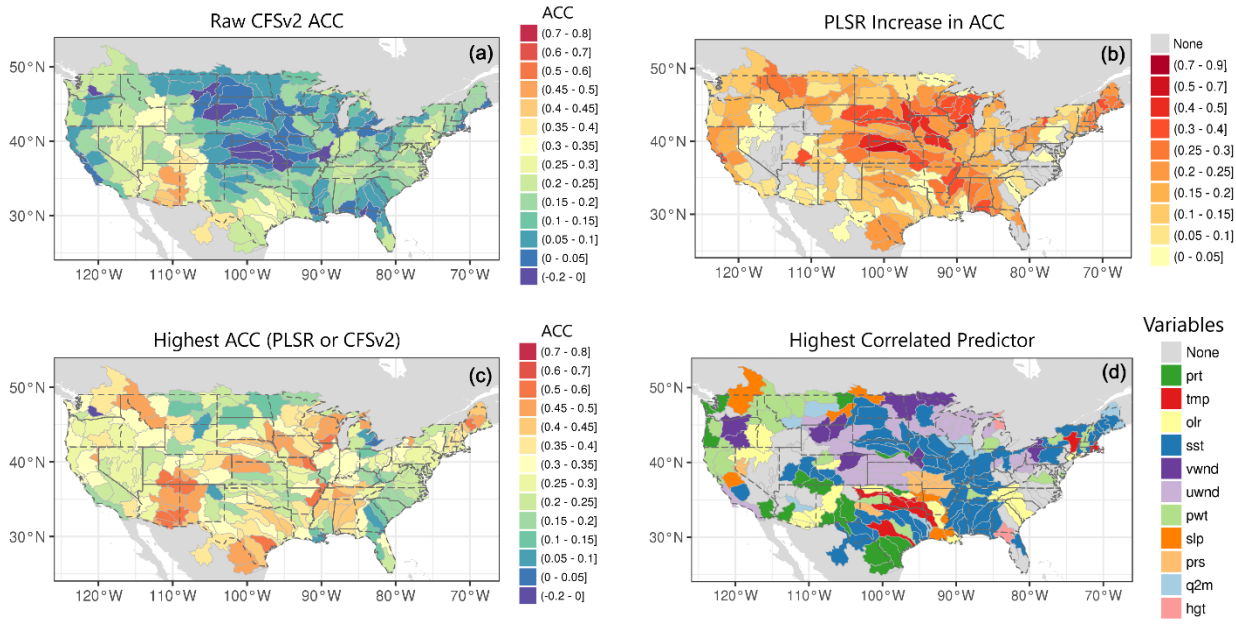

Figure 3-1. Visual analysis of predictor performance for forecasts of July weeks 2-3 precipitation. The figure panels are (a) raw CFSv2 ACC, (b) increase in ACC from raw CFSv2 to PLSR with best predictor, (c) maximum ACC from either the raw CFSv2 forecast or PLSR, and (d) CFSv2 predictor corresponding to the highest ACC from PLSR; the predictor variables are summarized in Table 3-1. The gray watersheds did not show improvement for PLSR models for any predictors.

In the July weeks 2-3 precipitation forecast example, the raw CFSv2 forecast performed poorly over most of the CONUS domain except in the four corners region and areas to the south. Since the raw skill is low, there is a potential for large increases in ACC for many watersheds. The highest increases in skill are found over regions with the lowest raw CFSv2 skill, for example in the Great Plains. The predictors resulting in the highest skill improvements were SST, precipitation, temperature, and meridional and zonal winds (Figure 3-1d). The best predictor varies from watershed to watershed, with limited regional consistency.

Hypothesizing that multiple predictors may perform better than the raw forecast and the ACC differences between predictors are driven to some extent by sample noise (despite the cross-validation), we assessed whether individual predictors may show more regional consistency. To do this, we calculate if the predictors are among the top three predictors (and have skill above the raw forecasts). If so, this would provide a rationale for more general, rather than watershed-specific predictors. The top 3 predictors for the July weeks 2-3 precipitation forecasts at each watershed are shown in Figure 3-2. Watersheds are displayed in color if the predictor ranks in the top 3 predictors for a watershed. The color is solid if it provides higher ACC than the raw CFSv2 forecast, and transparent if not. For this example, the best predictors measured by ACC are SST, followed by wind speeds (both meridional and zonal), outgoing longwave radiation (OLR), and temperature.

Exploratory data analysis of the type described above was repeated for precipitation and temperature for all leads in January and July, confirming the lack of regional consistency for best-performing predictors. The results suggest that for almost any watershed of interest, there may be an optimal set of atmospheric predictors that can be harnessed to augment the skill of CFSv2 model output. We opt here to assess whether a conservative baseline of using only the most frequently strong predictors (SST) in combination with the forecasted variable of interest (temperature or precipitation), is sufficient to enhance forecast skill in all study watersheds. SST, temperature, and precipitation were among the best predictors in

all instances. The focus on these three predictors also facilitates interpreting the validity of PLSR components in representing circulation dynamics that are consistent with variability in precipitation and temperature.

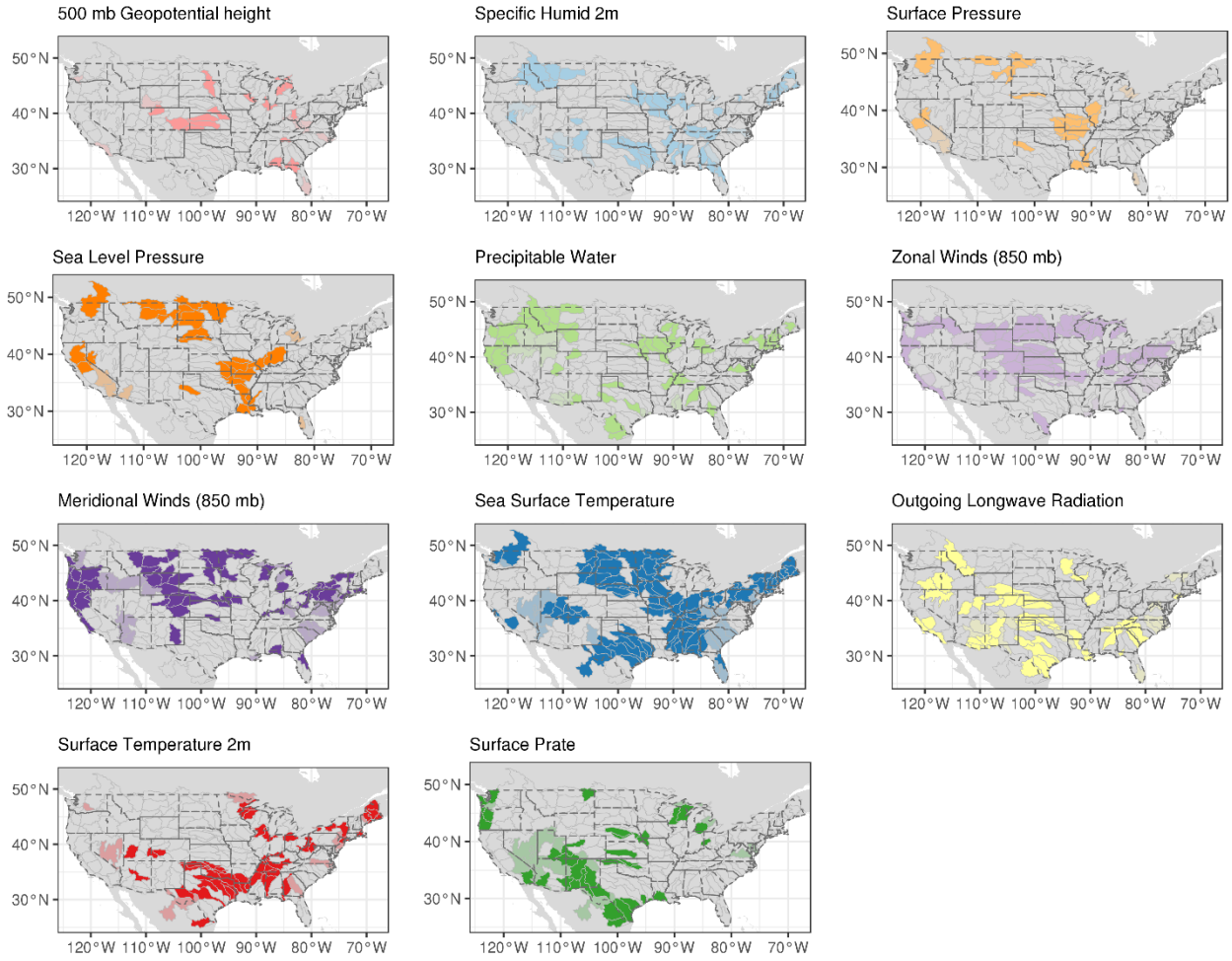

Figure 3-2. Top 3 PLSR predictors for July week 2-3 precipitation. The watershed is colored if the mapped variable has the top 3 ACC the watershed. If the raw CFSv2 forecast has a higher ACC than the PLSR model, the color is not solid.

3.5 Results

3.5.1 Individual watershed example

The results of post-processing with PLSR varied across watersheds, with some watersheds performing well using PLSR with predictors of SST and precipitation or temperature, while other watersheds performed poorly. Before turning to CONUS-

wide results, we illustrate the performance of PLSR for a single watershed using scatterplots observations versus raw CFSv2 and PLSR forecasted values, and maps of PLSR loadings for each predictor and component. PLSR loadings provide insight into the regions of the predictor fields that explain the highest covariance between the predictors and the predictand, which in turn is indicative of the climate dynamics informing the PLSR forecast.

We focus here on a watershed where PLSR performed well with the baseline predictors of SST and temperature. Figure 3-3 shows the raw CFSv2 (a) and PLSR (b) forecasts for June 3-4 week temperature in the Neosho & Verdigris watershed in southeastern Kansas. The raw CFSv2 forecast does not differentiate between hot and cold temperature event with an ACC of 0.03 and MAE of 1.4 degrees C over the bi-weekly period. The PLSR forecast reduces the forecast spread considerably and captures the extreme events much better. The ACC of the baseline PLSR model is 0.54 and the MAE is 0.95 degrees C.

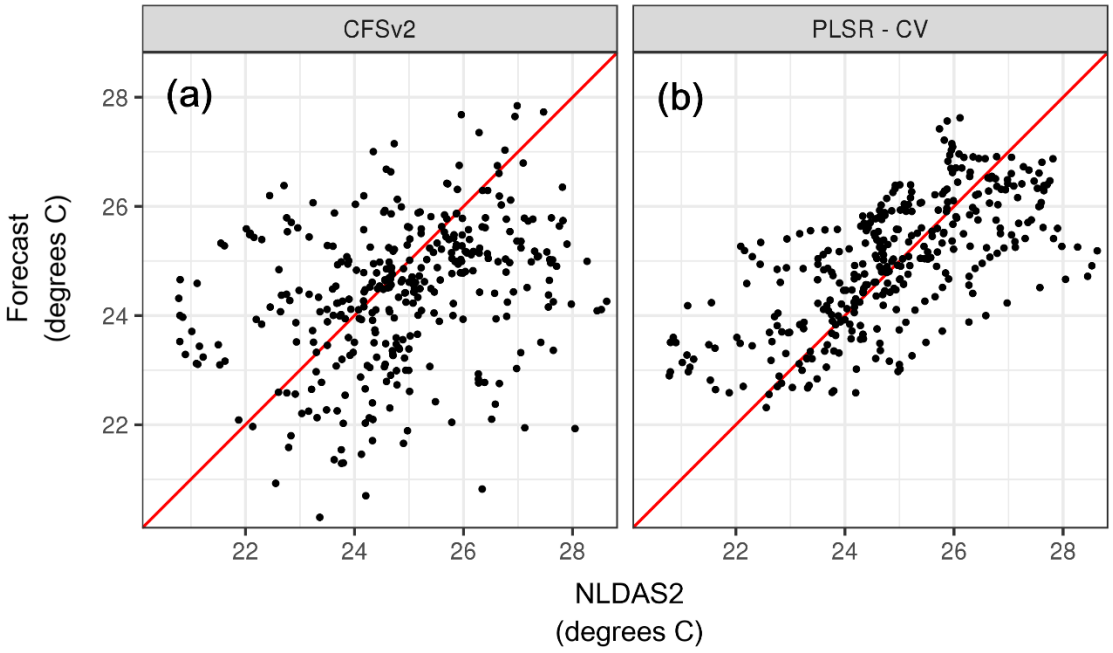

Figure 3-3. June 3-4 week temperature forecasts are plotted versus NLDAS-2 observations for the Neosho & Verdigris watershed in southeastern Kansas. The raw CFSv2 forecast is shown in panel (a) and the PLSR forecast is shown in panel (b).

The loadings for the PLSR model are shown in Figure 3-4 for SST and temperature. The first component loading patterns in the SST field has high magnitude loadings in the northern regions of the Pacific and Atlantic oceans. The first component of temperature has high loadings over most of the domain except in southwestern Texas and northeastern Mexico where there is an area of lower loadings. The second components for both predictors have lower magnitudes of loadings. For component 2, SST has highest loadings in equatorial regions and temperature has high loadings in the regions with lower loadings in component 1.

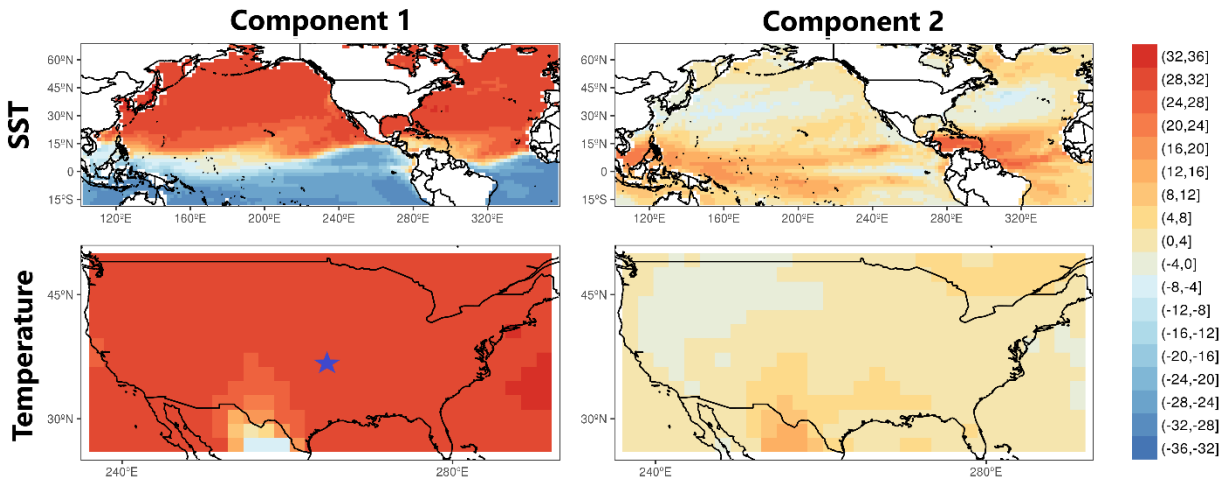

Figure 3-4. Mean cross-validated loadings for PLSR model of June week 3-4 temperature forecast for the Neosho & Verdigris watershed. The predictors used in the PSLR model are SST and temperature, which are represented as rows. The two components are shown as columns. The star represents the location of the watershed in the domain.

For a second example, we look at a watershed where PLSR did not perform well with the baseline predictors of SST and precipitation. Figure 3-5 shows the raw CFSv2 (a) and PLSR (b) forecasts for July 3-4 week precipitation in the Upper Pecos watershed in New Mexico. Similar to the first example, the raw CFSv2 forecast has difficulties differentiating between large and small precipitation event with an ACC of -0.01 and MAE of 1.15 mm/d over the bi-weekly period. The baseline PLSR forecast does not perform much better than the raw CFSv2 forecast. PLSR reduces the forecast spread considerably, but still does not differentiate well between large and small precipitation events. The ACC of the baseline PLSR model is 0.11 and the MAE is 0.98 mm/d.

We explored other PLSR predictors to determine if other predictors would perform better for forecasting July precipitation in the Upper Pecos watershed or if there was a lack predictability in this time period and watershed. We found that

PLSR with zonal and meridional winds and precipitation as predictors performed much better than the baseline PLSR predictors. The alternative PLSR model forecast is plotted against NLDAS observations in Figure 3-5(c). The alternative PLSR forecast performs much better than the raw CFSv2 and the baseline PLSR forecasts. The ACC was increased to 0.36 and the MAE decreased to 0.90 mm/d. The alternative PLSR forecast has a larger spread than the baseline forecast but still does not capture the extremely large events well.

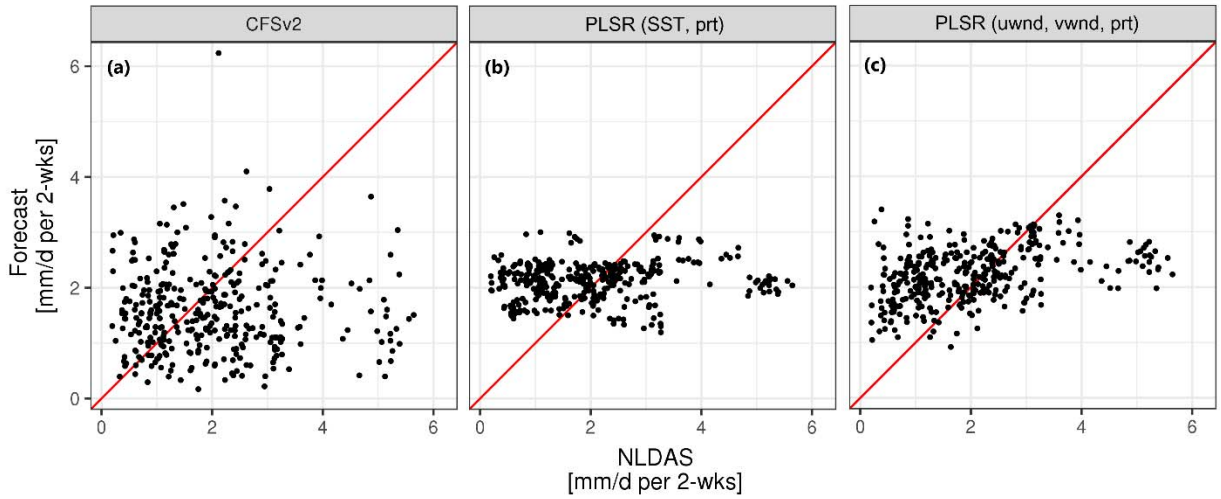

Figure 3-5: July 3-4 week temperature forecasts are plotted versus NLDAS-2 observations for the Upper Pecos watershed. The raw CFSv2 forecast is shown in panel (a), the PLSR forecast with baseline predictors is shown in panel (b), and the PLSR forecast with alternative predictors is in panel (c).

The loadings for the alternative PLSR model are shown in Figure 3-6 for zonal winds, meridional winds and precipitation. The first component loading patterns in the wind fields have many distinct areas of high magnitude loadings, more than any other predictor analyzed. Regions of the wind fields in the Pacific Ocean show dipole patterns where very low and high loadings right next to each other. For zonal winds, the high loading is over the Aleutian Islands directly north

of a low loading in the central Pacific. For the meridional winds, the high magnitude loadings are east and west of each other in the central Pacific. There are also regions of high magnitude loadings over North America. The wind circulation patterns in the tropical Pacific could be affected by MJO which is known to affect summer precipitation in the southwestern US during the North American Monsoon (Lorenz & Hartmann, 2006). The wind loadings for component 2 are less distinct and don't have as large of magnitude loadings compared to the first component. The precipitation loadings show a region of high loadings over the watershed and over the northwestern CONUS domain. The second precipitation component does not have as high of loadings as the first component.

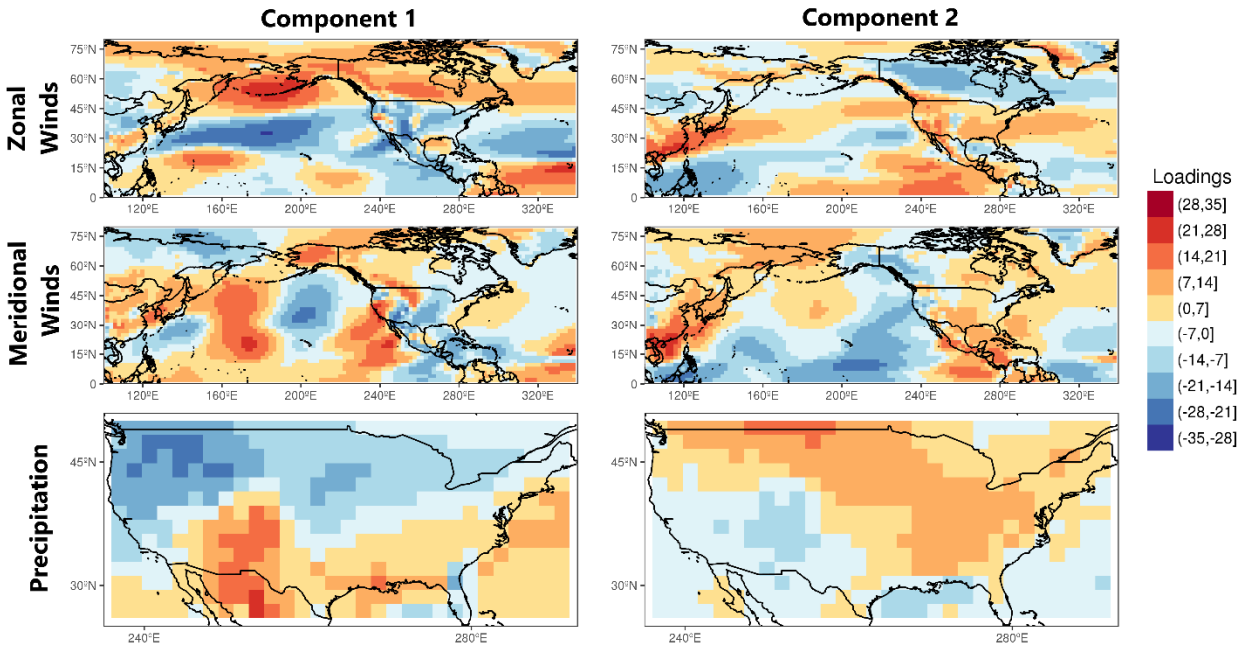

Figure 3-6: Mean cross-validated loadings for PLSR model of July week 3-4 precipitation forecast for the Upper Pecos watershed. The predictors used in the alternative PSLR model are zonal winds, meridional winds and precipitation which are represented as rows and the two components are shown as columns.

These watershed examples show two different examples of PLSR performance. The Neosho & Verdigris watershed performed well with baseline

PLSR predictors when forecasting week 3-4 January temperature. The second watershed explored, the Upper Pecos, did not perform well with baseline predictors when forecasting week 3-4 July precipitation. Through further analysis we showed that better forecasts could be made with alternative predictors. The improvements in ACC in both examples were modest, but enough to find some usability in the forecast. There could still be improvements in forecasting extreme events.

3.5.2 Seasonal CONUS domain analysis

The PLSR post-processing approach was applied to all CONUS HUC-4 watersheds for bi-weekly periods of weeks 2-3 and 3-4. The PLSR model predictors are concurrent gridded SST and either precipitation or temperature, depending which is the predictand. Verification metrics were calculated for raw CFSv2 and PLSR forecast on a seasonal basis for DJF, MAM, JJA, and SON, respectively Dec-Jan-Feb, Mar-Apr-May, Jun-Jul-Aug, and Sept-Oct-Nov.

Figure 3-7 shows the ACC for weeks 3-4 temperature forecasts. The raw CFSv2 ACC shown in the left column varies seasonally and spatially over the CONUS domain. The highest raw CFSv2 skill occurs during DJF in the eastern half of the US, while the western US doesn't exhibit skill. The lowest raw CFSv2 skill for 3-4 week temperature forecast is during MAM in the southwest and Rocky Mountains regions. JJA and SON show mostly lower forecast skill over the entire domain, except for in JJA where there is skill present in eastern Texas and Louisiana.

The far right column in Figure 3-7 shows the improvements in ACC through post-processing. The largest increases in ACC occur during MAM in watersheds in Utah and Colorado and during SON in Florida. The center column in Figure 3-7 shows the new ACC with the best model, either raw CFSv2 or PLSR, whichever exhibit the maximum ACC. This column shows that PLSR increases the ACC to above 0.3 in some cases, which is a potential usability threshold used by the Climate Prediction Center (O'Lenic et al., 2008). This illustrates that post-processing can increase skill enough to allow watersheds to exhibit usable skill where there was not any skill with the raw CFSv2 forecast. In general, however, it must be acknowledged that weeks 3-4 precipitation forecast skill from CFSv2 is not encouraging.

The MAE for the weeks 3-4 temperature forecasts is shown in Figure 3-8. The best models for each watersheds are based on the maximum ACC and are the same for the MAE and ACC analyses. The raw CFSv2 MAE is highest in DJF in the northern US where MAE can be upwards of 4 degrees C for the 2-week mean temperature. The lowest MAE is during JJA in watersheds along the east coast. We note that the seasonal and spatial patterns of the best and worst performing watersheds according to MAE do not correspond to the performance of these watersheds with respect to ACC. For example, the lowest skill in the raw CFSv2 forecast (Figure 3-7) during MAM does not correspond to the highest regions of MAE during MAM. Biases in the precipitation a region receives during the season

could affect the relative magnitudes of MAE without similarly impacting correlation.

The reduction in MAE through post-processing can be seen in the right column of Figure 3-8. The largest improvements in MAE occur along the west coast, especially in the Sierra Nevada mountain range in California during JJA illustrating that CFSv2 may not be accurately accounting for mountains in this region. This also corresponds to an increase in ACC (Figure 3-7), though it was not as drastic as the decrease in MAE. Most watersheds only show decreases in MAE of 0 to 1 degrees C for the 2-week mean temperature. When comparing the reduction in bias with PLSR to that of quantile mapping in Baker et al. (2019), we see less reduction in bias with PSLR.

For the ACC and MAE analysis for the 2-3 week temperature forecasts (not shown), the raw CFSv2 forecasts perform equal to or better than post-processed PLSR results in most watersheds. This is because the 2-3 week raw CFSv2 temperature skill is fairly high over the CONUS domain for all seasons.

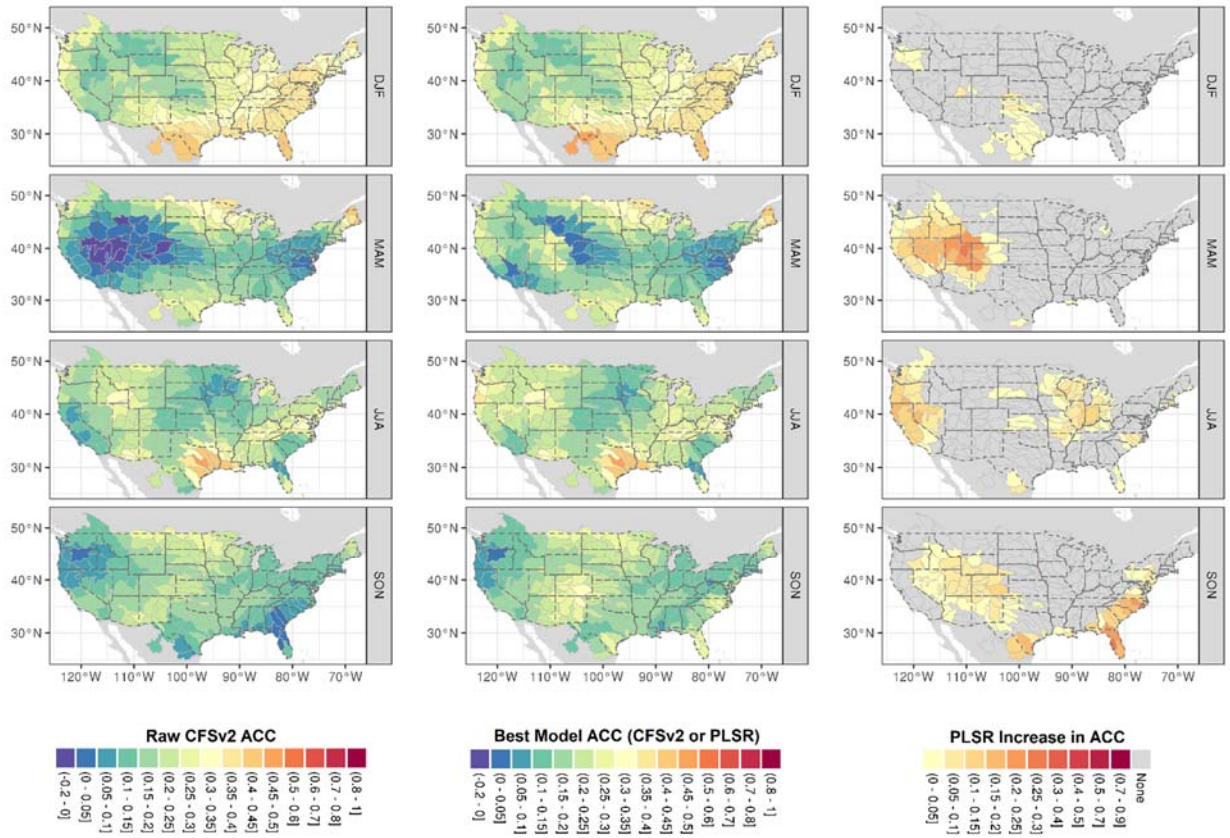

Figure 3-7. ACC results for 3-4 week temperature forecasts on a seasonal basis. The raw CFSv2 ACC is presented in the left column, the best model (either PLSR or raw CFSv2) ACC is shown in the center column, and the increase in ACC from PLSR is shown in the right column. The watersheds that are not colored (visible as gray) in the right column are where the raw CFSv2 forecast performed better than the PLSR forecast.

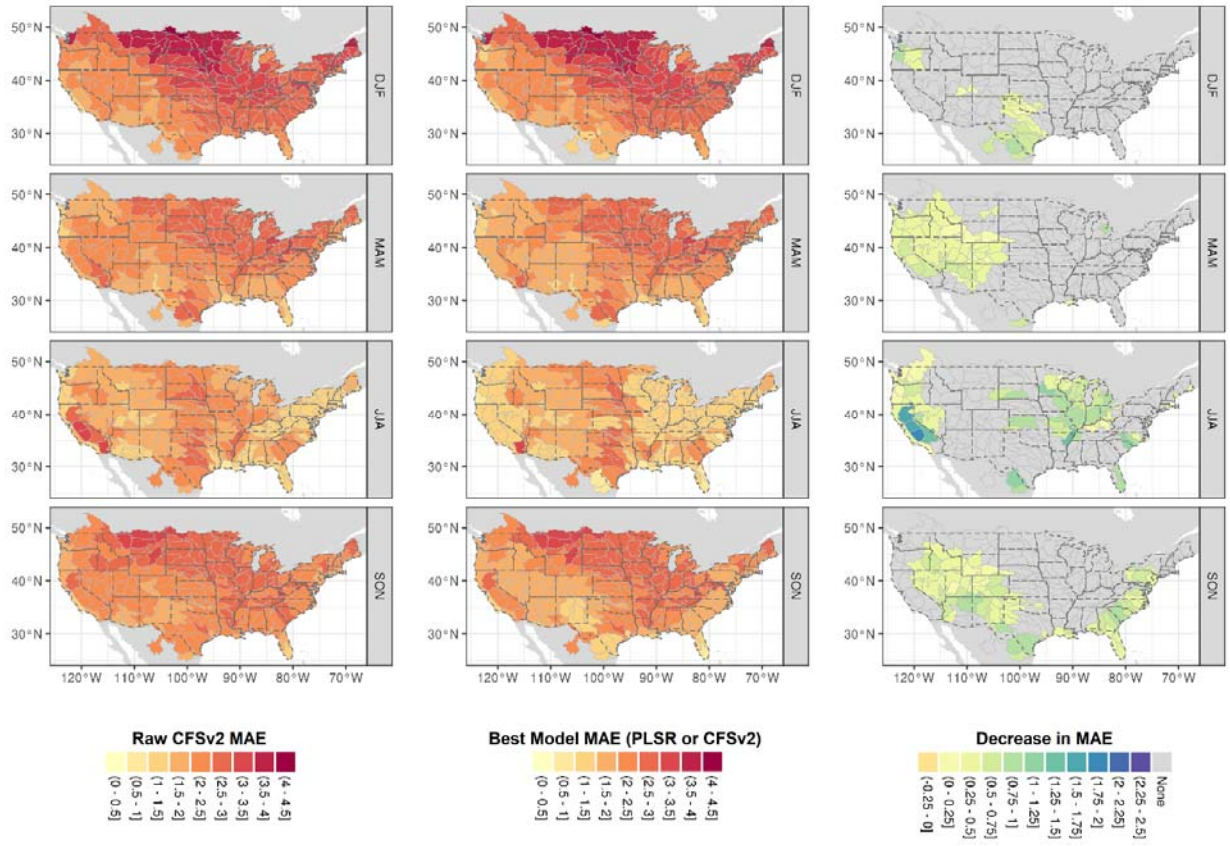

Figure 3-8. MAE results for 3-4 week temperature forecasts on a seasonal basis. The raw CFSv2 MAE is presented in the left column, the best model (either PLSR or raw CFSv2) MAE is shown in the center column, and the increase in MAE from PLSR is shown in the right column. The watersheds that are not colored (visible as gray) in the right column are where the raw CFSv2 forecast performed better than the PLSR forecast.

The seasonal ACC results for the 2-3 week precipitation forecasts (Figure 3-9) show that the raw CFSv2 forecasts have usable skill in watersheds in the western US and in the Great Lakes region during DJF. Lower skill values are shown in Texas, Louisiana, Alabama, and Kansas during MAM. During JJA, many watersheds show lower skill except a few watersheds in southern Idaho, northern Utah, and Nevada, which show areas with skill above 0.35. The increase in ACC with PLSR as shown by the right column of Figure 3-9 varies with season and watershed location. All seasons have watersheds that exhibit skill increases through post-processing. There is a consistent increase in ACC in the northeastern US over most seasons, with the largest increase during JJA. Watersheds in the north central US also show increases in skill over most seasons. In some of these watersheds in the north central US, the ACC increase is high enough to have usable skill with the post-processed PLSR models.

The spatial pattern of watersheds that show an increase in skill with PLSR for precipitation are quite different than that of the temperature results. The watersheds highlighted in the right column in Figure 3-9 are more spread out with more watersheds highlighted alone. For the temperature results in the right column of Figure 3-7, there was more regional consistency to where PLSR performed well. This could indicate that the predictors in component form, such as precipitation, may be poor predictors for modeling precipitation, which is consistent with the fact that the raw watershed precipitation forecast performs poorly.

The MAE for the 2-3 week precipitation forecasts are shown in Figure 3-10. The raw CFSv2 forecasts has the largest MAE along the west coast during SON and DJF, probably due to the prediction errors of large rain events. The lowest MAE is in the western half of the US for most seasons. The decrease in MAE, as shown in the right column, is relatively uniform over most watersheds. The post-processed PLSR models show some watersheds where the MAE increases slightly. This occurs in a few watersheds during JJA and SON. Overall, this increase in MAE is small (< 0.25 mm/d) and occurs in watersheds where the ACC increase is relatively large (around 0.1).

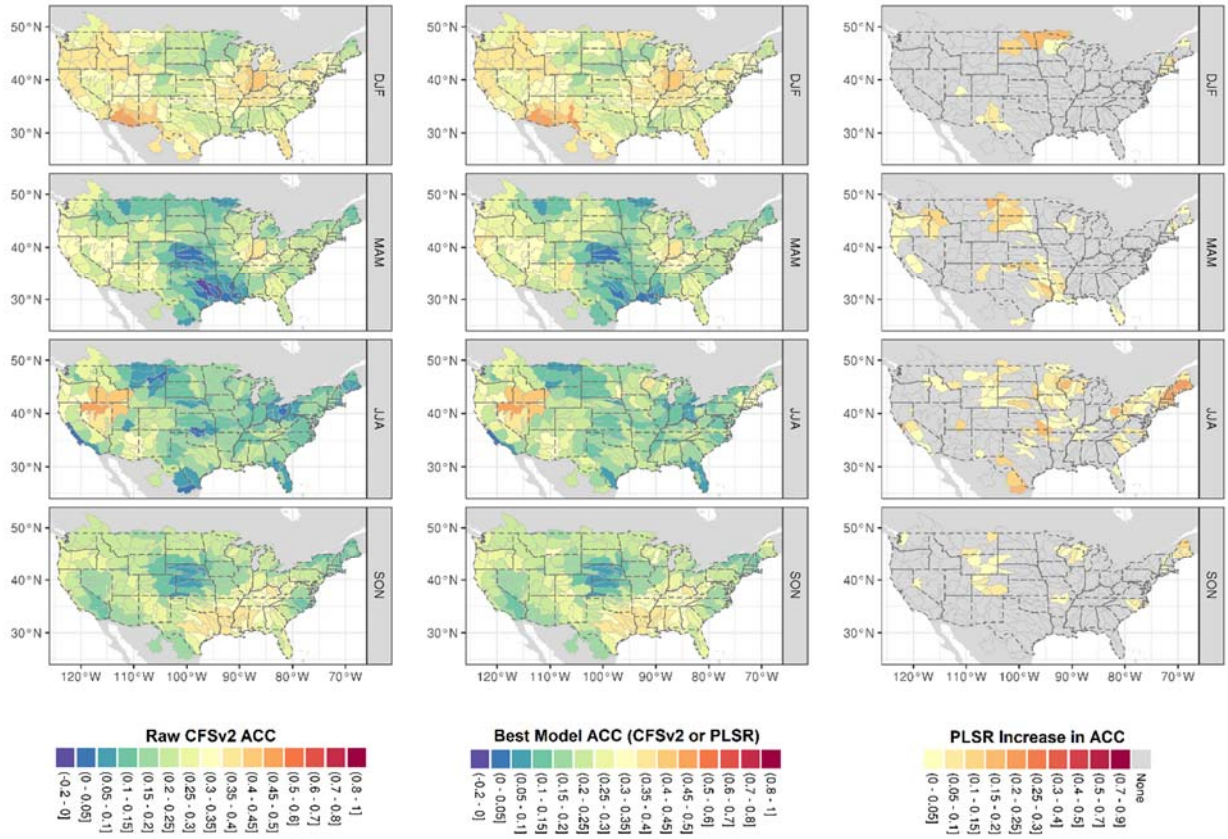

Figure 3-9. ACC results for 2-3 week precipitation forecasts on a seasonal basis. The format is the same as Figure 3-7.

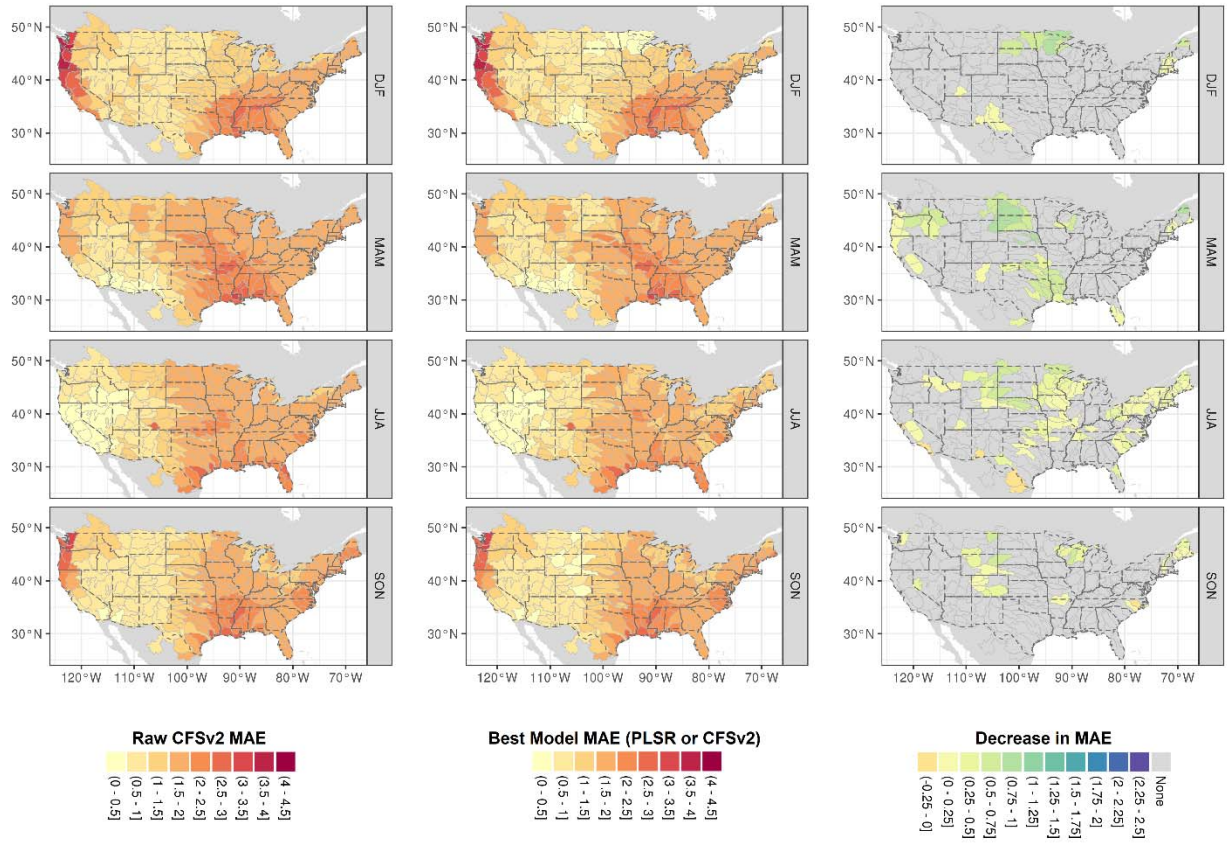

Figure 3-10. MAE results for 2-3 week precipitation forecasts on a seasonal basis. The format is the same as Figure 3-8.

The ACC for the 3-4 week precipitation forecasts are shown in Figure 3-11. The raw CFSv2 ACC is very low with little to no skill for the 3-4 week precipitation forecasts for most watersheds and seasons. The raw CFSv2 forecasts are lowest during MAM and JJA over the CONUS domain. The post-processed PLSR ACC is higher than the raw CFSv2 forecast for many watersheds and seasons, especially during JJA. The post-processed results do show increases in skill exceeding 0.3 in some watersheds. Specifically, a few watersheds in Texas, New Mexico, and North Dakota during DJF show large increases in ACC where watershed PLSR forecasts become usable. Other watersheds show large increases in skill, but very few have an ACC increase to a usable level.

The MAE for the 3-4 week precipitation forecasts is shown in Figure 3-12. The raw CFSv2 MAE for the 3-4 precipitation forecast is similar to the 2-3 week precipitation forecast. The magnitudes and spatial patterns of the error are comparable with some watersheds showing lower MAE in the 3-4 week time period. The decrease in MAE is relatively small for most watersheds that used the post-processed PLSR model for their forecast. The largest decrease in MAE are during DJF and MAM since these are the seasons that receive higher precipitation for most of the CONUS domain. There are a couple of watersheds, similar to those in the 2-3 week forecasts, which show a small increase in MAE.

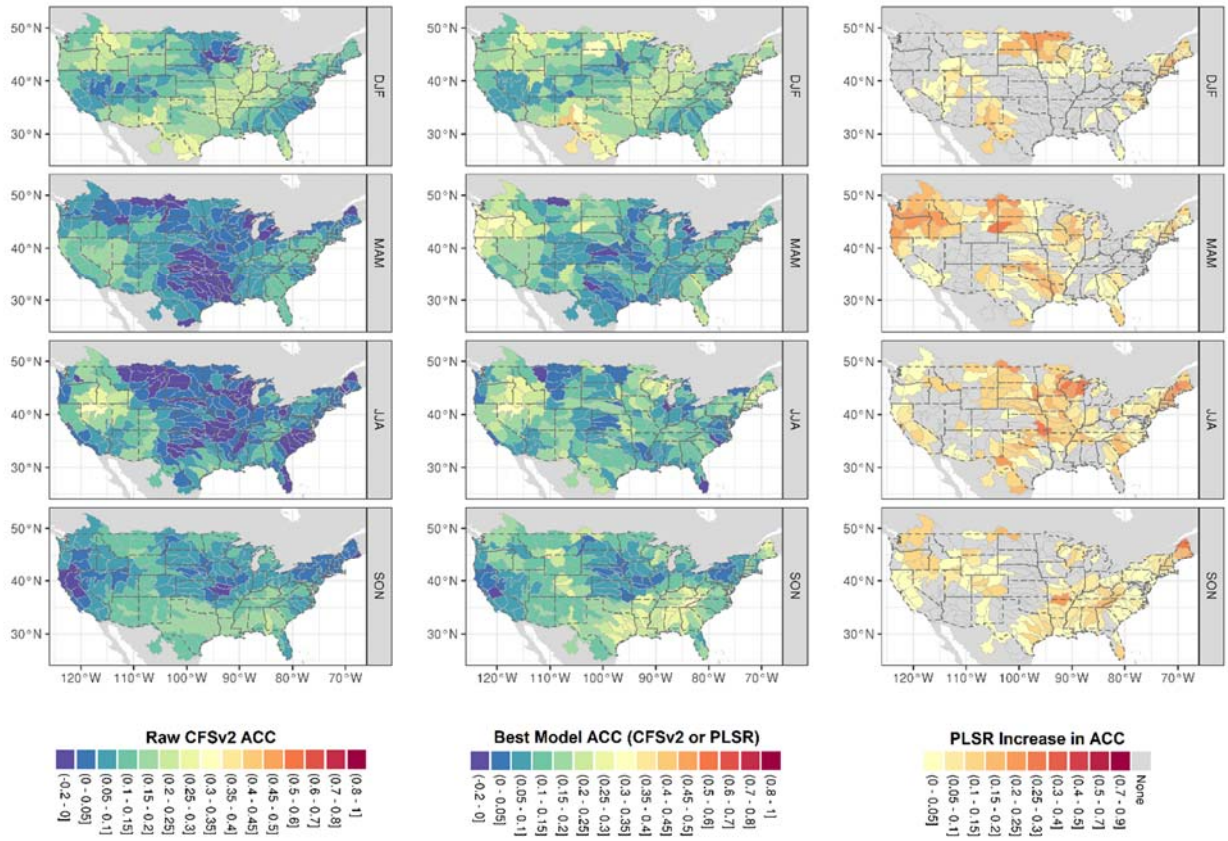

Figure 3-11. ACC results for 3-4 week precipitation forecasts on a seasonal basis. The format is the same as Figure 3-7.

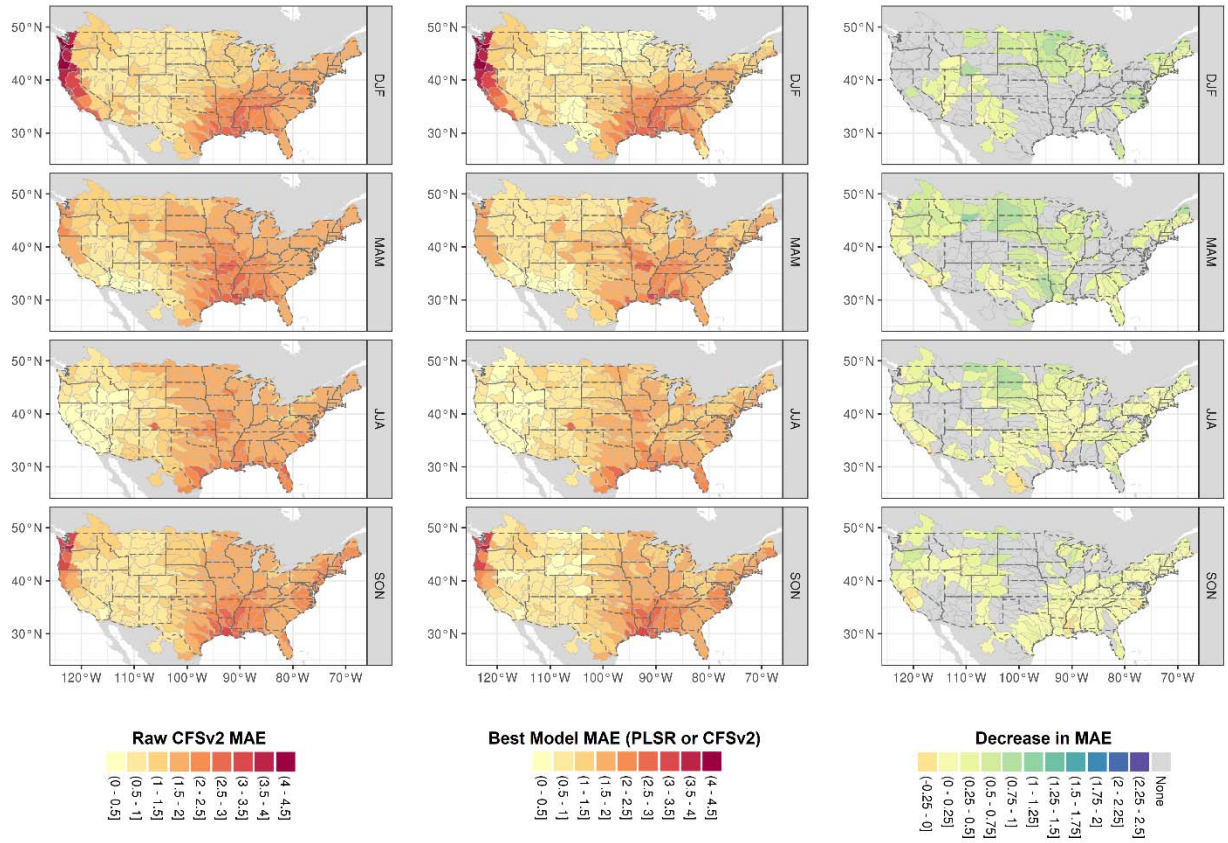

Figure 3-12. MAE results for 3-4 week precipitation forecasts on a seasonal basis. The format is the same as Figure 3-8.

3.6 Discussion and Conclusions

The post-processing of watershed scale sub-seasonal climate forecasts via PLSR demonstrates that there are opportunities to improve sub-seasonal forecast skill. The sub-seasonal forecast time period has received increasing attention in both the climate forecast and applications communities (U.S. Bureau of Reclamation, 2019). Both national project such as the NOAA S2S Task Force (Mariotti et al., 2018) and international efforts such as the S2S prediction project (Vitart et al., 2016; Vitart & Robertson, 2018) are working to improve forecast skill through enhancements of dynamical models and though techniques such as improved data assimilation, as well as through statistical post-processing of dynamical model output. Some of these

studies have used component based empirical regression methods to predict seasonal rainfall, but none have detailed an effort to enhance sub-seasonal bi-weekly climate forecasts from dynamical model forecasts by post-processing to watershed scales.

This study's objective was to assess the potential of post-processing in this sectoral context. Improvements to climate forecast skill would allow for more confidence and potential use by stakeholders. The water management sector has long demonstrated interest in climate forecasts, but in key applications of economic value such as operational streamflow prediction, the use of climate forecast information remains relatively limited, relying for instance on ENSO-based empirical conditioning of streamflow expectations. Our focus on the skill of current operational sub-seasonal climate forecasts on a watershed scale is intended to familiarize potential stakeholders with their raw performance as well as provided an indication of the potential for post-processing to enhance this performance. Baker et al. (2019) earlier presented a real-time demonstration of climate forecasts on watershed scales through an operational S2S Climate Outlooks for Watersheds web-based platform. Improvements to climate forecasts on such scales would help water managers improve decisions regarding reservoir operations, water allocation, flood control, hydropower generation, water treatment, and in-stream supported releases (Bolson et al., 2013).

Post-processing of watershed scale bi-weekly climate forecasts showed that even with a conservative incorporation of additional climate predictors, forecast

skill improvements are possible in many watersheds. The baseline predictors of forecasted CFSv2 SST and precipitation or temperature fields performed well in many, but not the majority, of watershed in the CONUS domain. We showed that using these predictors improved the ACC and MAE in some watersheds and seasons. PLSR contributed to large enough skill increases to produce usable forecasts in some cases where the raw forecasts fell below this threshold. It is important to note that post-processing did not performed well in many watersheds in this analysis. It is unclear whether different methods or input climate datasets other than those used in this analysis would improve performance, or whether precipitation and temperature variability in such locations is not systematically forced by identifiable or predictable circulation patterns.

Through this study of post-processing of climate forecasts at watershed scales, we offer a proof of concept rather than an exhaustive study. Further research could hone in on specific predictors for sub-seasonal climate forecasts, test different predictor domains for specific watersheds, use different lagged or pooled ensembles to reduce noise in raw forecast, or regionalizing predictors within the CONUS domain. The method we selected, PLSR, may be inferior to newer machine learning methods that can represent nonlinear and thresholded relationships between variables (Jones, 2017). After finding that post-processing did not capture extreme precipitation events well and often generated overly narrow ranges of forecasted values, we explored whether training PLSR models on only the extreme quantiles of precipitation events within a training sample would increase forecast skill. We

found slight improvements in forecasted value range and skill, which suggests that conditional training may a fruitful avenue of further study. Overall, however, we recommend the addition of post-processing techniques as part of climate services based on operational climate forecasts because, notwithstanding the limitations of this study, it provided evidence of benefit from the perspective of watershed scale sub-seasonal climate predictions.

4 CHAPTER IV: Enhancing ensemble seasonal streamflow forecasting in the Upper Colorado River Basin using multi-model climate forecasts

4.1 Abstract

Operational streamflow forecasts in the United States are predominately driven by the Ensemble Streamflow Prediction (ESP) method. ESP forecasts are produced by a hydrologic model initialized with current basin conditions and driven with historical temperature and precipitation traces to create probabilistic streamflow forecasts. In the Colorado River Basin (CRB), ESP forecasts drive operational planning models that project basin conditions out-multiple years. Any improvements to streamflow forecasts would help CRB stakeholders who depend of these operational projections for decision-making. With recent improvements in seasonal climate forecasts, we explore incorporating climate forecast information into the streamflow forecast through an ESP trace weighting scheme. The k-nearest neighbors (kNN) technique is employed to weight ESP traces based on North American Multi-model Ensemble (NMME) 1-month and 3-month temperature and precipitation forecasts, and preceding 3-month average observed streamflow. Two kNN weighting techniques are explored: (1) Basin-wide kNN uses the same ESP weights over the entire basin and (2) 4-Basin kNN separates the basin into four sub-basins, calculating ESP weighting in each, then recombining traces to calculate a new Lake Powell unregulated inflow forecast. Through analysis of the runoff season Lake Powell unregulated inflow, we find that kNN based forecasts have higher skill in the fall and winter and are more accurate at all leads compared to ESP. The 4-Basin kNN

method is more accurate than Basin-wide kNN through all leads, and more skillful at most leads.

4.2 Introduction

Many operational streamflow forecasts in the United States are provided by the National Weather Service River Forecasting Centers and National Resource Conservation Service (NRCS) (Pagano, Robertson, et al., 2014). 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 model initialized with current basin conditions and forced with historical temperature and precipitation traces (Day, 1985; Franz et al., 2003).

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 Mid-term Operations Model (MTOM) is one of these operational planning models that uses ESP to project monthly reservoir operations and basin conditions out 5 years (Daugherty, 2013). The results from MTOM can be used for decision making and risk assessment of potential shortage or surplus basin conditions that affect many communities and economies throughout the CRB (Bureau of Reclamation, 2015). Improving the skill of streamflow forecasts used in MTOM would benefit stakeholder who use projections of future basin conditions in decision making and planning.

Numerous techniques have been proposed to improve ESP. Methods include improvements to initial conditions or other inputs to the hydrology models that produce ESP; these methods are commonly referred to as pre-ESP techniques. The skill of ESP streamflow forecasts are at first highly dependent on the initial conditions, but at leads longer than one month, the skill is more dependent on climate forcings (Li et al., 2009; Shukla & Lettenmaier, 2011). Many studies have explored using climate forecasts to improve streamflow forecasts at sub-seasonal to seasonal (S2S) leads through pre-ESP methods. Wood and Lettenmaier (2006) showed improvement to ESP forecasts in the western US through the use of a land-surface model driven with climate forecasts ensembles from NASA's Seasonal-to-Interannual Prediction Project and other seasonal climate forecast. Mo and Lettenmaier (2014) completed a similar study over the contiguous US using the North American Multi-model Ensemble (NMME) climate forecasts. They compared runoff at a 1- to 3-month lead simulated from a hydrology model forced with NMME and historical temperature and precipitation traces. Their results showed skill improvements with NMME to be seasonally and regionally dependent, with NMME forecasts contributing to runoff skill after a 1-month lead, before which initial conditions dominated the forecast skill. NMME was found to improve runoff skill over ESP in the Upper CRB in April at a 2- and 3-month lead time. These studies illustrate the potential improvements to streamflow forecasts through the use of S2S climate forecasts such as NMME with pre-ESP methods.

Other studies have explored improving streamflow forecasts through post-processing of ESP. This is frequently referred to as post-ESP. Common post-ESP techniques include weighting ESP traces based on teleconnections or large scale climate information. Werner et al. (2004) showed that ESP weighting has potential in the CRB. Methods for weighting CBRFC ESP forecasts in three sub-basins in the CRB used climate indices (e.g., Nino-3.4) and CFSv2 reanalysis predictor components 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 ensemble members. Post-ESP methods were found to outperform pre-adjustment methods. Many other studies have compared post-ESP methods in watersheds in North America (Bazile, Boucher, Perreault, & Leconte, 2017; Beckers, Weerts, Tijdeman, & Welles, 2016; Gobena & Gan, 2010; Grantz et al., 2005; Pablo A. Mendoza et al., 2014; Najafi, Moradkhani, & Piechota, 2012; Andrew W. Wood & Schaake, 2008). 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 that were informed by climate and watershed initial condition information. Results showed that climate predictors added seasonal forecast skill, with the best results in watersheds with stronger teleconnections. These studies show that climate information has the potential to benefit seasonal streamflow forecasts through post-ESP methods.

Advancements in S2S climate forecast skill in recent decades can help improve streamflow forecast skill which could improve projections from operational planning

models such as MTOM (Raff et al., 2013). Better projections from MTOM would benefit stakeholders in the CRB who depend on predictions of shortage or surplus in the basin to drive operational decisions. We propose using a simple post-ESP k-nearest neighbors (kNN) trace weighting technique to improve streamflow forecasting in the Upper CRB using S2S temperature and precipitation forecasts from NMME.

This study is organized as follows. We first discuss MTOM ESP and the watershed scale S2S climate forecasts in the Background & Data section. The Methods section discusses the different predictors, kNN weighting methods, and verification metrics. The Results & Discussions section compares kNN weighting methods on the runoff season scale for different sub-basins, followed by a summary of the conclusions from this study.

4.3 Background & Data

4.3.1 MTOM & ESP

MTOM is a Reclamation CRB mid-term operational projection model built in RiverWare, a generalized river basin modeling software platform (Zagona et al., 2001). MTOM runs inflow traces through a decision making framework using rule logic that models system and reservoir operations. The model produces probabilistic monthly 5-year operational projections of 12 major reservoirs (9 Upper Basin and 3 Lower Basin) in the CRB. There are 12 Upper Basin forecast locations where ESP forecasts are input to MTOM. A schematic of the CRB with the important reservoirs and forecast locations is shown in Figure 4-1.

The ESP forecasts used in MTOM are unregulated inflow forecast. Unregulated flows are the streamflow that would have flowed through a location if there had not been dams located upstream of the forecast point. The CBRFC produced ESP forecasts with a Sacramento Soil Moisture Accounting (Sac-SMA) model calibrated with historical conditions and climate traces from 1981-2010. The ESP forecasts have 30 traces driven by 30 years of historical precipitation and temperature traces from the climatological record (1981-2010). The ESP forecast used in this work are a combination of reforecasts and operational forecasts. The reforecasts were provided by the CBRFC for 1981-2011 and operational forecasts were available from 2012-2016. For the 1981-2010 ESP reforecasts, 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 1981-2010 have 29 traces, while the forecasts form 2011-2016 have 30 traces.

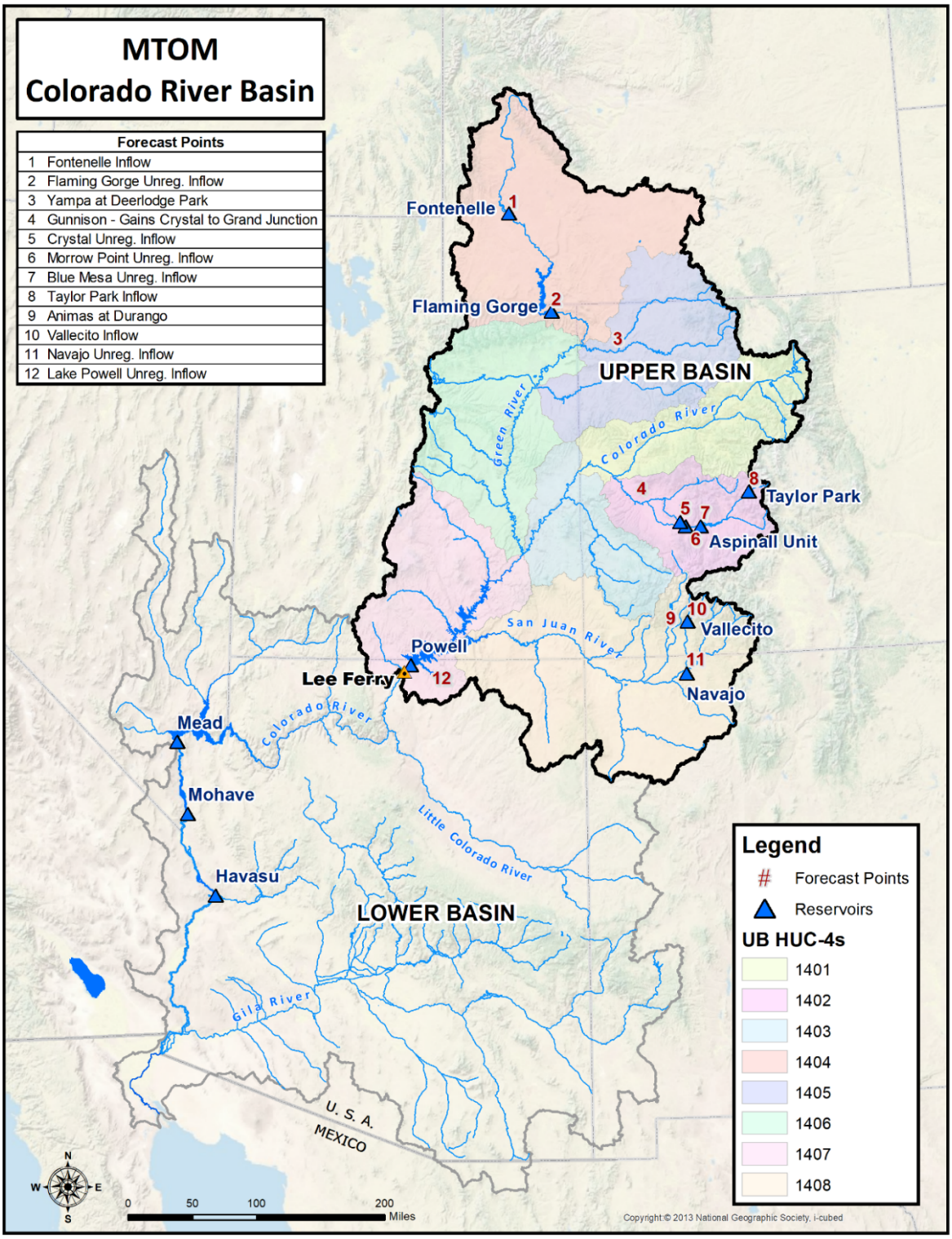

Figure 4-1: Schematic of the Colorado River Basin as setup in MTOM. The forecast locations are numbered from 1-12 and detailed in the top left table in the figure. The eight Upper Basin HUC-4 watersheds are symbolized by different colors of shading. The names of the HUC-4s can be found in Table 4-2. The reservoirs are represented by blue triangles with the Aspinall Unit, which includes Blue Mesa, Morrow Point, and Crystal Reservoirs.

4.3.2 S2S Climate Forecasts

NMME Phase 2 climate forecasts are used to inform the ESP trace weighting method for this study. NMME is a multi-model ensemble that combines seven global climate model forecasts at a monthly time step for leads up to 7 months (Kirtman et al., 2014). Reforecasts and operational forecast are available from 1982-2017. Baker et al. (2019) transitioned raw gridded NMME temperature and precipitation forecast to a United States Geological Survey (USGS) hydrologic unit code 4 (HUC-4) watershed scale. NMME forecasts were verified at monthly and seasonal leads to calculate the forecast skill compared to North American Land Data Assimilation System (NLDAS; Xia et al. 2012). Real-time watershed scale forecasts and benchmark assessments of hindcasts were made available online (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).

4.4 Methods

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

4.4.1 Feature Vectors

The feature vectors (also referred to as predictors) used in the kNN trace weighting scheme are the mean NMME temperature and precipitation watershed scale forecasts with associated NLDAS observations, and observed flow for different numbers of lagged months. The feature vectors were weighted based on prescribed weights (**W**) that were held constant for the entire simulation period.

Four feature vectors were derived from the mean NMME watershed scale forecasts: 1-month temperature forecast, 1-month precipitation forecast, 3-month temperature forecast, and 3-month precipitation forecast. The NLDAS observed temperature and precipitation for the same time periods were used to find the NMME forecasts' nearest neighbors. NMME watershed scale forecasts can provide information about future S2S climate conditions in the basin, allowing the forecast to be nudged one way or another through weighting.

Three feature vectors were tested using observed flow from the previous months: 3-month, 6-month, and 9-month average flows. The observed flow throughout the basin could help provide information about antecedent conditions in the basin, such as the amount of baseflow that relates to soil moisture and is important for projections in the fall prior to the runoff season. The feature vector weights are summarized as follows:

$$\boldsymbol{W} = \left(w_{T,1-mon}, w_{P,1-mon}, w_{T,3-mon}, w_{P,3-mon}, w_{Q,3-mon}, w_{Q,6-mon}, w_{Q,9-mon}\right), \quad \text{Eq. 4-1}$$

where $\sum \boldsymbol{W} = 1$

The skill of NMME forecasts differ depending on the variable and lead. To include information about the skill of NMME forecasts, we distributed the weights for each climate forecast lead time between temperature and precipitation based on their anomaly correlation. For example, say the anomaly correlation of the 1-month lead temperature and precipitation forecasts were 0.4 and 0.2, respectively. The 1-month temperature forecast would receive 2/3 of the total prescribed weight for the combined 1-month NMME forecast, and the 1-month precipitation forecast would

receive 1/3 of the total weight. Therefore, we attempt to capture the differences in skill of the NMME forecast in a useful manner.

We tested other antecedent conditions such as observed temperature and precipitation but found they did not add skill to the forecasts. These observed values should be accounted for through the initial conditions in the hydrologic model producing ESP.

4.4.2 kNN Trace Weighting Scheme

We have adapted the post-ESP trace weighting scheme from Werner et al. (2004) for ESP forecasts in the Upper CRB. The kNN method is executed for each month with feature vectors of NMME forecasts and observed flow to weight the monthly ESP forecasts for January 1982 – September 2016. The technique is described as follows:

1. The feature vectors are organized and standardizes for the given start month. Since ESP traces are informed by historical years (j) of 1981-2010, the feature vectors only need to include these years. The forecasted year is removed from these vectors, excluding any ESP traces with perfect knowledge of the future climate.
2. A distance vector, **D**, is computed that calculates the weighted Euclidean distance from the n feature vectors in the training period (x_i) to the vector for the forecast date (x_t). The feature vector weights, **W**, calculated from the weighted Euclidean distance are prescribed for the entire extent of that forecast date's ESP forecast.

$$\boldsymbol{D} = \left(d_1, d_2, \dots, d_j\right) \quad , \quad \text{where} \qquad \text{Eq. 4-2}$$

$$d_i = \sqrt{w_1\left(x_t - x_{1,i}\right)^2 + w_2\left(x_t - x_{2,i}\right)^2 + \cdots + w_n\left(x_t - x_{n,i}\right)^2} \quad \text{Eq. 4-3}$$

3. Sort the distance vector from lowest to highest values.

$$\widehat{\boldsymbol{D}} = \left(d_{\hat{1}}, d_{\hat{2}}, \ldots, d_{\hat{j}}\right), \quad d_{\hat{1}} \le d_{\hat{2}} \le \cdots \le d_{\hat{j}} \quad \text{Eq. 4-4}$$

4. The weights for each ensemble member are calculated using the following equations:

$$w_i = \left[1 - \frac{d_{\hat{i}}}{d_{\hat{k}}}\right]^{\lambda}, \ \ where \ \ d_{\hat{i}} \le d_{\hat{k}} \quad \text{Eq. 4-5}$$

$$w_i = 0, \ \ d_{\hat{i}} > d_{\hat{k}} \quad \text{Eq. 4-6}$$

$$k = NINT\left(\frac{n}{\alpha}\right) \quad \text{Eq. 4-7}$$

where λ is the distance sensitive weighting parameter and α defines the k nearest neighbor traces used from ESP. The NINT is the nearest neighbor operator. In this study, we set λ = 2.5 and α = 1 based on experiments not shown. This means that all ESP traces are included in the kNN forecast.

5. The weights are then normalized so that the sum of the weights is equal to 1.
6. ESP traces are resampled based on the normalized ensemble weights to obtain a 100-member ensemble.

4.4.3 Spatial Evaluation Scenarios

The kNN trace weighting method is applied for two different spatial scales described in the following sections: (1) the Basin-wide method and (2) the 4-Basin method.

4.4.3.1 Case 1. Basin-wide method

The basin-wide method weights ESP forecasts members based on NMME forecasted temperature and precipitation aggregated over the entire Upper CRB. The Upper

Basin aggregated NMME forecasts were calculated from a flow weighted average of individual HUC-4 NMME forecasts. The weights from each HUC-4 were based on the contributing inflows to Lake Powell on an annual scale from each HUC-4 basin. Individual HUC-4 annual flows were calculated based on USGS stream gages, where available. The ESP trace weights based on the basin-wide method were applied uniformly to all forecast locations.

4.4.3.2 Case 2. 4-Basin method

The 4-Basin method splits the Upper Basin into 4 different sub-basins based on the major tributaries in the Upper CRB and MTOM forecast locations. The kNN weighting scheme is performed at each of the 4 basins separately, producing different ESP trace weights for each sub-basin. This results in new ESP ensembles that may span different ranges of forecasted inflows to Lake Powell since different ESP traces in the 4 separate basins may be combined to form new Lake Powell inflow.

The four sub-basins are the (1) Main Stem, (2) Green, (3) Gunnison, and (4) San Juan. The eight HUC-4 are categorized into the 4-basins as shown in

Table 4-1. The decision of where to categorize each HUC-4 is based on the forecast locations used in MTOM. 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 MTOM forecast location on the Green River is above the Lower Green HUC-4.

Table 4-1: Table of HUC4 assignment in 4-Basin method.

HUC-4 ID	Label	4-Basin Assignment
1401	Colorado Headwaters	Main Stem
1402	Gunnison	Gunnison
1403	Upper Colorado-Delores	Main Stem
1404	Great Divide-Upper Green	Green
1405	White-Yampa	Green
1406	Lower Green	Main Stem
1407	Upper Colorado-Dirty Devil	Main Stem
1408	San Juan	San Juan

The NMME temperature and precipitation forecasts for the 4-basins are weighted based on contributing flow, if the sub-basin has multiple HUC-4 watersheds. The twelve MTOM forecast locations are also split into the 4 sub-basins as shown in Table 4-2. The observed preceding 3-month, 6-month and 9-month averaged flows used in the feature vector are the total intervening flow for each of the 4-basins.

Table 4-2: Table of forecast location assignment in 4-Basin method.

Forecast Locations		4-Basin Assignment
1	Fontenelle Inflow – Q_{Font}	Green
2	Flaming Gorge Unregulated Inflow – Q_{FG}	Green
3	Yampa at Deerlodge Park – Q_{Yampa}	Green
4	Gunnison - Gains Crystal to Grand Junction – Q_{CryGJ}	Gunnison
5	Crystal Unregulated Inflow – Q_{Cry}	Gunnison
6	Morrow Point Unregulated Inflow – Q_{MP}	Gunnison
7	Blue Mesa Unregulated Inflow – Q_{BM}	Gunnison
8	Taylor Park Inflow – Q_{TP}	Gunnison
9	Animas at Durango – Q_{Animas}	San Juan
10	Vallecito Inflow – Q_{Val}	San Juan
11	Navajo Unregulated Inflow – Q_{Nav}	San Juan
12	Lake Powell Unregulated Inflow– Q_{Powell}	Main Stem

This requires calculating new total unregulated flow volumes for each sub-basin. 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 slightly different since the Green, Gunnison, and San Juan River's flows must be subtracted from the Lake Powell unregulated inflow forecast location. The flows for each sub-basin are described in the following equations:

$$Q_{Green} = Q_{FG} + Q_{Yampa} \quad \text{Eq. 4-8}$$

$$Q_{Gunnison} = Q_{Cry} + Q_{CryGJ} \quad \text{Eq. 4-9}$$

$$Q_{SanJuan} = Q_{Nav} + Q_{Animas} \quad \text{Eq. 4-10}$$

$$Q_{MainStem} = Q_{Powell} - (Q_{Green} + Q_{Gunnison} + Q_{SanJuan}) \quad \text{Eq. 4-11}$$

with variables defined in Table 4-2.

Once kNN is performed on each of the four sub-basins separately, the forecasts are recombined to calculate a new MTOM ensemble for the 12 forecast locations. For all forecast locations except the Lake Powell unregulated inflow location, the resampled 100-member ensembles are combined. For the Lake Powell unregulated inflow forecast location, a new forecast is calculated based on the 100-member ensembles from all four sub-basins as follows:

$$Q_{Powell} = Q_{MainStem} + Q_{Green} + Q_{Gunnison} + Q_{SanJuan} \quad \text{Eq. 4-11}$$

The 4-basin method results in a new ensemble with a different spread and distribution than the original ESP forecast and the Basin-wide method.

4.4.4 Verification Metrics

Streamflow forecasts from ESP and kNN methods are compared to observed flows at the inflow to Lake Powell and the four sub-basins used in the 4-basin method. Verification metrics are calculated for the runoff season volume, April-July, for leads up to 12-months prior to the last forecast in July. The forecasts in May, June, and July have observed flows for the months already observed in the runoff season. These forecast leads should have higher skill and lower errors compared to other forecasts since there are fewer months in the runoff season to forecast.

The continuous ranked probability skill score (CRPSS) is used to calculate the skill probabilistic forecasts. CRPSS is a measure of the accuracy of a forecast relative to that of a reference forecast. The reference forecast used here is climatology from

1981-2010. CRPSS ranges from 1 to -∞, with a perfect skill score equal to 1. A skill score of 0 means the skill of the forecast is equal to that of climatology, and a negative skill score means the forecast is less skillful than climatology.

Error relative to observations is measured as the root mean squared error (RMSE) which is the square root of the average of the squared differences between projections and observations. Since the errors are squared, larger errors have a greater influence on RMSE than smaller errors.

4.5 Results

The first step taken in the kNN analysis was to determine the feature vector weights. We performed a heuristic optimization in which we manually varied the weighting scheme to find the best combination of weights. We found that all NMME forecasts, which includes the 1-month and 3-month temperature and precipitation forecasts, and the 3-month lagged flow contributed positively to predicting seasonal runoff. The best weights found through our experiments are listed in Table 4-3.

Table 4-3: Weights of feature vectors.

$w_{T,1\text{-mon}}$ & $w_{P,1\text{-mon}}$ *	$w_{T,3\text{-mon}}$ & $w_{P,3\text{-mon}}$*	$w_{Q,3\text{-mon}}$	$w_{Q,6\text{-mon}}$	$w_{Q,9\text{-mon}}$
0.275	0.275	0.45	-	-

*Weights for individual precipitation and temperature forecasts for each NMME lead are split based on forecast skill. See the Methods section 'Feature Vectors' for details.

The ESP, Basin-wide kNN, and 4-Basin kNN streamflow forecasts were analyzed for their skill and accuracy when forecasting runoff season inflow to Lake Powell. The CRPSS at leads of 12-months to 1-month for the three forecasts are shown in left panel of Figure 4-2. Since we are computing a seasonal flow, the forecast includes observed flows once the lead is less than 4-months since the April flow would

have been observed. 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. At these leads, the 4-Basin kNN forecast has a wider range of skills than the other two forecasts with some forecasts exhibiting high CRPSS, but overall the median CRPSS is lower than the other forecasting methods at earlier leads. This is likely due to the narrowing of the ensemble.

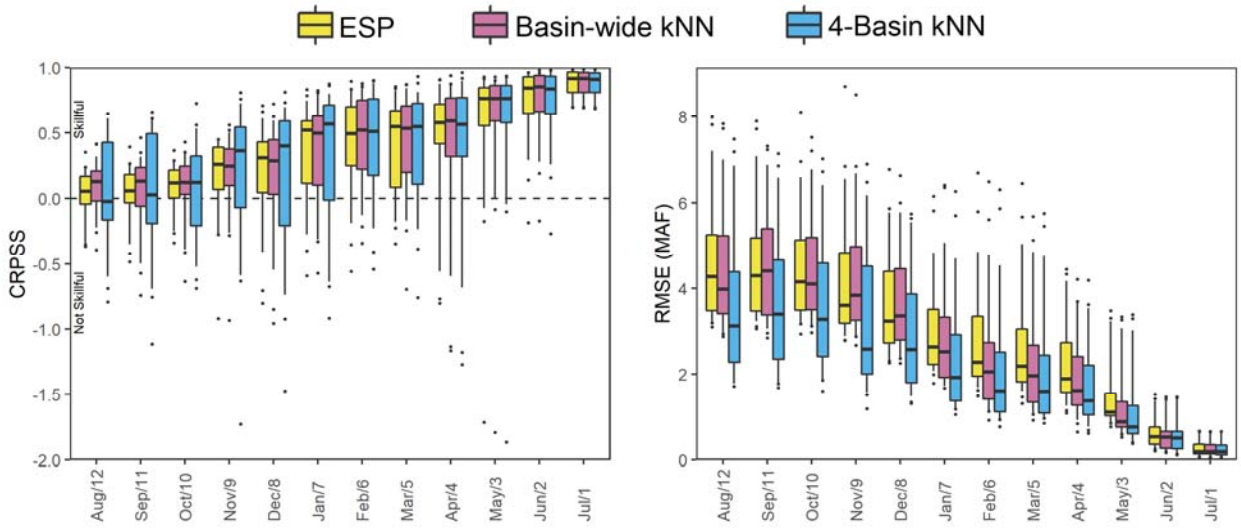

Figure 4-2: CRPSS and RMSE for runoff season streamflow forecasts of Lake Powell unregulated inflow. The streamflow forecasts ESP, Basin-wide kNN, and 4-Basin kNN are compared at leads of 12- to 1-month.

As leads decrease, the skill of all the forecasts increase. The forecast with the highest median skill changes with lead. In November through January, 4-Basin kNN has a higher median skill with a larger range of skills compared to the other two forecasting methods. By March, the median skill of all forecasts are relatively the same, with only small differences in the range of CRPSS as shown by the boxes of each boxplot.

The RMSE for ESP, Basin-wide kNN, and 4-Basin kNN is shown in the right panel of Figure 4-2. The RMSE is large at longer leads and decreases with lead. The 4-Basin kNN forecast has a lower RMSE compared to the other forecasts in all instances. The median RMSE for 4-Basin kNN is about 1 MAF (million acre-ft) lower than the other forecasts at longer leads and decrease as all forecasts start to perform better. The Basin-wide kNN forecast has lower errors than ESP at most leads, especially at shorter leads when the NMME forecasts have greater impact on the analyzed flow.

The performance of ESP and the 4-Basin kNN method for each of the four sub-basins are compared in Figure 4-3. For CRPSS in the top row of Figure 4-3, the skill of 4-basin kNN performs better than ESP in the Main Stem and Green with more lead dependent results in the Gunnison and San Juan sub-basins. Overall, the 4-bains kNN method reduces error when compared to ESP for most leads in most reaches, though not by a significant amount. These results aligns well with the CRPSS results in Figure 4-2.

Each sub-basin contributes different proportions of the seasonal inflow volume to Lake Powell. This is apparent in the RMSE plots in the bottom row of Figure 4-3, which have varying magnitudes of error. The highest error is in the Main Stem and Green since most of the total inflow to Lake Powell originates in these sub-basins. The reduction in RMSE from ESP to 4-Basin kNN is not as apparent in the individual reaches compared to the inflow to Lake Powell (Figure 4-2). This is due to the different forecast magnitudes. Overall, most leads in each reach exhibit a decrease in

RMSE for the 4-Basin kNN method compared to ESP, though this is less obvious at earlier leads.

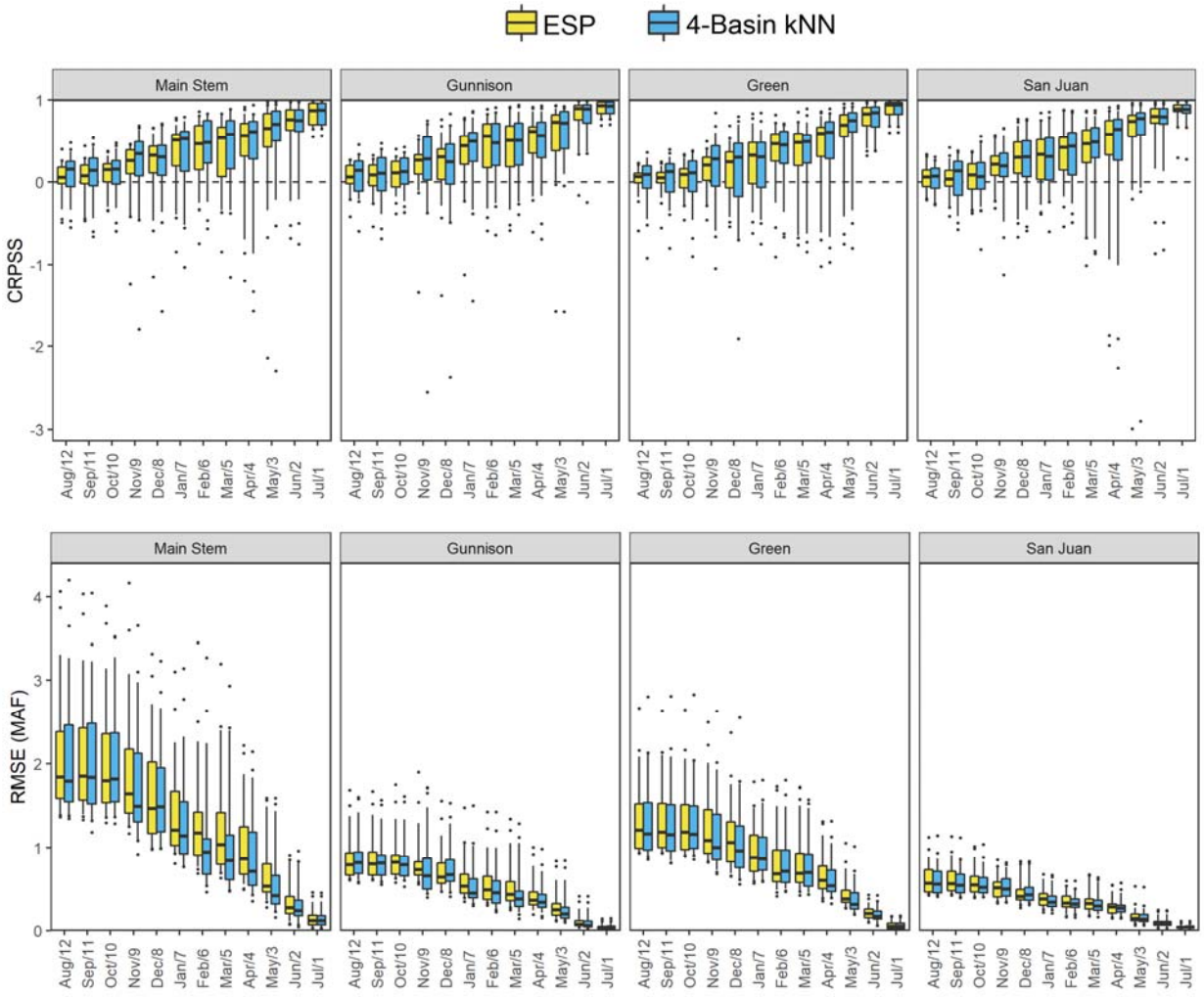

Figure 4-3: CRPSS and RMSE from the 4-Basin kNN and ESP methods for the four sub-basins. The top row is CRPSS and the bottom row is RMSE at a lead of 12-months to 1-month for the runoff season forecast.

Another way to compare forecasts is by looking at the streamflow forecasts for each year compared to observations. Figure 4-4 illustrates the ESP and 4-Basin kNN forecasts at a 7-month lead in January ranked from lowest to highest observed flow for 1982-2016. The 4-Basin kNN forecasts have a narrower range of flows compared to ESP. The higher years of observed flow are captured well by both forecasts since

these forecasts are likely driven by early snow accumulation which is represented by initial conditions in Sac-SMA. The 4-Basin kNN forecasts tend to have median flows that are slightly higher than ESP forecasts, illustrating that NMME forecasts are likely nudging the forecast towards wetter conditions based on the 1-month and/or 3-month forecasts. This topic is explored further in Appendix 8.2.

For lower observed flows, both the ESP and the 4-Basin kNN forecasts do not capture the observed flow well. There is a slight bump down in the forecast median of the 4-Basin kNN forecast compared to ESP in many of the lower years, though not all. In some years, both forecasts perform very poorly (e.g. Rank 30). Since the 4-Basin kNN forecast is based on the ESP forecast, it cannot correct for large errors in the ESP forecast. Thus, if ESP performs very poorly, the 4-Basin kNN forecast will also perform poorly.

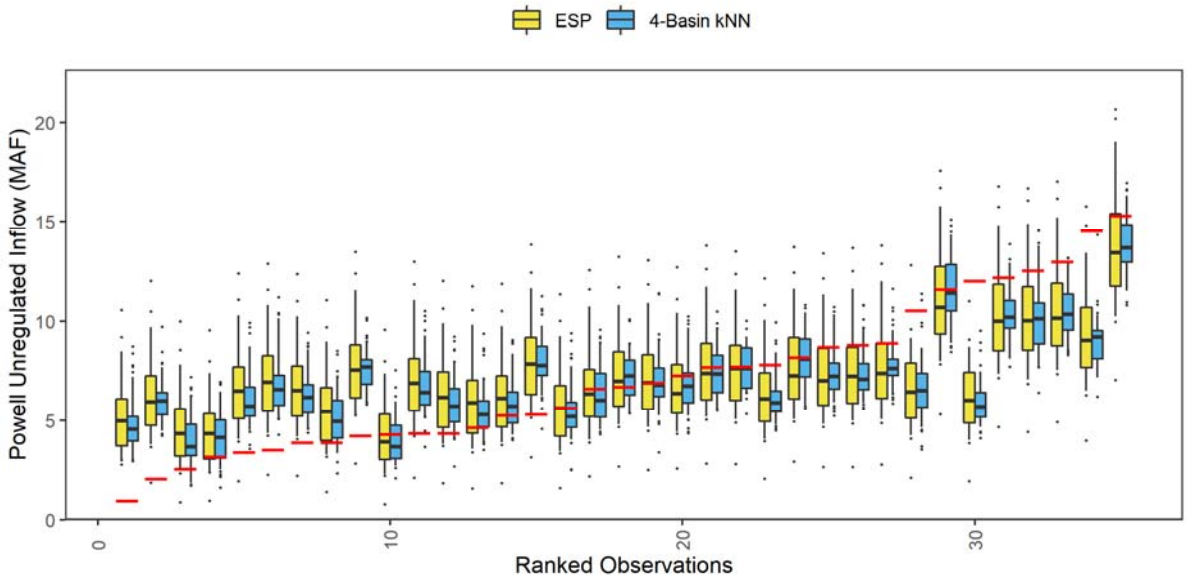

Figure 4-4: Runoff season ensemble forecasts for 1982-2016 compared to observations arranged by ranked observations. ESP and 4-bains kNN forecasts of Lake Powell April-July unregulated inflow are compared for each year at a 7-month lead in January. Observed inflow is represented by a red horizontal line.

4.6 Discussion & Conclusion

The post-ESP method explored in this work illustrates the opportunity for climate forecasts use to inform streamflow forecasts through a post-processing framework in the Upper CRB. NMME 1-month and 3-month temperature and precipitation forecasts, along with 3-month preceding averaged flow, were shown to be useful predictors for weighting ESP traces. We found that a skillful NMME forecast translated into improved streamflow forecasts, showing that this method was able to exploit the skill available from NMME (Appendix 8.2).

Two kNN weighting methods were explored, Basin-wide kNN and 4-Basin kNN. Basin-wide kNN weighted ESP traces through an aggregated method over the entire Upper CRB, while 4-Basin performed the weighting method over four sub-basins. Both kNN methods showed smaller errors compared to ESP for all leads. The 4-Basin kNN has lower errors than the other forecasts at all leads through May at a 3-month lead. The results from the analysis of skill differ, showing better performance by 4-Basin kNN in the fall through winter, but not at all leads. The CRPSS results are worse than RMSE likely because the spread of the 4-Basin kNN forecasts can be too narrow compared to the climatology forecast.

This work was not meant to be an exhaustive study of all post-ESP methods, instead we intended to explore potential uses of climate forecasts through post-ESP methods. Further research could explore other possible weighting schemes, distance calculations, predictors, or post-processing calibration techniques such as those discussed in Wood et al. (2008). Overall, this work recommends the use of climate

forecasts to inform streamflow forecasting at certain leads, such as the fall and winter when ESP forecasts are less skillful and NMME forecasts can project wetter or drier winters. As S2S climate forecasts become more skillful, kNN post-ESP weighting techniques would likely show improvements to streamflow forecasts.

The proposed methodology could be applied to other basins throughout the US that depend on ESP streamflow forecasts. The process is general enough to be applied to other watersheds since the S2S watershed scale climate forecasts are available. This method could be very useful in watersheds where NMME climate forecast skill is high, such as the southern California in winter and spring or the southeastern US during fall and winter. Watersheds that are not snowmelt dominated could have improved streamflow forecasts using this method, though the feature vectors and weights would need to be altered.

The streamflow forecasts analyzed in this work are on the scale used in Reclamation's mid-term operations and planning model, MTOM. The improved streamflow forecasts with the kNN weighting schemes could be applied to MTOM to assess how improved streamflow forecasts translate into operational projections. Improvements to streamflow forecasts, even if relatively small, could improve projections of future basin conditions in the CRB that would benefit stakeholders who depend of operational projections of future basin conditions for their decision-making. This topic will be explored in the next chapter for streamflow forecasts from ESP and 4-Basin kNN.

5 CHAPTER V: A Testbed for Assessing streamflow forecasts and operational projections in the Colorado River Basin

5.1 Abstract

Water managers depend on streamflow forecasts to make many operational decisions, including decisions regarding reservoir operations, water allocation, flood control, and in-stream supported releases. In the Colorado River Basin, management decisions would benefit from streamflow forecasts with improved skill that extends beyond short-term (1-year) leads to provide assessments of future basin risk, such as the probability of water shortage or flood control. The Colorado Basin Streamflow Forecast Testbed creates a framework for assessing the performance of streamflow forecasts and operational projections in the Colorado River Basin using the Bureau of Reclamation's Mid-Term Operations Model (MTOM) for a 2-year outlook. The current operational streamflow forecasting technique used in MTOM, Ensemble Streamflow Prediction (ESP), is compared to experimental forecasting method, 4-Basin k-nearest neighbor (kNN), which weights ESP traces using climate forecast information and observed streamflow. The streamflow projections are evaluated on an annual water year scale for the unregulated inflow to Lake Powell. MTOM operational projections from these forecasts are compared at Lake Powell and Lake Mead for the performance of projecting pool elevation, operating tier, and releases. The 4-Basin kNN method outperformed ESP for most leads in the winter and spring of the forecasted water year when comparing streamflow forecast skill and accuracy, see Section 4.5. The 4-Basin kNN method produced more accurate projections of pool

elevation and categorical scores for operating tiers and releases compared to ESP, though the differences in projected operating tiers were smaller.

5.2 Introduction

Streamflow forecasts provide valuable information regarding the quantity and timing of streamflow through a river system. Many water management decisions are made using streamflow forecasts including reservoir operations, water allocation, flood control, and in-stream releases. For large water resource agencies such as the Bureau of Reclamation (Reclamation), most operational streamflow forecasts are produced by the National Weather Service River Forecasting Centers (RFCs) and National Resource Conservation Service (NRCS) (Pagano et al., 2014). Streamflow forecasts produced by the RFCs rely on land-surface models initialized with current hydrologic conditions and forced with historical climate information (Raff et al., 2013). This method, Ensemble Streamflow Prediction (ESP), is widely used throughout the water management community (Day, 1985; Franz et al., 2003). The skill of short-term ESP forecasts are highly dependent on the initial conditions, but at leads longer than one month, the skill is more dependent on climate forcings (Li et al., 2009; Shukla & Lettenmaier, 2011). This leaves room for improvements to forecasts through the use of climate forcing. Statistical methods are also used to produce seasonal water supply forecasts (Garen, 1992; Pagano et al., 2014). These forecasts traditionally use principal component regression models trained on historical data such as precipitation and snow water equivalent (SWE). Both ESP and statistical water supply forecasting methods provide skill when initial conditions,

such as observed SWE or soil moisture, drive forecasted streamflow, but lack skill when climate forcings drive forecast skill (Andrew W. Wood et al., 2016).

In the Colorado River Basin (CRB), streamflow forecasts produced by the Colorado Basin RFC (CBRFC) are used by Reclamation to drive operations and planning models that are inform decision making and risk assessment (Bureau of Reclamation, 2015). Reclamation's Mid-term Operations Model (MTOM) is one of these operational planning models, which projects 5 years of monthly mid-term operations. ESP forecasts produced by the CBRFC are input to MTOM which uses operating rules to drive reservoir operations in the model. A skillful streamflow forecast, which extends beyond the current year, would be valuable to CRB stakeholders who rely on projections of reservoir operations to provide an outlook of potential shortage or surplus basin conditions.

Many studies have explored improvements to streamflow forecasting methods, but few have analyzed how these improvements translate into enhanced water resources decision making. Regonda et al. (2011) evaluated how increased streamflow forecast skill translates to improvements in operations and decision variables in the Gunnison River Basin. A nonlinear regression with different predictor combinations was used to create a multi-model ensemble streamflow forecast informed by large-scale climate information. Streamflow forecasts were then run through an operations model that projected outflow, storage, and power production at Blue Mesa Reservoir. The study found that streamflow forecast skill transferred to operational variable skill at long lead times, though nonlinearly.

Another study by Sankarasubramanian et al. (2009) investigated streamflow forecasts produced by principle component regression and informed by monthly updated precipitation forecasts and their performance in a reservoir simulation model in the Philippines. They found that using streamflow forecasts reduced spill, increased allocation for hydropower during above-normal years, and helped meet end of season storage targets for below-normal years. These studies show that streamflow forecasts do not translate linearly into improved reservoir operations and should be investigated based on the specific needs of the studied basin.

This study seeks to create a testbed to provide an organized, objective approach to compare current and experiment streamflow forecasting methods in the CRB. The testbed will establish a protocol for evaluating hydrologic forecast skill, as well as how the performance of streamflow forecasts translates into improved operational projections. We use the Colorado Basin Streamflow Forecast Testbed (now referred to as 'testbed') to create a framework for assessing the performance of streamflow forecasts and operational projections in the CRB using Reclamation's Mid-Term Operations Model (MTOM).

This paper is organized as follows. Section 5.2 will provide an overview of the models used for mid-term operations and reservoir operating policies in the CRB. Section 5.3 will describe the RiverWare model, streamflow forecasting methods assessed in the testbed, and the protocol for analyzing the forecasts through a set of defined metrics. Section 5.4 will present the results of different streamflow

forecasting methods through the defined metrics, and Section 5.5 will discuss results and conclusions.

5.3 Background

5.3.1 Operational MTOM

MTOM is the primary model used for evaluating mid-term probabilistic operations in the CRB. MTOM is built in RiverWare, a generalized river basin modeling software platform (Zagona et al., 2001). The model produces monthly, 5-year probabilistic operational projections for twelve major reservoirs in the CRB. The CRB is split into two basins, the Upper Basin and Lower Basin at Lee Ferry, a point located below Lake Powell. Nine of the twelve MTOM reservoirs are located in the Upper Basin and three are in the Lower Basin.

Streamflow forecasts are ingested by MTOM at twelve forecast locations in the Upper Basin. The streamflow forecasts for the forecast locations are unregulated, meaning the forecasted flows are the streamflows that would have occurred provided there was no regulation due to dams upstream of the forecast point. These forecasted unregulated flows include Upper Basin demands, which are projected by the CBRFC, except for three tunnel diversions. Figure 5-1 shows a map of the CRB that includes forecast locations, major reservoirs, and Upper Basin diversions.

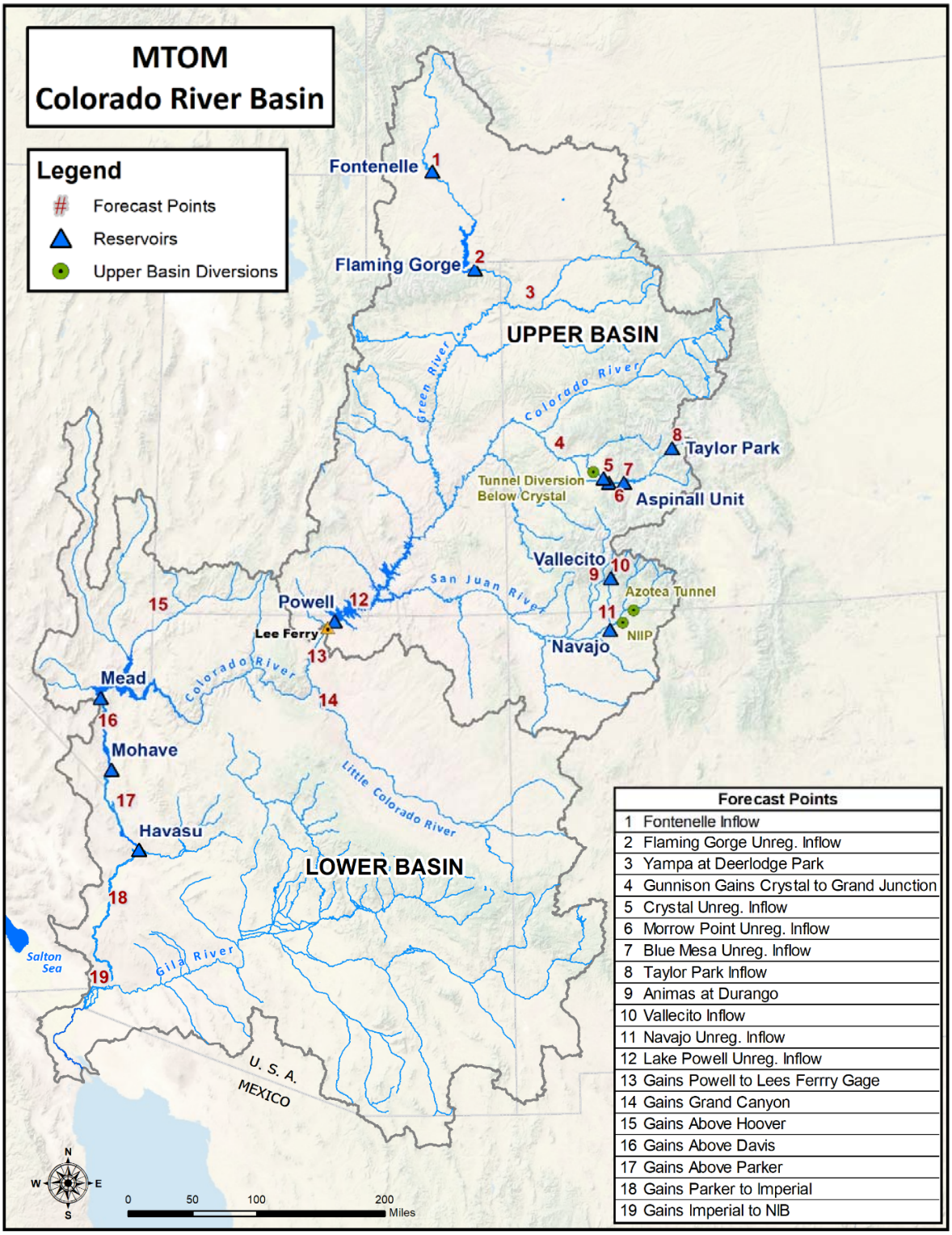

Figure 5-1. Map of the Colorado River Basin with important locations defined in MTOM. The map includes MTOM reservoirs, forecast points, and Upper Basin diversions. A table in the bottom right describes the names of each numbered forecast location; forecast points 1-12 are in the Upper Basin and 13-19 are in the Lower Basin. The Aspinall Unit is a series of three reservoirs: Blue Mesa, Morrow Point, and Crystal.

Lower Basin depletions are input to MTOM from official schedules of future water use. The Lower Basin has 7 intervening flow locations (13-19 in Figure 5-1). Intervening flows are determined by computing unaccounted for flows in each river reach based on a water balance of historical stream gages, water use, and reservoir operations in the Lower Basin. Lower Basin MTOM forecast points use the historical intervening flows since there are presently no skillful forecasts for these locations at the MTOM timescales.

5.3.2 Reservoir Operations in the CRB

The decision making framework that determines reservoir releases in MTOM is based on rule-based "if-the" logic scripted using the RiverWare software. The reservoir operations are in accordance with the "Law of the River", which includes the 2007 Colorado River Interim Guidelines for Lower Basin Shortages and Coordinated Operations of Lake Powell and Lake Mead (2007 Interim Guidelines). The 2007 Interim Guidelines specify coordinated annual operations between Lakes Powell and Mead to avoid curtailment of water use in the Upper Basin and to minimize shortages in the Lower Basin (U.S. Department of Interior, 2007). This is done through prescribed operating tiers as seen in Figure 5-2. Lake Powell has four operating tiers that are determined based on projected end of year reservoir elevations. Lake Mead has three main operating tiers, shortage, surplus, and normal conditions, which are also based on projected end of year reservoir elevations.

The projected reservoir elevations for determining the operating tier are produced by the 24-Month Study (24MS), a deterministic Reclamation model that

simulates reservoir operations for a 2-year period. A single deterministic streamflow forecast, the Most Probable forecast, is used in the 24MS. For the current year, the Most Probable forecast is a combination of the 50% exceedance ESP trace and forecaster judgement. The out-year of the forecast is climatology, the average flows from 1981-2010.

Instead of solving reservoir operations using rule logic as is done in MTOM, reservoir operators manually input reservoir outflows and operations in the 24MS. The 24MS is used for official CRB operational projections and to support decisions for the CRB Annual Operating Plan (AOP). The AOP provides a plan for CRB reservoir operations for the upcoming year using reservoir end of calendar year (EOCY) pool elevations projected by the August 24MS. Once reservoir operations for Lakes Powell and Mead are set by the AOP for the next year, they can only change if there is an April Adjustment or if flood control is implemented. An April Adjustment for Lake Powell can occur if the April 24MS projections of end of water year (EOWY) Lake Powell pool elevations are projected to be at or higher than 3,575 feet and Lake Mead pool elevation is less than 1,075 feet. An April Adjustment can also occur if Lake Powell is in the Upper Elevation Balancing Tier and the April 24MS projected Lake Powell EOWY pool elevation above the Equalization line as defined in the 2007 Interim Guidelines. The April Adjustment allows for operational changes when there are large increases in Most Probable streamflow forecast throughout the winter and spring. For more information regarding the coordinated operation of Lake Powell and Lake Mead, see the 2007 Interim Guidelines.

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 5-2. 2007 Interim Guidelines operating tiers. Schematic of the 2007 Interim Guidelines for the operating tiers of Lake Powell and Lake Mead with reservoir elevations, storage, and description of releases. The elevation between the Equalization and Upper Elevation Balancing Tiers in Lake Powell increases each year between 2007 and 2026.

5.4 Data & Methods

5.4.1 Testbed Framework

The framework for the Colorado Basin Streamflow Forecast Testbed is summarized in Figure 5-3. The testbed ingests the available streamflow forecasts that are run through MTOM to output operational projections for the CRB reservoirs. MTOM simulates on a monthly time step for 2-3 years, to the end of the second water year (WY; October - September). The full 5-year MTOM simulations were not analyzed in this project since the focus here was to analyze streamflow forecasts and possible

forecast improvements that are more feasible for the 1 to 2-year range. MTOM simulates operational projections such as reservoir releases and operating tiers. The testbed processes streamflow forecasts and operational projections separately according to a protocol of designated performance metrics. The hydrology metrics and operational projection metrics are discussed in detail in the following section.

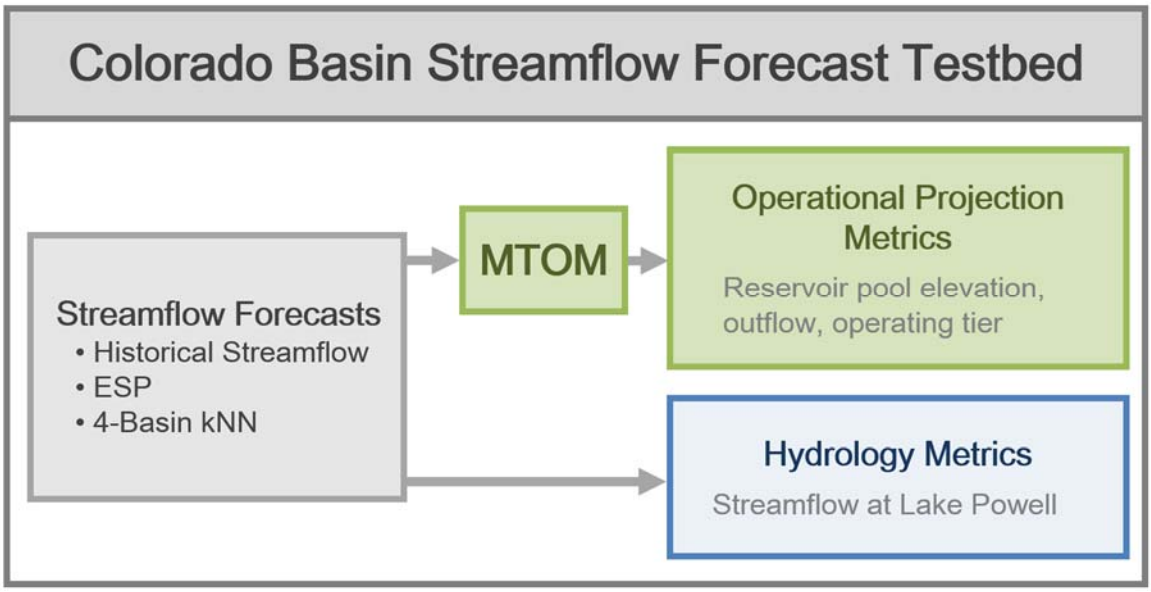

Figure 5-3. Colorado Basin Streamflow Forecast Testbed framework. Streamflow forecasts are analyzed and run through MTOM to output operational projections. The streamflow forecasts and operational projections are analyzed separately.

Part of the testbed framework depicted in Figure 5-3 is processed using the RiverWare Study Manager and Research Tool (RiverSMART). RiverSMART facilitates in the execution of RiverWare studies and can simulate several hydrology scenarios, demand scenarios, and operating policies. The testbed uses the capabilities of RiverSMART to simulate several streamflow forecast ensembles with varying numbers of traces to produce operational projections for the major reservoirs in MTOM.

The setup of the testbed in RiverSMART is illustrated in Figure 5-4. A combination of Run Range, DMI (Data Management Interface), and MRM (Multiple Run Management) events allow RiverSMART to simulate forecasts with different run lengths, number of traces, and input format. The scenarios use one model, MTOM, and one ruleset to simulate reservoir operation according to the 2007 Interim Guidelines. The basin-wide conditions and reservoir operations from each simulation are output to CSV files that are read into R scripts to analyze the streamflow forecasts and operational projections for hydrologic and operational skill.

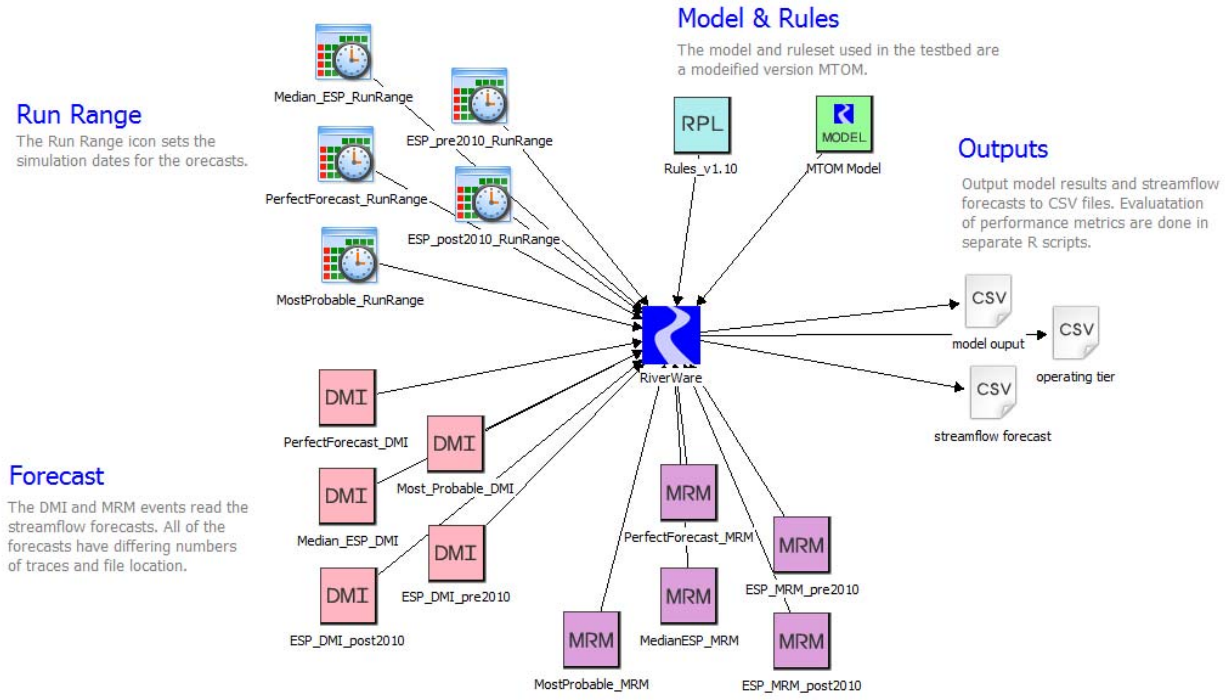

Figure 5-4. Setup of the testbed in RiverSMART. The forecasts are input to MTOM using the Run Range, DMI, and MRM events. Arrows depict the direction of data flow and process.

5.4.2 MTOM Research Model

The version of MTOM used in this study was adapted from the operational version of MTOM. Rule logic, which simulates CRB operations based on the 2007 Interim

Guideline, was extended back to 1981. Since the MTOM research model simulates prior to the implementation of the 2007 Interim Guidelines, Lake Powell and Lake Mead reservoir operations cannot be compared to observations until 2008 when the 2007 Interim Guidelines were implemented. A pseudo-historical projection of reservoir operations was created by running historical streamflow through MTOM. This simulation will be referred to as 'Historical Streamflow Projections' allows for comparison of the streamflow forecast's operational projections prior to 2008 and will be discussed further in the following section.

To run hindcasts through MTOM, historical reservoir conditions and inflows were added to the modeling framework to provide initial conditions for each simulation. Minor changes were also necessary in the rule logic. This included constraints on water use to match historical events. For example, the Central Arizona Project cannot divert water before it historically started diverting water. To isolate the effects of streamflow forecasts on reservoir operations, the model assumes perfect knowledge of Upper and Lower Basin depletions and Lower Basin intervening flows.

5.4.3 Streamflow Forecasts

The testbed was used to analyze current and experimental streamflow forecasting methods for 1981-2016 water years. The streamflow forecasts assessed in the testbed are as follows:

- **Historical Streamflow** (1981-2016): The historical streamflow is the observed or calculated historical flow at the 12 MTOM streamflow forecast location. The historical streamflow will be used for three purposes in this work:
    - Historical streamflow is used as the reference flows for other streamflow forecasts when calculating the hydrologic performance of each forecast.
    - Historical streamflow is used to create a baseline of pseudo-historical reservoir operations for the full simulation period. Prior to the 2007 Interim Guidelines, the CRB was operated under different reservoir operating rules. To analyze reservoir operations prior to 2007, historical streamflow was run through MTOM to produce reservoir operations as if the Interim Guidelines were implemented. These '**historical streamflow projected**' operations are used in place of historical operations. The historical streamflow projected operations allows for analysis of the effects of streamflow forecasts on operational projections, but does not allow for analysis of the error associated with the model.
    - Historical streamflow is used to assess potential model logic and parameterization errors from 2008-2016 after the 2007 Interim Guidelines were implemented. The observed reservoir operations should match the historical streamflow projected operations closely.
- **Climatology** (1981-2016): Climatology is the observed streamflow at each forecast location from 1981-2010. Using the historical flows from the climatology period, an ensemble of 30 members is created through an index

sequential method (ISM). ISM traces sample the historical flow, with a trace starting each year and extending to the end of the 2-3 year run period. For simulations at the end of the record, the forecast wraps back to the beginning of the record (e.g. 2010 to 1981). ISM is a common method used in Reclamation's long-term planning model, the Colorado River Simulation System (Prairie, Rajagopalan Balaji, Fulp Terry J., & Zagona Edith A., 2006).

- **ESP** (1981-2016): ESP reforecasts were provided by the CBRFC. ESP forecasts are produced with the Sacramento Soil Moisture Accounting (Sac-SMA) model calibrated with historical conditions and climate traces from the period of record, 1981-2010 (Burnash, Ferral, & McGuire, 1973; Miller et al., 2012). This period of record rolls forward every 5 years to include more years of climate variability. ESP ensembles have 30 traces corresponding to the 30 years of historical precipitation and temperature traces. For the 1981-2010 ESP forecasts, the trace from the forecasted year's climate forcings was removed from the ensemble to avoid a trace with perfect knowledge of the temperature and precipitation that drive the Sac-SMA model; therefore, the ESP forecasts from 1981-2010 have 29 traces. ESP forecasts for 2011-2016 have 30 traces.
- **4-Basin k-NN Forecast:** The 4-Basin k-nearest neighbors (kNN) method was created and evaluated in the previous chapter of this dissertation. This experimental forecast will be evaluated further in the testbed. The 4-Basin kNN trace weighting method weights ESP traces using North American Multi-model Ensemble (NMME) 1-month and 3-month watershed scale temperature

and precipitation forecasts, along with the preceding 3-month average observed streamflow. The method is evaluated on four separate basins in the CRB (Main Stem, Green, Gunnison, and San Juan) and is then recombined to create a full ensemble that includes the Lake Powell unregulated inflow forecast location. NMME is only available from 1982-2016, so the forecast record has one less year than ESP. For more details, see the previous chapter, Section 4.4.

5.4.4 Performance Metrics

The testbed uses a specified set of performance metrics to analyze each streamflow forecast. The metrics are split into two categories: hydrology metrics and operational projection metrics. Metrics are normally evaluated for a 24-month (2-year) to 1-month lead time from a projected date that is normally the EOWY (end of September).

5.4.4.1 Hydrology Metrics

Forecast performance can be measured through different attributes. Here we provide a brief summary of forecast attributes from Wilks (1995). Many of these forecast performance attributes can be described by more than one hydrology metric.

- ***Accuracy*** is a measure of overall correspondence between the forecasts and observations and can be interpreted as the overall quality of a set of forecasts. Many of the below attributes can be interpreted as components of accuracy
- ***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 differs from accuracy in that it measures the

correspondence between individual forecast pairs opposed to the average correspondence.

- ***Reliability***, or conditional bias, is a measure of the agreement between the forecast probabilities and 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. It is also referred to as the degree the forecasts sort the observed events into a subset of different groups. 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. In probability forecasts, forecasts with no resolution have no discrimination and vice versa (Bröcker, 2015).
- ***Sharpness*** is a measure of the degree of spread of a forecast and is a measure of the forecast alone. Forecasts that do not deviate from climatology are said to have low sharpness and are therefore under-confident. Forecasts that that are frequently different than climatology are sharp. Forecasts that are too sharp, meaning they are sharp, but at the expense of missing the observed event are over-confident and under-dispersed.

The hydrology metrics listed below measure the attributes of forecast performance based on Lake Powell's annual unregulated inflow. This study focuses on the Lake

Powell forecast location since it is an aggregate of all Upper Basin streamflow forecasts. The hydrology performance metrics are described as follows:

- **Root Mean Square Error (RMSE):** RMSE is the square root of the average of the squared differences between projections and observations. Since the errors are squared, larger errors have a greater influence on RMSE than smaller errors.
- **Continuous Ranked Probability Skill Score (CRPSS):** CRPSS is the accuracy of a forecast relative to the accuracy of a reference forecast such as climatology (Hersbach, 2000). Continuous ranked probability score represents the integrated squared difference between the cumulative distribution function of the forecasts and the corresponding distribution of the observations. CRPSS is similar to the ranked probability skill score (RPSS) except it uses a continuous distribution instead of categories. CRPSS ranges from 1 to $-\infty$. A perfect CRPSS is equal to 1, a score of 0 means the skill of the forecast is equal to that of climatology, and a negative score means the forecast is less skillful than climatology.
- **Forecast Spread vs. Observations:** For this research, the spread of the forecast is viewed as a plot of forecast ensemble versus observations, where the forecast ensemble is represented as boxplots. Each boxplot denotes one probabilistic forecast. The closer the boxplots fall to the 1:1 line, the closer the forecast is to the observed streamflow.

The boxplots allows for visualization of forecast spread or sharpness at multiple lead times. The confidence of a forecast describes if the forecast is under-dispersed meaning the forecast range is too narrow (over-confident) or if the forecast range is too large and there is no differentiation of where the observations might fall in the forecasted values (under-confident). A good forecast should span the observed streamflow and be able to discriminate between a high or low flow events.

- **Reliability Diagrams:** Reliability diagrams provide a measure of how closely the forecast probability is to the actual frequency of an observed event. Reliability diagrams describe the full joint distribution of forecast and observation probabilities (Wilks, 1995). A reliability diagram groups forecast probabilities into bins (x-axis), where the bin width is determined based on the number of observations (similar to determining the bin width of a histogram). Here, we use 5 bins of 0.2 width. The frequency of an observed event's occurrence within the climatological record is determined for each forecast and plotted in the respective forecast probability bin (y-axis). A forecast with perfect reliability would result in forecasts with an X% probability occurring X% of the time on average over all forecasts, resulting with points plotted on a diagonal 1:1 line. A sharpness histogram is plotted within each reliability diagram to show the number of forecast probabilities in each bin.

  Reliability Diagrams provide information about the reliability, resolution, and sharpness (confidence) of a forecast. Figure 5-5, a figure from Hamill (1997),

illustrates different forecast attributes in reliability diagrams. Diagram *a* shows a forecasts that is climatology, falling on the 1:1 line. Diagram *b* has minimal resolution and therefore cannot differentiate between different observed events. Diagram *c* is a forecast showing a conditional bias of the forecast probability being lower than the observed event's relative frequency within the climatological record. This is sometimes referred to as conditional 'under-forecasting', which is confusing when we normally refer to under-forecasting as it relates to unconditional bias (error). Therefore, we will refer to diagram *c* as having conditional bias of the forecast probability being lower than the observed relative frequency of the event. Diagram *d* is a forecast that has good resolution but at the expense of reliability. Diagram *e* shows a reliable rare event. Diagram *f* is a forecast where the sample size is too small or the forecast is too sharp.

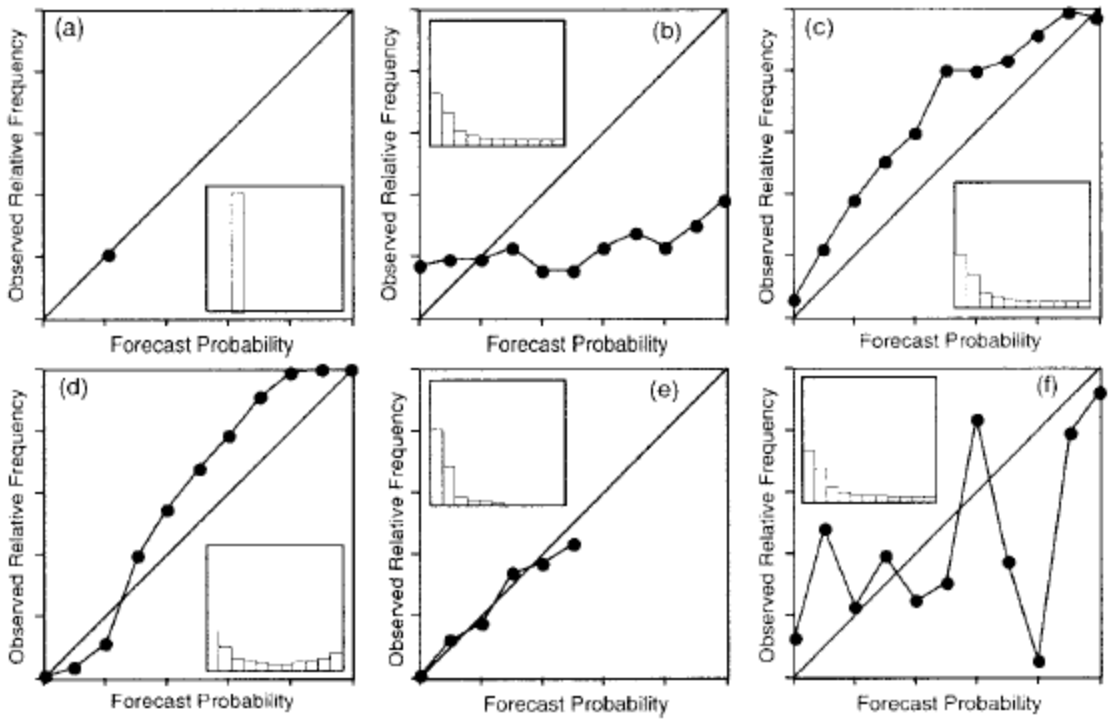

Figure 5-5. Examples of different reliability diagrams and their associated forecast performance. This figure was taken from Hamill (1997). The analysis of each reliability diagram are described in the paragraph above

5.4.4.2 Operational Projection Metrics

The operational projection metrics are used for two purposes: (1) to measure how well MOTM performs when simulating reservoir operations according to the 2007 Interim Guidelines and (2) assessing how streamflow forecast skill effects the performance of operational projections. The MTOM outputs, which are evaluated by operational projection metrics, include annual outflow, EOWY storage, EOWY pool elevation, and operating tiers for Lake Powell and Lake Mead. The specific metrics are described below:

- **Evolution of Pool Elevation:** The evolution of pool elevation projections for each model run allows for comparison of the projections to observed pool elevation. This visualization technique allows for trends in pool elevation to be

observed throughout the year as opposed to only analyzing the end of year value.

- **RMSE of Pool Elevation:** See RMSE description in the Hydrology Metrics section. Since operating tier determination is based on pool elevations, it is important to analyze the errors in pool elevation at various lead times.
- **Model Error:** Model error is calculated using a mass balance of known quantities to calculate any unaccounted for water present within Lake Powell or Lake Mead. The water mass balance for each reservoir is:

$$Storage_t = Storage_{t-1} - Outflow_t + Inflow_t - Evaporatoin_t + BankStorage_t + \varepsilon rror \quad \text{Eq. 5-1}$$

For this calculation, the historical streamflow is run through MTOM. The difference between the observed and MTOM calculated values are evaluated for each term in the mass balance. Since there are no observed values for evaporation or bank storage, the potential errors assocaiated with these terms are lumped into a total error term. The resulting equation is:

$$\left(S_t^{MTOM} - S_t^{obs}\right) = \left(S_{t-1}^{MTOM} - S_{t-1}^{obs}\right) - \left(O_t^{MTOM} - O_t^{obs}\right) + \left(I_t^{MTOM} - I_t^{obs}\right) + \varepsilon \quad \text{Eq. 5-2}$$

where S is the storage, O is the annual outflow, I is the annual inflow, and ε is the annual total unaccounted for error in the mass balance. The subscript t is the end of the projected year and *t-1* is the beginning of the simulation or the beginning of the year analyzed (i.e. April 2013 projection of WY 2014 years error would use *t-1* of October 2013). The superscripts *MTOM* and *obs*

represent the MTOM projected and observation values, respectively. This analysis assumes that there are no errors in the inflow, since historical inflow was input to the model. Rearranging produces the following equation:

$$\varepsilon = \left(S_t^{mtom} - S_t^{obs}\right) - \left(S_{t-1}^{MTOM} - S_{t-1}^{obs}\right) + \left(O_t^{mtom} - O_t^{obs}\right) \quad \text{Eq. 5-3}$$

This equation is used to calculate the error in Lake Powell's mass balance. The Lake Mead mass balance has one extra term that accounts for error in the inflow to Lake Mead as a result of errors in Lake Powell's outflow.

- **Percent Correct:** Percent Correct is a categorical score that measures the percent of forecasts in each category that are correct. Percent Correct ranges from 0% to 100%, representing the percentage of time the correct operating tier is projected by the model.
- **Heidke Skill Score:** The 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 Heidke Skill Score is a measure of skill for categorical events. The score ranges from 1 to -∞, where 1 is a perfect skill score, 0 indicates skill equal to random chance, and negatives values have no skill.

5.5 Results

5.5.1 Hydrology Metrics

5.5.1.1 CRPSS and RMSE of Ensemble Streamflow Forecasts

Streamflow forecasts are compared by analyzing the annual WY unregulated inflow to Lake Powell. The first set of hydrology metrics, CRPSS and RMSE, are measures

of the skill and accuracy of a probabilistic forecast relative to climatology (1980-2010). Figure 5-6 displays the CRPSS for ESP, 4-Basin kNN, and Climatology at a 24- to 1-month lead. The 'month/number' descriptor on the x-axis represents the month the forecast was created and the number of lead months to the end of the WY. Since we are computing an annual flow, the forecast includes observed flows once the lead is less than 12-months. For example, at an 11-month lead in November, October flows have been observed and no longer need to be forecasted. Therefore, as the lead decreases beyond a 12-months, more months of observed flows are included in the forecast and the skill increases towards 1.

At long leads, from 24- to 13-months, the CRPSS of most forecasts are close to climatology (zero line), showing that ESP converges towards climatology at longer leads. The 4-Basin kNN forecast has a larger range of skills at longer leads with median close to zero. This illustrates that there is no benefit to using the ESP or 4-Basin kNN forecast over climatology during this period.

The skill increases above climatology in the fall of the out-year, starting in October at a 12-month lead. The increase in skill is likely due to the knowledge of antecedent basin conditions such as soil moisture. Early season soil moisture often contributes significantly to streamflow forecast skill, especially in the early season, even though streamflow forecast skill is highly dependent on knowledge of snow in the high mountain regions in the late season (Randal D. Koster, Mahanama, Livneh, Lettenmaier, & Reichle, 2010). This forecast skill relationship is apparent for ESP and 4-Basin kNN forecasts in the CRB.

As the season progresses through winter and spring, the CRPSS increases as more information about snow storage in the basin is observed. The median skill of the 4-Basin kNN method is slightly higher during this period than the ESP forecast. The range of skill with the 4-Basin kNN method is larger in some months than ESP. 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.

By April at a 6-month lead and at the beginning of the runoff season (April – July), the skill has increased much above the climatology. The skill continues to increase through the runoff season as more months in the annual inflow are observed. During the late summer months, ESP and 4-Basin kNN forecasts can be too tightly constrained, while Climatology has a larger range of forecasted inflows. This causes Climatology to perform slightly better the other two forecasts.

The RMSE of WY Lake Powell unregulated inflow is shown in Figure 5-7. The 4-Basin kNN forecast outperforms the other forecasts at all leads. This is a different result compared to CRPSS in Figure 5-6. The RMSE stays relatively constant in the out-year and decreases starting in November at an 11-month lead as more information is known about initial conditions in the basin. The RMSE continues to decrease into the spring and to the end of the water year. It is important to take both metrics, CRPSS and RMSE, into account when assessing the streamflow forecasts.

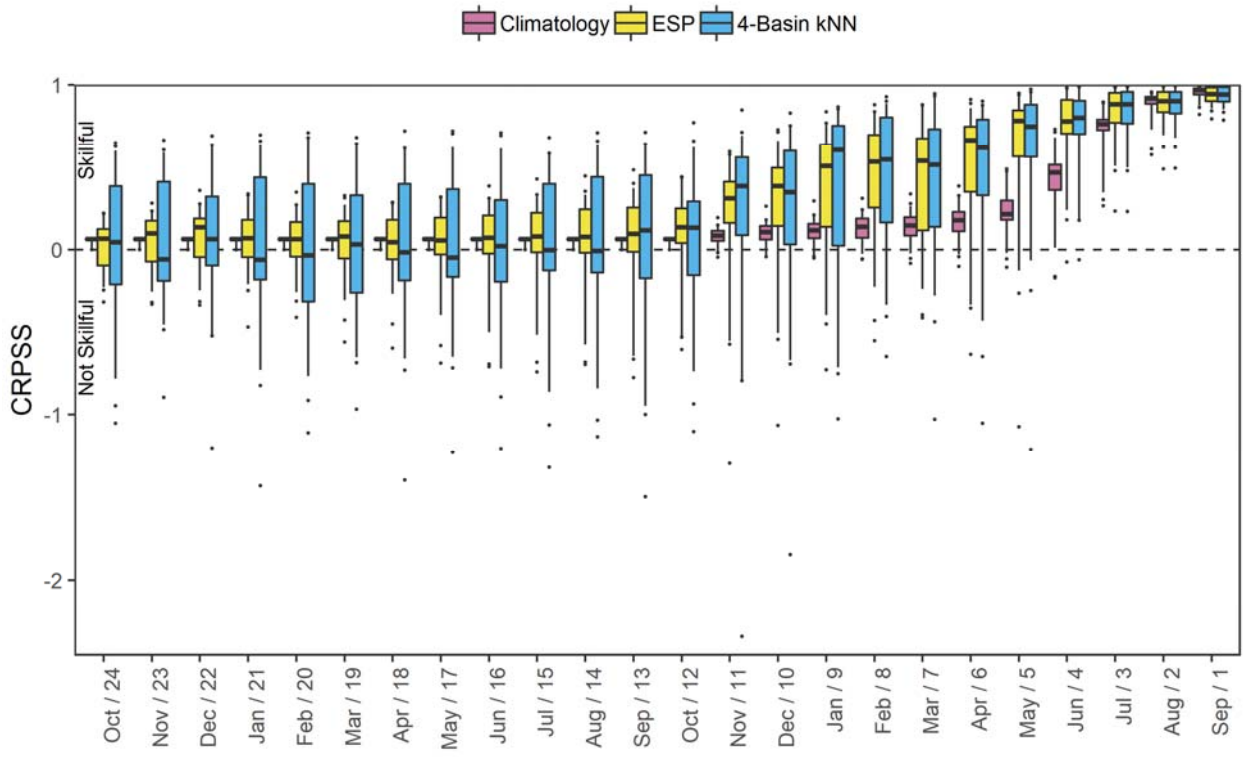

Figure 5-6. CRPSS of annual WY Lake Powell unregulated inflow. CRPSS at a 24- to 1-month lead is compared for Climatology, ESP, and 4-Basin kNN. The forecasts are available from 1982-2016. Each boxplot includes one data point for each year with 35 data points in each boxplot. The x-axis shows the 'month/number of lead months' to the end of the WY.

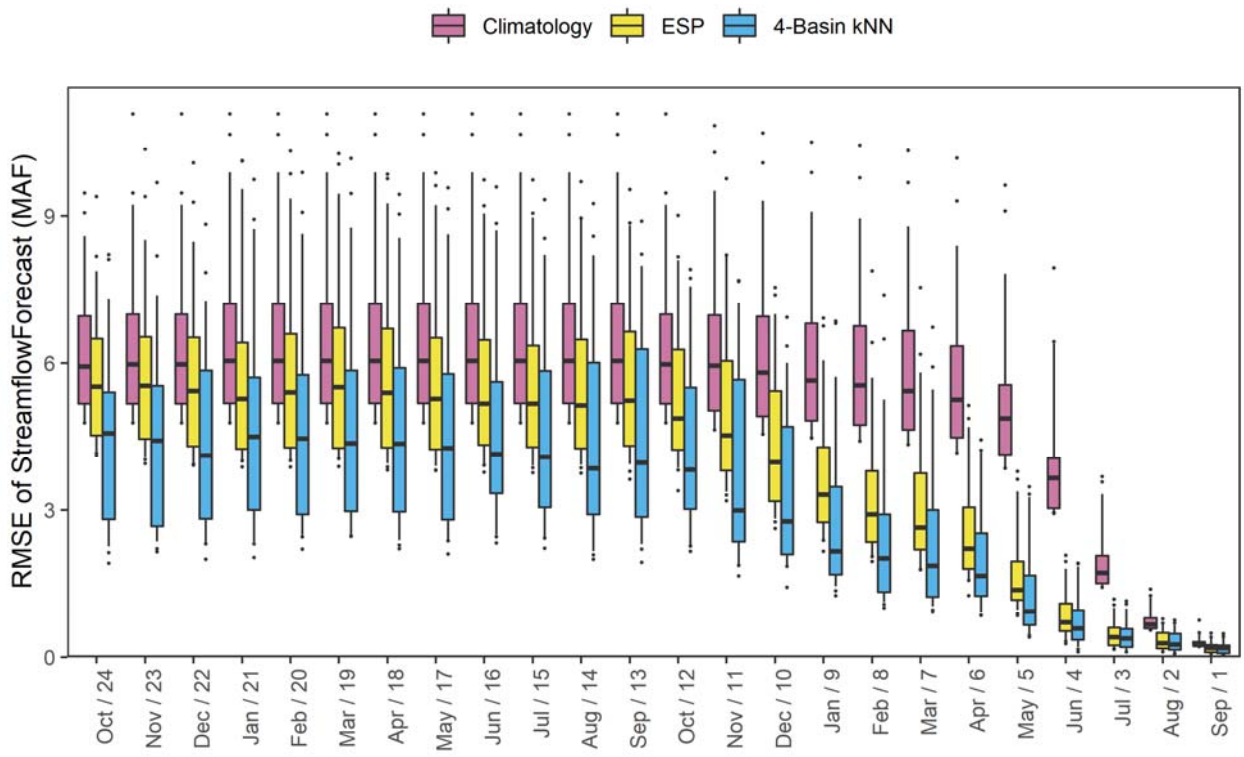

Figure 5-7: RMSE of annual WY Lake Powell unregulated inflow. The RMSE at a 24- to 1-month lead is compared for Climatology, ESP, and 4-Basin kNN. The forecasts are available from 1982-2016. Each boxplot includes one data point for each year with 35 data points in each boxplot. The x-axis shows the 'month/number of lead months' to the end of the WY.

5.5.1.2 Visualization of Spread: Scatter Plot of Forecast vs. Observations

The second hydrology metric is a visualization of spread of the streamflow forecasts. The spread of an ensemble forecast indicates information about bias, confidence, and discrimination of the forecast. A visualization of annual WY Lake Powell unregulated inflow spread for ESP and 4-Basin kNN are shown in Figure 5-8 and Figure 5-9, respectively. The Climatology figure was omitted because it does not portray useful information since it forecasts the same values though the out-year and is only informed by observed flow in the current forecasted year. Separate plots depict one lead time and contain 35 boxplots, each representing a single ensemble forecast of annual inflows. The location of the boxplot on the x-axis represents the observed

annual flow value versus the ensemble spread forecast on the y-axis. The boxplot's whiskers represent the full range of the forecast, the box is the 25th and 75th percentile, and the mid-line representing the forecast mean. The boxplot should span the 1:1 line indicating the forecasts range contains the observed flow.

At longer leads, the ESP forecasts have converged to climatology and lacks discrimination since all forecasts project relatively the same flows. The 4-Basin kNN forecasts have a smaller spread with some of the higher observed years forecasted at slightly higher flows than average. The spread of the 4-Basin kNN forecasts are also smaller, showing that the forecast may be exhibiting too much sharpness at such a long lead. The ESP forecasts start to shift from climatology in the summer of the out-year, while the 4-Basin kNN forecasts don't show as much change. In the fall of the out-year (October and November, 12- and 11-month leads), the ESP and 4-Basin kNN forecasts start to tighten, with most forecasts spanning the 1:1 line. The highest observed years start to discriminate from lower and average flow years.

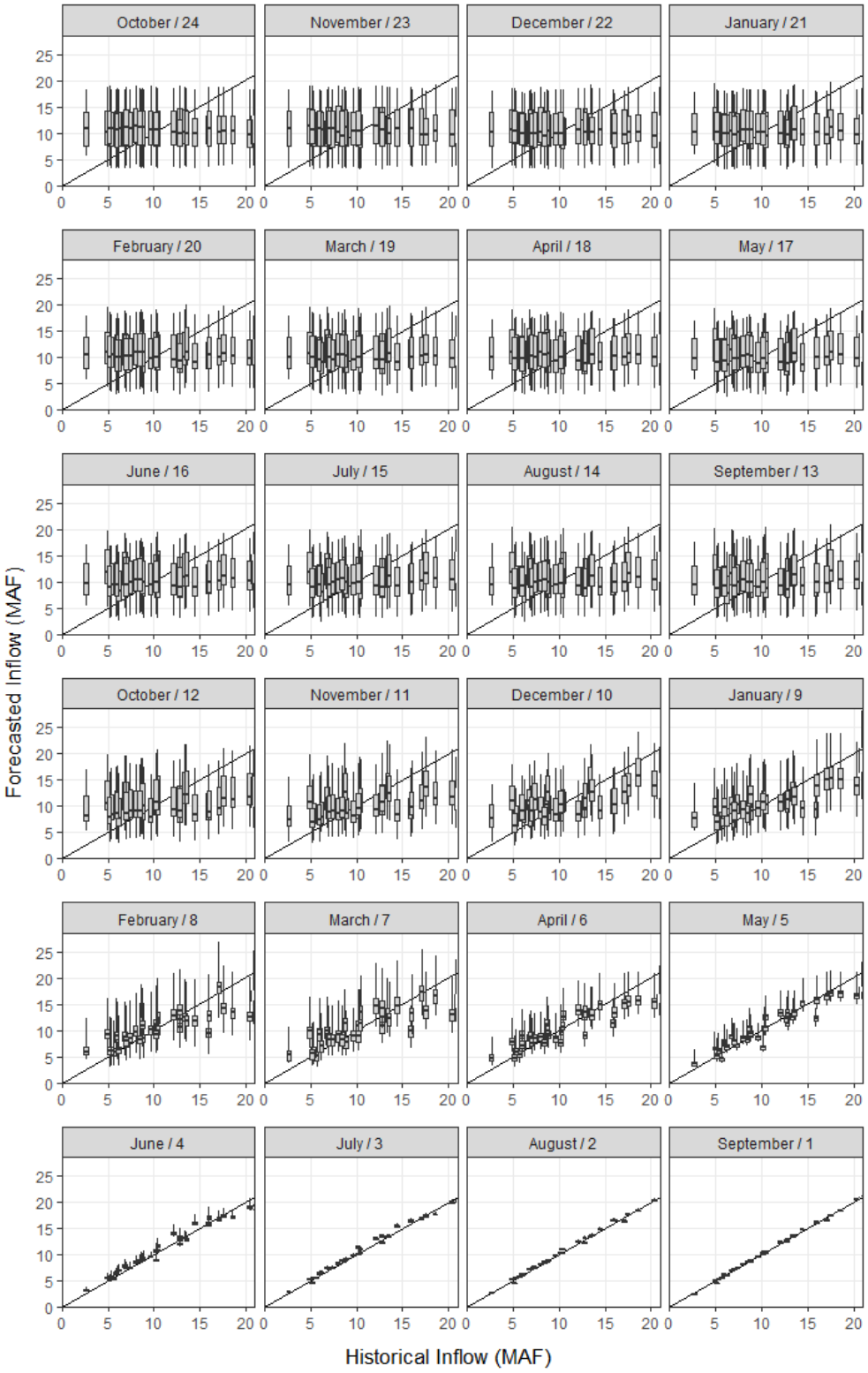

Figure 5-8. Spread visualization of annual Lake Powell unregulated inflow for ESP. Boxplots of ESP forecast versus the historical Lake Powell annual unregulated inflow (1981-2016) for leads of 24- to 1-month. Boxplots represent the 25th-75th quantile and whiskers show the full range of the forecast.

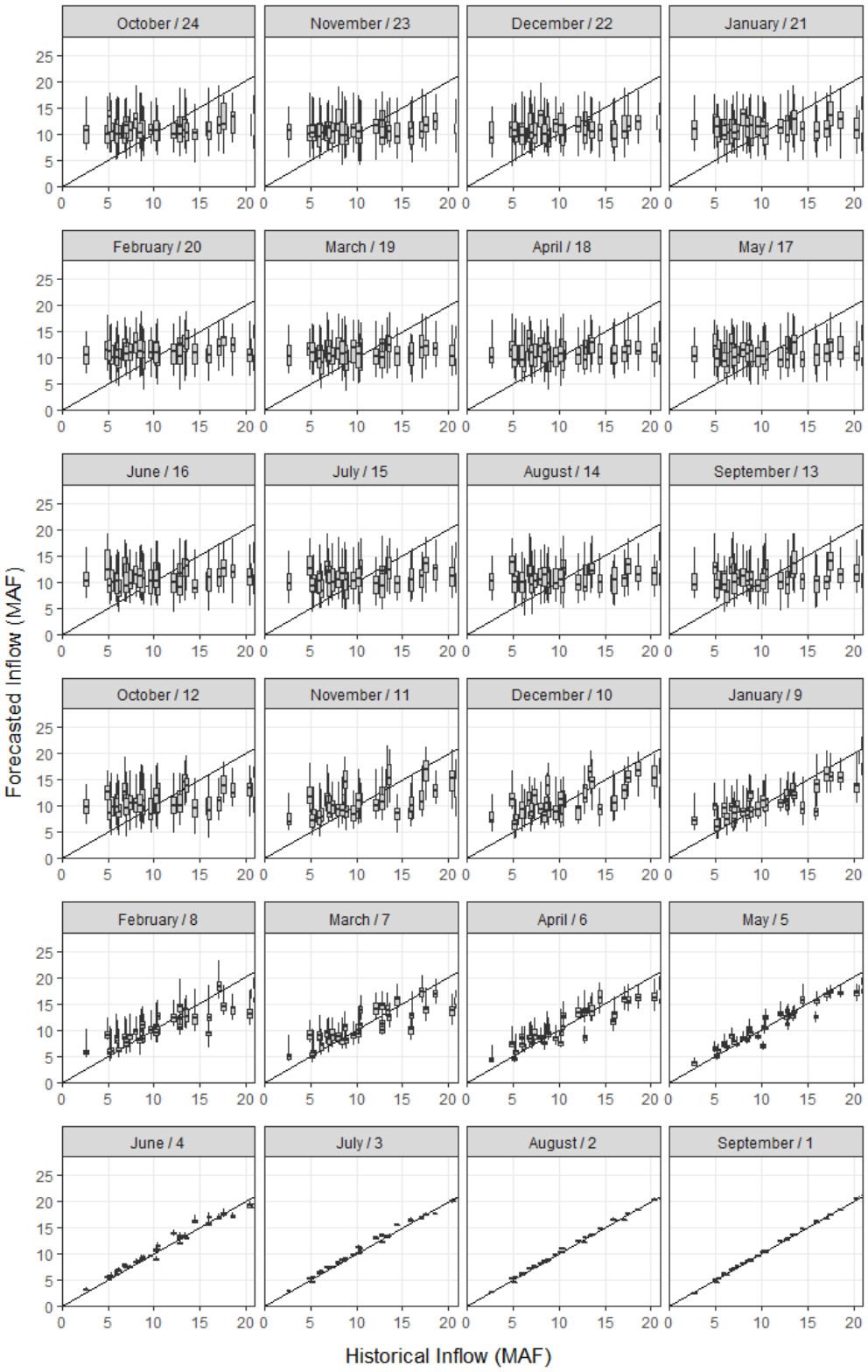

Figure 5-9. Spread visualization of annual Lake Powell unregulated inflow for 4-Basin kNN. Boxplots of 4-Basin kNN forecast versus the historical Lake Powell annual unregulated inflow (1981-2016) for leads of 24- to 1-month. Boxplots represent the 25th-75th quantile and whiskers show the full range of the forecast.

The extreme years in this hindcast period are hard to capture at extended lead times since the trace containing the historical weather that produced the extreme streamflow was removed from the ESP forecast. This is done so the forecast doesn't have perfect knowledge for 1 of the 30 traces. For instance, 2002 has the lowest streamflow in the analyzed period. The 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 won't produce as low of streamflow until the initial basin conditions drive the forecast, as opposed to precipitation and temperature.

By January (9-month lead), the forecast spread narrows, especially the 25th to 75th quartiles of the ensembles. The 4-Basin kNN forecast remains narrower than the ESP forecast through the end of the WY. Both forecasts capture the wet years well. In April at a 6-month lead, both forecasts' spread have narrowed significantly, in some instances to exclude the observed streamflow. The 4-Basin kNN forecasts have a narrower spread, with more forecasted flows closer to the 1:1 line. By June (4-month lead), forecasts show over-confidence; many ensembles are too narrow and do not capture the observed streamflow. At shorter lead than June, it is hard to discern the spread of the forecasts since they have converged on the 1:1 line.

Reliability Diagram

The third hydrology metric is reliability diagrams, which measure the relationship between the observed relative frequency of an event and the forecasted probability. Reliability diagrams describe the reliability, confidence, and resolution of the

forecast. The reliability diagrams for ESP and 4-Basin kNN are shown in Figure 5-10 and Figure 5-11, respectively, for annual Lake Powell unregulated inflow forecasts at leads from 24- to 1- month. The reliability diagram for Climatology was omitted; the forecast is reliable through most of the record, which is expected since the forecast covers a wide range of possible inflows.

At longer leads, the ESP and 4-Basin kNN forecasts have good reliability, as the line falls close to the 1:1 line. The 4-Basin kNN forecast is slightly less reliable and over-confident shown by the reliability line being above the 1:1 line for the lower forecast probability. The histogram within each plot shows where the observations fall within the forecasts. ESP has good sharpness and resolution at longer leads as shown by a histogram in each plot, which are fairly evenly distribution in each bin. The histograms for 4-Basin kNN shows that the forecasts have a conditional bias where the forecast probabilities are lower than the observed probabilities.

As leads decrease to the winter of the forecasted year, both ESP and 4-Basin kNN show worse reliability with the reliability line higher than the 1:1 line at lower forecast probabilities. This is because the observed flows are falling in the lower range of the forecast ensembles more often. As lead decreases into the spring and summer, both forecasts become much too sharp and over-confident. Forecasts are consistently too high relative to the event's relative frequency, meaning the average forecast probability is larger than the average observed frequency. As seen in the histogram, most observed frequencies fall in the lowest bin of the forecast probabilities, many of these falling outside the forecast ensemble.

The limited number of forecasts in specific bins give a reliability diagram a jagged look. This can be clearly seen in August at a 2-month lead in the ESP figure where one bin has no forecasts. The forecasts at shorter leads have poor reliability and are forecasting streamflows higher than observations in many cases. This matches the findings from visualization of spread in Figure 5-8 and Figure 5-9.

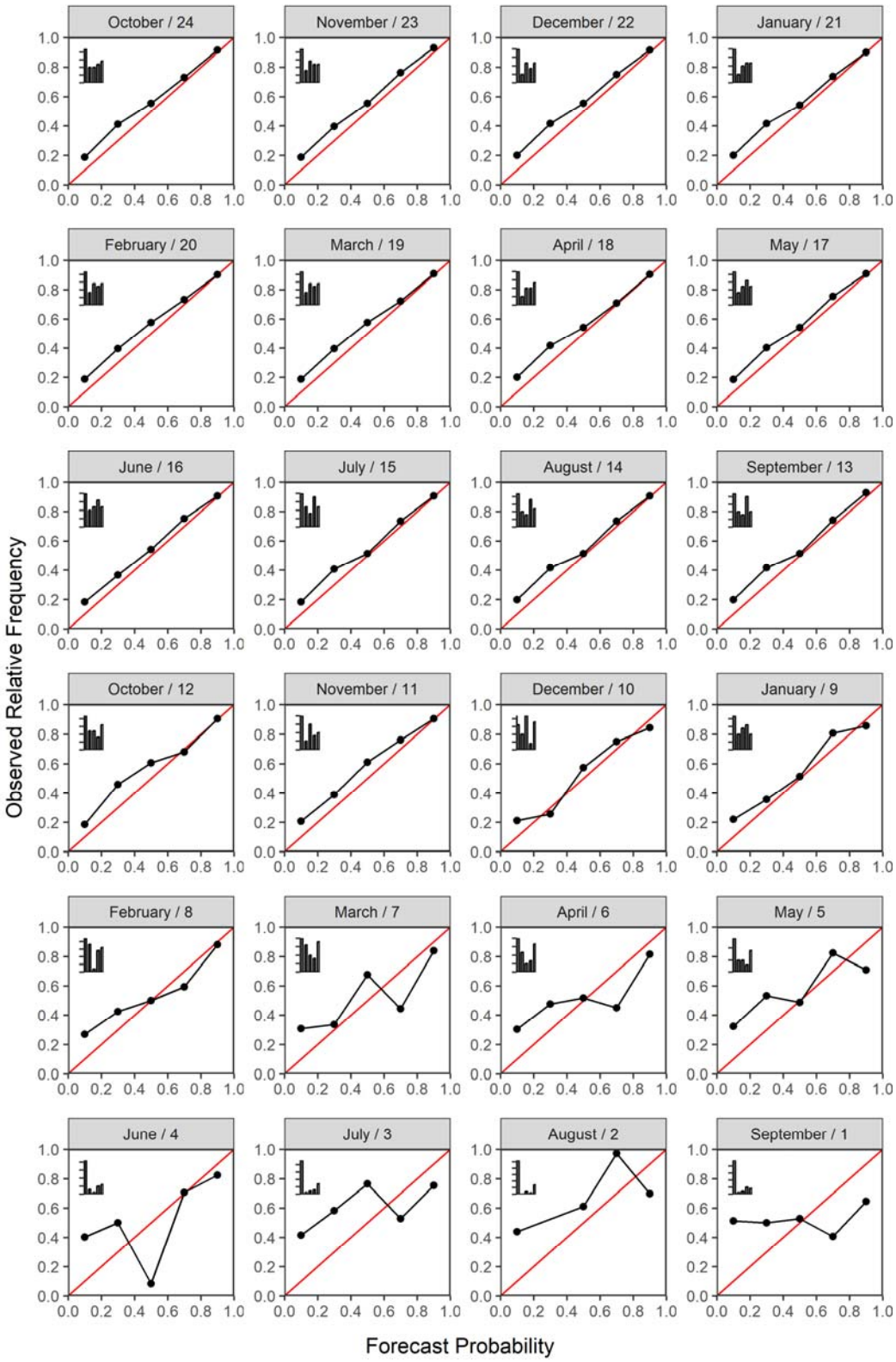

Figure 5-10. Reliability diagram of Lake Powell WY unregulated inflow for ESP. Diagrams are shown for a 24- to 1- month lead for Lake Powell annul unregulated inflow from 1982-2016.

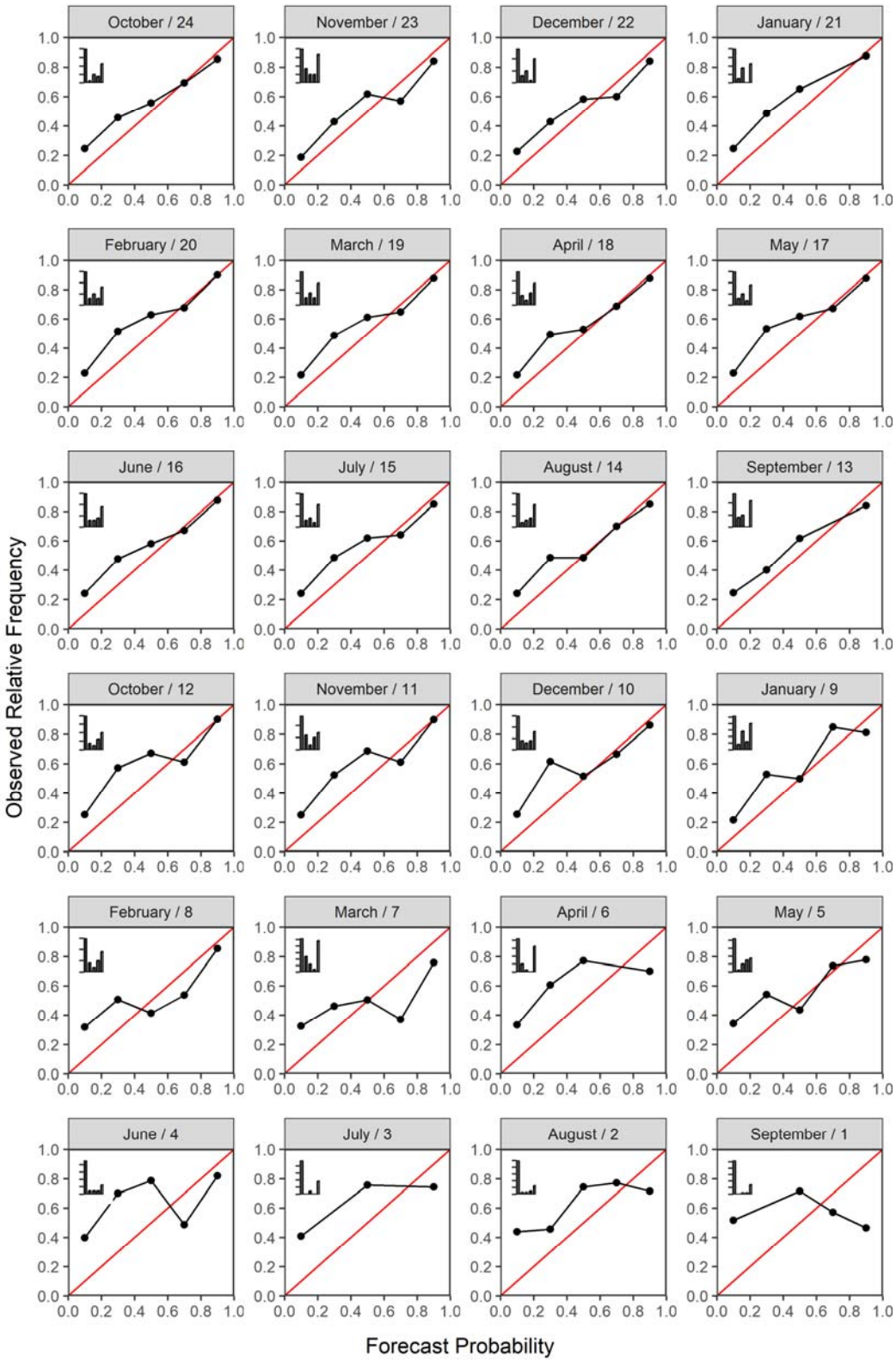

Figure 5-11. Reliability diagram of Lake Powell WY unregulated inflow for 4-Basin kNN. Diagrams are shown for a 24- to 1- month lead for Lake Powell annul unregulated inflow from 1982-2016.

5.5.2 Operational Projection Metrics

The operational projections are assessed in two separate sections that evaluate the performance of the MTOM through (1) historical streamflow simulations and (2) streamflow forecast simulations. In the first section, metrics enable better understanding of potential MTOM modeling errors through evaluation of model simulations with historical streamflows. In the second section, metrics allow for evaluation of how different streamflow forecasts affect operation projections made by MOTM. The streamflow forecasts compared are Climatology, ESP, and 4-Basin kNN. Operational projections assess errors in annual outflow, EOWY storage, and EOWY pool elevation, as well as categorical skill scores of operational tiers and releases.

5.5.2.1 Historical Streamflow Operational Projections

Errors in Pool Elevation & Outflow for Lake Powell and Lake Mead

Historical streamflow was used to analyze differences in historical and projected operations of Lakes Powell and Mead, as well as errors caused by MTOM assumptions or parameterization. Figure 5-12 illustrates differences between the historical and MTOM projected annual outflow and EOWY storage of Lakes Powell and Mead for each year from 2008-2016 (post-Interim Guidelines). Errors in the outflow of Lake Powell are due to incorrectly projected operating tier or special case operations. The Lake Mead outflow errors are very small and not discernable at the scale of Figure 5-12.

The historical operating tier and release for each year are compared to the MTOM projections. The historical operating tier and release are listed at the beginning of

each year's description (see Figure 5-2 for details about the operating tiers). The operating tier and outflow errors cause the larger errors in Lake Powell and Lake Mead EOWY storage. Since outflow from Lake Powell flows into Lake Mead, there is an inverse relationship between their storage errors. There are also smaller errors visible in EOWY storage for both reservoirs. These errors are due to model assumptions and parameterization, and will be discussed in the 'MTOM Errors' section below.

- *2008 – Upper Elevation Balancing – April Adjustment to Equalization at 8.98 MAF*

  At longer leads from 24- to 7-month, MTOM simulates Upper Elevation Balancing with no April Adjustment and a Lake Powell release of 8.23 MAF. The historical inflow to Lake Powell was lower than the Most Probable forecasted inflow, which is used to determine the annual release and potential April Adjustment. Due to the over-forecasting of the Most Probable forecast, the operating tier was changed to Equalization with an April Adjustment in the April 24MS. If the actual inflow was known, like in this simulation, there would have been no April Adjustment. MTOM was forced into an April Adjustment in April 2008 (6-month lead), since the decision to adjust had been made by Reclamation. Shorter leads also have error in the outflow and storage as MTOM balances the contents of Lake Powell and Lake Mead slightly differently than historically observed.

- *2009 – Upper Elevation Balancing at 8.23 MAF*

  The 24- to 19-month leads simulate Equalization above 8.23 MAF (~9.55 MAF) in 2009 due to projections of no 2008 April Adjustment, leading to a lower Lake Powell release (see above description of 2008). The storage in Lake Powell is under-projected and the storage in Lake Mead is over-projected due to the higher release from Lake Powell.

- *2010 – Upper Elevation Balancing at 8.23 MAF*

  There are no errors in 2010 annual outflow from Lake Powell and Lake Mead.

- *2011 – Upper Elevation Balancing – April Adjustment to Equalization at 13.74 MAF with a carryover release of 1.24 MAF in WY 2012*

  The historical reservoir operations in WY 2011 were more complex due to a large increase in inflow to Lake Powell in late winter and early spring. The increase in inflow moved the projected Lake Powell pool elevation above the Equalization line, resulting in an April Adjustment to Equalization. The full target release volume for 2011 could not be released by the end of the WY due to power plant capacity constraints at Glen Canyon Dam. Therefore, a portion of the 2011 release (1.24 MAF) was carried over and released in 2012. This special operation caused errors in projecting annual outflow and storage at longer leads

  The annual outflow from Lake Powell was over-projected for all leads since MTOM projects that all or most of the target annual release volume for WY 2011 can be released with no carryover. MTOM projects Equalization from the

beginning of the WY since the model has perfect knowledge of the historical inflow to Lake Powell and did not have to make the operating tier change in April. At shorter leads in WY 2011, MTOM starts to constrain the releases from Lake Powell due to the power plant capacity constraint. As fewer months are left to release in the WY, the outflow error decreases and outflow is carried over to WY 2012.

- *2012 – Equalization at 8.23 MAF with carryover release of 1.24 MAF*

  Lake Powell outflow errors at 24- to 15-month leads were caused by carryover errors in 2011 (see notes above).

- *2013 – Upper Elevation Balancing at 8.23 MAF*

  There are no errors in 2013 annual outflow from Lake Powell and Lake Mead.

- *2014 – Mid-Elevation Release at 7.48 MAF*

  For the leads of 24- to 15-months, MTOM projected Upper Elevation Balancing at 8.23 MAF for WY 2014. This is because MTOM had knowledge of the observed inflow to Lake Powell from August to December 2013, which was well above average (~2.4 MAF). The August 2013 24MS, which is used to determine the 2014 Lake Powell operating tier, used average inflow to Lake Powell (~1.2 MAF) that resulted in different EOCY pool elevations. Since MTOM had the observed inflow, Upper Elevation Balancing was projected for 2014 until August 2013 when the 24MS projected EOCY Lake Powell pool elevation was input to MTOM.

- *2015 – Upper Elevation Balancing at 9.0 MAF*

  There are no errors in 2015 annual outflow from Lake Powell and Lake Mead.

- *2016 – Upper Elevation Balancing at 9.0 MAF*

  There are no errors in 2016 annual outflow from Lake Powell and Lake Mead.

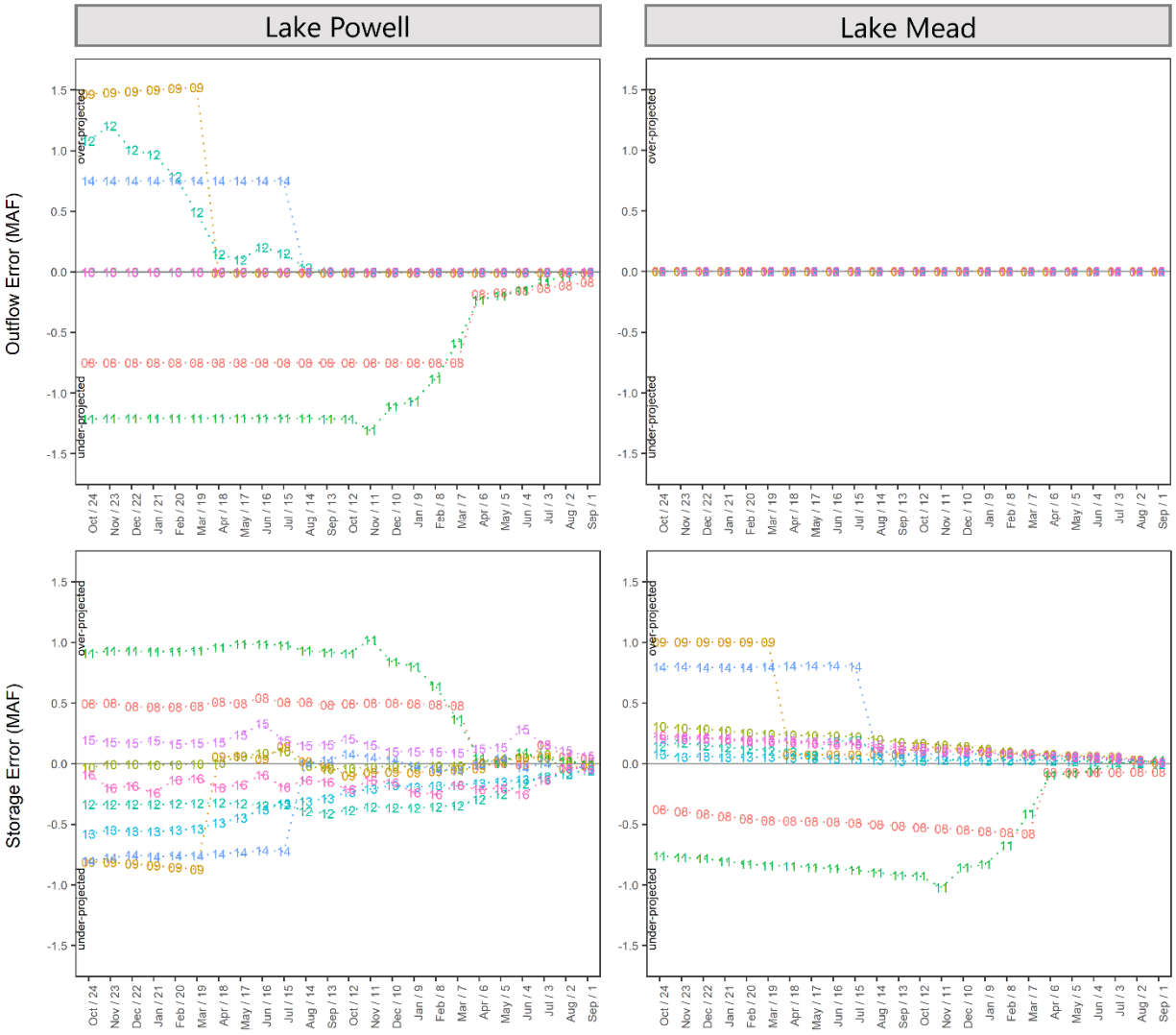

Figure 5-12. Lake Powell and Lake Mead errors in annual outflow and EOWY storage for 2008-2016. Columns represent Lake Powell (right) and Lake Mead (left). The top row is the annual outflow error and the bottom row is the EOWY storage error. Lines represent the error in a specific year from a 24- to 1- month lead with the two-digit number denote the year.

From the above year-by-year analysis we found that most of the large errors in annual outflow and EOWY storage were due to the errors in the deterministic Most Probable forecast used to make tier determinations in the 24MS. These errors in operating tiers were exemplified in 2008 when there was an April Adjustment to Equalization that wouldn't have been made if there was knowledge of the actual streamflow; these errors extended into the 2009 forecast. The errors due to carryover were also a result of streamflow forecasting as well as release constraints at Glen Canyon Dam. Similar to 2008, errors in 2014 were a result of MTOM having knowledge of the actual streamflow into Lake Powell that caused MTOM to project a different operating tier than the 24MS. Overall, this analysis illustrates how errors in streamflow forecasts can cause incorrect annual operating tier determinations and releases; with better streamflow forecasting, Reclamation could make better projections of operational projections at longer leads.

Pool Elevation Evolution

The evolution of pool elevations for MTOM simulations with input historical streamflow are shown in Figure 5-13. The simulations start in January, April, and August of 2008-2016. The solid black line represents observed pool elevation in Lake Powell and Lake Mead.

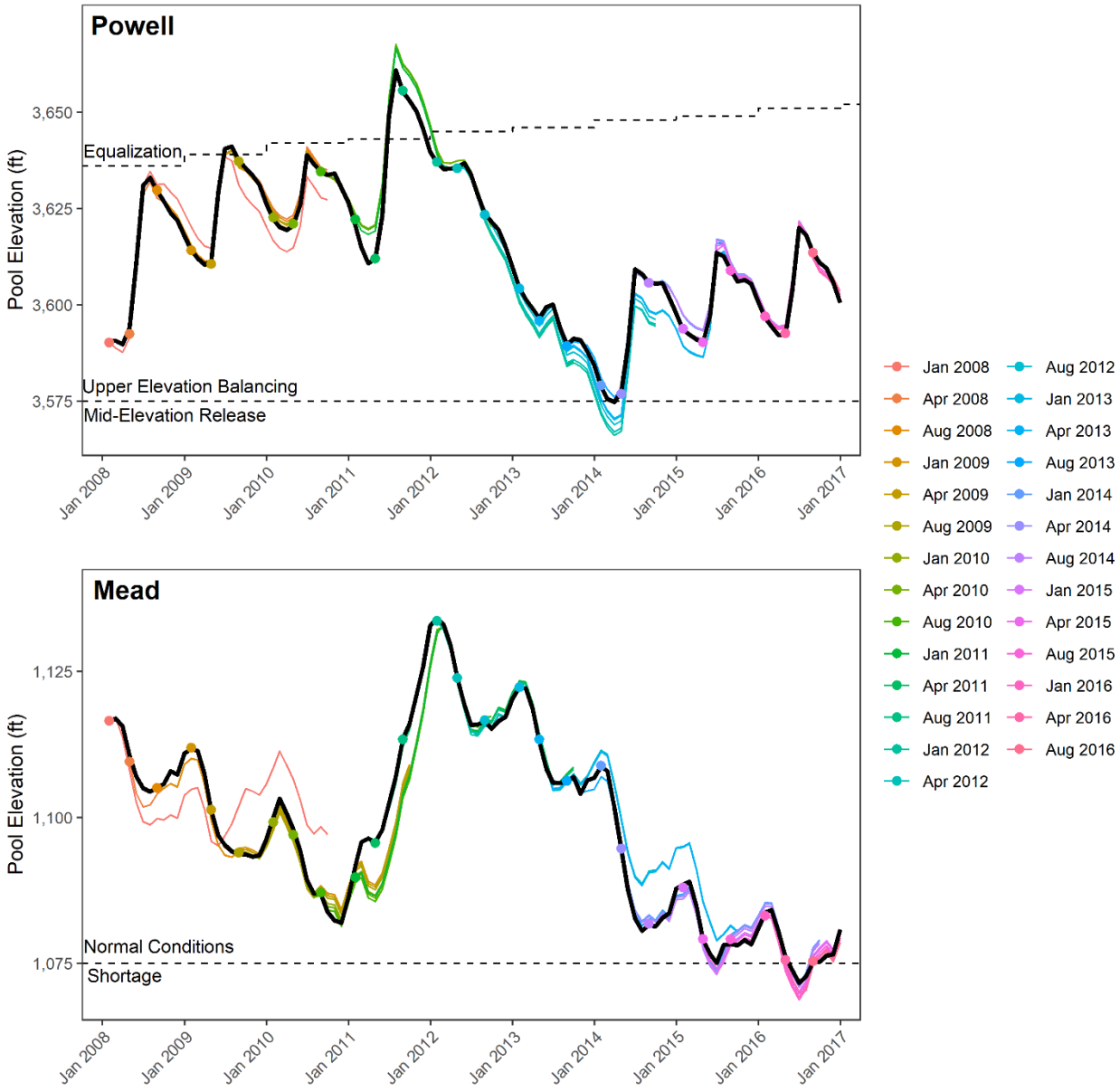

Figure 5-13. Pool elevation evolution for simulations starting in January, April, and August of 2008-2016 for Lake Powell and Lake Mead. The start month's pool elevation is represented by colored dots, followed by a line of the same color which is the pool elevations during the simulation. The solid black line shows observed pool elevation. The elevation of operating tiers are shown by annotated dashed lines.

The observed streamflow projected pool elevations follow observed pool elevations well. The analysis of Figure 5-12 describes many of the errors in the Figure 5-13 that are due to incorrectly projected operating tiers. For instance, in the January 2008 simulation, Lake Powell released 8.23 MAF instead of 8.98 MAF due to an April

Adjustment to Equalization that would not have occurred if the actual inflow to Lake Powell was known at the time of the adjustment in April. The Lake Powell pool elevation is over-projected starting at the end of 2008 through the remainder of the simulation. Lake Mead's pool elevation is under-projected then over-projected for this simulation due to the lasting effects of incorrect releases in the following year's operating tier.

Simulations in 2011 through 2012 show a different pool elevation evolution in Lake Mead and Lake Powell due to the incorrect carryover release timing. Simulations starting at the end of 2012 and 2013 have incorrect projected pool elevations at long leads due to MTOM projecting an Upper Elevation Balancing tier (8.23 MAF release) for 2014 instead of the observed Mid-Elevation Release tier (7.48 MAF). This caused over-projected pool elevations in Lake Mead and under-projected pool elevations in Lake Powell. Errors due to Lake Powell's operating tier projection were discussed in the section above labeled 'Reservoir Operations in the CRB'.

MTOM Error (2008-2016)

Besides errors in operational projections between historical and MTOM simulated reservoir operations, there are also errors in the physical process modeling (parameterization) and assumptions in MTOM. To quantify these errors, a basic mass balance was performed on Lake Powell and Lake Mead. The errors in storage and outflow were removed. The annual mass balance calculation was described in Section 'Operational Projection Metrics'.

Reservoir mass balance errors could be due to incorrect assumptions about evaporation, precipitation, bank storage, unaccounted side inflows into either reservoir, or adaptive management in Upper Basin reservoirs. In MTOM, evaporation and precipitation are assumed monthly volumes based on the area of the reservoir surface. Bank storage is determined by calculation from the change in pool elevation every month. These model parameters can cause errors in the mass balance of the reservoirs since these values are more dynamic than MTOM assumes.

There could also be errors in Lake Powell inflow since the inflow forecast location is for unregulated inflow where Upper Basin reservoir operations are removed. This allows the releases from Upper Basin reservoirs to affect the inflow to Lake Powell, and therefore the mass balance; further investigation would be needed to quantify these errors in Upper Basin reservoir operations. Errors in the intervening flows between Lake Powell and Lake Mead could also exist, though these intervening flows should be correct since they are calculated using a mass balance for the Lower Basin. Historical unregulated inflow input to MTOM could also have errors, since the unregulated inflow to Lake Powell is a calculated value, not a measured value.

Figure 5-14 shows the errors in the mass balance for Lake Powell and Lake Mead. For Lake Powell, errors in the water balance were at maximum 500 kaf at longer leads (~5% of annual Lake Powell unregulated inflow). MTOM is predominantly under-projecting storage in Lake Powell, except a couple years such as 2014 and 2015; therefore, MTOM is simulating less water than observed in Lake

Powell. This could be due to overestimating evaporation, unaccounted for side inflows to Lake Powell, errors in bank storage calculations, errors in the observed streamflow input to MTOM, or lower releases from Upper Basin reservoirs projected by MTOM compared to actual releases. Errors decrease with lead due to fewer months to project. Errors of the Lake Powell water balance at a 24-month lead are 201 kaf on average and decrease to about of 42 kaf at a 1-month lead. These errors are quites small, especially when compared to other known mass balance related errors. For example, the error of the Glen Canyon Dam penstock's measured outflow is 2%, which results in an annual error of 160 kaf with a 8.23 MAF release.

Errors in the Lake Mead water balance have a more linear trend than Lake Powell errors. MTOM over-projects the storage in Lake Mead in all years. This could be due to underestimating evaporation, over-projecting bank storage, or overestimating intervening flows between Lake Powell and Lake Mead. At a 24-month lead, the errors of the Lake Mead water balance are on average 127 kaf and decrease to 14 kaf at a 1-month lead. More work is needed to provide a more accurate assessment of the source of these errors.

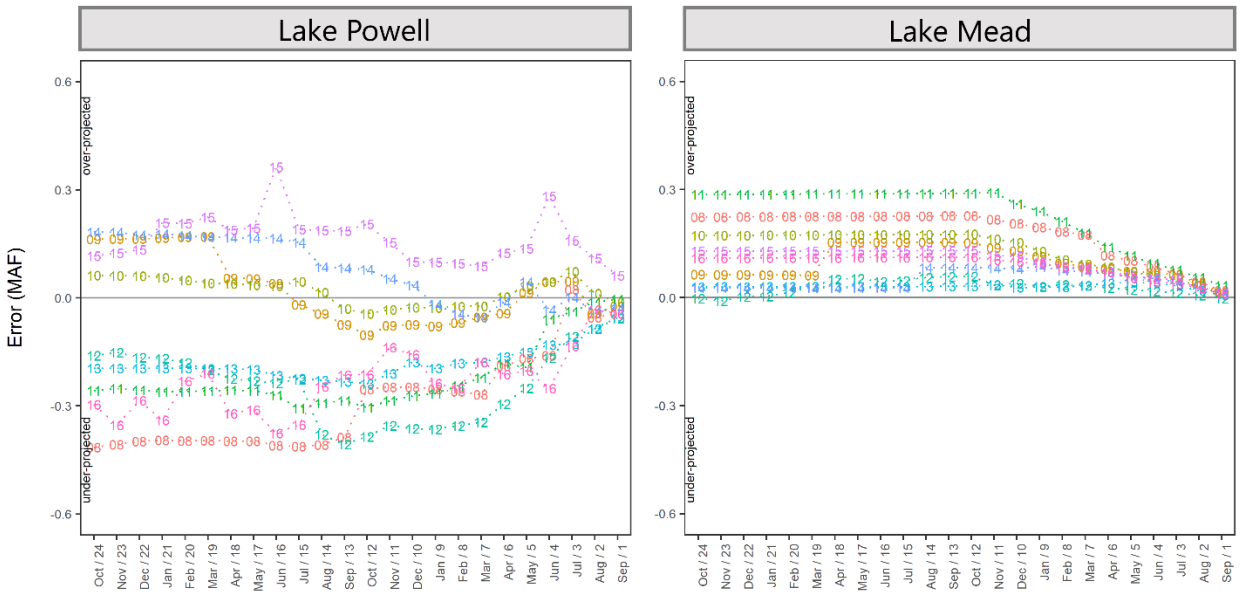

Figure 5-14. MTOM annual water balance error for Lake Powell and Lake Mead from a 24- to 1-month lead (2008-2016).

5.5.2.2 Ensemble Streamflow Forecast Operational Projections

RMSE of Pool Elevations

The RMSE of projected Lakes Powell and Mead EOWY pool elevations for Climatology, ESP, and 4-Basin kNN forecasts are compared in Figure 5-15 for 1982-2016. The streamflow forecasts are compared to the 'historical streamflow projected' pool elevations.

The streamflow forecasts have a wide range of possible flows, especially when forecasting from longer leads. Lake Powell has a larger RMSE 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 in Lake Powell inflow. At all leads, 4-Basin kNN outperforms ESP, while Climatology performs worse than both other forecasts. For Lake Powell, the RMSE for the forecasts are quite large with a median RMSE of 45 ft, 36.8 ft, and 26.6 ft for Climatology, ESP,

and 4-Basin kNN, respectively. The RMSE slowly decreases with lead, especially in January at a 9-month lead when the forecasts gain more skill and sharpness. The RMSE for ESP and 4-Basin kNN is much smaller than Climatology from January through July, to the end of the runoff season. By April at a 6-month lead, the median RMSE for forecasts have decreased to 25.4 ft, 11.9 ft, and 9.4 ft for Climatology, ESP, and 4-Basin kNN, respectively. The RMSE continues to decrease through the end of the WY.

For Lake Mead, RMSE decreases relatively linearly with median RMSE values at a 24-month lead of 30.7 ft, 24.2 ft, and 19.1 ft to a 6-month lead of 13.1 ft, 5.5ft, and 2.9 ft for Climatology, ESP, and 4-Basin kNN, respectively. Climatology has a slightly different trend in reduction in RMSE with not much change in RMSE until the runoff seasons for the two years. This is because Climatology doesn't have skill until the runoff has been observed. By the runoff season of the forecasted year, ESP and 4-Basin kNN have very small errors, since the release from Lake Powell has been set by the operating tier.

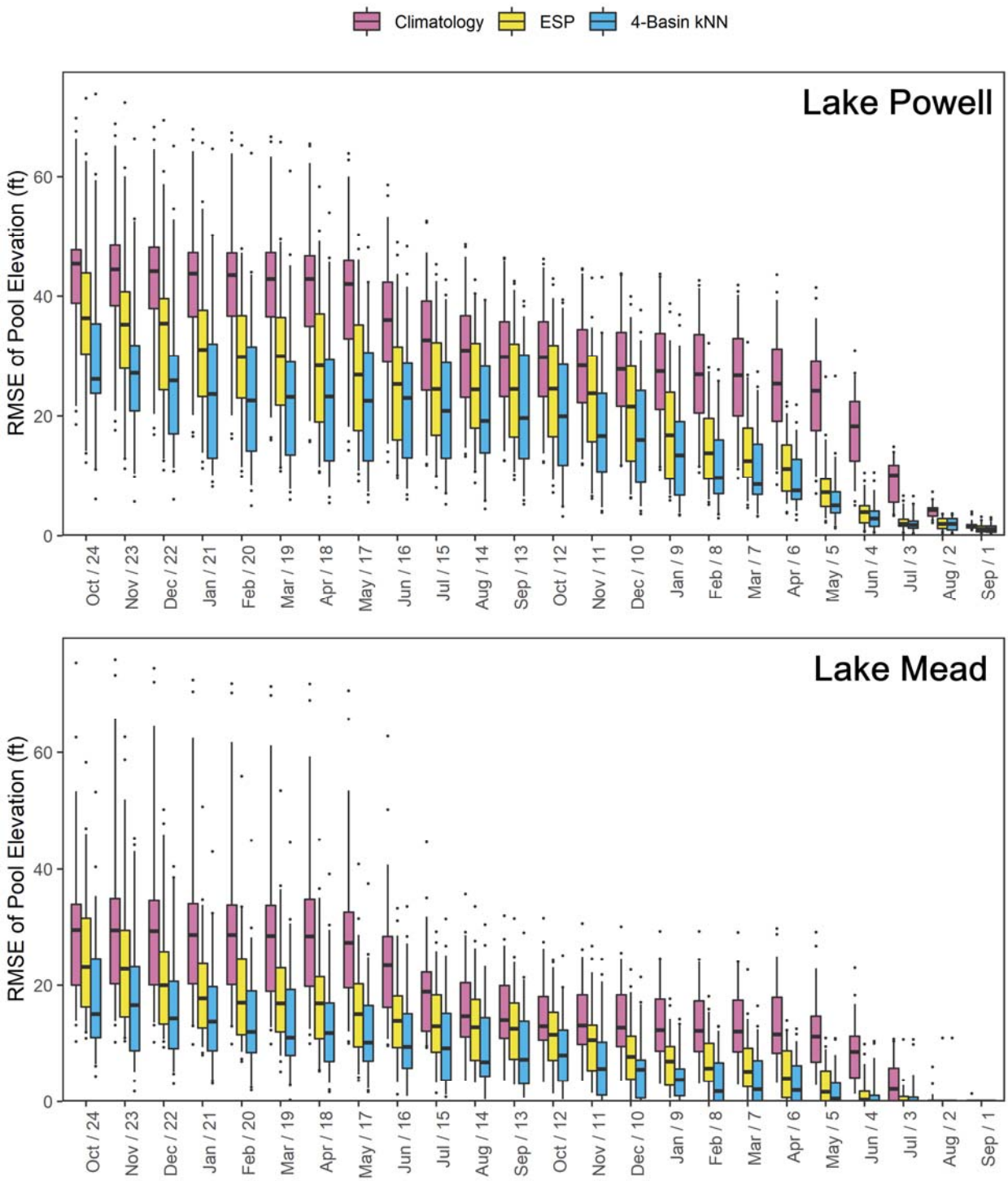

Figure 5-15. RMSE of EOWY pool elevation of Lake Powell and Lake Mead. Climatology, ESP, and 4-Basin kNN are compared to historical streamflow projected pool elevations (1982-2016).

Categorical Scores of Operational Projections

Categorical Scores are used to analyze the projected operating tiers of Lakes Powell and Mead. The categories for this analysis are the operating tiers and releases based on the 2007 Interim Guidelines which are summarized in Table 5-1. Categorical scores are evaluated based on the tier alone, or the combined tier and release.

Table 5-1. Operating tiers and releases used in categorical scores based on the 2007 Interim Guidelines.

Reservoir	Tier	Release
Lake Powell	Equalization	*Annual release > 8.23 maf*
		Annual release = 8.23 maf
	Upper Elevevation 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	*1st level (Mead ≤ 1,075 and ≥ 1,050 ft)*
		2nd level (Mead < 1,050 and ≥ 1,025 ft)
		3rd level (Mead < 1,025 ft)
	Surplus	*Any except Flood Control*
		Flood Control
	Normal	*Normal or ICS Surplus Condition*

Percent Correct and Heidke Skill Score are categorical verification metrics used here to analyze probabilistic operating tier and release projections for Lake Powell and Lake Mead in Table 5-2 and Table 5-3. The Percent Correct is the percent of forecast traces that projected the correct category for operating tier and/or release. The Heidke Skill Score is accuracy of the forecast in predicting the correct tier relative to that of random chance. The metrics are evaluated on two different scales. The

'Tier' results are broader categories and therefore have higher scores than the 'Tier & Release' categories, which require the simulation to correctly determine the release as well as the operating tier. The categorical scores for Climatology, ESP, and 4-Basin kNN streamflow forecasts are compared to historical streamflow projected operations for leads in January, April, and August of the out-year for 1982-2016. See Figure 5-2 for a detailed description of operating tier determination.

The Percent Correct in Table 5-2 shows that Lake Mead performs better than Lake Powell when forecasting the tier and release. This is expected since Lake Mead has fewer tiers and potential releases, and since the pool elevation has smaller errors at longer leads compared to Lake Powell (see Figure 5-15). All forecast projections of Lake Mead perform well. Even climatology performs well at the longest lead in January. For Lake Powell, the tier only determinations perform relatively well, especially by August. The 'Tier & Release' projections are lower. This is expected since it is harder to get these categories both correct when there are a wide variety of different forecasts at longer leads. The Heidke Skill Score in Table 5-3 shows similar results to the Percent Correct, except with lower values since we are comparing the forecasts to random chance.

When comparing the forecasts, 4-Basin kNN does equal to or better than the other forecasts. Climatology performs the worst since it is not an informed forecast. There is less difference between ESP and 4-Basin kNN. The differences between ESP and 4-Basin kNN operational projection performance occurs when projecting the operating tier of Lake Powell, especially when including the release. Specifically, 4-

Basin kNN performs better at Lake Powell when projecting Equalization and Upper Elevation Balancing, especially when the observed tier is an Equalization release equal to 8.23 MAF and an Upper Elevation Balancing release above 8.23 MAF.

Table 5-2. Percent Correct for Climatology, ESP, 4-Basin kNN versus historical streamflow projected operating tiers from the out-year at various months.

Reservoir	Streamflow Forecast	Tier			Tier & Release		
		Jan	*Apr*	*Aug*	*Jan*	*Apr*	*Aug*
Lake Powell	Climatology	68%	69%	83%	54%	55%	67%
	ESP	71%	75%	83%	57%	61%	69%
	4-Basin kNN	71%	77%	86%	60%	63%	71%
Lake Mead	Climatology	94%	95%	100%	94%	95%	100%
	ESP	99%	99%	100%	99%	99%	100%
	4-Basin kNN	100%	100%	100%	100%	100%	100%

Table 5-3. Heidke Skill Score for Climatology, ESP, 4-Basin kNN versus historical streamflow projected operating tiers from the out-year at various months.

Reservoir	Streamflow Forecast	Tier			Tier & Release		
		Jan	*Apr*	*Aug*	*Jan*	*Apr*	*Aug*
Lake Powell	Climatology	0.37	0.41	0.67	0.31	0.35	0.52
	ESP	0.39	0.50	0.68	0.34	0.41	0.55
	4-Basin kNN	0.39	0.54	0.73	0.38	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
	4-Basin kNN	0.99	1.00	1.00	0.99	1.00	1.00

5.6 Discussion & Conclusion

The Colorado Basin Streamflow Testbed provides a protocol for evaluating streamflow forecasts for the hydrologic and operational projection skill for a 2-year period. The testbed provides a framework for analyzing streamflow forecasts through metrics assessing the error, skill, spread, and reliability of the Lake Powell annual

unregulated inflow. Streamflow forecasts are run through MTOM to simulate operational projections with metrics including MTOM projected pool elevation, storage, and operating tiers. The testbed is built to process various streamflow forecasts with a specific protocol that allows for an objective comparison of current operation streamflow forecasting method, ESP, and experimental streamflow forecasting method, 4-Basin kNN.

Three ensemble streamflow forecasts were compared from 1982-2016. At long leads, all forecasts have good resolution, sharpness, and reliability, but lacks discrimination and skill. 4-Basin kNN forecasts are narrower than ESP and Climatology, and may be exhibiting too much sharpness at such a long lead. ESP and 4-Basin kNN outperformed Climatology starting in the fall of the out-year, with skills much better than climatology by April of the forecasted year when there is better information about basin initial conditions such as snowpack. 4-Basin kNN performs slightly better than ESP at most leads through the fall, winter, and spring of the forecasted year. At shorter leads, ESP and 4-Basin kNN are over-confident, too sharp, and has poor resolution and reliability. The forecasts can be too narrow, which sometimes excludes the observed annual streamflows. Overall, 4-Basin kNN seems to slightly outperform ESP in the winter and spring of the forecasted year, though it may not perform as well at longer leads. 4-Basin kNN has information about future climate forecasts that increases the forecast skill.

The historical streamflow projected operations were compared to observed operating tiers for 2008-2016. This comparison illustrates how errors in streamflow

forecasts cause errors in the annual operating tier determination and releases. If better streamflow forecasts were available for the 24MS Most Probable simulations, Reclamation could make better projections of operational projections at longer leads. This may be a difficult task to surmount since deterministic forecasts can only provide so much information about the potential future operations and how certain the forecast may be.

Errors in the MTOM model were analyzed with a mass balance on Lake Powell and Lake Mead. Lake Powell predominantly under-projected storage and Lake Mead over-projected storage. Errors are likely due to incorrect modeling assumptions or parameterization that include evaporation, precipitation, bank storage, unaccounted side inflows into either reservoir, or adaptive management in Upper Basin reservoirs. More work should be done to assess the sources of these modeling errors.

Operational projection metrics were used to evaluate how streamflow forecasts affect operating tier projections. Pool elevations showed larger errors at longer leads that decreased, especially by April for both Lakes Powell and Mead. 4-Basin kNN outperformed ESP for all leads, illustrating the even slightly better performing streamflow forecasts can translate into reduced error in operational projections. The categorical scores showed that 4-Basin kNN performed better than ESP, though in some leads the performance was only slightly better.

The testbed results for streamflow forecasts and operational projections in the CRB illustrates that improved streamflow forecasts can enhance operational projections. The experimental forecast, 4-Basin kNN, exemplifies the use of climate

informed streamflow forecasts in an operational projection model. This experiment shows that climate forecasts can be useful by nudging streamflow forecasts in the correct direction. The resulting operational projections were also shown to be more accurate. In the future, other experimental streamflow forecasts can be run through the testbed to assess the hydrologic skill and how the skill translates into operational projections in MTOM.

6 CHAPTER VI: Conclusions

6.1 Summary & Discussion

This dissertation presented S2S climate forecast products on a watershed scale and explored applications of climate forecasts use for water management in the Colorado River Basin. In many water sectors, climate forecast products are underutilized due to a perceived poor skill, difficulties injecting the data, mismatched spatial or temporal resolution, and institutional reasons. To overcome some of these hurdles, this work transitioned and post-processed climate forecasts on a watershed scale. In Chapter 2, raw CFSv2 and NMME forecasts were aggregated to a USGS HUC-4 watershed scale over the CONUS domain and bias corrected through quantile mapping. These S2S climate forecasts were made available in real-time on the S2S Climate Outlooks for Watersheds web-based tool that displays forecasts and baseline skill assessments of bi-weekly CFSv2 and monthly NMME temperature and precipitation forecasts. In Chapter 3, we explored the potential of the post-processing method PLSR for improving the raw CFSv2 bi-weekly forecasts. The PLSR method utilizes climate and land surface information from concurrent CFSv2 field components. We found PLSR resulted in marginal to moderate increases in skill for 2-3 and 3-4 week precipitation forecasts and 3-4 week temperature forecasts for some watersheds.

In Chapters 4 and 5, we explored watershed scale S2S climate forecast applications through streamflow forecasting in the Upper Colorado River Basin. In the Colorado River Basin, management decisions would benefit from streamflow

forecasts with improved skill to provide assessments of future basin risk, such as the probability of water shortage or flood control. An experimental streamflow forecasting technique, 4-Basin kNN, was compared to the operational forecasting method, ESP, to determine how improved streamflow forecasts would affect operational projections of future basin conditions. The 4-Basin kNN trace weighting scheme was found to increase streamflow forecast accuracy and skill at most leads through weighting ESP traces in four sub-basins in the Upper Basin using 1-month and 3-month NMME climate forecasts, and the proceeding 3-month average observed flow. These streamflow forecasts were then run through the Colorado Basin Streamflow forecast testbed that uses Reclamation's Mid-term Operations Model, MTOM, to evaluate streamflow forecasting on a water year scale, along with operational projections at Lakes Powell and Mead. The 4-Basin kNN method was found to outperform ESP at leads in the winter and spring when comparing the accuracy of projecting pool elevation, operating tiers, and releases. The increases in streamflow forecasting skill translated to improved accuracy of operational projections, though nonlinearly.

The climate forecast applications in the management of the Colorado River Basin illustrated the potential of climate forecast use to inform operational projections such as reservoir operations and projections of future basin conditions. The climate forecasts used in the 4-Basin kNN method are available in real-time, allowing for this experimental method to be operationalized. Since NMME and CFSv2 climate forecasts are already translated to a watershed scale at a useful

temporal time frame, they can be more easily used by water managers who lack the time or tools to process gridded climate forecasts for their specific watershed needs. The post-processed CFSv2 or NMME climate forecasts could be used in other basins for water management applications such as to inform timing of reservoir releases during runoff season based on temperature forecast or to provide projections of precipitation anomalies to inform the possibility of flood control operations at longer leads.

6.2 Future Directions

There are several extensions of possible work based on the techniques and findings presented here. In the future, we could add additional utility to the S2S Climate Outlooks for Watersheds web-based tool presented in Chapter 2. Users have shown interest in time series plots of forecasts and a comparison to observations for individual watersheds. We could also solicit additional feedback from water managers through formal surveys.

In Chapter 3, we assessed the PLSR post-processing technique with baseline predictors over the CONUS domain. Our selected technique showed the potential of post-processing climate forecasts to improve forecast skill on the S2S timescale. Since this study took a conservative stance to predictor selection with a single set of predictors applied over the entire CONUS domain, other studies could select other predictors or predictor domains tailored to a specific watershed of interest. Different post-processing techniques could be explored such as recently popularized

techniques in machine learning that have the potential to capture nonlinear relationships between variables.

The post-ESP method in Chapter 4 showed the opportunity for weighting ESP forecast members with climate forecast information. Other trace weighting techniques have been explored and compared in other studies of basins in the US and could be applied to ESP in the Upper Colorado River Basin. Since ESP produced unreliable forecasts at shorter leads, calibration methods could be explored to improve forecast reliability and skill. The Colorado Basin Streamflow Forecast Testbed presented in Chapter 5 is intended for future use and developments beyond this study. When other streamflow forecasts become available, especially if they can show skill improvements into the out-year, they could be run through the testbed.

7 References

Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(1), 97–106. https://doi.org/10.1002/wics.51

Abudu, S., King J. Phillip, & Pagano Thomas C. (2010). Application of Partial Least-Squares Regression in Seasonal Streamflow Forecasting. *Journal of Hydrologic Engineering*, *15*(8), 612–623. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000216

Baker, S. A., Wood, A. W., & Rajagopalan, B. (2019). Developing Subseasonal to Seasonal Climate Forecast Products for Hydrology and Water Management. *JAWRA Journal of the American Water Resources Association*, *0*(0). https://doi.org/10.1111/1752-1688.12746

Barnston, A. G., & Mason, S. J. (2011). Evaluation of IRI's Seasonal Climate Forecasts for the Extreme 15% Tails. *Weather and Forecasting*, *26*(4), 545–554. https://doi.org/10.1175/WAF-D-10-05009.1

Bazile, R., Boucher, M.-A., Perreault, L., & Leconte, R. (2017). Verification of ECMWF System 4 for seasonal hydrological forecasting in a northern climate. *Hydrology and Earth System Sciences*, *21*(11), 5747–5762. https://doi.org/10.5194/hess-21-5747-2017

Becker, E., den Dool, H. van, & Zhang, Q. (2014). Predictability and Forecast Skill in NMME. *Journal of Climate*, *27*(15), 5891–5906. https://doi.org/10.1175/JCLI-D-13-00597.1

Becker, E., & van den Dool, H. (2016). Probabilistic Seasonal Forecasts in the North American Multimodel Ensemble: A Baseline Skill Assessment. *Journal of Climate*, *29*(8), 3015–3026. https://doi.org/10.1175/JCLI-D-14-00862.1

Beckers, J. V. L., Weerts, A. H., Tijdeman, E., & Welles, E. (2016). ENSO-conditioned weather resampling method for seasonal ensemble streamflow prediction. *Hydrology and Earth System Sciences*, *20*(8), 3277–3287. https://doi.org/10.5194/hess-20-3277-2016

Black, J., Johnson, N. C., Baxter, S., Feldstein, S. B., Harnos, D. S., & L'Heureux, M. L. (2017). The Predictors and Forecast Skill of Northern Hemisphere Teleconnection Patterns for Lead Times of 3–4 Weeks. *Monthly Weather Review*, *145*(7), 2855–2877. https://doi.org/10.1175/MWR-D-16-0394.1

Bolinger, R. A., Gronewold, A. D., Kompoltowicz, K., & Fry, L. M. (2017). Application of the NMME in the Development of a New Regional Seasonal Climate Forecast Tool. *Bulletin of the American Meteorological Society*, *98*(3), 555–564. https://doi.org/10.1175/BAMS-D-15-00107.1

Bolson, J., Martinez, C., Breuer, N., Srivastava, P., & Knox, P. (2013). Climate information use among southeast US water managers: Beyond barriers and toward opportunities. *Regional Environmental Change*, *13*(1), 141–151. https://doi.org/10.1007/s10113-013-0463-1

Bracken, C. W. (2011). *Seasonal to Inter-Annual Streamflow Simulation and Forecasting on the Upper Colorado River Basin and Implications for Water Resources Management* (University of Colorado Boulder). Retrieved from https://www.colorado.edu/cadswes/sites/default/files/attached-files/bracken-ms_thesis-2011.pdf

Bröcker, J. (2015). Resolution and discrimination–two sides of the same coin. *Quarterly Journal of the Royal Meteorological Society*, *141*(689), 1277–1282. https://doi.org/10.1002/qj.2434

Bureau of Reclamation. (2015, May). *Colorado River Basin Mid-Term Probabilistic Operations Model (MTOM) Overview and Description*.

Burnash, R. J. C., Ferral, R. L., & McGuire, R. A. (1973). *A generalized streamflow simulation system, conceptual modeling for digital computers*. Sacramento, CA: Joint Federal State River Forecasts Center.

Callahan, B., Miles, E., & Fluharty, D. (1999). Policy implications of climate forecasts for water resources management in the Pacific Northwest. *Policy Sciences*, *32*(3), 269–293. https://doi.org/10.1023/A:1004604805647

Cannon, A. J., Sobie, S. R., & Murdock, T. Q. (2015). Bias Correction of GCM Precipitation by Quantile Mapping: How Well Do Methods Preserve Changes in Quantiles and Extremes? *Journal of Climate*, *28*(17), 6938–6959. https://doi.org/10.1175/JCLI-D-14-00754.1

Clark, M. P., Serreze, M. C., & McCabe, G. J. (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), 741–757. https://doi.org/10.1029/2000WR900305

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*. University of Colorado Boulder.

Day, G. (1985). Extended Streamflow Forecasting Using NWSRFS. *Journal of Water Resources Planning and Management*, *111*(2), 157–170. https://doi.org/10.1061/(ASCE)0733-9496(1985)111:2(157)

de Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, *18*(3), 251–263. https://doi.org/10.1016/0169-7439(93)85002-X

DelSole, T., & Banerjee, A. (2016). Statistical Seasonal Prediction Based on Regularized Regression. *Journal of Climate*, *30*(4), 1345–1361. https://doi.org/10.1175/JCLI-D-16-0249.1

DelSole, T., Trenary, L., Tippett, M. K., & Pegion, K. (2017). Predictability of Week-3–4 Average Temperature and Precipitation over the Contiguous United States. *Journal of Climate*, *30*(10), 3499–3512. https://doi.org/10.1175/JCLI-D-16-0567.1

Dilling, L., & Lemos, M. C. (2011). Creating usable science: Opportunities and constraints for climate knowledge use and their implications for science policy. *Global Environmental Change*, *21*(2), 680–689. https://doi.org/10.1016/j.gloenvcha.2010.11.006

Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P., & Rodrigues, L. R. L. (2013). Seasonal climate predictability and forecasting: Status and prospects. *Wiley Interdisciplinary Reviews: Climate Change*, *4*(4), 245–268. https://doi.org/10.1002/wcc.217

Doblas-Reyes, F. J., Hagedorn, R., & Palmer, T. N. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting – II. Calibration and combination. *Tellus A*, *57*(3), 234–252. https://doi.org/10.1111/j.1600-0870.2005.00104.x

Feldman, D. L., & Ingram, H. M. (2009). Making Science Useful to Decision Makers: Climate Forecasts, Water Management, and Knowledge Networks. *Weather, Climate, and Society*, *1*(1), 9–21. https://doi.org/10.1175/2009WCAS1007.1

Franz, K. J., Hartmann, H. C., Sorooshian, S., & Bales, R. (2003). Verification of National Weather Service Ensemble Streamflow Predictions for Water Supply Forecasting in the Colorado River Basin. *Journal of Hydrometeorology*, *4*(6), 1105–1118. https://doi.org/10.1175/1525-7541(2003)004<1105:VONWSE>2.0.CO;2

Garen, D. (1992). Improved Techniques in Regression-Based Streamflow Volume Forecasting. *Journal of Water Resources Planning and Management*, *118*(6), 654–670. https://doi.org/10.1061/(ASCE)0733-9496(1992)118:6(654)

Glahn, H. R., & Lowry, D. A. (1972). The Use of Model Output Statistics (MOS) in Objective Weather Forecasting. *Journal of Applied Meteorology*, *11*(8), 1203–1211. https://doi.org/10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2

Gobena, A. K., & Gan, T. Y. (2010). Incorporation of seasonal climate forecasts in the ensemble streamflow prediction system. *Journal of Hydrology*, *385*(1), 336–352. https://doi.org/10.1016/j.jhydrol.2010.03.002

Grantz, K., Rajagopalan, B., Clark, M., & Zagona, E. (2005). A technique for incorporating large-scale climate information in basin-scale ensemble streamflow forecasts. *Water Resources Research*, *41*(W10410). https://doi.org/10.1029/2004WR003467

Gutmann, E., Pruitt, T., Clark, M. P., Brekke, L., Arnold, J. R., Raff, D. A., & Rasmussen, R. M. (2014). An intercomparison of statistical downscaling methods used for water resource assessments in the United States. *Water Resources Research*, *50*(9), 7167–7186. https://doi.org/10.1002/2014WR015559

Hagedorn, R., Doblas-Reyes, F. J., & Palmer, T. N. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept. *Tellus A*, *57*(3), 219–233. https://doi.org/10.1111/j.1600-0870.2005.00103.x

Hamill, T. M. (1997). Reliability Diagrams for Multicategory Probabilistic Forecasts. *Weather and Forecasting*, *12*(4), 736–741. https://doi.org/10.1175/1520-0434(1997)012<0736:RDFMPF>2.0.CO;2

Hamill, T. M., & Whitaker, J. S. (2006). Probabilistic Quantitative Precipitation Forecasts Based on Reforecast Analogs: Theory and Application. *Monthly Weather Review*, *134*(11), 3209–3229. https://doi.org/10.1175/MWR3237.1

Hamilton, E., Eade, R., Graham, R. J., Scaife, A. A., Smith, D. M., Maidens, A., & MacLachlan, C. (2012). Forecasting the number of extreme daily events on seasonal timescales. *Journal of Geophysical Research: Atmospheres*, *117*(D3), D03114. https://doi.org/10.1029/2011JD016541

Hartmann, H. C., Pagano, T. C., Sorooshian, S., & Bales, R. (2002). Confidence Builders: Evaluating Seasonal Climate Forecasts from User Perspectives. *Bulletin of the American Meteorological Society*, *83*(5), 683–698. https://doi.org/10.1175/1520-0477(2002)083<0683:CBESCF>2.3.CO;2

Hersbach, H. (2000). *Decomposition of the Continuous Ranked Probability Score fore Ensemble Prediction Systems*.

Jones, N. (2017). How machine learning could help to improve climate forecasts. *Nature*, *548*(7668), 379–380. https://doi.org/10.1038/548379a

Kirchhoff, C. J., Lemos, M. C., & Engle, N. L. (2013). What influences climate information use in water management? The role of boundary organizations and governance regimes in Brazil and the U.S. *Environmental Science & Policy*, *26*, 6–18. https://doi.org/10.1016/j.envsci.2012.07.001

Kirtman, B. P., Min, D., Infanti, J. M., Kinter, J. L., Paolino, D. A., Zhang, Q., … Wood, E. F. (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

Koster, R. D., Betts, A. K., Dirmeyer, P. A., Bierkens, M., Bennett, K. E., Déry, S. J., … Yuan, X. (2017). Hydroclimatic variability and predictability: A survey of recent research. *Hydrol. Earth Syst. Sci.*, *21*(7), 3777–3798. https://doi.org/10.5194/hess-21-3777-2017

Koster, Randal D., Mahanama, S. P. P., Livneh, B., Lettenmaier, D. P., & Reichle, R. H. (2010). Skill in streamflow forecasts derived from large-scale estimates of soil moisture and snow. *Nature Geoscience*, *3*(9), 613–616. https://doi.org/10.1038/ngeo944

Krakauer, N. Y. (2017). Temperature trends and prediction skill in NMME seasonal forecasts. *Climate Dynamics*, 1–13. https://doi.org/10.1007/s00382-017-3657-2

Lawrence, D. M., Oleson, K. W., Flanner, M. G., Fletcher, C. G., Lawrence, P. J., Levis, S., … Bonan, G. B. (2012). The CCSM4 Land Simulation, 1850–2005: Assessment of Surface Climate and New Capabilities. *Journal of Climate*, *25*(7), 2240–2260. https://doi.org/10.1175/JCLI-D-11-00103.1

Lehner, F., Wahl, E. R., Wood, A. W., Blatchford, D. B., & Llewellyn, D. (2017). Assessing recent declines in Upper Rio Grande runoff efficiency from a paleoclimate perspective. *Geophysical Research Letters*, *44*(9), 4124–4133. https://doi.org/10.1002/2017GL073253

Lehner, F., Wood, A. W., Llewellyn, D., Blatchford, D. B., Goodbody, A. G., & Pappenberger, F. (2017). Mitigating the Impacts of Climate Nonstationarity on Seasonal Streamflow Predictability in the U.S. Southwest. *Geophysical Research Letters*, 2017GL076043. https://doi.org/10.1002/2017GL076043

Li, H., Luo, L., Wood, E. F., & Schaake, J. (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

Lorenz, D. J., & Hartmann, D. L. (2006). The Effect of the MJO on the North American Monsoon. *Journal of Climate*, *19*(3), 333–343. https://doi.org/10.1175/JCLI3684.1

Luo, L., & Wood, E. F. (2008). Use of Bayesian Merging Techniques in a Multimodel Seasonal Hydrologic Ensemble Prediction System for the Eastern United States. *Journal of Hydrometeorology*, *9*(5), 866–884. https://doi.org/10.1175/2008JHM980.1

Madadgar, S., AghaKouchak, A., Shukla, S., Wood, A. W., Cheng, L., Hsu, K.-L., & Svoboda, M. (2016). A hybrid statistical-dynamical framework for meteorological drought prediction: Application to the southwestern United States. *Water Resources Research*, *52*(7), 5095–5110. https://doi.org/10.1002/2015WR018547

Maraun, D., Wetterhall, F., Ireson, A. M., Chandler, R. E., Kendon, E. J., Widmann, M., … Thiele-Eich, I. (2010). Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, *48*(3). https://doi.org/10.1029/2009RG000314

Maraun, Douglas. (2013). Bias Correction, Quantile Mapping, and Downscaling: Revisiting the Inflation Issue. *Journal of Climate*, *26*(6), 2137–2143. https://doi.org/10.1175/JCLI-D-12-00821.1

Mariotti, A., Ruti, P. M., & Rixen, M. (2018). Progress in subseasonal to seasonal prediction through a joint weather and climate community effort. *Npj Climate and Atmospheric Science*, *1*(1), 4. https://doi.org/10.1038/s41612-018-0014-z

McIntosh, P. C., Ash, A. J., & Smith, M. S. (2005). From Oceans to Farms: The Value of a Novel Statistical Climate Forecast for Agricultural Management. *Journal of Climate*, *18*(20), 4287–4302. https://doi.org/10.1175/JCLI3515.1

Mendoza, P. A., Wood, A. W., Clark, E., Rothwell, E., Clark, M. P., Nijssen, B., … Arnold, J. R. (2017). An intercomparison of approaches for improving operational seasonal streamflow forecasts. *Hydrol. Earth Syst. Sci.*, *21*(7), 3915–3935. https://doi.org/10.5194/hess-21-3915-2017

Mendoza, Pablo A., Rajagopalan, B., Clark, M. P., Cortés, G., & McPhee, J. (2014). A robust multimodel framework for ensemble seasonal hydroclimatic forecasts. *Water Resources Research*, *50*(7), 6030–6052. https://doi.org/10.1002/2014WR015426

Merryfield, W. J., Lee, W.-S., Boer, G. J., Kharin, V. V., Scinocca, J. F., Flato, G. M., … Polavarapu, S. (2013). The Canadian Seasonal to Interannual Prediction

System. Part I: Models and Initialization. *Monthly Weather Review*, *141*(8), 2910–2945. https://doi.org/10.1175/MWR-D-12-00216.1

Miller, W. P., Butler, R. A., Piechota, T., Prairie, J., Grantz, K., & DeRosa, G. (2012). Water Management Decisions Using Multiple Hydrologic Models within the San Juan River Basin under Changing Climate Conditions. *Journal of Water Resources Planning and Management*, *138*(5), 412–420. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000237

Mo, K. C., & Lettenmaier, D. P. (2014). Hydrologic Prediction over the Conterminous United States Using the National Multi-Model Ensemble. *Journal of Hydrometeorology*, *15*(4), 1457–1472. https://doi.org/10.1175/JHM-D-13-0197.1

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

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

Moradkhani, H., & Meier, M. (2010). Long-Lead Water Supply Forecast Using Large-Scale Climate Predictors and Independent Component Analysis. *Journal of Hydrologic Engineering*, *15*(10), 744–762. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000246

Murphy, A. H., & Epstein, E. S. (1989). Skill Scores and Correlation Coefficients in Model Verification. *Monthly Weather Review*, *117*(3), 572–582. https://doi.org/10.1175/1520-0493(1989)117<0572:SSACCI>2.0.CO;2

Najafi, M. R., Moradkhani, H., & Piechota, T. C. (2012). Ensemble Streamflow Prediction: Climate signal weighting methods vs. Climate Forecast System Reanalysis. *Journal of Hydrology*, *442–443*, 105–116. https://doi.org/10.1016/j.jhydrol.2012.04.003

Ning, L., Riddle, E. E., & Bradley, R. S. (2015). Projected Changes in Climate Extremes over the Northeastern United States. *Journal of Climate*, *28*(8), 3289–3310. https://doi.org/10.1175/JCLI-D-14-00150.1

O'Lenic, E. A., Unger, D. A., Halpert, M. S., & Pelman, K. S. (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., Hartmann, H. C., & Sorooshian, S. (2001). Using Climate Forecasts for Water Management: Arizona and the 1997–1998 El Niño. *JAWRA Journal of the American Water Resources Association*, *37*(5), 1139–1153. https://doi.org/10.1111/j.1752-1688.2001.tb03628.x

Pagano, T. C., Robertson, A. W., Werner, K., & Tama-Sweet, R. (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

Pagano, T. C., Wood, A. W., Ramos, M.-H., Cloke, H. L., Pappenberger, F., Clark, M. P., … Verkade, J. S. (2014). Challenges of Operational River Forecasting. *Journal of Hydrometeorology*, *15*(4), 1692–1707. https://doi.org/10.1175/JHM-D-13-0188.1

Panofsky, H. A., & Brier, G. W. (1968). *Some applications of statistics to meteorology*. Retrieved from https://trove.nla.gov.au/version/22778128

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

Quan, X., Hoerling, M., Whitaker, J., Bates, G., & Xu, T. (2006). Diagnosing Sources of U.S. Seasonal Forecast Skill. *Journal of Climate*, *19*(13), 3279–3293. https://doi.org/10.1175/JCLI3789.1

Raff, D. A., Brekke, L., Werner, K., Wood, A., & White, K. (2013). *Short-Term Water Management Decisions -User Needs for Improved Climate, Weather, and Hydrologic Information* (Technical Report No. CWTS 2013-1; p. 256). Retrieved from U.S. Army Corps of Engineers; Bureau of Reclamation; National Oceanic and Atmospheric Administration website: http://www.ccawwg.us/docs/Short-Term_Water_Management_Decisions_Final_3_Jan_2013.pdf

Rayner, S., Lach, D., & Ingram, H. (2005). Weather Forecasts are for Wimps: Why Water Resource Managers Do Not Use Climate Forecasts. *Climatic Change*, *69*(2–3), 197–227. https://doi.org/10.1007/s10584-005-3148-z

Regonda, S. K., Rajagopalan, B., Clark, M., & Zagona, E. (2006). A multimodel ensemble forecast framework: Application to spring seasonal flows in the Gunnison River. *Water Resources Research*, *42*(W09404). https://doi.org/10.1029/2005WR004653

Regonda, S., Zagona, E., & Rajagopalan, B. (2011). Prototype Decision Support System for Operations on the Gunnison Basin with Improved Forecasts. *Journal of Water Resources Planning and Management*, *137*(5), 428–438. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000133

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., … Becker, E. (2014). The NCEP Climate Forecast System Version 2. *Journal of Climate*, *27*(6), 2185–2208. https://doi.org/10.1175/JCLI-D-12-00823.1

Sankarasubramanian, A., Lall, U., Devineni, N., & Espinueva, S. (2009). The Role of Monthly Updated Climate Forecasts in Improving Intraseasonal Water Allocation. *Journal of Applied Meteorology and Climatology*, *48*(7), 1464–1482. https://doi.org/10.1175/2009JAMC2122.1

Santosa, F., & Symes, W. (1986). Linear Inversion of Band-Limited Reflection Seismograms. *SIAM Journal on Scientific and Statistical Computing*, *7*(4), 1307–1330. https://doi.org/10.1137/0907087

Scaife, A. A., Arribas, A., Blockley, E., Brookshaw, A., Clark, R. T., Dunstone, N., … Williams, A. (2014). Skillful long-range prediction of European and North American winters. *Geophysical Research Letters*, *41*(7), 2514–2519. https://doi.org/10.1002/2014GL059637

Schepen, A., Wang, Q. J., & Robertson, D. E. (2014). Seasonal Forecasts of Australian Rainfall through Calibration and Bridging of Coupled GCM Outputs. *Monthly Weather Review*, *142*(5), 1758–1770. https://doi.org/10.1175/MWR-D-13-00248.1

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Shukla, S., & Lettenmaier, D. P. (2011). Seasonal hydrologic prediction in the United States: Understanding the role of initial hydrologic conditions and seasonal climate forecast skill. *Hydrol. Earth Syst. Sci.*, *15*(11), 3529–3538. https://doi.org/10.5194/hess-15-3529-2011

Slater, L. J., Villarini, G., & Bradley, A. A. (2016). Evaluation of the skill of North-American Multi-Model Ensemble (NMME) Global Climate Models in predicting average and extreme precipitation and temperature over the continental USA. *Climate Dynamics*, 1–16. https://doi.org/10.1007/s00382-016-3286-1

Slater, L. J., Villarini, G., & Bradley, A. A. (2017). Weighting of NMME temperature and precipitation forecasts across Europe. *Journal of Hydrology*, *552*, 646–659. https://doi.org/10.1016/j.jhydrol.2017.07.029

Slater, L. J., Villarini, G., Bradley, A. A., & Vecchi, G. A. (2017). A dynamical statistical framework for seasonal streamflow forecasting in an agricultural watershed. *Climate Dynamics*, 1–17. https://doi.org/10.1007/s00382-017-3794-7

Smoliak, B. V., Wallace, J. M., Lin, P., & Fu, Q. (2015). Dynamical Adjustment of the Northern Hemisphere Surface Air Temperature Field: Methodology and Application to Observations. *Journal of Climate*, *28*(4), 1613–1629. https://doi.org/10.1175/JCLI-D-14-00111.1

Smoliak, B. V., Wallace, J. M., Stoelinga, M. T., & Mitchell, T. P. (2010). Application of partial least squares regression to the diagnosis of year-to-year variations in Pacific Northwest snowpack and Atlantic hurricanes. *Geophysical Research Letters*, *37*(3). https://doi.org/10.1029/2009GL041478

Tian, D., Martinez, C. J., Graham, W. D., & Hwang, S. (2014). Statistical Downscaling Multimodel Forecasts for Seasonal Precipitation and Surface Temperature over the Southeastern United States. *Journal of Climate*, *27*(22), 8384–8411. https://doi.org/10.1175/JCLI-D-13-00481.1

Tian, D., Wood, E. F., & Yuan, X. (2017). CFSv2-based sub-seasonal precipitation and temperature forecast skill over the contiguous United States. *Hydrology and Earth System Sciences*, *21*(3), 1477–1490. http://dx.doi.org/10.5194/hess-21-1477-2017

Tootle, G. A., Singh, A. K., Piechota, T. C., & Farnham, I. (2007). Long Lead-Time Forecasting of U.S. Streamflow Using Partial Least Squares Regression. *Journal of Hydrologic Engineering*, *12*(5), 442–451. https://doi.org/10.1061/(ASCE)1084-0699(2007)12:5(442)

U.S. Bureau of Reclamation. (2019). Teams complete Bureau of Reclamation's Sub-Seasonal Climate Forecast Rodeo — outperforming the baseline forecasts. Retrieved May 7, 2019, from https://www.usbr.gov/newsroom/newsrelease/detail.cfm?RecordID=64969

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*. United States Bureau of Reclamation.

van Dijk, A. I. J. M., Peña-Arancibia, J. L., Wood, E. F., Sheffield, J., & Beck, H. E. (2013). Global analysis of seasonal streamflow predictability using an ensemble prediction system and observations from 6192 small catchments worldwide. *Water Resources Research*, *49*(5), 2729–2746. https://doi.org/10.1002/wrcr.20251

Vecchi, G. A., Delworth, T., Gudgel, R., Kapnick, S., Rosati, A., Wittenberg, A. T., … Zhang, S. (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., Rienecker, M., Kovach, R., & Keppenne, C. (2012). *The GEOS-iODAS: description and evaluation. In: Technical Report Series on Global Modeling and Data Assimilation, Volume 30* (No. NASA/TM-2012-104606/Vol 30). Retrieved from http://gmao.gsfc.nasa.gov/pubs/docs/Vernieres589.pdf

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., … Zhang, L. (2016). The Subseasonal to Seasonal (S2S) Prediction Project Database. *Bulletin of the American Meteorological Society*, *98*(1), 163–173. https://doi.org/10.1175/BAMS-D-16-0017.1

Vitart, Frédéric, & Robertson, A. W. (2018). The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. *Npj Climate and Atmospheric Science*, *1*(1), 3. https://doi.org/10.1038/s41612-018-0013-0

Wanders, N., & Wood, E. F. (2016). Improved sub-seasonal meteorological forecast skill using weighted multi-model ensemble simulations. *Environmental Research Letters*, *11*(9), 094007. https://doi.org/10.1088/1748-9326/11/9/094007

Ward, M. N., & Folland, C. K. (1991). Prediction of seasonal rainfall in the north nordeste of Brazil using eigenvectors of sea-surface temperature. *International Journal of Climatology*, *11*(7), 711–743. https://doi.org/10.1002/joc.3370110703

Werner, K., Brandon, D., Clark, M., & Gangopadhyay, S. (2004). Climate Index Weighting Schemes for NWS ESP-Based Seasonal Volume Forecasts. *Journal of Hydrometeorology*, *5*(6), 1076–1090. https://doi.org/10.1175/JHM-381.1

Werner, K., Brandon, D., Clark, M., & Gangopadhyay, S. (2005). Incorporating Medium-Range Numerical Weather Model Output into the Ensemble Streamflow Prediction System of the National Weather Service. *Journal of Hydrometeorology*, *6*(2), 101–114. https://doi.org/10.1175/JHM411.1

White, C. J., Carlsen, H., Robertson, A. W., Klein, R. J. T., Lazo, J. K., Kumar, A., … Zebiak, S. E. (2017). Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteorological Applications*, *24*(3), 315–325. https://doi.org/10.1002/met.1654

Wilks, D. S. (1995). *Statistical Methods in the Atmospheric Sciences* (2nd ed., Vol. 95). Retrieved from http://www.jstor.org/stable/2669579?origin=crossref

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In *Multivariate analysis* (pp. 391–420). New York, Academic Press: P. R. Krishnajah (Ed.).

Wood, A. W., & Werner, K. (2011). *Development of a Seasonal Climate and Streamflow Forecasting Testbed for the Colorado River Basin*. Presented at the 36th NOAA Annual Climate Diagnostics and Prediction Workshop, Fort Worth, TX.

Wood, Andrew W., Hopson, T., Newman, A., Brekke, L., Arnold, J., & Clark, M. (2016). Quantifying Streamflow Forecast Skill Elasticity to Initial Condition and Climate Prediction Skill. *Journal of Hydrometeorology*, *17*(2), 651–668. https://doi.org/10.1175/JHM-D-14-0213.1

Wood, Andrew W., & Lettenmaier, D. (2006). A test bed for new seasonal hydrologic forecasting approaches in the western US. *BAMS*, *87*(12), 1699–1712. https://doi.org/10.1175/BAMS-87-12-1699

Wood, Andrew W., Leung, L. R., Sridhar, V., & Lettenmaier, D. P. (2004). Hydrologic Implications of Dynamical and Statistical Approaches to Downscaling Climate Model Outputs. *Climatic Change*, *62*(1–3), 189–216. https://doi.org/10.1023/B:CLIM.0000013685.99609.9e

Wood, Andrew W., Maurer, E. P., Kumar, A., & Lettenmaier, D. P. (2002). Long-range experimental hydrologic forecasting for the eastern United States. *Journal of Geophysical Research: Atmospheres*, *107*(D20), 4429. https://doi.org/10.1029/2001JD000659

Wood, Andrew W., & Schaake, J. C. (2008). Correcting Errors in Streamflow Forecast Ensemble Mean and Spread. *Journal of Hydrometeorology*, *9*(1), 132–148. https://doi.org/10.1175/2007JHM862.1

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., … Mocko, D. (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

Xing, W., Wang, B., & Yim, S.-Y. (2016). Long-Lead Seasonal Prediction of China Summer Rainfall Using an EOF–PLS Regression-Based Methodology.

Journal of Climate, *29*(5), 1783–1796. https://doi.org/10.1175/JCLI-D-15-0016.1

Yoon, J.-H., Mo, K., & Wood, E. F. (2011). Dynamic-Model-Based Seasonal Prediction of Meteorological Drought over the Contiguous United States. *Journal of Hydrometeorology*, *13*(2), 463–482. https://doi.org/10.1175/JHM-D-11-038.1

Yuan, X., & Wood, E. F. (2012). Downscaling precipitation or bias-correcting streamflow? Some implications for coupled general circulation model (CGCM)-based ensemble seasonal hydrologic forecast. *Water Resources Research*, *48*(12), W12519. https://doi.org/10.1029/2012WR012256

Yuan, X., Wood, E. F., & Liang, M. (2014). Integrating weather and climate prediction: Toward seamless hydrologic forecasting. *Geophysical Research Letters*, *41*(16), 2014GL061076. https://doi.org/10.1002/2014GL061076

Yuan, X., Wood, E. F., Luo, L., & Pan, M. (2011). A first look at Climate Forecast System version 2 (CFSv2) for hydrological seasonal prediction. *Geophysical Research Letters*, *38*(13), L13402. https://doi.org/10.1029/2011GL047792

Yuan, X., Wood, E. F., Roundy, J. K., & Pan, M. (2013). CFSv2-Based Seasonal Hydroclimatic Forecasts over the Conterminous United States. *Journal of Climate*, *26*(13), 4828–4847. https://doi.org/10.1175/JCLI-D-12-00683.1

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

Zhang, S., Harrison, M. J., Rosati, A., & Wittenberg, A. (2007). System Design and Evaluation of Coupled Ensemble Data Assimilation for Global Oceanic Climate Studies. *Monthly Weather Review*, *135*(10), 3541–3564. https://doi.org/10.1175/MWR3466.1

Zhao, T., Bennett, J. C., Wang, Q. J., Schepen, A., Wood, A. W., Robertson, D. E., & Ramos, M.-H. (2017). How Suitable is Quantile Mapping For Postprocessing GCM Precipitation Forecasts? *Journal of Climate*, *30*(9), 3185–3196. https://doi.org/10.1175/JCLI-D-16-0652.1

8 Appendices

8.1 Appendix 1: Chapter 2 Quantile Mapping Supplemental Discussion

In addition to the issues with capturing extreme events, QM can alter the modeled covariance of temperature and precipitation by QM treating them independently. In downscaling of daily weather data, it is common (and important) to preserve interrelationships between precipitation, temperature, and other fields because there are strong observable relationships linked by synoptic atmospheric dynamics. For instance, wet/precipitating days typically have a compressed temperature range versus clear days. At the sub-seasonal timescale, this covariance is typically weaker. We nonetheless assess the impact of QM on cross-correlations between precipitation and temperature for sub-seasonal bi-weekly CFSv2 predictands for all of the US HUC4s, in comparison to observations from NLDAS. We find that for QM the impact varies by season and lead time. Figure 8-1 shows these cross-correlations for NLDAS, and CFSv2 forecasts before and after QM, with the HUCs in each subplot sorted from low to high values for observed correlation (with samples sizes for each statistic between X and Y of ~360). QM does not significantly affect cross-correlation for January or July forecasts, but has a larger impact, and one that brings cross-correlations of the CFSv2 forecasts into closer agreement with observations for the April and October forecasts. The disagreement between raw CFSv2 and NLDAS grows slightly with lead time. These results suggest that treating temperature and precipitation independently may be acceptable when using QM at the sub-seasonal

timescale, and may even improve cross-correlations where the model is biased relative to the observations.

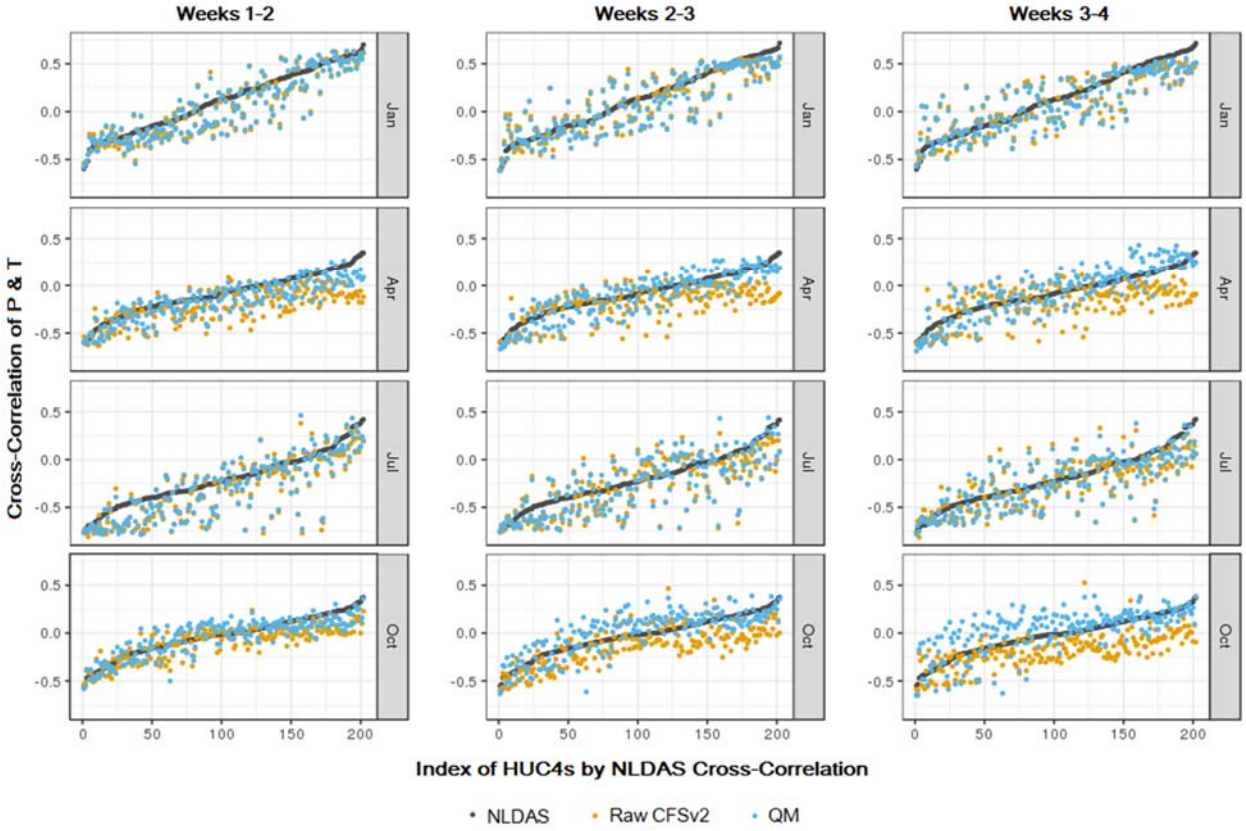

Figure 8-1. Cross-correlation of NLDAS, raw CFSv2, and quantile mapped CFSv2. The bi-weekly cross-correlation of temperature and precipitation for NLDAS, raw CFSv2, and QM CFSv2 for the forecast months January, April, July, and October. The x-axis is the index of the HUC4s in ascending order of NLDAS cross-correlation.

8.2 Appendix 2: Comparison of NMME Forecasts and 4-Basin kNN Performance

We analyze how NMME climate forecast skill translated into streamflow forecast skill. To do this, we looked at a specific example of the December 1983 streamflow forecast of the 1984 April-July runoff volume. Table 8-1 shows the NMME 3-month and 1-month forecast anomalies for temperature (Temp.) and precipitation (Precip.) compared to observations for the four sub-basins. The colors indicate the direction and magnitude of the forecast anomalies and observations, with red showing illustration higher values and green lower values.

Table 8-1: Sub-basin precipitation and temperature NMME forecast versus observations.

Sub-Basin	Fcst Period	Precip. Obs	Precip. Fcst	Temp. Obs	Temp. Fcst
Main Stem	3-month	0.46	0.99	-1.10	-1.38
Gunnison	3-month	1.29	1.04	-0.86	-1.29
Green	3-month	0.71	1.07	-1.69	-1.54
San Juan	3-month	-0.28	0.48	-0.64	-1.03
Main Stem	1-month	2.80	0.58	-0.33	-0.98
Gunnison	1-month	3.22	0.53	-0.12	-0.86
Green	1-month	2.56	0.53	-1.25	-1.06
San Juan	1-month	1.21	0.54	0.56	-0.80

In Table 8-1, the NMME forecasts for both precipitation and temperature perform relatively well for all sub-basins except the San Juan. The climate forecasts are for above average precipitation and below average temperatures for the month and season. This translates into a higher 4-Basin kNN ensemble median compared to ESP as shown in Figure 8-2. This shows that the 4-Basin kNN method is able to extract additional information from the NMME climate forecasts. There remains

issues with the narrowing of the 4-Basin kNN ensemble spread, since in some instances, especially those shown here, the streamflow forecast becomes too narrow. Future work will investigate calibration of the forecast to improve issues with ensemble spread that are exemplified here.

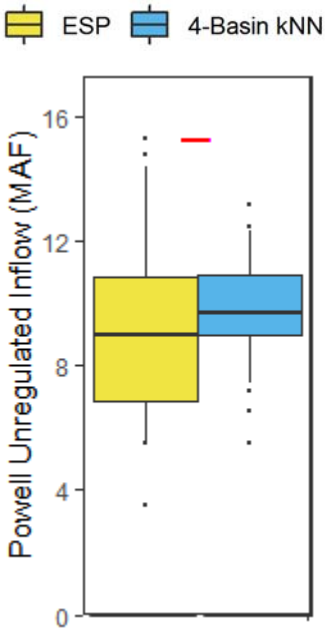

Figure 8-2: Runoff season ensemble forecasts for 1982-2016 compared to observations arranged by ranked observations. ESP and 4-bains kNN forecasts of Lake Powell April-July unregulated inflow are compared at an 8-month lead in December 1983. Observed inflow is represented by a red horizontal line.