A new methodology to produce more skillful United States cool season precipitation forecasts

Matthew Blaise Switanek[1] and Thomas Hamill[2]

[1]CIRES / NOAA Physical Sciences Laboratory
[2]NOAA/CDC

November 24, 2022

Abstract

The water resources of the western United States have enormous agricultural and municipal demands. At the same time, droughts like the one enveloping the West in the summer of 2021 have disrupted supply of this strained and precious resource. Historically, seasonal forecasts of cool season (November-March) precipitation from dynamical models such as North American Multi-Model Ensemble (NMME) and the SEAS5 from the European Centre for Medium-Range Weather Forecasts have lacked sufficient skill to aid in Western stakeholders' and water managers' decision making. Here, we propose a new empirical-statistical framework to improve cool season precipitation forecasts across the contiguous United States (CONUS). This newly developed framework is called the Statistical Climate Ensemble Forecast (SCEF) model. The SCEF framework applies a principal component regression model to predictors and predictands that have undergone dimensionality reduction, where the predictors are large-scale meteorological variables that have been prefiltered in space. The forecasts of the SCEF model captures 12.0% of the total CONUS-wide standardized observed variance over the period 1982/1983-2019/2020, while NMME captures 7.2%. Over the more recent period 2000/2001-2019/2020, the SCEF, NMME and SEAS5 models respectively capture 11.8%, 4.0% and 4.1% of the total CONUS-wide standardized observed variance. Importantly, much of the improved skill in the SCEF, with respect to models such as NMME and SEAS5, can be attributed to better forecasts across most of the western United States.

A new methodology to produce more skillful United States cool season precipitation forecasts

Matthew B. Switanek[1,2], Thomas M. Hamill[2]

[1]CIRES, University of Colorado Boulder
[2]NOAA/Physical Sciences Laboratory, Boulder, Colorado

Key Points:

- We develop a weighted ensemble of statistical models that improves precipitation forecast skill in the cool season of November-March.

Corresponding author: Matthew Switanek, `matt.switanek@noaa.gov`

Abstract

The water resources of the western United States have enormous agricultural and municipal demands. At the same time, droughts like the one enveloping the West in the summer of 2021 have disrupted supply of this strained and precious resource. Historically, seasonal forecasts of cool season (November-March) precipitation from dynamical models such as North American Multi-Model Ensemble (NMME) and the SEAS5 from the European Centre for Medium-Range Weather Forecasts have lacked sufficient skill to aid in Western stakeholders' and water managers' decision making. Here, we propose a new empirical-statistical framework to improve cool season precipitation forecasts across the contiguous United States (CONUS). This newly developed framework is called the Statistical Climate Ensemble Forecast (SCEF) model. The SCEF framework applies a principal component regression model to predictors and predictands that have undergone dimensionality reduction, where the predictors are large-scale meteorological variables that have been prefiltered in space. The forecasts of the SCEF model captures 12.0% of the total CONUS-wide standardized observed variance over the period 1982/1983-2019/2020, while NMME captures 7.2%. Over the more recent period 2000/2001-2019/2020, the SCEF, NMME and SEAS5 models respectively capture 11.8%, 4.0% and 4.1% of the total CONUS-wide standardized observed variance. Importantly, much of the improved skill in the SCEF, with respect to models such as NMME and SEAS5, can be attributed to better forecasts across most of the western United States.

1 Introduction

Widespread international collaboration and model-development efforts have noticeably improved precipitation forecasts at lead-times of days to weeks (Brunet et al., 2010; Doblas-Reyes et al., 2013; Alley et al., 2019; Benjamin et al., 2019). Bauer et al. (2015) termed this advancement as the "quiet revolution in weather forecasting." Despite the gains observed in short-term weather forecasts, broad-scale skillful numerical seasonal forecasts remain elusive. The El Niño Southern Oscillation (ENSO), however, continues to remain the dominant driver of large-scale teleconnections and predictability on the global scale (Ropelewski & Halpert, 1987; Redmond & Koch, 1991; Cayan et al., 1999; Power et al., 2013; Capotondi et al., 2015; Hoell et al., 2016; Guo et al., 2017; Kumar & Chen, 2017; Nigam & Sengupta, 2021). ENSO teleconnective patterns can persist for months, and as a result, can modulate precipitation with ENSO phase and provide some seasonal forecast skill relative to its unconditional distribution (Quan et al., 2006; Manzanas et al., 2014).

Over the last decade, substantial resources have been put into ensemble seasonal prediction systems such as North American Multi-Model Ensemble (NMME) (Kirtman et al., 2014b) and the SEAS5 model from the European Centre for Medium-Range Weather Forecasts (ECMWF) (Johnson et al., 2019b). These dynamical models have demonstrated skillful forecasts across regions of the contiguous United States (CONUS) where concurrent ENSO teleconnections are strongest (Becker et al., 2014; Gubler et al., 2020; Roy et al., 2020). Despite the success of these dynamical models in forecasting cool season precipitation in those regions, they often fail to provide skill in the most water-critical regions such as the western United States.

Across the western United States, the cool season has a profound impact on water resources (Udall & Overpeck, 2018; Zengchao et al., 2018; Broxton et al., 2019). The cool season, which in this paper we define between the months of November and March, is the the primary snow accumulation period across the mountainous West. Snow accumulation in the cool season can then be used to provide more accurate estimates of streamflow and water resources for the spring and summer seasons.

Building on existing ENSO teleconnections, Switanek et al. (2020) showed a robust statistical relationship between ENSO and cool season precipitation at surprisingly long lead times across much of the western United States. For some regions such as northern California through the American Rocky Mountains, this statistical relationship was found to be greatest at lead/lagged (ENSO/precipitation) times of greater than one year. The authors subsequently built a simple statistical forecast model (the combined lead sea surface temperature (CLSST) model) that exploits the statistical teleconnections between ENSO and precipitation, at multiple lead-times of up to 18 months, using the NINO3.4 sea surface temperature (SST) time series as a sole predictor. The CLSST statistical model from Switanek et al. (2020) was shown to provide moderately more skillful forecasts across CONUS than either NMME or ECMWF's SEAS5 model. Importantly, the CLSST model was shown to substantially improve the forecast skill across much of the West.

In this paper, we extend the work of Switanek et al. (2020) and develop a statistical modeling framework to further improve CONUS precipitation forecasts for the cool season November-March. The forecast product that we develop herein can be used directly, or as a reference standard for other dynamically based forecast systems.

2 Data

Accumulated monthly precipitation was obtained from PRISM (2021). This data was first upscaled from its native 1/24° degree resolution to 1/8° using arithmetic averaging. Next, we summed precipitation at each 1/8° grid cell over the November-March cool season. Then, we calculated areal averages for the 204 division 4 hydrologic unit codes (HUC) across CONUS (Seaber et al., 1987). HUCs use six levels of spatial hierarchy to parse watersheds, represented by numeric codes 2 through 12 (where divisions 2 and 12 delineate the most coarse-scale to the most fine-scale resolutions, respectively). Given our own discussions with water managers across the western United States and the general lack of spatial and temporal precision of seasonal forecasts, we have deemed precipitation cool season forecasts at the division 4 HUC resolution as most appropriate and useful for many large-scale decisions concerning water resources. Henceforth, we use HUC to refer to this division 4 level of spatial resolution (refer to Figure 2, for example, to observe the division 4 HUCs across CONUS).

Sea surface temperature (SST) time series were computed using the NOAA Extended Reconstructed Sea Surface Temperature (ERSST) version 5 (Huang & coauthors, 2020). The SST dataset contains monthly averages at a 2° resolution. We used this data set to subsequently calculate the monthly NINO3.4 (5N-5S, 170W-120W) time series.

Sea-level pressure (SLP), in addition to, zonal and meridional wind speeds (UWND, VWND) were extracted from the NCEP/NCAR Reanalyis dataset at different pressure heights (Kalnay & coauthors, 1996). We obtained global fields of SLP, UWND, and VWND at a temporal resolution of 2.5°.

Historical reforecasts of ensemble mean precipitation were obtained for NMME (Kirtman et al., 2014b, 2014a) in addition to the more recent years of real-time forecasts (Kirtman et al., 2014c). The reforecast data and the real-time forecasts correspond to the years 1982-2010 and 2011-2020, respectively. These reforecasts and the real-time forecasts were obtained for the individual months using an October initialization date. We then calculated precipitation sums for the November-March cool season and spatially averaged the forecasts across each HUC. To be consistent with the procedure we used to obtain observed cool season precipitation at each HUC, the NMME ensemble mean values were resampled to 1/8°, prior to averaging, where the 64 finer resolution grid cell anomaly values are simply equal to that of the containing 1 degree value. Then, spatially averaged precipitation amounts were calculated at each HUC as the average of the 1/8° precipitation amounts that were contained by each respective HUC shapefile.

Seasonal forecasts from ECMWF's long-range SEAS5 model were obtained for the years 1993-2020 (Johnson et al., 2019b, 2019a). Ensemble monthly averages for the individual months between November-March were computed where the model was initialized in October, then summed over the cold season. As with NMME, the data was resampled to 1/8° and averaged across the individual HUCs.

3 Validation and skill metrics

In this study, we make forecasts using two different cross validation approaches. With the first, we use a split sample test case where only the data up through and including 1999/2000 is used in calibration, and we predict and validate model performance over the 20 cool seasons in the period 2000/2001-2019/2020. In the second test, we perform a ten-fold cross validation. We subsequently compare our cool season forecasts to those made by the NMME and ECMWF-SEAS5 models.

The performance of the forecasts are evaluated using anomaly correlation and root mean square error (RMSE) (Eqs. 8.68 and 8.30 respectively from Wilks (2006)). We use throughout the paper the terms CONUS-average and CONUS-wide anomaly correlation or RMSE. CONUS-average anomaly correlation (or RMSE) is the result of first calculating the anomaly correlation for each of the 204 HUCs, then averaging these anomaly correlations across all 204 HUCs. In contrast, CONUS-wide anomaly correlation first standardizes the forecasts and observations, then calculates one anomaly correlation value (or RMSE) between the entire set of our forecasts and observations. For example, if we are forecasting the 20 cool seasons over the period, 2000/2001-2019/2020, for the 204 HUCs, we have 4080 (i.e., 20 x 204) samples that are used to calculate our CONUS-wide anomaly correlation.

4 Methods

Similarly to other ensemble predictions, such as NMME, we developed a modeling framework that uses an ensemble of models. In contrast to the dynamical models of NMME or the SEAS5, however, we have developed a set of statistical models. The forecasts we produce ultimately result from a weighted mean of four different statistical models. Our proposed modeling framework outlines the methods used to develop and combine these statistical models. We term this modeling framework, the Statistical Climate Ensemble Forecast (SCEF) system or the SCEF model. In this paper, we focus on the development and the application of the SCEF model to make cool season (November-March) forecasts of precipitation.

4.1 The SCEF model

The SCEF modeling framework is a three-step process. First, the user develops a set of potentially skillful statistical forecast models using filtered data from key predictors such as SST, sea-level pressure, u-component wind, and v-component wind. Second, each individual statistical model is optimized over the calibration period. Lastly, the individual model forecasts are merged or combined into a weighted ensemble mean. The SCEF model was implemented using principal component regression (PCR) and partial least squares regression (PLSR, similar to canonical correlation analysis (Wilks (2006), chapter 12)). We will show in Section 5 that both of these methods produce similar levels of skill.

4.2 Prescreening the SCEF

We began by exploring a range of potential predictors. Switanek et al. (2020) showed that a simple statistical forecast model that employs the NINO3.4 index as a sole predictor, at multiple lead-times, provides moderately more skillful forecasts than either the NMME or ECMWF's SEAS5 model over much of the US. That model, which is called the CLSST model, and it is one of the statistical models that we use in the SCEF. Additionally, we explored potential predictor variables that were taken from the NCEP/NCAR reanalysis data set. We compared the skillfulness of different potential predictors using leave-one-out cross validation in the calibration period. Through this approach, we selected three additional predictors to be used in the SCEF; these were sea level pressure (SLP), and zonal and meridional winds (UWND and VWND) at a pressure level of 850 hPa. These four statistical forecast models (i.e., CLSST, SLP, UWND, VWND) together comprise our SCEF modeling framework.

During our exploratory analysis, we observed that averages of August-September values of SLP, UWND, and VWND provided better forecasts in our calibration period than using September alone. Additionally, we found better skill in our calibration period by upscaling the resolution of our SLP, UWND, and VWND data from 2.5° latitude by 2.5° longitude to 5.0° latitude by 7.5° longitude. This upscaling was performed using arithmetic averaging, and it removes a level of variability at the smallest scales which we expect are not predictable at seasonal time scales anyway.

4.3 PCR implementation of the SCEF

The CLSST is used very similarly to how it is outlined in Switanek et al. (2020). Here, we provide a very brief overview of the CLSST model. However, for more details, please refer to Switanek et al. (2020). The CLSST model uses the NINO3.4 index as a predictor at different lead times between 1 and 18 months prior. For each preceding month, $m \in (1...18)$, a multiple linear regression model is fit between that month's NINO3.4 SST value and the number of leading principal components of precipitation which we are trying to predict. This model fit is performed during the calibration period, and then the fitted model is used to make forecasts for both the calibration and validation periods. The forecasts in the validation period, at each HUC, are then the weighted mean of the forecasts from these preceding 18 months as a function of their skill in the calibration period. We had experimented with using fields of SSTs as predictors, in place of solely using the NINO3.4 predictor time series. However, that approach did not yield better forecasts than the CLSST model. Here we make a few small modifications to the default implementation of CLSST. These are:

1. We use the respective calibration periods for our two cross validated cases. This is in contrast to 1901/1902-1980/1981 period used in the Switanek et al. (2020) study.
2. The forecasts of each of the preceding 18 months, at each HUC, are weighted by historical skill (i.e., skill in the calibration period) alone and not with an additional linearly decaying weighted function. Adding the linearly decaying weighted function was found not to improve the CONUS-wide forecast skill during the calibration period. Therefore, we have opted to reduce model complexity and weight the CLSST forecasts by historical skill alone.
3. The leading five principal components (PCs) of precipitation are being predicted, in contrast to the leading three. This is to be consistent with the number of principal components we found to be optimal for the SLP, UWND, and VWND statistical models. The leading PCs, in our case, find the spatial patterns (eigenvectors) of precipitation across all HUCs which produce the greatest variability with respect to time.

Next, the three different statistical models (SLP, UWND, and VWND) are independently calibrated. We started by treating four adjustable parameters as ones that could potentially be optimized through calibration. These are, 1) the northern-most latitude of our predictor field, 2) the southern-most latitude of our predictor field, 3) the number of predictor principal components (PCs) to use in our multiple linear regression model, and 4) the number of predictand PCs to use in our multiple linear regression model.

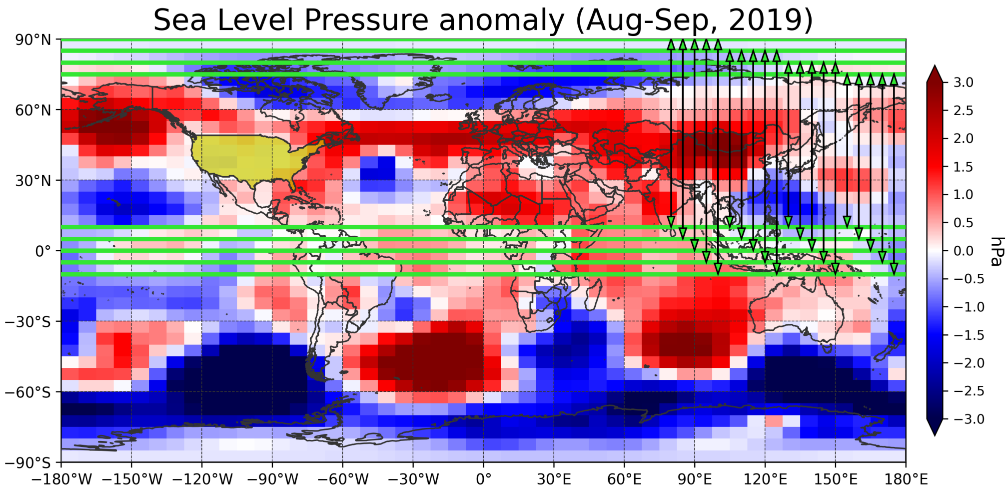

Figure 1. Sea-level pressure anomalies are plotted using the red-to-blue colorbar, where the anomalies were calculated with respect to the period 1948-1999. The horizontal green lines show the northern-most and southern-most latitudinal bounds that we use to constrain our predictor data. The range of possible iterative combinations of these two parameters, given a specified number of predictor PCs, is depicted by the black lines with green arrows on the right side of the plot.

In an effort to reduce the number of parameters that we optimize, we fixed parameter 4 (the number of leading predictand PCs) to five, since that number consistently produced better results than other numbers of PCs. As a result, we now have the other three parameters which require optimization. The prespecified ranges we chose for the three parameters were [87.5°N, 82.5°N, 77.5°N, 72.5°N, where these are the latitudinal centroids] for the northern-most latitude, [12.5°N, 7.5°N, 2.5°N, 2.5°S, 7.5°S, where again these are the latitudinal centroids] for the southern-most latitude (see Figure 1), and [1,...,25] for the number of predictor PCs. We decided at the start, that we would include all longitudinal data in our predictor fields. Therefore, we have not included any additional parameters governing the East-West boundaries of our predictor field.

We begin with our predictor matrix $\mathbf{X}$, whose columns are samples in time and rows are grid points ($\mathbf{X}$ matrix has 72 rows by a variable number of columns), and our predictand matrix $\mathbf{Y}$, whose columns are samples in time and rows are HUCs ($\mathbf{Y}$ matrix is 72 x 204). $\mathbf{X}$ is a subset of the global field of August-September data (SLP, UWND, or VWND), where parameters 1 and 2 control the latitudinal bounds from which we constrain the predictor field. $\mathbf{Y}$ contains our November-March precipitation amounts in the 204 HUC basins. Prior to performing any calibration, we first remove the mean from $\mathbf{Y}$ with

$$\mathbf{Y}_j = \mathbf{Y}_j^{raw} - \mathbf{1}\bar{\mathbf{y}}_j \tag{1}$$

where $\mathbf{Y}_j$ contains our precipitation anomalies at HUC, j, $\mathbf{Y}_j^{raw}$ are our raw precipitation amounts, $\mathbf{1}$ is a 72 x 1 column vector of ones, and $\bar{\mathbf{y}}_j$ is a 1 x 204 row vector containing our mean precipitation amounts with respect to our calibration period (e.g., 1948/49-1999/2000 when using the split sample test case). For our predictors, we remove any existing historical trends,

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i^{raw} - \mathbf{x}_i^{trend} \tag{2}$$

where $\tilde{\mathbf{x}}_i$ and $\mathbf{x}_i^{raw}$ are respectively our detrended and raw time series of predictor values (SLP, UWND, or VWND) at grid cell, i, and $\mathbf{x}_i^{trend}$ is the least-squares trend line fitted with respect to the period of calibration. Next, the predictor data is weighted by latitude,

$$\mathbf{X}_i = \tilde{\mathbf{X}}_i \mathbf{D} \tag{3}$$

where $\mathbf{D}$ is a diagonal matrix with the diagonal elements filled with $cos(\phi_i)$ and ϕ is the latitude of grid cell i. Then, $\mathbf{X}$ is decomposed over the calibration period, using singular value decomposition with the Python package **numpy**,

$$\mathbf{X} = \mathbf{U}_1 \mathbf{S}_1 \mathbf{V}_1 \tag{4}$$

where $\mathbf{S}_1$ is the diagonal matrix containing the singular values of $\mathbf{X}$ and $\mathbf{U}_1$ and $\mathbf{V}_1$ are the left-singular and right-singular vectors, respectively. Similarly, decompose $\mathbf{Y}$ over the calibration period such that,

$$\mathbf{Y} = \mathbf{U}_2 \mathbf{S}_2 \mathbf{V}_2 \tag{5}$$

where $\mathbf{S}_2$ is the diagonal matrix containing the singular values of $\mathbf{Y}$ and $\mathbf{U}_2$ and $\mathbf{V}_2$ are the left-singular and right-singular vectors, respectively. Next, we calculate our principal components of $\mathbf{X}$,

$$\mathbf{X}_{PCS} = \mathbf{X}\mathbf{V}_1^T \tag{6}$$

and similarly, we calculate our principal components (PCs) of $\mathbf{Y}$,

$$\mathbf{Y}_{PCS} = \mathbf{Y}\mathbf{V}_2^T \tag{7}$$

Thus, we can now define our PCR model as a mulitple linear regression,

$$\mathbf{y}_{PCS_k} = \mathbf{X}_{PCS}^{p3} \boldsymbol{\beta} + \boldsymbol{\beta}_0 \tag{8}$$

where $\mathbf{y}_{PCS_k}$ is our leading principal component, k, of our precipitation, where $k \in (1...5)$, $\mathbf{X}_{PCS}^{p3}$ is our matrix of leading principal components of $\mathbf{X}$ using the leading PCs specified by parameter 3, where $p3 \in (1...25)$, and $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$ respectively contain the regression coefficients and intercept obtained through a least-squares fit. The calibration period is used to fit the regression coefficients of Eq. 8. Lastly, we back-transform the data from PC space to precipitation anomaly space at each of the HUCs. This is done with

$$\mathbf{Y}^{fcst} = \mathbf{Y}_{PCS}^{5} \tilde{\mathbf{V}}_2 \tag{9}$$

where $\mathbf{Y}^{fcst}$ are the forecasted precipitation anomalies for the HUCs across CONUS, $\mathbf{Y}_{PCS}^{5}$ are our leading five forecasted PCs, and $\tilde{\mathbf{V}}_2$ are the leading five eigenvectors from our decomposition in Eq. 5.

Our goal, at this point, is to establish for each of the three models (i.e., SLP, UWND, and VWND) which sets of parameters yield the best CONUS-average anomaly correlation forecast skill in our calibration period. Therefore, we use observed precipitation anomalies, $\mathbf{Y}$, and forecasted precipitation anomalies, $\mathbf{Y}^{fcst}$, to calculate the anomaly correlations of each parameter combination at each HUC. These values are calculated over the calibration period. Then, CONUS-average anomaly correlations, for a specified parameter combination, is calculated as

$$r_{p1,p2,p3} = \frac{1}{n} \sum_{j=1}^{204} r_{j,p1,p2,p3} \tag{10}$$

where $r_{p1,p2,p3}$ is our CONUS-average anomaly correlation at HUC, j, $p1$ is our parameter governing the northern-most latitude ($p1 \in (1...4)$ [i.e., 87.5°N, 82.5°N, 77.5°N, 72.5°N]), $p2$ is our parameter governing the southern-most latitude ($p2 \in (1...5)$ [i.e., 12.5°N, 7.5°N, 2.5°N, 2.5°S, 7.5°S]), and $p3$ is our parameter governing the number of leading predictor PCs ($p3 \in (1...25)$).

Next, we want to find which parameter sets are optimal in producing the most skillful out-of-sample forecasts. Therefore, in addition to the cross validated cases that we have already outlined, we also implement leave-one-out cross validation over the calibration period itself. Here, we outline an example implementation of the SLP model with the split sample case:

1. Prior to Eq. 1, we choose values for parameters 1 and 2. In the first iteration, we use the northern-most latitude of each of these (i.e., 87.5°N and 12.5°N, respectively). Then, the global field of SLP data is constrained by our chosen latitudinal bounds.
2. Specify the value of parameter 3 which controls the number of leading PCs to use from our predictor matrix. In our initial iteration, only the first leading PC is used.
3. Proceed with Eqs. 1-7.
4. Use Eqs. 8-9 with leave-one-out cross validation to forecast the years in the calibration period. For example, data from the years 1949/50-1999/2000 is used to fit the model in Eq. 8, and use Eq. 9 to make retrospective forecasts for the HUCs in the season November-March 1948/49. Next, the season 1949/50 is left out and the other 51 calibration years are used to forecast that season. Then, proceed in the same manner until all of the calibration years have been reforecasted. Lastly, fit the model in Eq. 8 to the entire calibration period (all 52 years), and use Eq. 9 to make forecasts for the years 2000/01-2019/20.

The steps enumerated above are repeated until we have iterated over all possible combinations of our three parameters (4x5x25 = 500 possible scenarios). And Eq. 10 is then used to find the sets of parameters which produced the greatest cross-validated skill in our calibration period. The parameter combinations that produced the top 1% of CONUS-average anomaly correlations (the 5 best performing parameter combinations in the calibration period) are subsequently averaged to calculate ensemble mean forecasts. This process is performed independently for each of the three SLP, UWND, and VWND statistical models.

At this point, we have produced four sets of forecasts. These are the CLSST model forecasts, and the forecasts resulting from our optimized ensemble mean PCR forecasts using the SLP, UWND, and VWND fields. Lastly, we obtain the weighted mean ensemble forecasts as

$$\mathbf{Y}_j^{fcst} = \frac{\mathbf{Y}_{1j}^{fcst} w_{1j} + \mathbf{Y}_{2j}^{fcst} w_{2j} + \mathbf{Y}_{3j}^{fcst} w_{3j} + \mathbf{Y}_{4j}^{fcst} w_{4j}}{w_{1j} + w_{2j} + w_{3j} + w_{4j}} \tag{11}$$

where our weighted ensemble mean forecasts, $\mathbf{Y}^{fcst}$, at HUC, j, are comprised of the forecasts of the CLSST model, $\mathbf{Y}_{1j}^{fcst}$, the SLP model, $\mathbf{Y}_{2j}^{fcst}$, the UWND model, $\mathbf{Y}_{3j}^{fcst}$, and the VWND model, $\mathbf{Y}_{4}^{fcst}$, and w_{1j}, w_{2j}, w_{3j}, and w_{4j} are the weights of those models, respectively. Prior to Eq. 11, the forecasts of $\mathbf{Y}_{1}^{fcst}$, $\mathbf{Y}_{2}^{fcst}$, $\mathbf{Y}_{3}^{fcst}$, and $\mathbf{Y}_{4}^{fcst}$, were each independently standardized for each HUC over the calibration period (e.g., 1949/50-1999/2000 using the split sample case). The weights are defined as

$$w_{1j} = \left(\frac{r_{1j}+1}{2}\right)^2, w_{2j} = \left(\frac{r_{2j}+1}{2}\right)^2, w_{3j} = \left(\frac{r_{3j}+1}{2}\right)^2, w_{4j} = \left(\frac{r_{4j}+1}{2}\right)^2 \tag{12}$$

where r_{1j}, r_{2j}, r_{3j}, and r_{4j} are the anomaly correlations of our four statistical models calculated over the calibration period for HUC, j. Through calculating the Akaike information criterion (AIC) (Akaike, 1974), we were able to confirm that the skill improvement using all four predictor models was better than any individual model or model combination.

In addition to the split sample case, which we have used to outline the methods above, we also performed a 10-fold cross validated test. In the 10-fold case, for each fold we leave out four consecutive years for a total of ten different times. This was done over the 40 year period 1980/1981-2019/2020. For example, we initially left out 1980/81-1983/84, and used the years 1948/49-1979/80 and 1984/85-2019/20 to fit the SLP, UWND and VWND models and make forecasts for those four years. Next, we did the same with the years 1984/85-1987/88, and so on. Otherwise, the model fitting and forecasting procedure is the same as outlined for the split sample test. However, in contrast to the split sample test, the standardization of the forecasts $\mathbf{Y}_1$, $\mathbf{Y}_2$, $\mathbf{Y}_3$, and $\mathbf{Y}_4$, for all HUCs, is performed over the period 1949/50-1979/1980.

4.4 PLSR implementation of the SCEF

PLSR has a potential advantage over PCR, insofar that PLSR can find statistical relationships between transformed predictors and predictands where the transformed predictors may explain a low amount of variance. Using PLSR allows us to check for: 1) How effectively can a method such as PLSR sift through the data and pull out relevant predictors without any prescreening? and 2) Do we gain anything by allowing predictor projections that potentially explain less variance than through a method such as PCR? We implement PLSR using the Python package **scikit-learn**. For a detailed explaination of PLSR, please refer to Wold et al. (2001).

Initially, we simply calculated the skill of the PLSR weighted mean forecasts using only the August-September average SLP, UWND, and VWND data. We leave out the CLSST model, since the CLSST model forecasts remain constant, and therefore, the difference lies in the PCR or PLSR implementation of the other three statistical models. This initial baseline forecast was performed using our split sample test with the default number of components (i.e., two components) in the PLS regression. The predictor data was the entire grid of global SLP, UWND, and VWND at the same 5.0° latitude by 7.5° longitude resolution.

Next, we added complexity to the PLSR model by fitting the same three parameters that we fit with PCR.

5 Results

The anomaly correlation forecast skill over the last 20 years for NMME, ECMWF-SEAS5, and the SCEF models can be seen in Figure 2. The optimized PCR and the PLSR implementations of the SCEF model, using the split sample cross validated case, both clearly outperform NMME and ECMWF-SEAS5 over the period 2000/2001-2019/2020. The CONUS-average anomaly correlation for the SCEF model is nearly double that of NMME and ECMWF-SEAS5. After accounting for field significance (Benjamini & Hochberg, 1995; Wilks, 2016), we found 10% of the 204 CONUS HUCs to have statistically significant forecast skill for NMME, 10% for ECMWF-SEAS5, 58% for SCEF (PCR), and 61% for SCEF (PLSR) (using a false discovery rate, α_{FDR}, of 0.10, please refer to Wilks (2016) for details). More specfically, the SCEF model has a more dramatic improvement in forecast skill across the western United States.

In the previous section, we discussed that one of the first things we did was to observe how well a baseline PLSR model performed. This is an implementation of the PLSR model using SLP, UWND, and VWND data with no preprocessing (i.e., we are not controlling the regional limits of our predictors, and we simply use the default number of components, which was two). Under that set of conditions, and predicting the last 20 years using the split sample case, the forecasts had a CONUS-average correlation of 0.230. That CONUS-average anomaly correlation is substantially less than what we achieve by fitting our three parameters across these three statistical models in the PCR framework, which is 0.369.

Through fitting the same three parameters discussed in Section 4, however, the PLSR implementation of the SCEF model is able to achieve similar performance to that of the

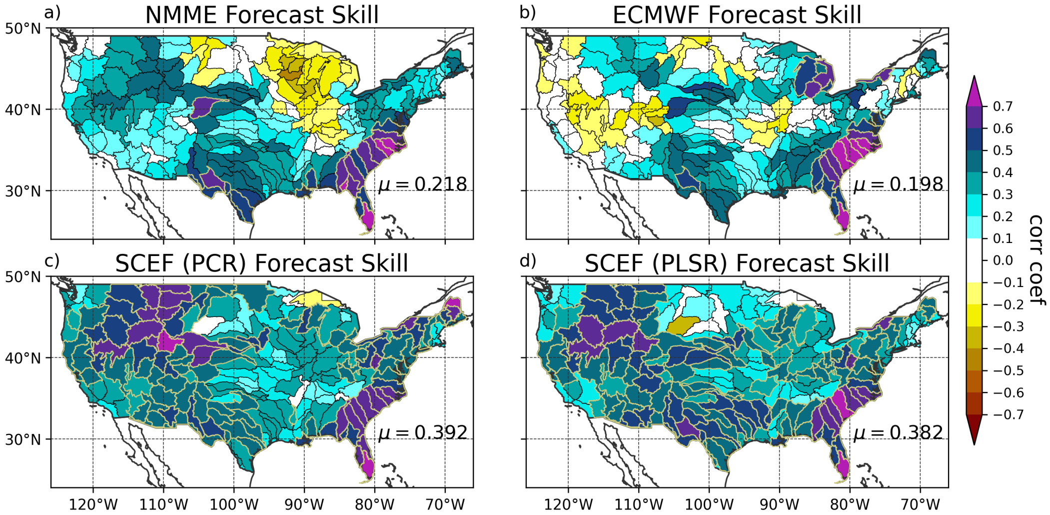

Figure 2. Anomaly correlation skill of the split validation forecasts for the period 2000/2001-2019/2020.

PCR implementation. This is true for our chosen skill metrics and cross validation schemes. Ultimately, the PCR implementation was found to perform modestly better, and as a result, we focus the duration of the paper on showing the SCEF model forecasts and associated forecast skill metrics using only the PCR implementation.

In Figure 3a, one can observe the similarity of the SCEF (PCR) forecasts themselves and the skill of these forecasts (Figure 3b) when using the two different validation cases. In the end, it is desirable to produce cross validated forecasts over a period greater than the 20 year period 2000/2001-2019/2020 (which is illustrated in Figure 2). That way, we can compare skill over a longer period of record like NMME's, for example, which is 1982/1983-2019/2020. Given the relatively small sample size of the NCEP/NCAR Reanalyis dataset (72 cool seasons or samples), though, it is not reasonable by default to expect a good fit of our model parameters if we attempt to perform a split sample test with a validation period equal to NMME's period of record. In that case, we would use the calibration period 1948/1949-1981-1982 to fit the model and we would validate over the period 1982/1983-2019/2020. Therefore, we needed to rely on a different cross validation scheme that allows us: 1) to have longer periods of calibration data for more robust model fitting, and 2) compare the forecasts over a longer period of record. We used 10-fold cross validation to overcome that challenge. However, prior to simply comparing the skill of the 10-fold cross validated SCEF model to NMME over a longer period, we want to be confident that the 10-fold case is not overfitting our model in such a way as to inflate our forecast skill with respect to the more robust split sample test. Figure 3a shows that we do not have any systematic bias in the forecasts themselves between the two cross validation cases, while Figure 3b then shows that the 10-fold case is not overestimating or inflating the forecast skill with respect to the split sample case (i.e., the scatter is well distributed about unity in Figure 3b). This now gives us the necessary confidence to move forward and compare the forecast skills of the 10-fold case of the SCEF model to those of NMME for the longer period of record 1982/1983-2019/2020.

In Figure 4, we show the sensitivity of our three model parameters for each of the individual statistical models comprising the SCEF (PCR) framework. One can observe that the models are most sensitive to the number of predictor PCs, where using only the first few predictor PCs (left sides of the individual plots) yields much less skill. The mod-

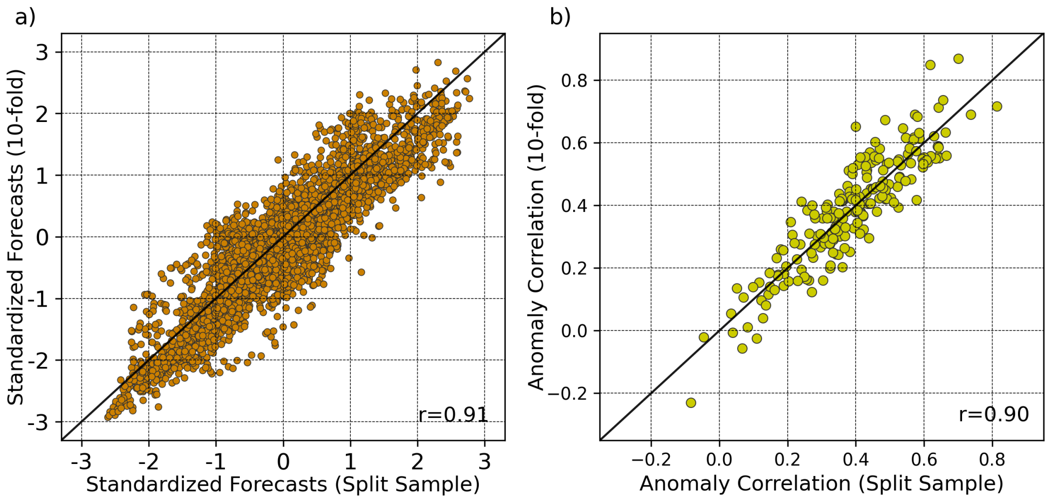

Figure 3. Similarity between the forecasts and the anomaly correlations over the same period of record, 2000/2001-2019/2020, using the split sample and 10-fold cross validation cases. a) plots the standardized forecasts, for all HUCs, using the split sample (x-axis) versus the 10-fold (y-axis) cross validation cases. b) compares the anomaly correlations between the split sample (x-axis) and the 10-fold (y-axis) cross validation cases.

els can be seen to exhibit less sensitivity to the parameters controlling the northern-most and southern-most latitudinal bounds. The best performing combination of model parameters are enclosed by the black boxes in Figure 4, where these are the top performing 1% of parameter sets as calculated using the calibration data. It is also evident for the UWND model that the parameters reach saturation at the upper limits of our prespecified boundary ranges. This appears to indicate that using larger ranges for our parameters could yield better performance. However, we did not want to influence the performance of our model by how skillful we found it to be during validation. Therefore, we stick with our original prespecified parameter ranges that were chosen prior to model implementation.

Figure 5 compares the anomaly correlation forecast skill of the NMME model to that of the SCEF model over the longer period of record 1982/1983-2019/2020. The CONUS-average anomaly correlation for the SCEF model is 0.361, while for NMME it is 0.271. Statistically significant forecast skill is observed for 52% and 77% of the basins across CONUS for NMME and SCEF, respectively. For the western United States, west of 100°W, 63% and 94% of basins have statistically significant forecast skill.

The reduction in RMSE with respect to climatology, for the NMME and SCEF forecasts, over the longer period of record, 1982/1983-2019/2020, is shown in Figure 6. RMSE is calculated using standardized forecasts and observations. Though, prior to calculating RMSE, we first obtain a constant scaling factor which we apply to the forecasts. This scaling factor is optimized to provide the greatest reduction in RMSE for the SCEF model in the calibration period 1948/1949-1981/1982. The scaling factor for the SCEF model forecasts was 0.40. It should be noted that this scaling factor is robust and the same value is obtained if we had optimized in-sample over the validation period 1982/1983-2019/2020. Similarly, we optimized the scaling factor for NMME. Though, we cannot calculate an out-of-sample scaling factor for NMME, and simply optimized this value in-sample over the validation period 1982/1983-2019/2020. NMME's scaling factor was 0.30. We then multiply all of the SCEF and NMME standardized forecasts, at all HUCs, in the vali-

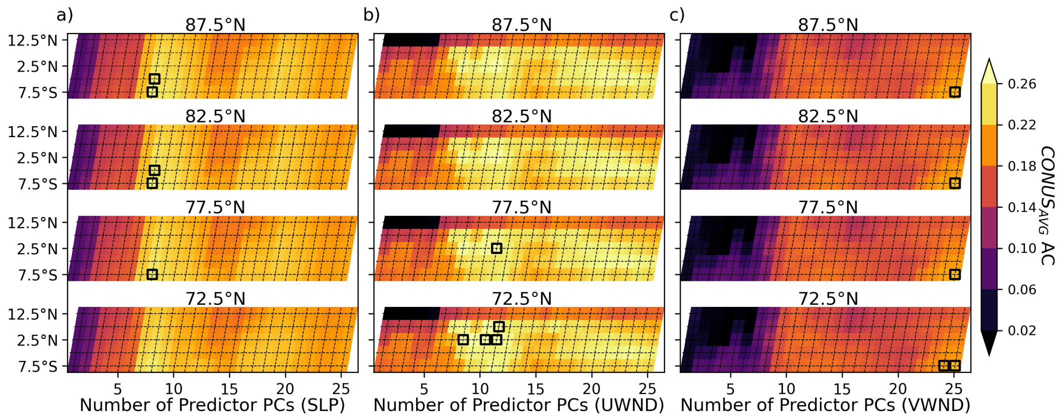

Figure 4. Anomaly correlations skill scores are shown for the different parameter combinations for the SLP, UWND and VWND statistical PCR models. These are averaged (averaged over each of the 10 folds) anomaly correlations calculated from the calibration period for each parameter combination. The x-axis shows the senstivity of the individual models to using different numbers of predictor PCs in our PCR model. Each panel from top to bottom illustrates the sensitivity of the model to using different northern-most latitudes. And the y-axis illustrates the sensitivity of the model to using different southern-most latitudes. The best performing combination of model parameters (i.e., the top performing 1%) are enclosed by the black boxes.

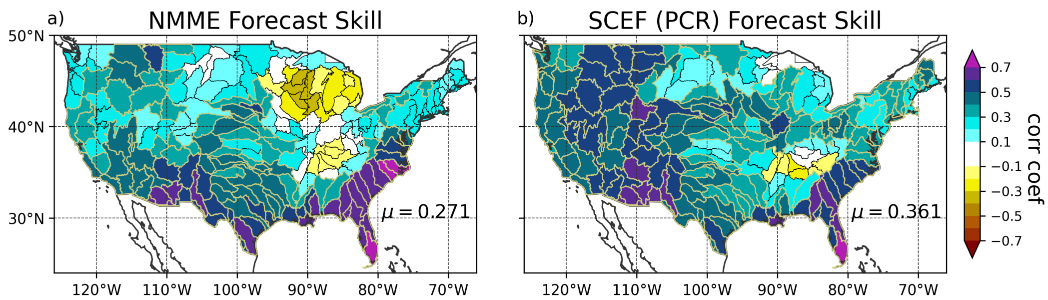

Figure 5. Anomaly correlation skill of the forecasts for the 38 year period between 1982/83-2019/20.

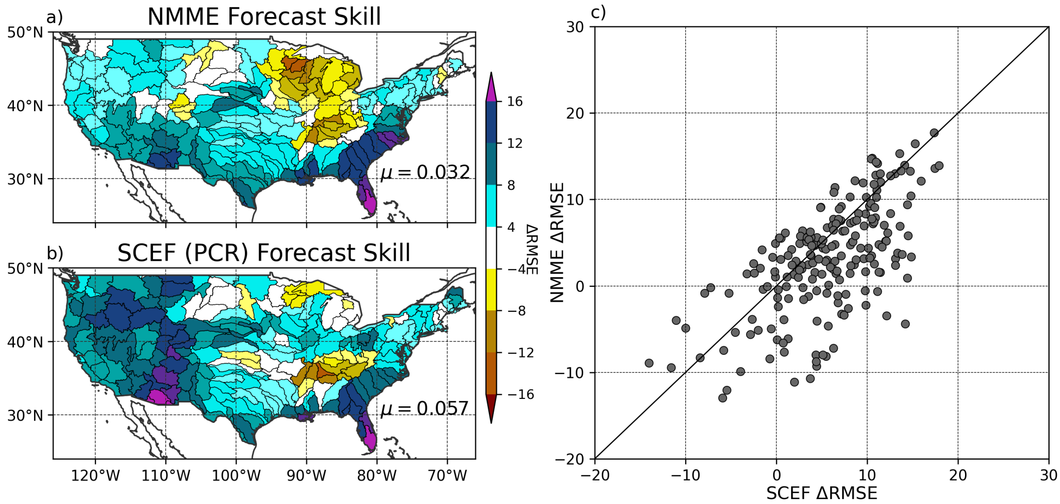

Figure 6. Subplots a) and b) show the percentage reductions in RMSE with respect to climatology. Positive values indicate forecasts that are a positive reduction, or forecasts that perform better than climatology. The CONUS-average RMSE percentage reduction can be seen in the bottom right of subplots a) and b). c) plots the percentage reductions in RMSE, at each HUC, of the SCEF model versus NMME.

dation period by 0.40 and 0.30, respectively. The reductions in RMSE are subsequently calculated using these scaled standardized forecasts. For the NMME forecasts over the period 1982/1983-2019/2020, there is a CONUS-average reduction in RMSE of 3.20% with respect to climatology. In contrast, the SCEF forecasts provide a CONUS-average reduction of 5.70% with respect to climatology over the same period. The SCEF model forecast error reductions again show a more dramatic improvement across the West. In Figure 6c, we can see that both models are capable of providing better forecasts in certain HUCs than the other model, while the SCEF model generally shows greater reductions (i.e., more of the scatter points are situated further to the right of unity than scatter points situated to the left).

Figure 7 shows the scatter points of the standardized forecasts versus observations, for all HUCs simultaneously. The relationship between NMME standardized forecasts and the standardized observations over the longer period of record, 1982/1983-2019/2020, are shown in Figure 7a. The standardized forecasts of the SCEF model versus standardized observations over the same period are shown in Figure 7b. The CONUS-wide percent reduction in RMSE with respect to climatology and the CONUS-wide anomaly correlations can be seen in the upper left-hand of the different subplots of Figure 7. Similarly to the CONUS-averaged results, the CONUS-wide SCEF model forecast skill clearly outperforms NMME. The forecasts of the SCEF and the NMME models respectively capture 12.0% and 7.2% of the total CONUS-wide standardized observed variance over the period 1982/1983-2019/2020. Likewise, the cool season SCEF forecast skill over the more recent period 2000/2001-2019/2020 shows an even greater improvement with respect to NMME (Figures 7c and 7d). Not shown are the ECMWF CONUS-wide results for this shorter period; ECMWF has an anomaly correlation of 0.202 with a reduction in RMSE of 2.2%. Over this more recent period 2000/2001-2019/2020, the SCEF, NMME and SEAS5 models respectively capture 11.8%, 4.0% and 4.1% of the total CONUS-wide standardized observed variance.Figures 7e and 7f compare the standardized forecasts of the SCEF and NMME models for the first 18 years of the record (i.e., 1982/1983-1999/2000). For this earlier period, we observe very similar forecast skill in the two models. It should be

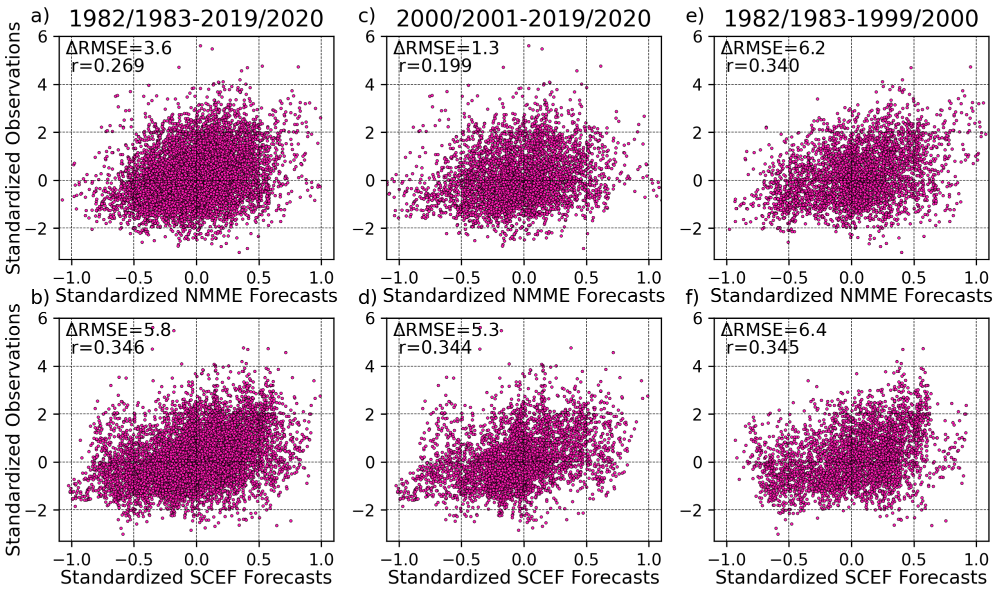

Figure 7. Standardized forecasts plotted against standardized observations for all HUCs simultaneously. The top and bottom rows plot the NMME and SCEF standardized forecasts along the x-axis, respectively, while the standardized observations are plotted on the y-axis. The columns show the impact of different validation periods on the forecast skill. The CONUS-wide percentage reduction in RMSE with respect to climatology and the CONUS-wide anomaly correlation values can be seen in the upper left of each subplot.

noted that the scales of the x and y axes in Figure 7 are different; the forecasted extremes are not nearly as extreme as some of the observed values.

Figure 8 shows the 10-fold cross validated anomaly correlation skill of each of the models that contribute to SCEF. Each model contributes skill in different regions. The CONUS-average skill of the SLP and UWND models generally outperform those of the CLSST and VWND models. Though, importantly, the CLSST model is observed to pick up on skill in the central (north-to-south) region of the West. This is due to the long-lead statistical relationship between NINO3.4 and precipitation (Switanek et al., 2020). What is obvious, when comparing to Figure 5, is that the cross validated weighted ensemble mean forecasts of the SCEF clearly outperform any of the individual models.

The average set of weights (Eq. 12) applied to each of the four models can be seen in Figure 9. Since the weights vary to some degree with respect to the chosen calibration period, the values illustrated in Figure 9 are calculated to be the averages of the weights across each of the 10 folds. As can be expected, the geographic distribution of weights aligns quite closely with the cross validated skill of the individual models from Figure 8.

6 Conclusions

This paper proposes a new statistical modeling framework, which we have called the Statistical Climate Ensemble Forecast (SCEF) model. The SCEF model is capable of producing more skillful cool season November-March precipitation forecasts than both the NMME and ECMWF SEAS5 models. These improvements in cool season forecast skill were shown for the validation periods 2000/2001-2019/2020 and 1982/1983-2019/2020

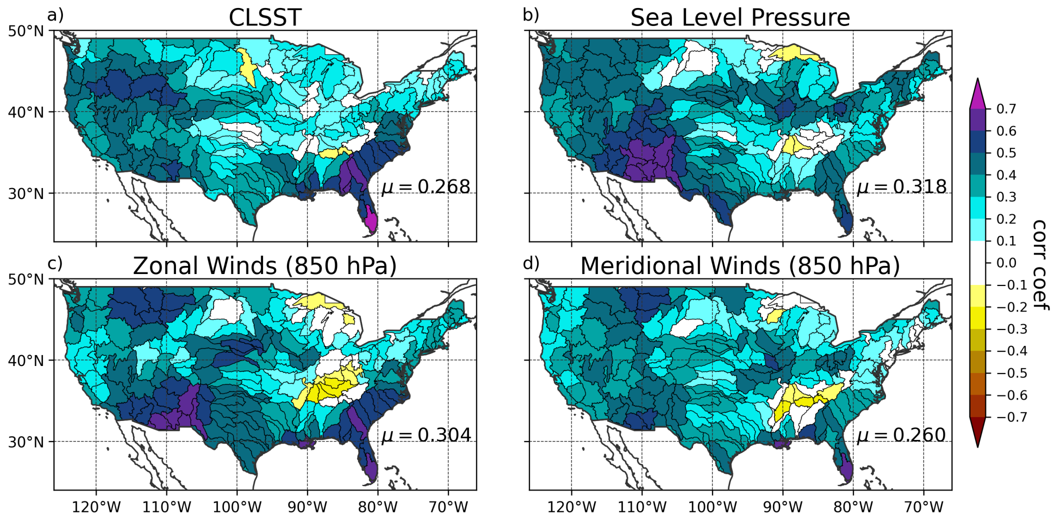

Figure 8. The skill of the individual models, using 10-fold cross validation, over the period 1982/1983-2019/2020.

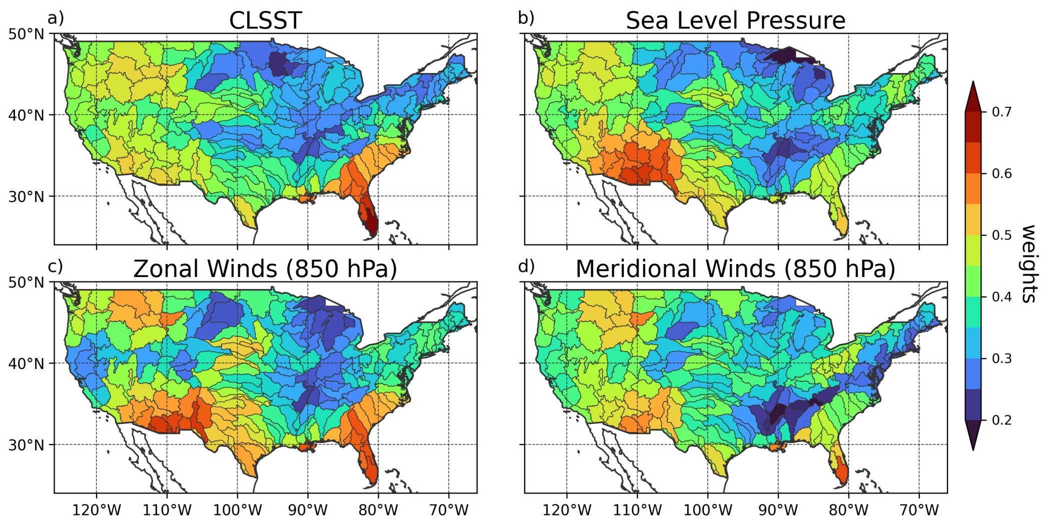

Figure 9. Model weights at each HUC established over the calibration period.

using split validation and 10-fold cross validation, respectively. In particular, the SCEF model most dramatically improves forecast skill across the western United States.

As new observational measurements add to the length of our historical records, more sophisticated empirical-statistical algorthims (Rasouli et al., 2012; Leng & Hall, 2020; Scheuerer et al., 2020) have the capacity to yield further improvements to forecast skill. Even with the simpler empirical-statistical techniques implemented in this paper, however, we can provide optimism for cool season precipitation forecasts across the West. The main contributions of this paper are summarized as: 1) Using statistical predictors at long-lead times of greater than 6 months has the potential to improve forecasts over relying solely on predictors at short-lead times of 1-6 months. 2) Better forecasts can be achieved by prescreening the predictor data. Examples of this can include constraining the spatial extent of our predictor field, in addition to reducing the dimensionality of our predictor and/or predictand data by using fewer leading principal components than our number of samples. 3) Increasing model complexity (NMME versus SCEF) does not necessarily lead to added value.

The results illustrated in Figure 7 raise a few intriguing questions. What explains the SCEF model performing so much better than NMME in the more recent period of 2000/2001-2019/2020? Is this a data quality issue, where better observational and reanalysis data can lead to better forecasts? Can the difference in skill be explained by something such as the magnitude of our predictor data during the validation period (Newman, 2017; Huang et al., 2021; Mariotti et al., 2020)? What could explain periods of greater or lesser forecast skill across the western United States? More effort and continued research is required to unravel some or all of these pertinent questions.

Compounding the difficulties presented by climate change, there has historically been limited forecast skill of cool season precipitation across the water-stressed western United States. As a result, improving these forecasts can provide invaluable decision-making assistance to water managers across the West. Given the devastating drought currently consuming the region in the summer of 2021, the West needs any and all additional tools to help navigate its many natural resource challenges.

Acknowledgments

This study was funded by the California Department of Water Resources through federal grant number 4BM9NCA-P00. The authors do not have any conflicts of interest. MS conceived of the study, performed the analysis, generated the figures, and wrote the paper. TH provided supervision and contributed to the writing of the paper. We would like to thank Michael Alexander, Michael Scheuerer, and Joseph Barsugli for their useful comments and feedback. The code and data required to run the SCEF model can be found at https://github.com/mswitanek/scef-model.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716-723. doi: https://doi.org/10.1109/TAC.1974.1100705

Alley, R. B., Emanuel, K. A., & Zhang, F. (2019). Advances in weather prediction. *Science*, 342-344. doi: https://doi.org/10.1126/science.aav7274

Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, *525*, 47–55. doi: https://doi.org/10.1038/nature14956

Becker, E., Van Den Dool, H., & Zhang, Q. (2014). Predictability and forecast skill in nmme. *J. Climate*, *27*, 5891–5906. doi: https://doi.org/10.1175/JCLI-D-13-00597.1

Benjamin, S. G., Brown, J. M., Brunet, G., Lynch, P., Saito, K., & Schlatter, T. W. (2019). 100 years of progress in forecasting and nwp applications. *Meteorol. Monogr.*, *59*, 13.1–13.66. doi: https://doi.org/10.1175/AMSMONOGRAPHS-D-18-0020.1

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Stat. Soc.*, *57B*, 289–300.

Broxton, P. D., van Leeuwen, W. J. D., & Biederman, J. A. (2019). Improving snow water equivalent maps with machine learning of snow survey and lidar measurements. *BWater Resour. Res.*, *55*, 3739–3757. doi: https://doi.org/10.1029/2018WR024146

Brunet, G., Shapiro, M., Hoskins, B., Moncrieff, M., Dole, R., Kiladis, G. N., … Shukla, J. (2010). Collaboration of the weather and climate communities to advance subseasonal-to-seasonal prediction. *Bull. Amer. Meteor. Soc.*, 1397–1406.

Capotondi, A., Wittenberg, A. T., Newman, M., Di Lorenzo, E., Yu, J. Y., Braconnot, P., … Yeah, S.-W. (2015). Understanding enso diversity. *Bull. Amer. Meteor. Soc.*, *96*, 921–938. doi: https://doi.org/10.1175/BAMS-D-13-00117.1

Cayan, D. R., Redmond, K. T., & Riddle, L. G. (1999). Enso and hydrologic extremes in the western united states. *J. Climate*, *12*, 2881–2893.

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

Gubler, S., Sedlmeier, K. S., Bhend, J., Avalos, G., Coelho, C. A. S., Escajadillo, Y., … Ch., S. (2020). Assessment of ecmwf seas5 seasonal forecast performance over south america. *Wea. Forecasting*, *35*, 561–584. doi: https://doi.org/10.1175/WAF-D-19-0106.1

Guo, Y., Ting, M., Wen, Z., & Lee, D. (2017). Distinct patterns of tropical pacific sst anomaly and their impacts on north american climate. *J. Climate*, *30*, 5221–5241. doi: 10.1175/JCLI-D-16-0488.1

Hoell, A., Hoerling, M., Eischeid, J., Wolter, K., Dole, R., Perlwitz, J., … Cheng, L. (2016). Does el niño intensity matter for california precipitation? *Geophys. Res. Lett.*, *43*, 819–825. doi: https://doi.org/10.1002/2015GL067102

Huang, B., & coauthors. (2020). *Noaa extended reconstruction sea surface temperature (ersst), version 5.* NOAA/National Centers for Environmental Information, accessed on 3 February 2021. Retrieved from `https://doi.org/10.7289/V5T72FNM`

Huang, B., Shin, C.-S., Kumar, A., L'Heureux, M., & Balmaseda, M. A. (2021). The relative roles of decadal climate variations and changes in the ocean observing system on seasonal prediction skill of tropical pacific sst. *Clim. Dyn.*, *56*, 3045-3063. doi: https://doi.org/10.1007/s00382-021-05630-1

Johnson, S. J., Stockdale, T. N., Ferranti, L., Balmaseda, M. A., Molteni, F., Magnusson, L., … Monge-Sanz, B. M. (2019a). *Seas5 data set.* Copernicus Climate Data Store, accessed on 20 December 2020. Retrieved from `https://cds.climate.copernicus.eu`

Johnson, S. J., Stockdale, T. N., Ferranti, L., Balmaseda, M. A., Molteni, F., Magnusson, L., … Monge-Sanz, B. M. (2019b). Seas5: the new ecmwf seasonal forecast system. *Geosci. Model Dev.*, *12*, 1087–1117. doi: https://doi.org/10.5194/gmd-12-1087-2019

Kalnay, E., & coauthors. (1996). The ncep/ncar 40-year reanalysis project. *Bull. Amer. Meteor. Soc.*, *77*, 437–471.

Kirtman, B. P., Min, D., Infanti, J. M., Kinter III, J. L., Paolino, D. A., Zhang, Q., … F., W. E. (2014a). *Hindcast data set of the north american multimodel ensemble: Phase-1 seasonal-to-interannual prediction; phase-2 to-*

ward developing intraseasonal prediction. NOAA National Centers for Environmental Prediction, accessed on 20 December 2020. Retrieved from https://ftp.cpc.ncep.noaa.gov/International/nmme

Kirtman, B. P., Min, D., Infanti, J. M., Kinter III, J. L., Paolino, D. A., Zhang, Q., ... F., W. E. (2014b). The north american multimodel ensemble: Phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bull. Amer. Meteor. Soc.*, *95*, 585-601. doi: https://doi.org/10.1175/BAMS-D-12-00050.1

Kirtman, B. P., Min, D., Infanti, J. M., Kinter III, J. L., Paolino, D. A., Zhang, Q., ... F., W. E. (2014c). *Real-time forecast data set of the north american multimodel ensemble: Phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction.* NOAA National Centers for Environmental Prediction, accessed on 20 December 2020. Retrieved from ftp://ftp.cpc.ncep.noaa.gov/NMME/realtime_anom/ENSMEAN

Kumar, A., & Chen, M. (2017). What is the variability in us west coast winter precipitation during strong el niño events? *Clim. Dyn.*, *49*, 2789–2802. doi: 10.1007/s00382-016-3485-9

Leng, G., & Hall, J. W. (2020). Predicting spatial and temporal variability in crop yields: an inter-comparison of machine learning, regression and process-based models. *Environ. Res. Lett.*, *15*, 044027. doi: https://doi.org/10.1088/1748-9326/ab7b24

Manzanas, R., Frías, M. D., Cofiño, A. S., & Gutiérrez, J. M. (2014). Validation of 40 year multimodel seasonal precipitation forecasts: The role of enso on the global skill. *J. Geophys. Res. Atmos.*, *119*, 1708–1719. doi: https://doi.org/10.1002/2013JD020680

Mariotti, A., Baggett, C., Barnes, E. A., Becker, E., Butler, A., Collins, D. C., ... Albers, J. (2020). Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. *Bull. Amer. Meteor. Soc.*, *101*, E608–E625. doi: https://doi.org/10.1175/BAMS-D-18-0326.1

Newman, P. D., M. and Sardeshmukh. (2017). Are we near the predictability limit of tropical sea surface temperatures? *Geophys. Res. Lett.*, *44*, 8520–8529. doi: https://doi.org/10.1002/2017GL074088

Nigam, S., & Sengupta, A. (2021). The full extent of el niño's precipitation influence on the united states and the americas: The suboptimality of the niño 3.4 sst index. *Geophys. Res. Lett.*, *48*, 1–12. doi: https://doi.org/10.1029/2020GL091447

Power, S., Delage, F., Chung, C., Kociuba, G., & Keay, K. (2013). Robust twenty-first-century projections of el niño and related precipitation variability. *Nature*, *502*, 543–545. doi: https://doi.org/10.1038/nature12580

PRISM, C. G. (2021). *Prism gridded climate data.* Oregon State University, accessed on 10 January 2021. Retrieved from http://prism.oregonstate.edu

Quan, X., Hoerling, M., Whitaker, J., Bates, G., & Xu, T. (2006). Diagnosing sources of u.s. seasonal forecast skill. *J. Climate*, *19*, 3279–3293.

Rasouli, K., Hsieh, W. W., & Cannon, A. J. (2012). Daily streamflow forecasting by machine learning methods with weather and climate inputs. *J. Hydrol.*, *414*, 284–293. doi: https://doi.org/10.1016/j.jhydrol.2011.10.039

Redmond, K. T., & Koch, R. W. (1991). Surface climate and streamflow variability in the western united states and their relationship to large scale circulation indices. *Water Resour. Res.*, *27*, 2381–2399.

Ropelewski, C. F., & Halpert, M. S. (1987). Global and regional scale precipitation patterns associated with el niño/southern oscillation. *Mon. Wea. Rev.*, *115*, 1606–1626. doi: 10.1175/1520-0493(1987)115⟨1606:GARSPP⟩2.0.CO;2

Roy, T., He, X., Lin, P., Beck, H. E., Castro, C., & Wood, E. F. (2020). Global evaluation of seasonal precipitation and temperature forecasts from nmme. *J. Hydrometeor.*, *21*, 2473–2486. doi: https://doi.org/10.1175/JHM-D-19-0095.1

Scheuerer, M., Switanek, M. B., Worsnop, R. P., & Hamill, T. M. (2020). Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over california. *Mon. Wea. Rev.*, *148*, 3489–3506. doi: https://doi.org/10.1175/MWR-D-20-0096.1

Seaber, P. R., Kapinos, F. P., & Knapp, G. L. (1987). Hydrologic unit maps. *USGS Publications Warehouse, Water Supply Paper 2294*. doi: https://doi.org/10.3133/wsp2294

Switanek, M. B., Barsugli, J. J., Scheuerer, M., & Hamill, T. M. (2020). Present and past sea surface temperatures: a recipe for better seasonal climate forecasts. *Wea. Forecasting*, *54*, 6739–6756. doi: https://doi.org/10.1029/2018WR023153

Udall, B., & Overpeck, J. (2018). The twenty-first century colorado river hot drought and implications for the future. *Water Resour. Res.*, *53*, 2404–2418. doi: 10.1002/2016WR019638

Wilks, D. S. (2006). *Statistical methods in the atmospheric sciences (second edition)*. Elsevier, Burlington, Massachusetts.

Wilks, D. S. (2016). "the stippling shows statistically significant grid points": How research results are routinely overstated and overinterpreted, and what to do about it. *Bull. Amer. Meteor. Soc.*, *15*, 2263–2273. doi: https://doi.org/10.1175/BAMS-D-15-00267.1

Wold, S., Sjöström, M., & Eriksson, L. (2001). Pls-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*(2), 109-130. doi: https://doi.org/10.1016/S0169-7439(01)00155-1

Zengchao, H., Singh, V. P., & Xia, Y. (2018). Seasonal drought prediction: Advances, challenges, and future prospects. *Rev. Geophys.*, *56*, 108–141. doi: https://doi.org/10.1002/2016RG000549