

A New Methodology to Produce More Skillful United States Cool-Season Precipitation Forecasts

MATTHEW B. SWITANEK^{a,b} AND THOMAS M. HAMILL^b

^a *Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, Colorado*

^b *NOAA/Physical Sciences Laboratory, Boulder, Colorado*

(Manuscript received 2 December 2021, in final form 11 March 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 Seasonal Forecasting System 5 (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 1982/83–2019/20, whereas NMME captures 7.2%. Over the more recent period 2000/01–2019/20, the SCEF, NMME, and SEAS5 models respectively capture 11.8%, 4.0%, and 4.1% of the total CONUS-wide standardized observed variance. An important finding is that 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.

KEYWORDS: ENSO; Hydrology; Hydrometeorology; Seasonal forecasting; Statistical forecasting

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, broadscale skillful numerical seasonal forecasts remain elusive. The El Niño–Southern Oscillation (ENSO) is the dominant driver of large-scale teleconnections and predictability on the global scale (Ropelewski and Halpert 1987; Redmond and Koch 1991; Cayan et al. 1999; Power et al. 2013; Capotondi et al. 2015; Hoell et al. 2016; Guo et al. 2017; Kumar and Chen 2017; Nigam and 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; Manzanos 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 Seasonal Forecasting System 5 (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 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 and Overpeck 2018; Hao 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 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 1 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 Niño-3.4 sea surface temperature (SST) time series as a sole predictor. The CLSST statistical

 Denotes content that is immediately available upon publication as open access.

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

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 purely statistical modeling framework to further improve CONUS precipitation forecasts for the cool season November–March. The modeling framework applies a principal component regression (PCR) 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 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 Climate Group (2021). These data were first upscaled from their native $1/24^\circ$ degree resolution to $1/8^\circ$ using arithmetic averaging. Next, we summed precipitation at each $1/8^\circ$ 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 2–12 (where divisions 2 and 12 delineate the most coarse-scale and 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 that concern water resources. Henceforth, we use HUC to refer to this division-4 level of spatial resolution (refer to Fig. 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 et al. 2020). The SST dataset contains monthly averages at a 2° resolution. We used this dataset to subsequently calculate the monthly Niño-3.4 (5°N – 5°S , 170° – 120°W) time series.

Sea level pressure (SLP), in addition to zonal and meridional wind speeds (UWND and VWND, respectively), were extracted from the NCEP–NCAR reanalysis dataset at different pressure heights (Kalnay et al. 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,a) 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 1982–2010 and 2011–20, 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 regridded to $1/8^\circ$, prior to averaging, where the 64 finer-resolution grid cell anomaly values are simply equal to that of the containing 1° value. Then, spatially averaged precipitation amounts were calculated at each HUC as the average of the $1/8^\circ$ 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,a). 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 were regridded to $1/8^\circ$ 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 are used in calibration, and we predict and validate model performance over the 20 cool seasons in the period 2000/01–2019/20. In the second test, we perform a 10-fold cross validation. We subsequently compare our cool-season forecasts with those made by the NMME and ECMWF-SEAS5 models.

The performances 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, 2000/01–2019/20, for the 204 HUCs, we have 4080 (i.e., 20×204) samples that are used to calculate our CONUS-wide anomaly correlation.

4. Methods

Similar 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 ECMWF-SEAS5, however, we have developed a set of statistical models. The forecasts we produce ultimately result from a weighted mean of four different purely 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.

a. The SCEF model

The SCEF modeling framework is a three-step process. First, the user develops a set of potentially skillful statistical forecast models, in our case 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. Last, the individual model forecasts are merged or combined into a weighted ensemble mean. The SCEF model was implemented using 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.

b. 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 Niño-3.4 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 United States. That model, which is called the CLSST model, and 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 dataset. 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, and VWND) together compose 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.

c. 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 Niño-3.4 index as a predictor at different lead times between 1 and 18 months prior. For each preceding month, $m \in (1, \dots, 18)$, a multiple linear regression model is fit between that month's Niño-3.4 SST value and the number of leading principal components of precipitation that 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 Niño-3.4 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:

- 1) We use the respective calibration periods for our two cross-validated cases. This is in contrast to the 1901/02–1980/81 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 that 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 northernmost latitude of our predictor field, 2) the southernmost latitude of our predictor field, 3) the number of predictor PCs to use in our multiple linear regression model, and 4) the number of predictand PCs to use in our multiple linear regression model. 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 that require optimization. The prespecified values for these three parameters, along with their associated ranges, are shown in Table 1. To find the optimal parameter combination in the calibration period, we iterate over the range of possible values, which in our case was 4, 5, and 25, respectively. 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 39 rows by a variable number of columns), and our predictand

TABLE 1. The values and possible integer ranges of the three model parameters.

Parameter	Values	Range
Northern lat	87.5°N, 82.5°N, 77.5°N, and 72.5°N	4
Southern lat	12.5°N, 7.5°N, 2.5°N, 2.5°S, and 7.5°S	5
Predictor PCs	1, 2, 3, ..., 25	25

matrix \mathbf{Y} , whose columns are samples in time and rows are HUCs (the \mathbf{Y} matrix is 72×204). Matrix \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. Matrix \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

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

where \mathbf{Y}_j contains our precipitation anomalies at HUC j , $\mathbf{Y}_j^{\text{raw}}$ are our raw precipitation amounts, and \otimes is the vector outer product between $\mathbf{1}$, which is a 72×1 column vector of numerical ones, and \bar{y}_j , which is a 1×204 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,

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

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

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

where \mathbf{V}_1^T is the transpose of \mathbf{V}_1 , and, similarly, we calculate our PCs of \mathbf{Y} ,

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

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

$$\mathbf{y}_{\text{PCS}_k} = \mathbf{X}_{\text{PCS}_{p3}} \boldsymbol{\beta} + \boldsymbol{\beta}_0, \tag{8}$$

where $\mathbf{y}_{\text{PCS}_k}$ is our leading principal component k of our precipitation, where $k \in (1, \dots, 5)$, $\mathbf{X}_{\text{PCS}_{p3}}$ is our matrix of leading principal components of \mathbf{X} using the leading PCs specified by parameter 3, where $p3 \in (1, \dots, 25)$, 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). Last, we back-transform the data from PC space to precipitation anomaly space at each of the HUCs. This is done with

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

where \mathbf{Y}^{fcst} are the forecast precipitation anomalies for the HUCs across CONUS, \mathbf{Y}_{PCS} are our leading five forecast 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 forecast 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, are 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 , and $p1$, $p2$, and $p3$ are our three parameters (refer to Table 1).

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 computation of Eq. (1), we choose values for parameters 1 and 2. In the first iteration, we use the northernmost latitude of each of these (i.e., 87.5° and 12.5°N, respectively). Then, the global field of SLP data is constrained by our chosen latitudinal bounds.
- 2) We 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) We proceed with Eqs. (1)–(7).
- 4) we use Eqs. (8) and (9) with leave-one-out cross validation to forecast the years in the calibration period. For

example, data from 1949/50–1999/2000 are used to fit the model in Eq. (8), and we 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, we proceed in the same manner until all of the calibration years have been reforecast. Last, we 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.

These steps are repeated until we have iterated over all possible combinations of our three parameters ($4 \times 5 \times 25 = 500$ possible scenarios). Equation (10) is then used to find the best-performing parameter combination, that is, the parameters that produced the greatest cross-validated skill in our calibration period. 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. Last, we obtain the weighted ensemble-mean forecasts as

$$\mathbf{Y}_j^{\text{fcst}} = \frac{\mathbf{Y}_{1j}^{\text{fcst}} w_{1j} + \mathbf{Y}_{2j}^{\text{fcst}} w_{2j} + \mathbf{Y}_{3j}^{\text{fcst}} w_{3j} + \mathbf{Y}_{4j}^{\text{fcst}} w_{4j}}{w_{1j} + w_{2j} + w_{3j} + w_{4j}}, \quad (11)$$

where our weighted ensemble-mean forecasts $\mathbf{Y}_j^{\text{fcst}}$ at HUC j are composed of the forecasts of the CLSST model $\mathbf{Y}_{1j}^{\text{fcst}}$, the SLP model $\mathbf{Y}_{2j}^{\text{fcst}}$, the UWND model $\mathbf{Y}_{3j}^{\text{fcst}}$, and the VWND model $\mathbf{Y}_{4j}^{\text{fcst}}$, and w_{1j} , w_{2j} , w_{3j} , and w_{4j} are the weights of those models, respectively. Prior to Eq. (11), the forecasts of $\mathbf{Y}_{1j}^{\text{fcst}}$, $\mathbf{Y}_{2j}^{\text{fcst}}$, $\mathbf{Y}_{3j}^{\text{fcst}}$, and $\mathbf{Y}_{4j}^{\text{fcst}}$ were each independently standardized for each HUC over the calibration period (e.g., 1948/49–1999/2000 using the split-sample case). The weights are defined as

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

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 (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 10 different partitions. This was done over the 40-yr period 1980/81–2019/20. For example, we initially left out 1980/81–1983/84 and used 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 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 1948/49–1979/80.

d. 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 explanation of PLSR, please refer to Wold et al. (2001).

Initially, we simply calculated the skill of the PLSR weighted ensemble-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 were 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

In Fig. 1, we show the sensitivity of our three model parameters for each of the individual statistical models composing the SCEF (PCR) framework. This is shown for the split-sample cross-validation case. 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 subplots) yields much less skill. The models can be seen to exhibit less sensitivity to the parameters controlling the northernmost and southernmost latitudinal bounds. The best-performing combination of model parameters are enclosed by the green boxes in Fig. 1, where these are the top performing parameter sets as calculated using the calibration data. It is also evident for both the UWND and VWND models 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. By this same argument, we initially chose to use all longitudes for our predictors, and therefore, we did not test the effectiveness of having additional parameters that govern the longitudinal extent.

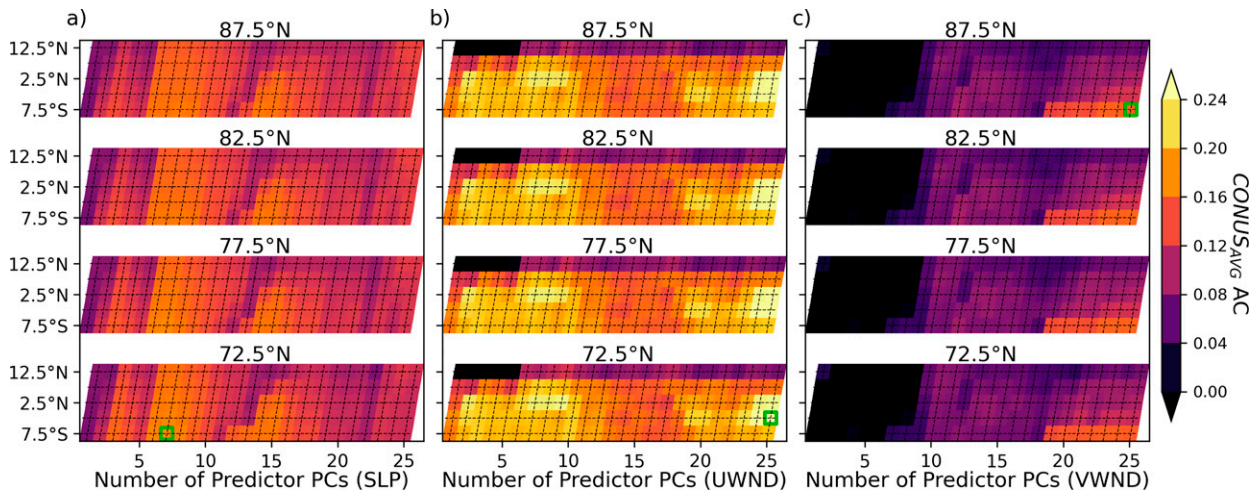


FIG. 1. Anomaly correlations skill scores for the different parameter combinations for the (a) SLP, (b) UWND, and (c) VWND statistical PCR models. These are anomaly correlations calculated from the calibration period, using the split-sample case, for each parameter combination. The x axis shows the sensitivity 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 northernmost latitudes. The y axis illustrates the sensitivity of the model to using different southernmost latitudes. The best-performing combinations of model parameters are enclosed by the green boxes.

The anomaly correlation forecast skill over the last 20 years for NMME, ECMWF-SEAS5, and the SCEF models can be seen in Fig. 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 2000/01–2019/20. The CONUS-average anomaly correlation for the SCEF model is nearly double that of NMME and ECMWF-SEAS5. After accounting for field significance (Benjamini and 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 two-tailed p values along with a false discovery rate, α_{FDR} , of 0.10; please refer to Wilks (2016) for details]. More specifically, the SCEF model has a more dramatic improvement in forecast skill across the western United States. Our approach to establish statistical significance, which is nicely covered in Wilks (2016), can be thought of as being similar to comparing the p value at each basin with a reference p value such as 0.05, except that it additionally accounts for field significance. Furthermore, we did not find temporal autocorrelation of the observed and forecast time series to be statistically significant, and therefore, our test for statistical significance does not account for temporal dependence.

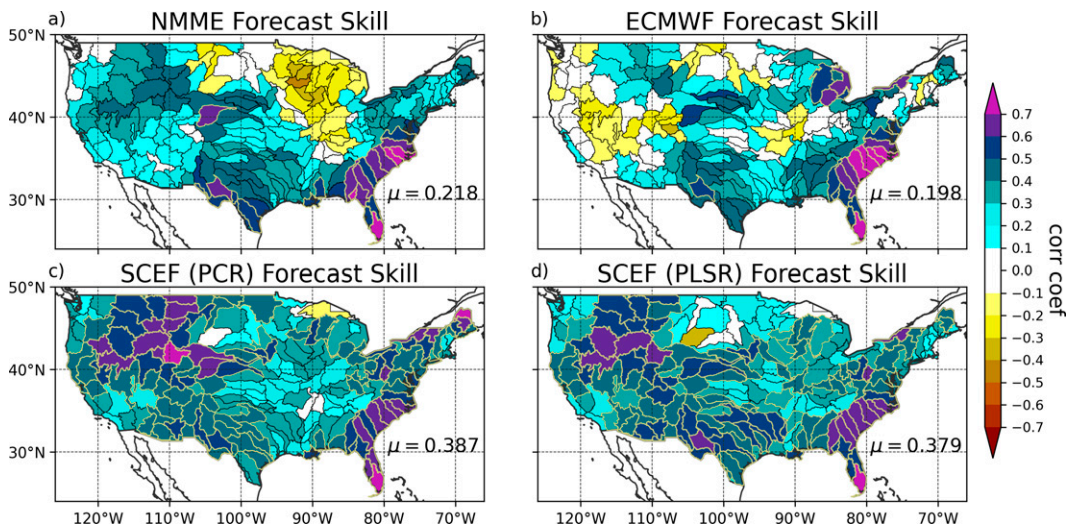


FIG. 2. Anomaly correlation skill of the split validation forecasts for the period 2000/01–2019/20. Statistically significant basins, or HUCs, are outlined by the light-yellow lines.

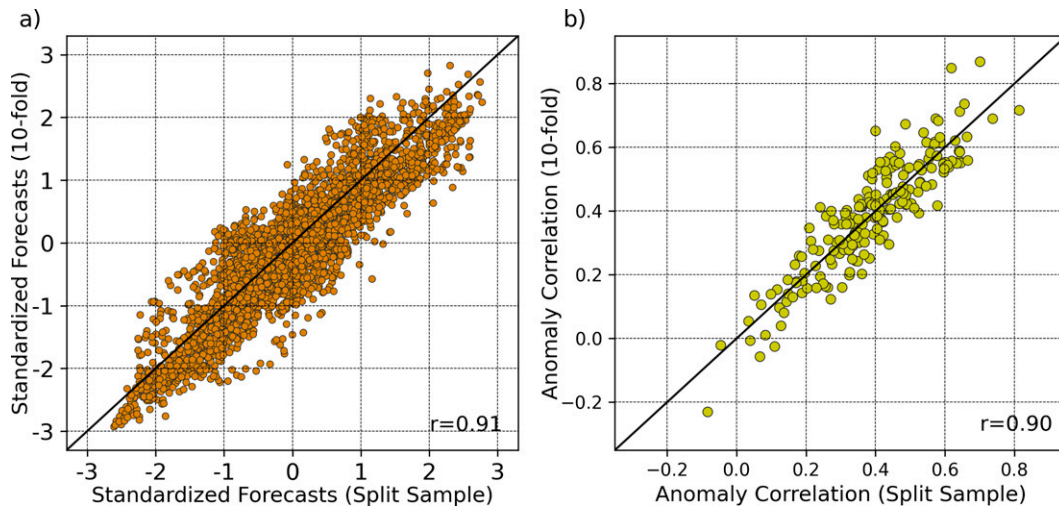


FIG. 3. Similarity between the forecasts and the anomaly correlations over the same period of record, 2000/01–2019/20, using the split-sample and 10-fold cross-validation cases: (a) The standardized forecasts, for all HUCs, using the split-sample (x axis) vs the 10-fold (y axis) cross-validation cases. (b) Comparison of the anomaly correlations between the split-sample (x axis) and the 10-fold (y axis) cross-validation cases.

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 0.370.

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 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 Fig. 3a, one can observe the similarity of the SCEF (PCR) forecasts themselves and the skill of these forecasts (Fig. 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-yr period 2000/01–2019/20 (which is illustrated in Fig. 2). That way, we can compare skill over a longer period of record like NMME's, for example, which is 1982/83–2019/20. Given the relatively small sample size of the NCEP–NCAR reanalysis 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 1948/49–1981/82 to fit the model and we would validate over

the period 1982/83–2019/20. 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) to 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 with 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 Fig. 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 Fig. 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 with those of NMME for the longer period of record 1982/83–2019/20.

Figure 4 compares the anomaly correlation forecast skill of the NMME model with that of the SCEF model over the longer period of record 1982/83–2019/20. The CONUS-average anomaly correlation for the SCEF model is 0.358, 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, 1982/83–2019/20, is shown in Fig. 5. RMSE is calculated using standardized forecasts and observations. First, we calculate these standardized forecasts and observations using 10-fold cross validation. For example, the Z scores (i.e., standard deviations from the mean) are calculated, at each HUC, for

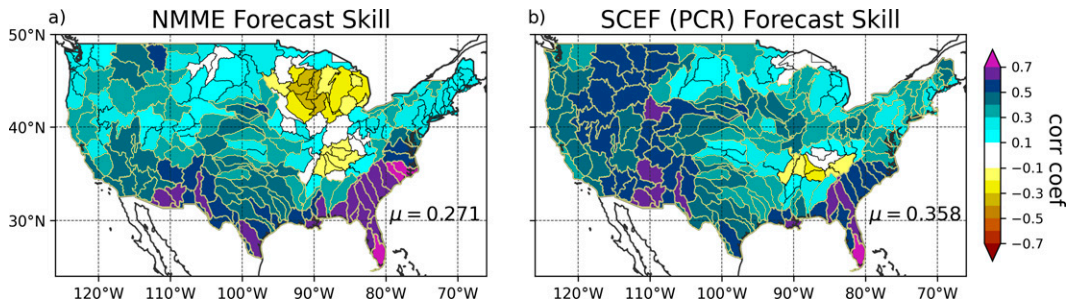


FIG. 4. Anomaly correlation skill of the forecasts for the 38-yr period between 1982/83 and 2019/20.

2016/17–2019/20 using the mean and standard deviations calculated over the period 1982/83–2015/16. Then, prior to calculating RMSE, we obtain a constant scaling factor that 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 1948/49–1981/82. 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 1982/83–2019/20. Similarly, we optimized the scaling factor for NMME. However, we cannot calculate an out-of-sample scaling factor for NMME and simply optimized this value in-sample over the validation period 1982/83–2019/20. NMME's scaling factor was 0.30. We then multiply all of the SCEF and NMME standardized forecasts, at all HUCs, in the validation 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 1982/83–2019/20, there is a CONUS-average reduction in RMSE of 3.2% with respect to climatology. In contrast, the SCEF forecasts provide a CONUS-average reduction of 5.7% with respect to climatology over the same period. The SCEF

model forecast error reductions again show a more dramatic improvement across the West. In Fig. 5c, 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 farther to the right of unity than scatter points situated to the left).

Figure 6 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, 1982/83–2019/20, are shown in Fig. 6a. The standardized forecasts of the SCEF model versus standardized observations over the same period are shown in Fig. 6b. The CONUS-wide percent reduction in RMSE with respect to climatology and the CONUS-wide anomaly correlations can be seen in the upper left of the different subplots of Fig. 6. 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 1982/83–2019/20. Likewise, the cool-

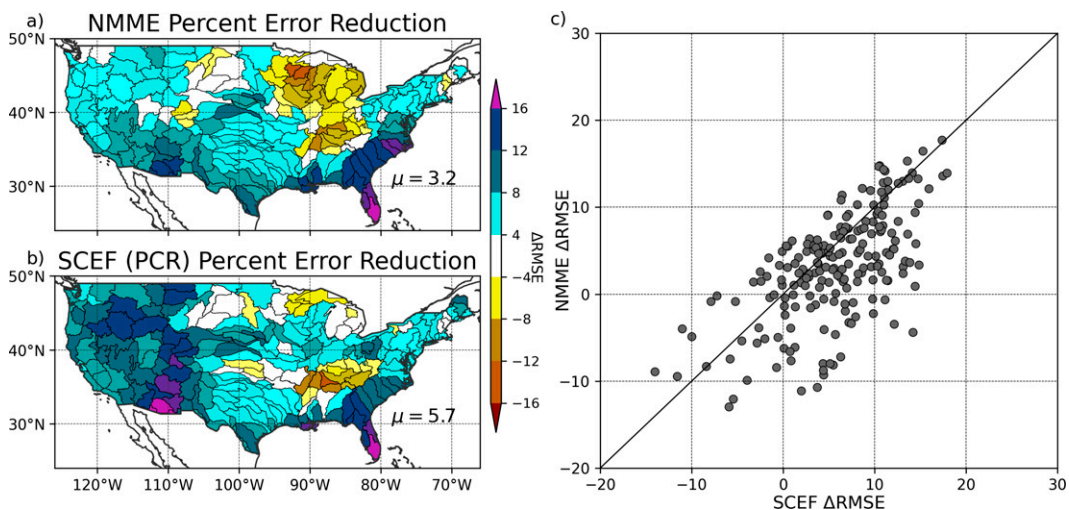


FIG. 5. (a),(b) The percentage reductions in RMSE with respect to climatology. Positive values indicate forecasts that are a positive reduction, or forecasts that perform better than climatology. The CONUS-average RMSE percentage reduction is given in the bottom right of (a) and (b). (c) The percentage reductions in RMSE, at each HUC, of the SCEF model vs NMME.

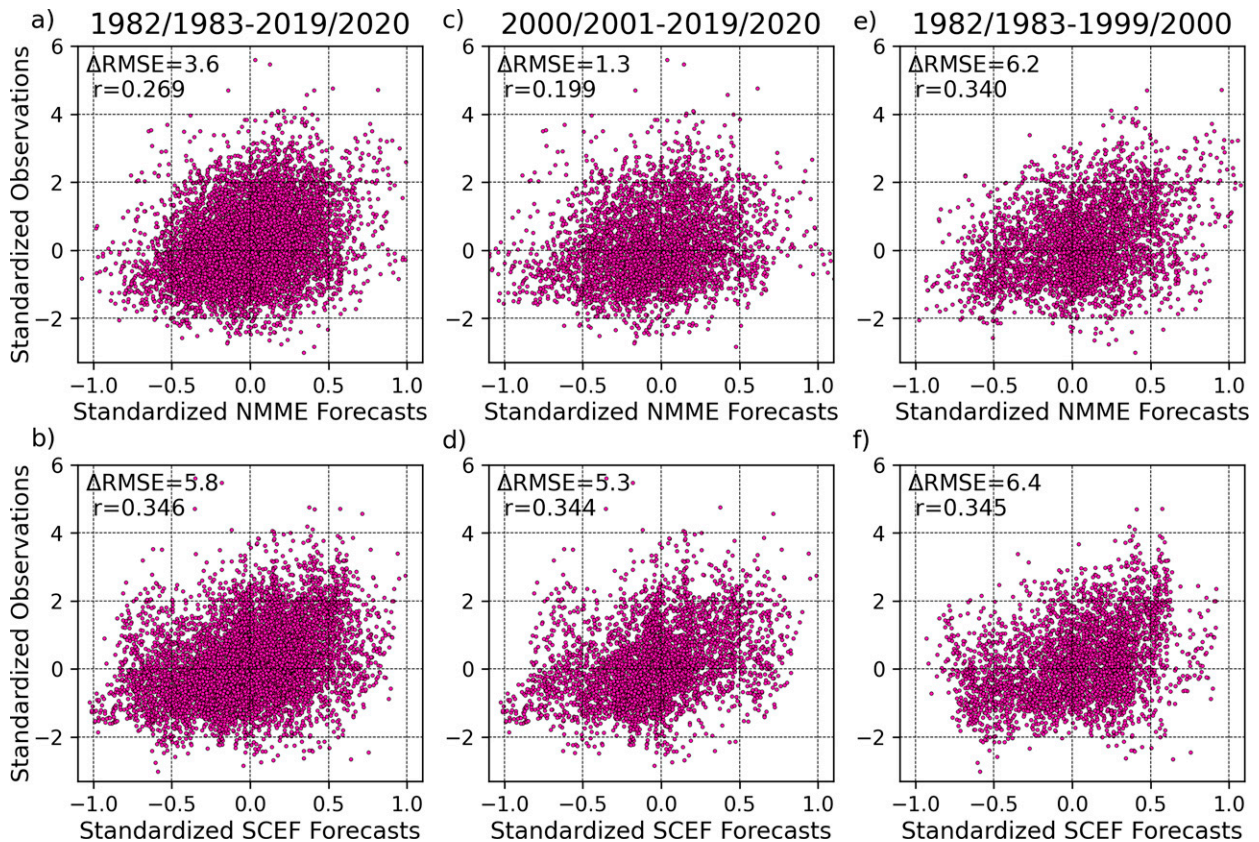


FIG. 6. Standardized forecasts plotted against standardized observations for all HUCs simultaneously, showing the (a),(c),(e) NMME and (b),(d),(f) SCEF standardized forecasts along the x axis and the standardized observations on the y axis. The columns show the impact of different validation periods on the forecast skill. The CONUS-wide percentage reduction in RMSE with respect to climatology and the CONUS-wide anomaly correlation values are shown in the upper left of each panel.

season SCEF forecast skill over the more recent period 2000/01–2019/20 shows an even greater improvement with respect to NMME (Figs. 6c,d). 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 2000/01–2019/20, the SCEF, NMME, and ECMWF-SEAS5 models respectively capture 11.8%, 4.0%, and 4.1% of the total CONUS-wide standardized observed variance. Figures 6e and 6f compare the standardized forecasts of the SCEF and NMME models for the first 18 years of the record (i.e., 1982/83–1999/2000). For this earlier period, we observe very similar forecast skill in the two models. It should be noted that the scales of the x and y axes in Fig. 6 are different; the forecast extremes are not nearly as extreme as some of the observed values.

Figure 7 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. Although, 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 Niño-3.4 and precipitation (Switaneck et al. 2020). What is obvious, when

comparing with Fig. 4, 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 Fig. 8. Since the weights vary to some degree with respect to the chosen calibration period, the values illustrated in Fig. 8 are calculated to be the averages of the weights across each of the 10 folds. As one might expect, the geographic distribution of weights aligns very closely with the cross-validated skill of the individual models from Fig. 7.

6. Discussion

In contrast to the NMME and ECMWF-SEAS5 models, the SCEF model is shown to produce better cool-season forecast skill across much of the contiguous United States, with particular improvements across the West. As a result, it is worth providing some insight as to why that is. We find that much of the skill improvement realized by the SCEF can be attributed to some key differences in model infrastructure.

First, there is a lagged SST statistical response that the SCEF model picks up on. This is done through the CLSST modeling component of the SCEF framework. The SCEF

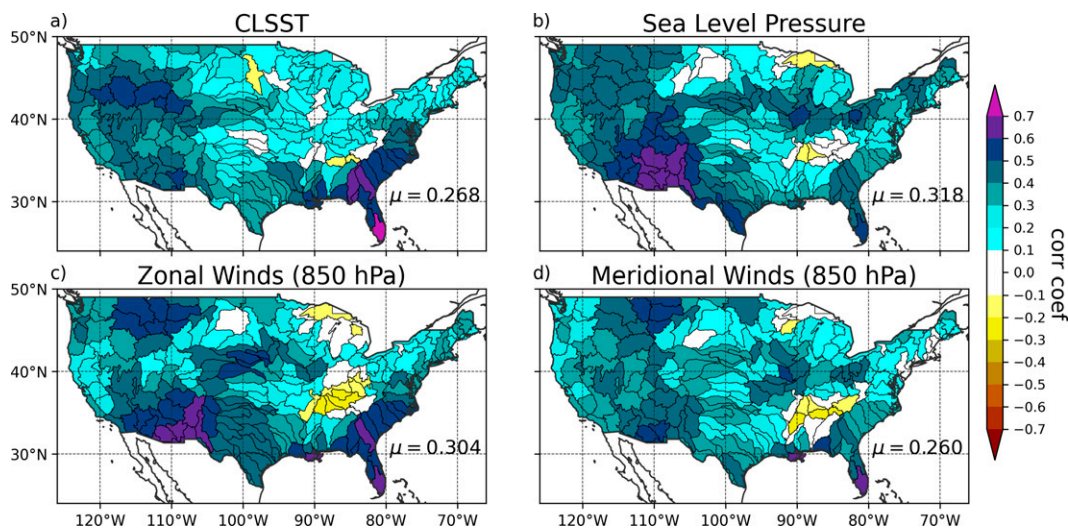


FIG. 7. The skill of the individual models, using 10-fold cross validation, over the period 1982/83–2019/20.

model has an infrastructure that can readily utilize predictors from any lead time. For each region or basin, the SCEF model finds the best predictors from different forecast lead times. For example, the CLSST forecasts, and hence the SCEF, for the northern California region weigh the Niño-3.4 predictor more heavily from 1 year prior than from 1 month prior. This is because the statistical relationship between Niño-3.4 and cool-season precipitation in northern California is stronger at a greater lead time such as 13 months prior. Figure 9 can help us gain some insight as to why that is. In Fig. 9a, we observe the concurrent (November–March) correlation, over the 1901/02–2019/20 period, between the detrended SST field and northern California precipitation. Within the purple box, we observe that conditions are most favorable to having greater than average cool-season precipitation in northern California when there are anomalously warm conditions south of 30°N

and centered between 210° and 240°E, while simultaneously there are anomalously cool conditions north of 30°N and centered between 200° and 230°E. Similarly, with inverse anomalous conditions, one can expect less than normal precipitation, on average. Figure 9b shows the composite difference between concurrent cool-season SSTs conditioned upon El Niño and La Niña events that occurred 13 months prior to the October forecast date. The El Niño and La Niña events were chosen at a threshold so as to have these events make up one-third of the total sample size. We used a threshold where an El Niño or La Niña event was chosen when the standardized anomaly was greater than 0.92. This threshold yields 40 events over the period 1901/02–2019/20. So, for example, we find an El Niño event in September 1997, then the SST field in our concurrent period November 1998–March 1999 will be used to compute our El Niño composite. We find all El

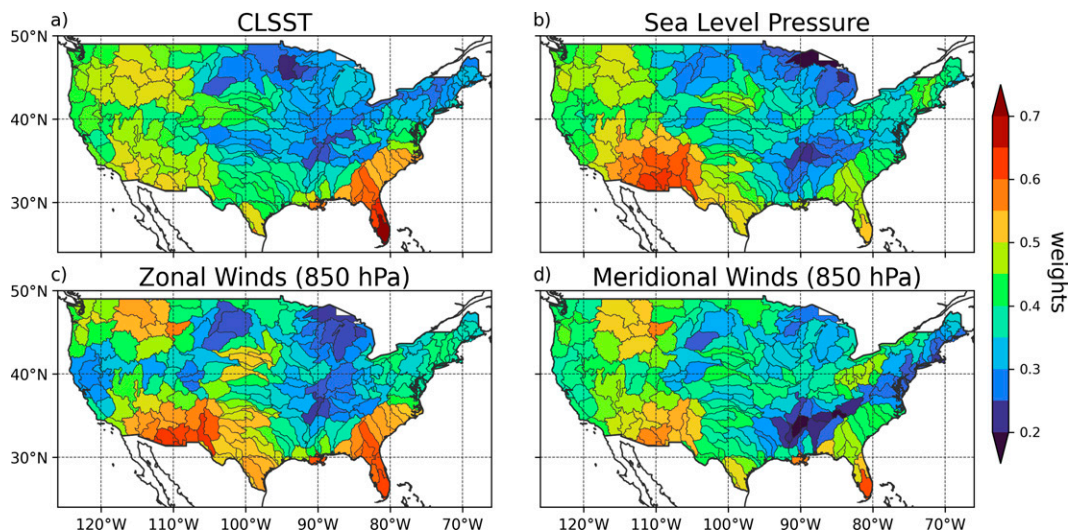


FIG. 8. Model weights at each HUC established over the calibration period.

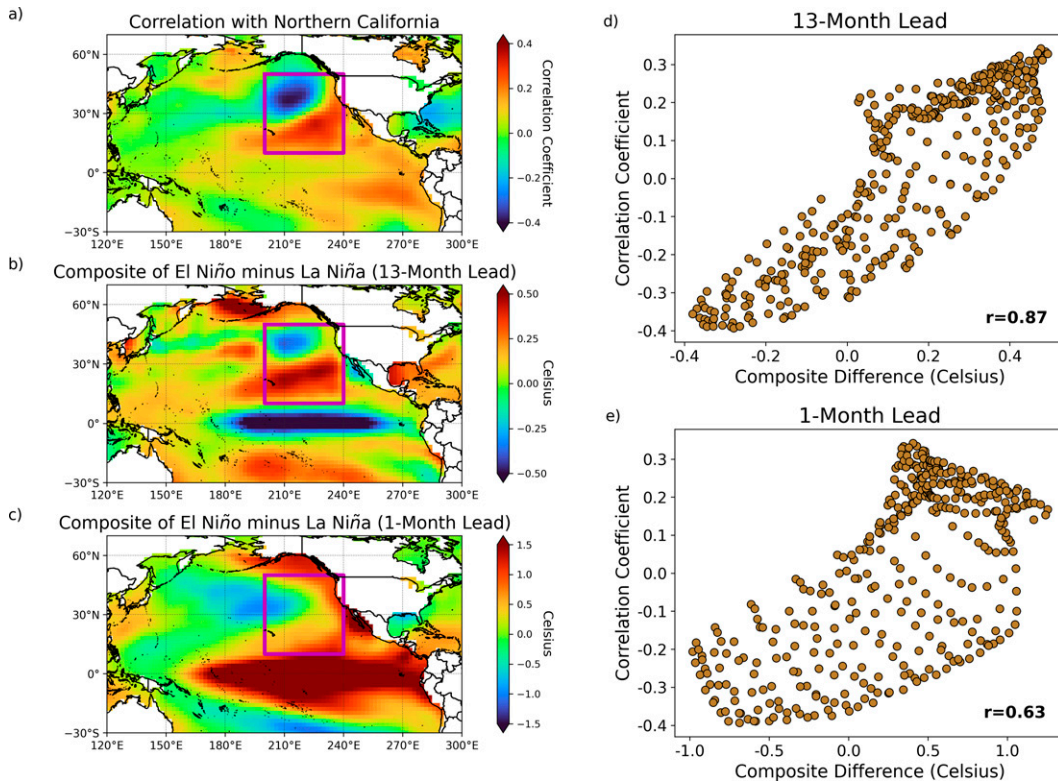


FIG. 9. (a) The concurrent (November–March) correlation between the detrended SST field and northern California precipitation. The composite difference between concurrent cool-season SSTs conditioned upon El Niño and La Niña events that occurred (b) 13 months and (c) 1 month prior to the October forecast date. (d) The composite differences in (b) plotted against the correlation coefficients from (a) for the grid cells that fall within the region outlined by the purple box. (e) Similarly, the composite differences in (c) plotted against the correlation coefficients from (a).

Niño and La Niña events from 13 months prior to our forecast and calculate our composites of the SSTs that follow more than 1 year later in the concurrent cool-season period. Similarly, Fig. 9c shows the composite difference of the SSTs conditioned on El Niño and La Niña events that occurred 1 month prior to our October forecast date. In Fig. 9d, we can observe the composite differences in Fig. 9b plotted against the correlation coefficients from Fig. 9a for the grid cells that fall within the region outlined by the purple box. And similarly, Fig. 9e plots the composite differences in Fig. 9c plotted against the correlation coefficients from Fig. 9a. One can clearly see by the correlations in Figs. 9d and 9e that the development of concurrent SSTs, which are more/less favorable to increased/decreased northern California precipitation, occur more frequently when conditioned upon El Niño/La Niña events with a 13-month lead time (Fig. 9b) than upon events with a 1-month lead time (Fig. 9c).

In Fig. 10, we directly compare the Niño-3.4 index itself, at these different lead times, with cool-season precipitation in northern California. Figures 10a and 10b show the Niño-3.4 index plotted against cool-season precipitation in northern California at 13- and 1-month lead times, respectively. These subplots are for all years (1901/02–2019/20), and we observe greater correlation at the longer lead time. Figures 10c and 10d

show the cool-season precipitation accumulations that occur when conditioned on the ENSO events that we previously defined for Fig. 9. For both lead times, there is an increase in correlation when only considering the stronger ENSO events, and their statistical significance increases as well (i.e., p values decrease) even after accounting for the change in sample size. This can be thought of as a forecast of opportunity, or times in which we can expect greater forecast skill. Figures 9e and 9f show the empirical cumulative distribution functions of the cool-season precipitation accumulations conditioned on these prior ENSO events from 13 months and 1 month, respectively. At these seasonal accumulation time scales, the distributions are approximately Gaussian. As a result, we use a Student’s t test to find that the mean difference of the two distributions is statistically significantly different when conditioned on El Niño and La Niña events from 13 months prior, while we fail to reject the null hypothesis for the precipitation conditioned upon events from 1 month prior.

The second key difference between the SCEF model and NMME or ECMWF-SEAS5 is the fact that the SCEF model does not attempt to perform numerical weather prediction, whereby an entire state system is advanced through a set of equations governed by physical laws. The SCEF model, in its

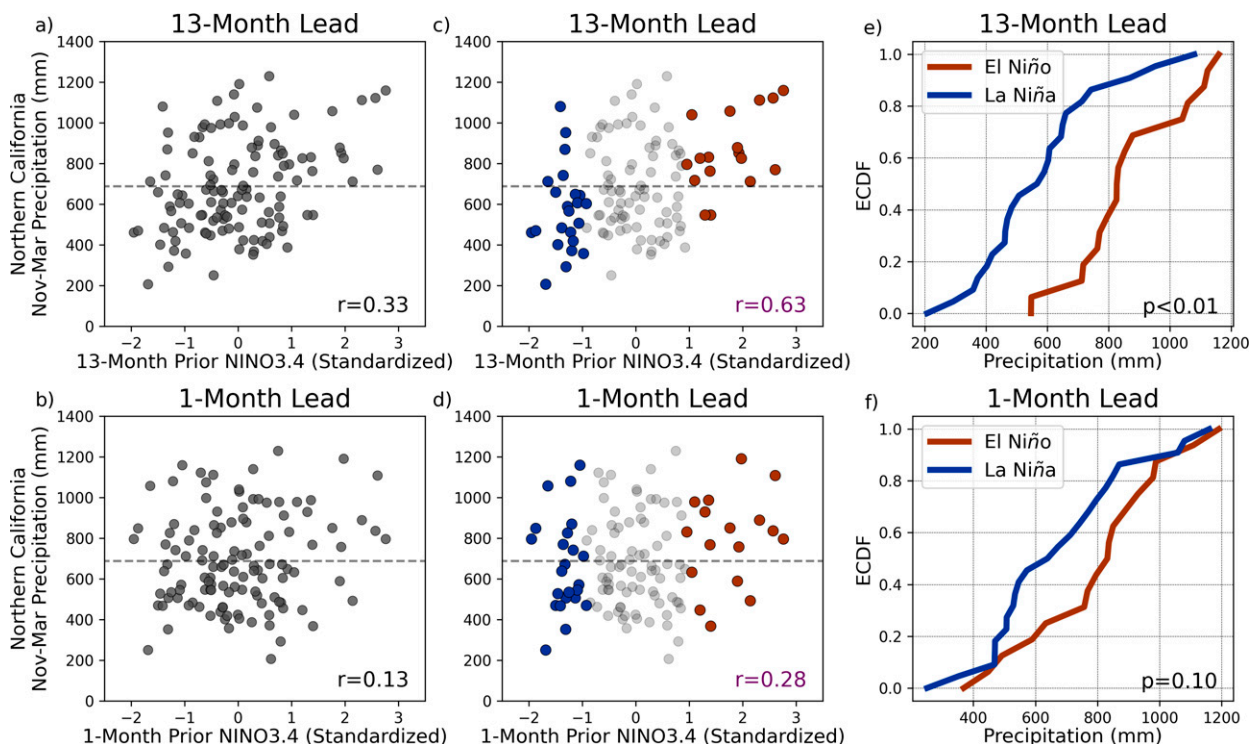


FIG. 10. The Niño-3.4 index plotted against cool-season precipitation in northern California at (a) 13-month and (b) 1-month lead times for all years. The cool-season precipitation accumulations that occur when conditioned on El Niño and La Niña events at (c) 13-month and (d) 1-month lead times. The empirical cumulative distribution functions of the cool-season precipitation accumulations conditioned on these prior ENSO events from (e) 13 months and (f) 1 month.

current format, attempts to exploit and use statistical relationships specifically geared toward predicting cool-season CONUS precipitation. In Fig. 11a, we observe the correlation between sea level pressure values in August–September and the precipitation in the following cool-season in the Puget Sound basin located in Washington State. Figure 11b shows the correlation between sea level pressure values in

August–September and our predicted or forecast time series for Puget Sound. Likewise, Fig. 11c shows the correlation between *u*-component wind values in August–September and the precipitation in the following cool-season in the South Florida Basin, and Fig. 11d shows the correlation between *u*-component wind values in August–September and our predicted or forecast time series for southern Florida. We can

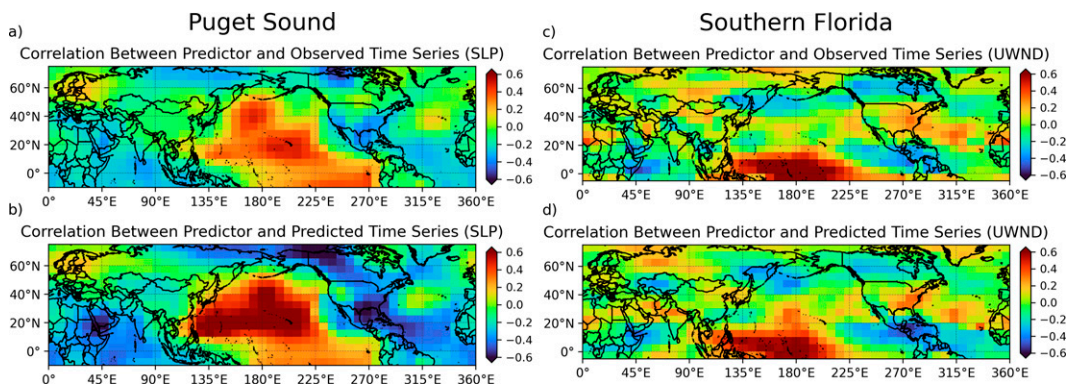


FIG. 11. (a) The correlation between the field of sea level pressure values in August–September and the precipitation in the following cool season in the Puget Sound basin. (b) The correlation between sea level pressure values in August–September and the predicted time series for Puget Sound. (c) The correlation between the field of August–September *u*-component wind values and the following cool-season precipitation in southern Florida. (d) The correlation between the August–September *u*-component wind values and the predicted time series for southern Florida.

observe that the forecast time series is largely picking up on the information from where the historical statistical relationship was strongest. This can be seen by how similar Figs. 11b and 11d look in comparison with Figs. 11a and 11c, respectively. Keep in mind, however, that there is not a perfect overlap between Figs. 11a and 11b (or Figs. 11c and 11d) because 1) we are fitting and applying the SCEF model in data-reduced space (i.e., we use the leading principal components of our predictors and our predictands) and 2) we obtain our forecasts through cross validation. A user of the SCEF model could apply an approach such as what we have presented here to empirically or statistically attribute the origin(s) of the forecasts for a particular basin or region.

7. Conclusions

This paper proposes a new statistical modeling framework, which we have called the Statistical Climate Ensemble Forecast model. The SCEF model is capable of producing more skillful cool-season November–March precipitation forecasts than either the NMME or the ECMWF-SEAS5 models. These improvements in cool-season forecast skill were shown for the validation periods 2000/01–2019/20 and 1982/83–2019/20 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 algorithms (Rasouli et al. 2012; Leng and 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 follows:

- 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.

Through our discussion concerning Figs. 9–11, we have provided greater insight into how the SCEF model is leveraging certain information to achieve improved forecast skill. However, questions still remain, such as those raised by the results of Fig. 6. What explains the skill-level discrepancy between the SCEF model and NMME for the more recent period 2000/01–2019/20 and the prior period 1982/83–1999/2000? 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 and Sardeshmukh 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 are 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 4BM9NCA-P00. The authors do not have any conflicts of interest. Author Switanek conceived of the study, performed the analysis, generated the figures, and wrote the paper. Author Hamill provided supervision and contributed to the writing of the paper. We thank Michael Alexander, Michael Scheuerer, and Joseph Barsugli for their useful comments and feedback.

Data availability statement. The code and data required to run the SCEF model can be found online (<https://github.com/mswitanek/scef-model>).

REFERENCES

- Akaike, H., 1974: A new look at the statistical model identification. *IEEE Trans. Autom. Contr.*, **19**, 716–723, <https://doi.org/10.1109/TAC.1974.1100705>.
- Alley, R. B., K. A. Emanuel, and F. Zhang, 2019: Advances in weather prediction. *Science*, **363**, 342–344, <https://doi.org/10.1126/science.aav7274>.
- Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, <https://doi.org/10.1038/nature14956>.
- Becker, E., H. Van Den Dool, and Q. Zhang, 2014: Predictability and forecast skill in NMME. *J. Climate*, **27**, 5891–5906, <https://doi.org/10.1175/JCLI-D-13-00597.1>.
- Benjamin, S. G., J. M. Brown, G. Brunet, P. Lynch, K. Saito, and T. W. Schlatter, 2019: 100 years of progress in forecasting and NWP applications. *A Century of Progress in Atmospheric and Related Sciences: Celebrating the American Meteorological Society Centennial*, Meteor. Monogr., No. 59, Amer. Meteor. Soc., <https://doi.org/10.1175/AMSMONOGRAPHS-D-18-0020.1>.
- Benjamini, Y., and Y. Hochberg, 1995: Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Stat. Soc.*, **57B**, 289–300, <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>.
- Broxton, P. D., W. J. D. van Leeuwen, and J. A. Biederman, 2019: Improving snow water equivalent maps with machine learning of snow survey and lidar measurements. *Water Resour. Res.*, **55**, 3739–3757, <https://doi.org/10.1029/2018WR024146>.

- Brunet, G., and Coauthors, 2010: Collaboration of the weather and climate communities to advance subseasonal-to-seasonal prediction. *Bull. Amer. Meteor. Soc.*, **91**, 1397–1406, <https://doi.org/10.1175/2010BAMS3013.1>.
- Capotondi, A., and Coauthors, 2015: Understanding ENSO diversity. *Bull. Amer. Meteor. Soc.*, **96**, 921–938, <https://doi.org/10.1175/BAMS-D-13-00117.1>.
- Cayan, D. R., K. T. Redmond, and L. G. Riddle, 1999: Enso and hydrologic extremes in the western United States. *J. Climate*, **12**, 2881–2893, [https://doi.org/10.1175/1520-0442\(1999\)012<2881:EAHEIT>2.0.CO;2](https://doi.org/10.1175/1520-0442(1999)012<2881:EAHEIT>2.0.CO;2).
- Doblas-Reyes, F. J., J. García-Serrano, F. Lienert, A. P. Biescas, and L. R. Rodrigues, 2013: Seasonal climate predictability and forecasting: Status and prospects. *Wiley Interdiscip. Rev.: Climate Change*, **4**, 245–268, <https://doi.org/10.1002/wcc.217>.
- Gubler, S., and Coauthors, 2020: Assessment of ECMWF SEAS5 seasonal forecast performance over South America. *Wea. Forecasting*, **35**, 561–584, <https://doi.org/10.1175/WAF-D-19-0106.1>.
- Guo, Y., M. Ting, Z. Wen, and D. Lee, 2017: Distinct patterns of tropical Pacific SST anomaly and their impacts on North American climate. *J. Climate*, **30**, 5221–5241, <https://doi.org/10.1175/JCLI-D-16-0488.1>.
- Hao, Z., V. P. Singh, and Y. Xia, 2018: Seasonal drought prediction: Advances, challenges, and future prospects. *Rev. Geophys.*, **56**, 108–141, <https://doi.org/10.1002/2016RG000549>.
- Hoell, A., M. Hoerling, J. Eischeid, K. Wolter, R. Dole, J. Perlwitz, T. Xu, and L. Cheng, 2016: Does El Niño intensity matter for California precipitation? *Geophys. Res. Lett.*, **43**, 819–825, <https://doi.org/10.1002/2015GL067102>.
- Huang, B., and Coauthors, 2020: NOAA Extended Reconstruction Sea Surface Temperature (ERSST), version 5. NOAA/National Centers for Environmental Information, accessed 3 February 2021, <https://doi.org/10.7289/V5T72FNM>.
- , C.-S. Shin, A. Kumar, M. L'Heureux, and M. A. Balmaseda, 2021: The relative roles of decadal climate variations and changes in the ocean observing system on seasonal prediction skill of tropical Pacific SST. *Climate Dyn.*, **56**, 3045–3063, <https://doi.org/10.1007/s00382-021-05630-1>.
- Johnson, S. J., and Coauthors, 2019a: SEAS5 data set. Copernicus Climate Data Store, accessed 20 December 2020, <https://cds.climate.copernicus.eu>.
- , and Coauthors, 2019b: SEAS5: The new ECMWF seasonal forecast system. *Geosci. Model Dev.*, **12**, 1087–1117, <https://doi.org/10.5194/gmd-12-1087-2019>.
- Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.*, **77**, 437–471, [https://doi.org/10.1175/1520-0477\(1996\)077<0437:TNYRP>2.0.CO;2](https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2).
- Kirtman, B. P., and Coauthors, 2014a: Hindcast 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 20 December 2020, <https://ftp.cpc.ncep.noaa.gov/International/nmme>.
- , and Coauthors, 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, <https://doi.org/10.1175/BAMS-D-12-00050.1>.
- , and Coauthors, 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 20 December 2020, ftp://ftp.cpc.ncep.noaa.gov/NMME/realtime_anom/ENSMEAN.
- Kumar, A., and M. Chen, 2017: What is the variability in its west coast winter precipitation during strong El Niño events? *Climate Dyn.*, **49**, 2789–2802, <https://doi.org/10.1007/s00382-016-3485-9>.
- Leng, G., and J. W. Hall, 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, <https://doi.org/10.1088/1748-9326/ab7b24>.
- Manzanas, R., M. D. Frías, A. S. Cofiño, and J. M. Gutiérrez, 2014: Validation of 40 year multimodel seasonal precipitation forecasts: The role of ENSO on the global skill. *J. Geophys. Res. Atmos.*, **119**, 1708–1719, <https://doi.org/10.1002/2013JD020680>.
- Mariotti, A., and Coauthors, 2020: Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. *Bull. Amer. Meteor. Soc.*, **101**, E608–E625, <https://doi.org/10.1175/BAMS-D-18-0326.1>.
- Newman, M., and P. D. Sardeshmukh, 2017: Are we near the predictability limit of tropical Indo-Pacific sea surface temperatures? *Geophys. Res. Lett.*, **44**, 8520–8529, <https://doi.org/10.1002/2017GL074088>.
- Nigam, S., and A. Sengupta, 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**, e2020GL091447, <https://doi.org/10.1029/2020GL091447>.
- Power, S., F. Delage, C. Chung, G. Kociuba, and K. Keay, 2013: Robust twenty-first-century projections of El Niño and related precipitation variability. *Nature*, **502**, 541–545, <https://doi.org/10.1038/nature12580>.
- PRISM Climate Group, 2021: Prism gridded climate data. Oregon State University, accessed 10 January 2021, <http://prism.oregonstate.edu>.
- Quan, X., M. Hoerling, J. Whitaker, G. Bates, and T. Xu, 2006: Diagnosing sources of U.S. seasonal forecast skill. *J. Climate*, **19**, 3279–3293, <https://doi.org/10.1175/JCLI3789.1>.
- Rasouli, K., W. W. Hsieh, and A. J. Cannon, 2012: Daily streamflow forecasting by machine learning methods with weather and climate inputs. *J. Hydrol.*, **414–415**, 284–293, <https://doi.org/10.1016/j.jhydrol.2011.10.039>.
- Redmond, K. T., and R. W. Koch, 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, <https://doi.org/10.1029/91WR00690>.
- Ropelewski, C. F., and M. S. Halpert, 1987: Global and regional scale precipitation patterns associated with El Niño/Southern Oscillation. *Mon. Wea. Rev.*, **115**, 1606–1626, [https://doi.org/10.1175/1520-0493\(1987\)115<1606:GARSPP>2.0.CO;2](https://doi.org/10.1175/1520-0493(1987)115<1606:GARSPP>2.0.CO;2).
- Roy, T., X. He, P. Lin, H. E. Beck, C. Castro, and E. F. Wood, 2020: Global evaluation of seasonal precipitation and temperature forecasts from NMME. *J. Hydrometeorol.*, **21**, 2473–2486, <https://doi.org/10.1175/JHM-D-19-0095.1>.
- Scheuerer, M., M. B. Switanek, R. P. Worsnop, and T. M. Hamill, 2020: Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over California. *Mon. Wea. Rev.*, **148**, 3489–3506, <https://doi.org/10.1175/MWR-D-20-0096.1>.

- Seaber, P. R., F. P. Kapinos, and G. L. Knapp, 1987: Hydrologic unit maps. USGS Water-Supply Paper 2294, 66 pp., http://pubs.usgs.gov/wsp/wsp2294/pdf/wsp_2294.pdf.
- Switanek, M. B., J. J. Barsugli, M. Scheuerer, and T. M. Hamill, 2020: Present and past sea surface temperatures: A recipe for better seasonal climate forecasts. *Wea. Forecasting*, **35**, 1221–1234, <https://doi.org/10.1175/WAF-D-19-0241.1>.
- Udall, B., and J. Overpeck, 2018: The twenty-first century colorado river hot drought and implications for the future. *Water Resour. Res.*, **53**, 2404–2418, <https://doi.org/10.1002/2016WR019638>.
- Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. Elsevier, 627 pp.
- , 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.*, **97**, 2263–2273, <https://doi.org/10.1175/BAMS-D-15-00267.1>.
- Wold, S., M. Sjöström, and L. Eriksson, 2001: PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.*, **58**, 109–130, [https://doi.org/10.1016/S0169-7439\(01\)00155-1](https://doi.org/10.1016/S0169-7439(01)00155-1).