Superensemble Statistical Forecasting of Monthly Precipitation over the Contiguous United States, with Improvements from Ocean-Area Precipitation Predictors

THOMAS M. SMITH

NOAA/NESDIS/STAR, and Cooperative Institute for Climate and Satellites, Earth System Science Interdisciplinary Center, University of Maryland, College Park, College Park, Maryland

SAMUEL S. P. SHEN

San Diego State University, San Diego, California

RALPH R. FERRARO

NOAA/NESDIS/STAR, and Cooperative Institute for Climate and Satellites, Earth System Science Interdisciplinary Center, University of Maryland, College Park, College Park, Maryland

(Manuscript received 11 January 2016, in final form 9 June 2016)

ABSTRACT

Extended precipitation forecasts, with leads of weeks to seasons, are valuable for planning water use and are produced by the U.S. National Weather Service. Forecast skill tends to be low and any skill improvement could be valuable. Here, methods are discussed for improving statistical precipitation forecasting over the contiguous United States. Monthly precipitation is forecast using predictors from the previous month. Testing shows that improvements are obtained from both improved statistical methods and from the use of satellite-based ocean-area precipitation predictors. The statistical superensemble method gives higher skill compared to traditional statistical forecasting. Ensemble statistical forecasting combines individual forecasts. The proposed superensemble is a weighted mean of many forecasts or of forecasts from different prediction systems and uses the forecast reliability estimate to define weights. The method is tested with different predictors to show its skill and how skill can be improved using additional predictors. Cross validation is used to evaluate the skill. Although predictions are strongly influenced by ENSO, in the superensemble other regions contribute more to the forecast skill. The superensemble optimally combines forecasts based on different predictor regions and predictor types. The contribution from multiple predictor regions improves skill and reduces the ENSO spring barrier. Adding satellite-based ocean-area precipitation predictors noticeably increases forecast skill. The resulting skill is comparable to that from dynamic-model forecasts, but the regions with best forecast skill may be different. This paper shows that the statistical superensemble forecasts may be complementary to dynamic forecasts and that combining them may further increase forecast skill.

1. Introduction

Accurate short- to medium-range climate forecasting, from weeks to seasons, is valuable for planning and preparing for situations that could have large economic or health impacts. Precipitation forecasting is particularly important because it is critical to agriculture, municipal water supplies and control, and disaster relief support. However, predicting precipitation tends to be more difficult than predicting temperature (see, e.g., Barnston and Smith 1996), and much effort has been made to improve those forecasts. One reason for that difficulty is the smaller spatial and time scales of precipitation. In the extratropics, precipitation is often concentrated along moving weather fronts. In the tropics and the warm-season extratropics, small-scale convective precipitation is common, which can be even more difficult to predict (Stensrud et al. 2000; dos Santos et al. 2013). However, statistics of precipitation such as

Supplemental information related to this paper is available at the Journals Online website: http://dx.doi.org/10.1175/JHM-D-16-0018.s1.

Corresponding author address: Thomas Smith, ESSIC, University of Maryland, College Park, 5825 University Research Ct., Suite 4001, College Park, MD 20740.
E-mail: tom.smith@noaa.gov

DOI: 10.1175/JHM-D-16-0018.1

the monthly average anomalies considered here tend to have larger spatial scales and may be more predictable (Krishnamurti et al. 2002). Monthly precipitation anomaly scales still tend to be smaller than monthly temperature anomaly scales, which are dominated by airmass movements often persistent for weeks to months. The difficulties of the short- to medium-range forecasts are partly due to the nonlinearity and chaotic limitations of the numerical weather prediction models whose forecasts can have good skill out to a week but lose skill at longer leads (e.g., Chen et al. 2013; Vitart 2014). Conventional nonensemble linear statistical prediction also has limited skill at this time scale.

This paper demonstrates how the superensemble statistical methods can improve the precipitation forecast skill for the contiguous United States (CONUS). The goal of this study is to demonstrate improvements of the U.S. monthly precipitation forecasting from two sources. One is the method of ensemble statistical forecasts using multiple predictors and multiple statistical models. These improvements occur because each predictor can best predict different parts of the forecast region. In addition, different statistical models can resolve the forecast relationship in distinct ways. We demonstrate these method-based improvements using sea surface temperature (SST) and land precipitation predictors.

Another improvement is from the use of the satellite-based oceanic-precipitation predictors. Since large-scale CONUS precipitation events over the oceans are connected to large-scale episodes spanning land–sea boundaries, the addition of ocean-area precipitation predictors should be able to improve forecasts. The contribution of individual events is smoothed out by monthly averaging, but monthly forecasts can be helped by the detection of the oceanic-precipitation tendencies, as shown by our results. These results show that oceanic precipitation gives some information for prediction that is independent of the information from SST.

Compared to nonensemble forecasts, the testing shows improved skill from the ensemble methods. In most of our discussion, we use anomaly correlation as a measure of skill, although in section 4c three-category forecast skill is also considered. The ensemble methods include separating predictor data into separate spatial regions and using multiple statistical models and superensemble weighting to combine individual forecasts. Superensembles are formed by weighting individual forecasts or different ensemble prediction systems by their relative reliability (Krishnamurti et al. 1999; Palmer et al. 2004). In the statistical ensembles considered here the superensemble inputs are forecasts from two statistical models that use a range of predictors.

Overall anomaly correlations for short-term precipitation forecasts using these methods are comparable or slightly better than those from a dynamic forecast model for CONUS. That suggests that these methods may be used to supplement and improve forecasts for the United States and to provide more reliable forecasts for other regions where resources may be limited. The next section describes the data used to develop and test the improved forecast methods, followed by a description of the methods and testing done, and a discussion of results.

2. Data

Several satellite- and in situ–based datasets are used for testing and demonstrating improved precipitation forecasting. The Global Precipitation Climatology Project One-Degree Daily (GPCP 1DD; Huffman et al. 2001) data are used for precipitation. Over land GPCP includes both satellite estimates and the available gauges, while over oceans only satellite estimates are available. We use GPCP data so that we can evaluate the impact of oceanic satellite data on predictions. The monthly GPCP begins in 1979, while the GPCP daily data begin in 1997. We use the daily data from the shorter record because its oceanic variations are based on more and better-quality satellite estimates. The daily data used for this study cover the years 1997–2014 (Huffman et al. 2001; Adler et al. 2003). For the monthly forecasts discussed, the daily data are averaged to one-degree monthly values. Monthly precipitation anomalies are computed by removing the annual cycle for the 18-yr period.

In addition to GPCP, one-degree optimum interpolation (OI) SST (Reynolds et al. 2002) is used to help predict precipitation. The OI SST is dominated by satellite inputs, but in situ data are important for correcting large-scale satellite biases and helping to fill consistently cloudy regions. As with the precipitation, SST anomalies are computed by removing the 18-yr annual cycle for the same period.

3. Methods

In this section statistical methods are described, including the statistical models for individual ensemble members, the use of cross validation to tune and evaluate models, and the ensemble statistical (ES) method.

a. Testing and tuning methods

All analyses use data from a fixed period to estimate forecast anomaly correlation skill using cross validation (e.g., Michaelsen 1987). In our cross validation we

construct statistical models that exclude data around the target forecast time. Data excluded extend for at least 3 months before and after the target forecast time to exclude dependent relationships with some persistence. The independent models are used to forecast the target time. This process is repeated for each forecast time. Using cross validation, we can evaluate the skill of each model and the ES combinations of individual models. The forecast skill is evaluated using temporal correlation between the cross-validation forecast anomalies with the withheld GPCP anomalies. For some comparisons, spatial averages of temporal correlation are used. For averaging correlations, regions with negative correlation have no skill and are assigned values of zero correlation.

Preliminary investigations were performed to test the ensemble method and to evaluate the differences between using one anomaly model for all seasons and separate anomaly models for different seasons. This initial testing used linear regression to predict the CONUS precipitation from individual climate-mode indices. Six climate-mode indices were obtained from the National Oceanic and Atmospheric Administration (NOAA) Climate Prediction Center (CPC), including the Southern Oscillation index (SOI); Niño-3.4 SST anomaly; SST anomaly indices for the North Atlantic, South Atlantic, and tropical Atlantic; and the North Atlantic Oscillation (NAO). In addition, the Mantua et al. (1997) Pacific decadal oscillation (PDO) monthly index was used.

Using these indices as predictors, we considered cross-validation regression forecasts using anomalies from all seasons and cross-validation regressions from individual 3-month seasons. Several regressions yielded noticeable skill, but there was no advantage to using separate regressions for different seasons. The all-month model has many more months available for model development, which may offset disadvantages from mixing anomalies from different seasons. Based on these initial results we use anomalies from all months in later model development.

For the rest of this study, we consider forecast models using both joint empirical orthogonal function (JEOF) analysis and canonical correlation analysis (CCA), described in the following subsections. Both the JEOF and CCA decompose predictors and predictand into spatial modes. Most prediction variance is accounted for by the leading modes, and including higher modes can increase noise. The number of modes to use is a model selection problem in statistical theory or communication engineering. Akaike information criterion (AIC) is sometimes used for the model selection (Burnham and Anderson 2002). Here, cross validation is used to determine the optimal number of modes for each.

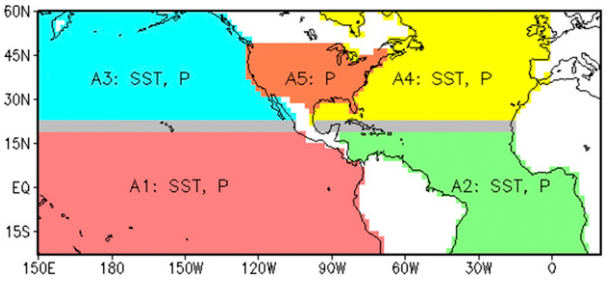

FIG. 1. Predictor areas using different colors for the five areas: tropical Pacific (A1), tropical Atlantic (A2), North Pacific (A3), North Atlantic (A4), and CONUS (A5). Gray shading in 20°–23°N indicates the regions where tropical and extratropical areas overlap. Both SST and precipitation predictors (i.e., P) from time $t - 1$ are used for oceanic regions. The CONUS predictor region uses precipitation from time $t - 1$ as a predictor.

b. JEOF analysis

Empirical orthogonal function (EOF) analysis (e.g., Davis 1976) is a statistical method for decomposing space–time data for a climate parameter, such as precipitation, into a set of spatial EOF modes weighted by associated time series, called principal components (PCs). Both EOFs and PCs for different modes are orthogonal, so each mode represents statistically independent variations. The JEOF analysis decomposes the space–time data for more than one climate parameter, for example, precipitation and temperature. Mathematical formality of the JEOF is the same as the conventional EOF, but a JEOF shows spatial variations associated with the multiple climate parameters, and the same can be said for the joint PCs. To compute a JEOF, the different climate fields are normalized and then stacked together and the EOF analysis is performed on the combined normalized fields. After the forecast is made, the normalization can be removed.

An analysis for a global region may require some statistical compromise to maintain orthogonality. Analysis using a smaller region can limit the compromise needed and better express relationships with fewer modes. For the tests described here there are two fields. One is the predictand $P_f(t + 1)$ which is the space–time field of precipitation anomalies for the CONUS. The other is the predictor fields for the previous month, which can be SST for one of four regions $\mathrm{SST}_k(t)$ or precipitation for the same regions or for the U.S. area $P_k(t)$. We construct separate JEOF analyses for each predictor field and $P_f(t + 1)$, for each of the k predictor regions, giving a number of different forecasts for U.S. precipitation anomalies. Among the k predictor regions are the ocean regions of the tropical Pacific, the tropical Atlantic, the North Pacific, and the North Atlantic (Fig. 1). Tropical regions extend over 23°S–23°N and extratropical regions extend over 20°–60°N. They

overlap in 20°–23°N, but are otherwise separate. These regions reflect variations in the climate indices considered for preliminary testing and are roughly the same four predictor regions used by Lau et al. (2002), who showed that the ensemble CCA can improve forecast for the United States and minimize the spring barrier, which can cause lower skill for U.S. precipitation prediction in March and April. There are four JEOF forecasts based on the previous month's SST, one for each ocean region. There are five precipitation-predictor areas tested, the four ocean regions and the U.S. area for the previous month. Cross-validation testing shows that JEOF correlation is highest for these predictor regions when about five modes are used.

c. *Canonical correlation analysis*

Barnett and Preisendorfer (1987) developed the CCA method in EOF-based spectral space and demonstrated the predictability of the U.S. area monthly and seasonal temperature based on the global SST field. The CCA predictor and predictand fields are both first decomposed using EOFs to remove noise that could contaminate the forecast and to simplify computation. Since Barnett and Preisendorfer (1987), the CCA has been used many times for prediction and climate analysis. The CPC has used CCA as one of its operational long-term climate prediction methods since 1990s (Barnston and Ropelewski 1992). The CCA also has potential for improving U.S. dynamic climate forecasts by correcting systematic model errors (Smith and Livezey1999). Shen et al. (2001) developed the method of ensemble CCA (ECCA) to predict the U.S. precipitation. Lau et al. (2002) divided the global ocean into different basins based on the climate dynamic analysis, applied the ECCA method to predict monthly and seasonal U.S. precipitation, and showed noticeable correlation improvement compared to other methods and success in overcoming the spring barrier. Mo (2003) introduced the ECCA method to CPC to improve the U.S. temperature, and CPC later adapted the ECCA method as one of its six operational seasonal forecast tools.

As with the JEOF, the CCA forecasts are computed separately for each predictor region and data type, for a total of nine CCA forecasts of U.S. area precipitation. Cross-validation tests show that the CCA correlation is highest when about 20 modes are used for these predictor regions. We use both JEOF and CCA forecasts based on the same predictors for two reasons. One is to test which method yields the best overall skill. The other is to see if combining the two methods yields a better forecast. Although we may expect that JEOF and CCA will give similar results because they are similar linear models, there are differences. The CCA optimizes correlation between predictor and predictand fields, while the JEOF optimizes the explained variance of the combined predictor–predictand joint field. Because of those differences, and because the CCA prefilters data using EOFs, their forecasts will be slightly different and testing them both is justified.

d. *Superensembles*

Superensemble forecasts are computed using the individual statistical forecasts. As discussed above, there are four ocean predictor areas (Fig. 1). For each ocean area a statistical model is developed using both SST and ocean precipitation to predict the CONUS precipitation for the next month. For each model, JEOF and CCA, there are four SST-based models and four ocean-precipitation-based models. In addition, the U.S. precipitation itself is used as a predictor for U.S. precipitation the next month. Thus, there are a total of nine possible JEOF predictions and nine possible CCA predictions for the next month, illustrated in Fig. 1. The superensemble finds the optimal combination of the models at each spatial location for each calendar month.

Ensemble forecasts are a weighted average of multiple forecasts, which tends to damp inconsistent results that may be more from chance than from model skill. The ensemble statistical method was developed by Shen et al. (2001) and the value of the method was further demonstrated by Lau et al. (2002) and Mo (2003). The basic idea behind the ensemble method is that combined results from many statistical models can yield a result superior to any of the individual forecasts and also better than may be expected from a nonensemble statistical model using the same predictor data. A single statistical model using all predictors together can lead to damping at some locations where different predictor regions have inconsistent relationships.

In some ensembles, results from different models are given equal weight. In a superensemble, individual models or forecast systems are weighted differently, depending on the reliability of their forecast (Krishnamurti et al. 2000). Our superensemble assumes that models are not all equally good for every forecast target region and that stable statistics can be computed for the unequal weighting of ensemble members.

Here, separate superensemble weights are computed for each forecast at each location and each calendar month. The optimal weights are computed assuming that the individual statistical forecasts are unbiased. An assumption of little or no systematic error, or bias, is reasonable for anomaly forecasts. Any systematic errors in the forecast may be retained in the superensemble. However, systematic errors are not apparent in the results and the assumption seems justified. To estimate

optimum weights, we use a method similar to OI. If we apply OI to a point, we may assume that spatial correlations between inputs are all 1. Since data are assumed to be unbiased, all errors are random. In that case the weights for each of the k estimates are proportional to

$$\frac{1}{1+\eta_k^2}. \tag{1}$$

Here, η_k^2 is the noise/signal variance ratio for estimate k, defined as the ratio of random-error variance for predictor k to the variance of the error-free signal.

To further simplify the estimate of the relative weights, we may express each model estimate as

$$v_k = v + \varepsilon_k, \tag{2}$$

where v is the error-free variable and ε_k is the unbiased random error, which may be different for each predictor k. Note that the random error is uncorrelated with v. If we define the error-free signal variance as σ^2, then the variance of the estimate can then be expressed as $\sigma_k^2 = \sigma^2 + \langle \varepsilon_k^2 \rangle$. Here $\langle \varepsilon_k^2 \rangle$ is the noise variance and the angle brackets denote averaging. Because the error is uncorrelated with zero mean, the covariance between v_k and v reduces to simply σ^2 and the correlation between the two reduces to σ/σ_k. Squaring the correlation gives

$$r_k^2 = \frac{\sigma^2}{\sigma^2 + \langle \varepsilon_k^2 \rangle}. \tag{3}$$

Multiplying the numerator and denominator of the right-hand side of (3) by $1/\sigma^2$ yields

$$r_k^2 = \frac{1}{1+\eta_k^2}. \tag{4}$$

Thus, the correlation squared is proportional to the optimal weight for each estimate. To avoid including forecasts with no skill for a region, values of $r_k < 0$ are set to 0.

To avoid damping of the weights, we can define $S = \sum_{k=1}^{N} r_k^2$ and define the optimal superensemble weight for each estimate as $w_k = r_k^2/S$.

The monthly cross-validation maps of correlation for each individual forecast are used to compute the weights as a function of space and calendar month, with negative correlation set to zero. Since the cross validation excludes data from the forecast time, the cross-validation correlations are independent of the forecasts and representative of the correlations that would be used in actual forecasts. Because the correlations are independent of the forecast time, they may be used to optimally combine the forecasts without overfitting. We demonstrated the stability and independence of the weights by comparison to double-cross-validation weights (discussed in the supplemental material).

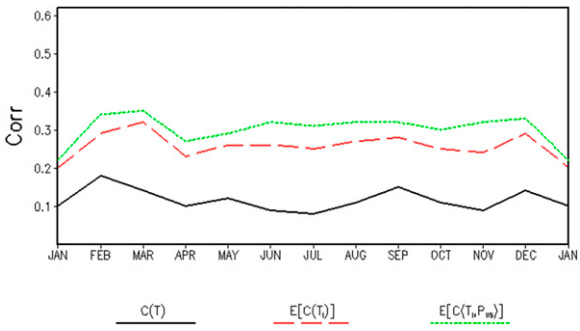

FIG. 2. Spatial averages of temporal cross-validation correlation from CCA prediction, using: all SST together $C(T)$, an ensemble of CCAs using the same SST divided into four regions $E[C(T_i)]$, and an ensemble that also adds U.S. precipitation predictors $E[C(T_i, P_{US})]$.

4. Results

Temporal correlations computed from cross-validation anomalies are used to evaluate the different forecast models. Both spatial averages over the U.S. area and maps of correlations are used for the evaluation. Correlations are computed using the 18-yr record for each month. Thus, for January correlations the 18 Januaries are used, and for each month a separate correlation is computed to evaluate the seasonal cycle of anomaly correlation. Additional evaluations are done using three-category validation and time series of cross-validation forecasts and validation data for several regions.

a. Average correlations

First, the value of the ensemble method is shown using spatial-average correlations of different CCA forecasts of the CONUS precipitation anomalies (Fig. 2). The nonensemble forecast is computed using SST from all regions together in one CCA, here called $C(T)$. For every month of the year $C(T)$ has systematically lower correlation than the superensemble CCA forecast using the same SST predictors divided into the four regions $E[C(T_i)]$. For the 18-yr record, a 95% significant correlation is 0.4 or higher. Here, the significance is estimated using a directional Student's t test with $n-1$ degrees of freedom, computed using online VassarStats tools (http://vassarstats.net/rsig.html). By that standard, the $C(T)$ forecast is significant over 7% of CONUS, averaged over all months of the year. The average correlation is much lower than the significant correlation because of the presence of many areas with much lower

correlation. By contrast, the $E[C(T_i)]$ forecast is significant over 22% of the region, on average. The $C(T)$ average U.S. correlation is comparable to the zero-lead CCA precipitation forecast average North American correlation from Barnston and Smith (1996). They used a longer record of global SST to evaluate lag relationships with precipitation. Although that area-average correlation is low, it can be useful for forecasting some parts of the region. This test shows that dividing the predictor region and forming a superensemble of multiple forecasts can more than double the area-average correlation and greatly expand the area with significant correlations. Next, the superensemble can be expanded using the prediction that uses the forecast-area precipitation from the prior month $E[C(T_i, P_{US})]$. That adds additional correlation skill in every month and increases the region with significant correlations to 31% of the total. These comparisons show that the ensemble method is clearly better than the nonensemble method and that adding additional predictors has the potential for increasing skill.

The annual spatial average of the monthly $E[C(T_i)]$ correlation is 0.26, while the average $C(T)$ correlation is 0.12 and the CCA for the tropical Pacific SSTs yields an average of 0.10. That suggests the dominance of El Niño–Southern Oscillation (ENSO) variations on $C(T)$. The monthly correlation for $C(T)$ is highest in February (Fig. 2), when correlation based on the tropical Pacific and the North Pacific SST subareas are also strongest. In March, the correlation based on the tropical Pacific is much lower, reflecting the spring barrier in ENSO-related predictions. The ENSO variations are stronger than those from other regions, leading to its dominance of the $C(T)$ correlation. The $E[C(T_i)]$ correlation is highest in March, when correlations based on the tropical and North Atlantic subregions are highest. In the superensemble, the March predictions from the tropical Pacific are given lower weight because of their lower skill in that month, and predictions from other regions are allowed to contribute more. It may be argued that, given enough modes, the $C(T)$ correlation could approach the $E[C(T_i)]$ correlation. In practice that does not occur, perhaps because the record length for computing relationships is limited and the data that the relationships are based on may contain errors. Cross-validation testing showed that including more modes does not increase CCA skill. The superensemble method overcomes these practical limitations on the $C(T)$ model.

Ensemble forecasts using SST from four regions and P_{US} are used to compare correlation skill from ensembles using both CCA and JEOF forecasts. Again, spatial averages are used to make comparisons (Fig. 3). For most months the ensemble of JEOF forecasts $E[J(T_i, P_{US})]$ has

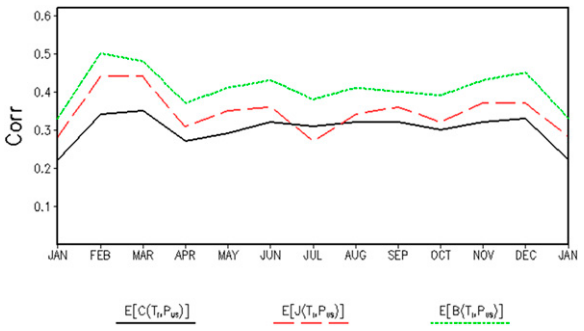

FIG. 3. Spatial averages of temporal cross-validation correlation from ensemble prediction SST divided into four regions and U.S. precipitation predictors, using: CCA $E[C(T_i, P_{US})]$, JEOF $E[J(T_i, P_{US})]$, and the ensemble of both CCA and JEOF models $E[B(T_i, P_{US})]$.

slightly higher average correlation than the ensemble of CCA forecasts. However, the ensemble that combines both CCA and JEOF forecasts using the same predictors $E[B(T_i, P_{US})]$ has systematically higher correlation than the ensembles using only one type of model. This shows that CCA and JEOF models do not resolve identical variations. Examination of individual correlation maps shows that the regions of highest correlation are not the same for both CCA and JEOF. Since the superensemble weights are a function of both month and spatial region, the ensemble allows the best forecast for each region to dominate.

Another potentially valuable predictor is ocean-area precipitation. Using both CCA and JEOF, forecasts are produced using ocean-area precipitation for the same four SST ocean regions, and those forecasts are added to the ensemble. With both CCA and JEOF models, the ensemble $E[B(T_i, P_i, P_{US})]$ has eighteen members. Comparisons show that adding oceanic-predictor members systematically increases the average correlations (Fig. 4). The annual average correlation without oceanic precipitation is 0.42, while with it the average is 0.50. The $E[B(T_i, P_{US})]$ forecast correlation is significant over 55% of the forecast region, on average, while the $E[B(T_i, P_i, P_{US})]$ forecast correlation is significant over 74% of the region, a valuable improvement in the significant area.

An average correlation of 0.50 for precipitation prediction is good, even for the short lead discussed here. However, the cross-validation correlations are computed over a relatively brief period. For each month there are only 18 years of cross-validation estimates. Although cross validation removes dependent information from the analysis statistics, it is possible that the number of important climate episodes in the period could bias the correlation estimate. For example, if there is good predictability associated with ENSO and the

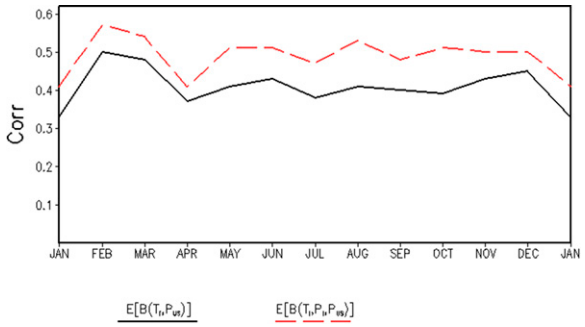

FIG. 4. Spatial averages of temporal cross-validation correlation from ensemble prediction using both CCA and JEOF models. One uses SST divided into four regions and U.S. precipitation predictors $E[B(T_i, P_{US})]$, and the other also includes oceanic-precipitation predictors $E[B(T_i, P_i, P_{US})]$.

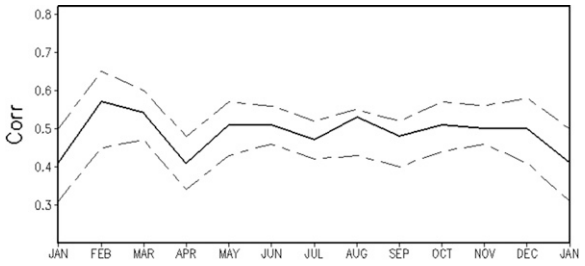

FIG. 5. Spatial averages of temporal cross-validation correlation from ensemble prediction using both models for all predictors $E[B(T_i, P_i, P_{US})]$. The solid line is the cross-validation estimate using all data and the dashed lines show the bootstrap confidence interval.

number of those episodes in the validation period is unusual, then the cross-validation correlation may not be representative of other periods. To test that sampling representativeness, for each month we apply bootstrap sampling to the 18 years of cross-validation estimates to evaluate how sampling of climate episodes may influence the correlation [see, e.g., Efron and Gong (1983) for a description of the bootstrap method].

For the bootstrap sampling, for each month, we randomly select 18 cross-validation forecasts and validation pairs from the full set of estimates. A random number generator is used to select years, with duplicates as needed to get a sample of 18 pairs. The randomly selected pairs are used to compute correlation, and the process is repeated 1000 times. Using those 1000 correlation samples, the 5th and 95th percentiles are used to define confidence intervals. For the U.S. average correlation of the best model $E[B(T_i, P_i, P_{US})]$, in each month the cross-validation average correlation varies by less than ± 0.10 for nearly all samples (Fig. 5). The spread is slightly smaller in the warm season, and the average confidence interval is about ± 0.07. For the annual average of the monthly correlations, the 5th and 95th percentiles are 0.42 and 0.56, respectively. A longer record with more sampling of climate episodes could yield more reliable correlations, but this test suggests that the main conclusions are unlikely to change. Using that confidence interval to compare area-average correlations from $E[B(T_i, P_{US})]$ to those from $E[B(T_i, P_i, P_{US})]$ suggests that in almost every month the increase in average correlation for $E[B(T_i, P_i, P_{US})]$ is significant, and the increase is more significant in the warm season.

b. *Monthly correlation maps*

Correlation maps show where the forecast tends to be reliable for each month and where and when it may not be useful. In considering anomaly correlation maps it is useful to compare them with anomaly standard deviation maps (Fig. 6) of four months: January, April, July, and October. Since regions with low standard deviation are typically close to climatology, the forecast correlation skill of those regions is less important. Anomaly correlation skill is more important in regions with high standard deviation. In much of the west, standard deviation is low, except on the West Coast in the cool season. High values also occur along the Gulf of Mexico coast and parts of the Northeast, especially in the cool season.

The correlation skill maps (Fig. 7) show that in January correlation is good on most of the West Coast, greater than 0.5, indicating that the improved forecast could help forecasting for the wet season in that region where standard deviation is high. But the high correlation in January does not extend to Southern California, where there is also high standard deviation. January correlation is also high on the southern part of the Gulf Coast, although it is lower just north of the coast. Standard deviation is high in both regions. There are large regions of low correlation, less than 0.2, including the Northeast. Useful correlation in April is in many places an extension of the January skill. In April correlation in the Gulf region is improved and expanded northward into areas with high standard deviation. Such skill could be useful for agriculture because it is in an important farming region and is early in the warm growing season. Improved estimates of rainfall for the month could improve planning for the growing season, including issues such as the timing of planting and how much water will be needed for irrigation. However, the Northeast, where standard deviation is also relatively high, has almost no skill in April.

For July correlation is weaker in the much of the Gulf of Mexico region but is still good along the Texas coast where standard deviation is high. Correlation is also relatively high southwest of the Great Lakes, an important region for agriculture with high standard

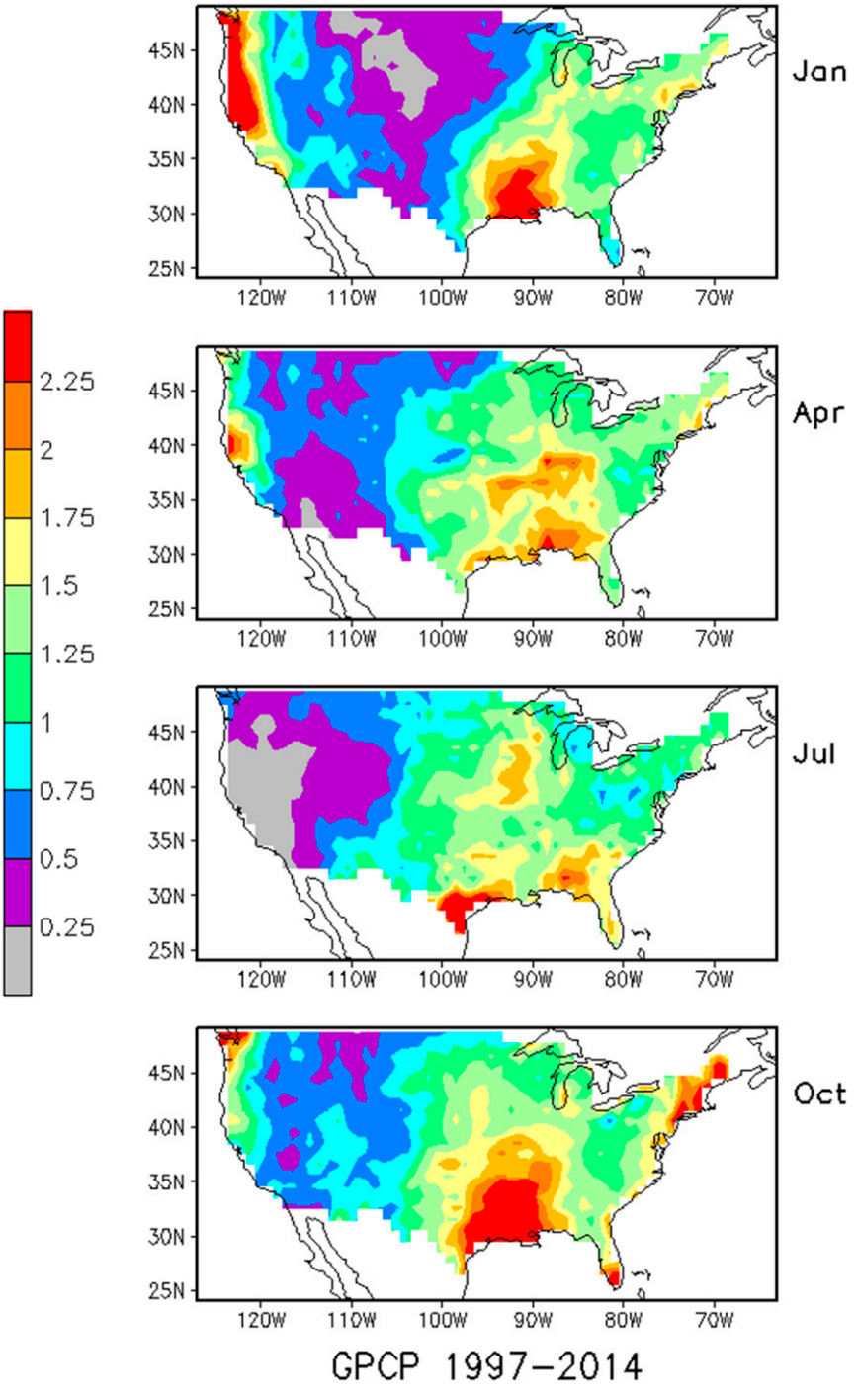

FIG. 6. Anomaly standard deviation for the indicated month from GPCP data.

deviation. In October correlation skill is high in the Gulf and extending north, where standard deviation is high. However, skill for the Northeast is low, in a region that also has high standard deviation.

The superensemble method allows any number of models to be included in the ensemble, as long as their relative skill can be estimated. For example, the National Multi-Model Ensemble (NMME) of dynamic models (Mo and Lettenmaier 2014) yields average correlation skill similar to that from the statistical superensemble. In the NMME, skill is about as good as the skill from the best model in that ensemble, but better

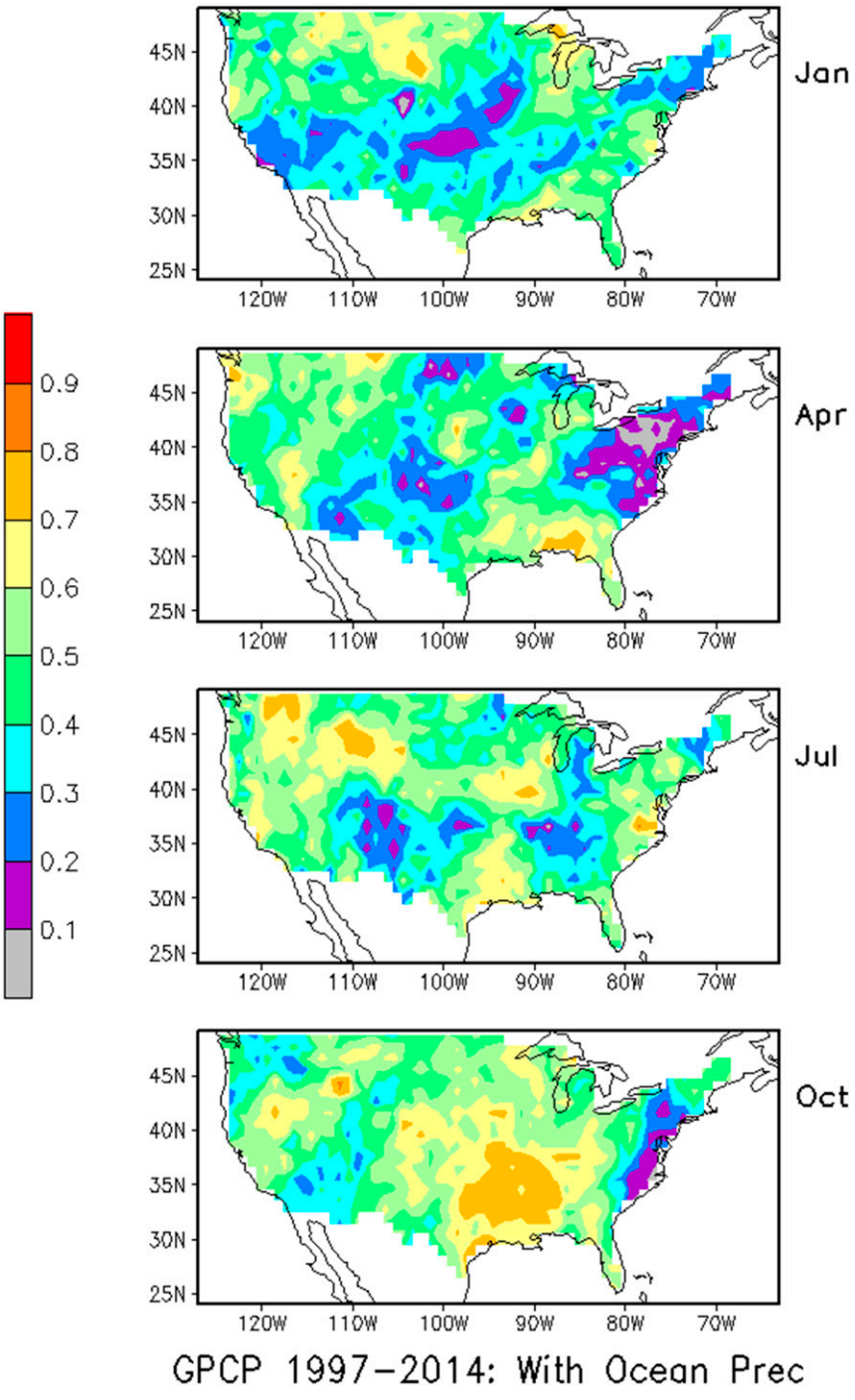

FIG. 7. Anomaly forecast correlation for the indicated month for the $E[B(T_i, P_i, P_{US})]$ model. Correlations above 0.4 are statistically significant at the 95% level.

than the skill from the worst model. No single model in NMME is consistently best, so the ensemble helps to find the highest skill. Compared to the statistical superensemble, the NMME highest correlation skill is often in different regions, such as Southern California in January and the Northeast in April. Thus, including those NMME predictions in an expanded superensemble with statistical models could further improve results. There is no guarantee that including more models will increase forecast skill (Kumar et al. 2001;

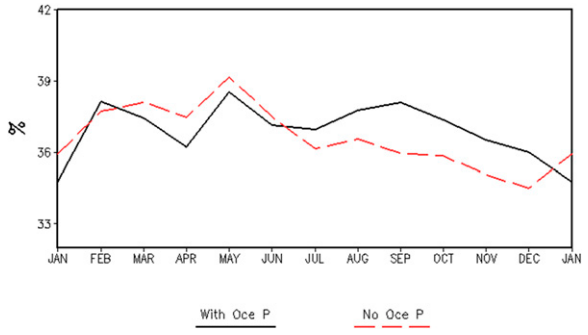

FIG. 8. The percent area where the forecast falls into the correct third for the indicated forecasts.

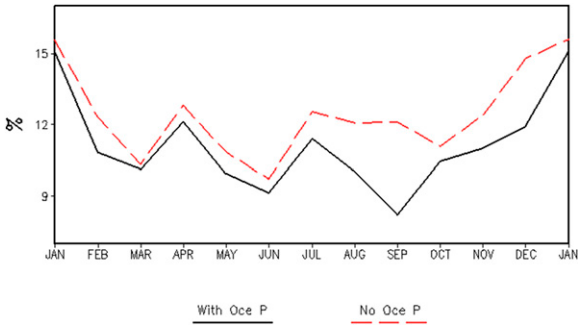

FIG. 9. The percent area where the forecast misses the correct third by two categories for the indicated forecasts.

Weigel et al. 2008). If the additional models only have high skill in the same place as existing models then they may not add useful information. But when an additional forecast has skill in regions where the existing forecast has low skill, the superensemble method described here will recognize and use the complementary skill.

c. *Three-category validation*

Short-range climate forecasts are often presented as the chance that conditions will fall into one of three categories: above normal, normal, or below normal (e.g., see forecast produced by CPC at http://www.cpcpara.ncep.noaa.gov/). Here we use the 18-yr record for each month to define the lower, middle, and upper third of precipitation anomalies and to evaluate how well the forecast predicts the correct category. For comparison, ensemble forecasts without and with oceanic-precipitation predictors are compared to show how the oceanic precipitation impacts the superensemble forecasting of the correct category.

Several comparisons are used to broadly evaluate the impact of oceanic predictors on hits and bad misses using three categories. The percent of hits is one comparison, defined as the forecast area where forecasts fall within the correct category. With random guessing, about 33% of the area would fall in the correct category. Another measure of interest is the percent of forecast bad misses, defined as the percent of the area where the forecast misses the correct category by two categories. A bad miss is counted if above average is forecast and below average occurs, or if below average is forecast and above average occurs. With random guessing bad misses would occur about 22% of the time. For each comparison the combined CCA and JEOF forecasts are used to form the superensemble, without and with oceanic-precipitation predictors. The percent of forecast area with the correct tercile (Fig. 8) is always above 33%. The percentage of hits is also higher in the warm season when it is near 38%. Including oceanic-precipitation predictors does not help forecast the correct category in the first half of the year. This shows that for forecasting the correct tercile, adding more predictors is not always useful. The value of the oceanic-precipitation predictors is clearer in the second half of the year when excluding it causes the correct category skill to drop. The relatively low hit scores in Fig. 8 shows the inability of forecasting extremes using either CCA or JEOF, a common problem of statistical forecasting methods.

The advantage of including oceanic-precipitation predictors is clearer in the percent of area that misses the correct tercile by two categories (Fig. 9). The percent of area with bad misses is always low, roughly 8%–15%, and including oceanic precipitation always reduces the percent of bad misses. Examination of area-average correlation for individual forecasts indicates that the oceanic-precipitation forecasts typically have higher correlations than SST forecasts when the JEOF prediction model is used with tropical Pacific and Atlantic predictors. For extratropical predictors using the JEOF model and for all regions using the CCA model, the average correlations from oceanic-precipitation predictors is about the same or less than from SST predictors. These comparisons suggest that the JEOF model makes better use of tropical oceanic-precipitation information that is partly independent of the SST prediction information. The lower CCA skill from those regions may be due to its smoothing of predictors and predictands using EOFs. In any case, it appears that the JEOF model is responsible for much of the decrease in bad misses when oceanic precipitation is added. Bad misses can have a large impact on forecast confidence. For example, if above average is predicted and normal occurs, it may not be a disaster for agriculture, but if below average occurs it could be a much larger problem. Therefore, including oceanic-precipitation predictor forecasts is justified.

d. *Spatial averages of forecasts and validation*

Averages of cross-validation forecast and the validation data are compared for several regions. For simplicity

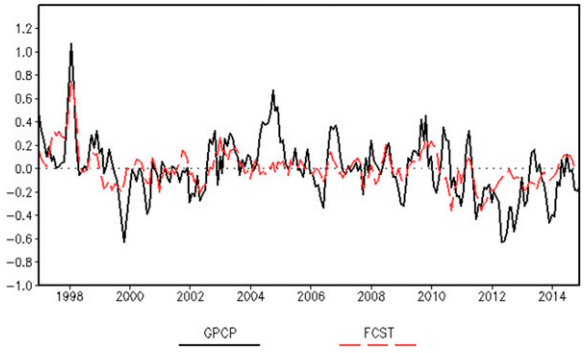

FIG. 10. Monthly forecast and GPCP validation anomalies averaged over the CONUS.

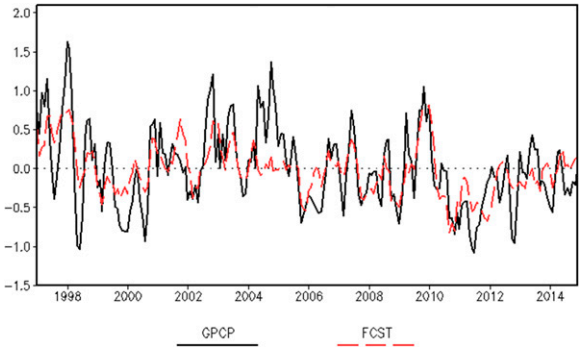

FIG. 11. Monthly forecast and GPCP validation anomalies averaged over the CONUS south of 35°S.

we use area averages of the best monthly forecasts, the superensemble using all predictor inputs for both CCA and JEOF models. The monthly spatial averages are smoothed with a 3-month running mean to damp subseasonal variations and make comparisons easier.

Averages over the entire forecast area (Fig. 10) show that the forecasts tend to predict large ENSO variations, as in early 1998, and it also tends to capture much of the lower-frequency variations in the period. However, there are times when the forecast fails, such as in 2004 and 2012. When the forecast fails it usually damps the anomaly toward zero, although there are times when it is stronger, such as in 2002 and 2010. The southern U.S. precipitation is more strongly impacted by ENSO and the forecast has higher skill in that region. Over the southern United States, anomaly variations are larger and the forecast clearly resolves much of the interannual and lower-frequency variation, although there are still times when the forecast fails (Fig. 11). Forecast failures tend to occur when there is no dominant climate process occurring, such as when there is no ENSO episode. When a strong ENSO occurs, it strongly influences climate anomalies over the forecast region, making forecasting more reliable. In the absence of a dominant process, a combination of weaker processes contribute to climate anomalies. The statistical models are less skillful at forecasting weaker processes and how they may interact to cause climate anomalies, accounting for forecast failures. In 2004/05 there was a modest-intensity central-Pacific warm ENSO episode, but the forecast failed to predict heavier-than-normal precipitation in the southern United States associated with the episode. Using a dynamic model, Garfinkel et al. (2013) showed that central- and eastern-Pacific warm ENSO episodes can have similar teleconnections in the cool season, but there is some variation in the responses, making the sample size important. The index of Kao and Yu (2009) indicates some extended periods of central-Pacific warmth in our base period, but they may not be sufficient for fully resolving the relationships statistically and a longer base period may be needed.

There are many more regional comparisons that could be useful for different applications, but the results of these examples are typical of what we found in evaluating other regional averages. Although these forecasts have good skill for precipitation, there are many variations that they do not resolve and there is room for improving the ensemble. In particular, the addition of models that resolve variations unrelated to dominant processes like ENSO could improve the superensemble forecast. For statistical models it may be possible to find smaller predictor subareas that better resolve weaker relationships, and it may be possible to improve forecasting of weaker relationships using additional predictors, such as total precipitable water. However, a longer base training period may also be needed to resolve the weaker climate relationships using statistical models. Another way to better resolve weaker relationships may come from dynamic models or ensembles of dynamic models. Such forecasts could be incorporated in a superensemble along with statistical model forecasts. Since dynamic forecasts are fundamentally different from statistical forecasts, it is possible that they may contain significant independent information.

5. Summary and discussion

We show that a superensemble of statistical models can improve the correlation skill of short-period precipitation forecasts over CONUS. We do this using cross-validation testing of precipitation for each month using predictors from the previous month. Including models that use oceanic-precipitation predictors further improves the skill, raising the average correlation to 0.5. This shows the value of the satellite-based oceanic precipitation for short-term precipitation forecasts.

The individual statistical models used in the ensemble are inexpensive, which makes it possible to develop and test models for many more regions using an expanded set of predictors. There is no limit on the number of models that may be included in the superensemble forecast, which weights individual model forecasts by their relative skill. In addition, both statistical and numerical model results can be included in a superensemble forecast in cases when both are available.

As Kumar et al. (2001) discuss, adding an unlimited number of models to an ensemble does not guarantee continued improvement of skill. If a superensemble is augmented with another run of the same type of model using the same kind of predictors, then it may not resolve anything new about the prediction relationship. However, if an additional model resolves a relationship that is not resolved in the existing set of models in a superensemble, then adding it should improve prediction. The difficulty is in identifying unique predictor–predictand relationships. In addition, including dynamic forecast in a superensemble could enhance skill if the dynamic models resolve relationships that are poorly resolved by the statistical models in the superensemble. Here, monthly forecasts were discussed for one lead over one region for precipitation.

Much of the predictability from the superensemble is from variations associated with ENSO, but the results show clearly that including models with predictors from outside the tropical Pacific greatly improves the average correlation skill of the prediction. In addition, including additional predictors reduces the spring barrier associated with ENSO-based prediction. Examination of average correlations for predictions from individual models and subregions shows that all models and subregions yield similar average correlations. For example, the North Pacific predictors yield about as much skill as the tropical Pacific predictor. Some of the relationship from the North Pacific is likely to be associated with ENSO because of teleconnections with the tropical Pacific, but other variations not directly associated with ENSO also influence its variations. Therefore, including models that use North Pacific predictors gives some independent information and improves the skill of the superensemble. The North Atlantic is influenced by other climate modes such as the North Atlantic Oscillation, and including that region yields additional independent information. Since the different regions are separated, information from the North Atlantic that may more strongly influence one region does not need to be balanced by information from the tropical Pacific, which may most strongly influence a different region. The superensemble weights, which are a function of calendar month and location, are used to form an optimal combination of the various forecasts.

Acknowledgments. We thank Kingtse Mo and William Lau for useful discussion and suggestions. The contents of this paper are solely the opinions of the authors and do not constitute a statement of policy, decision, or position on behalf of NOAA or the U. S. Government. S.S.P.S. acknowledges the financial support from the U.S. National Science Foundation's research Grants (Awards AGS-1015926 and AGS-1015957). The Cooperative Institute for Climate and Satellites is funded by the NOAA Cooperative Agreement 2014-1229, Award Number NA14NES4320003.

REFERENCES

Adler, R. F., and Coauthors, 2003: The version-2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979–present). *J. Hydrometeor.*, **4**, 1147–1167, doi:10.1175/1525-7541(2003)004<1147:TVGPCP>2.0.CO;2.

Barnett, T. P., and R. Preisendorfer, 1987: Origins and levels of monthly and seasonal forecast skill for United States surface air temperature determined by canonical correlation analysis. *Mon. Wea. Rev.*, **115**, 1825–1850, doi:10.1175/1520-0493(1987)115<1825:OALOMA>2.0.CO;2.

Barnston, A. G., and C. F. Ropelewski, 1992: Prediction of ENSO episodes using canonical correlation analysis. *J. Climate*, **5**, 1316–1345, doi:10.1175/1520-0442(1992)005<1316:POEEUC>2.0.CO;2.

——, and T. M. Smith, 1996: Specification and prediction of global surface temperature and precipitation from global SST using CCA. *J. Climate*, **9**, 2660–2697, doi:10.1175/1520-0442(1996)009<2660:SAPOGS>2.0.CO;2.

Burnham, K. P., and D. R. Anderson, 2002: *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd ed. Springer, 488 pp.

Chen, M., W. Wang, and A. Kumar, 2013: Lagged ensembles, forecast configuration, and seasonal predictions. *Mon. Wea. Rev.*, **141**, 3477–3497, doi:10.1175/MWR-D-12-00184.1.

Davis, R. E., 1976: Predictability of sea surface temperature and sea level pressure anomalies over the North Pacific Ocean. *J. Phys. Oceanogr.*, **6**, 249–266, doi:10.1175/1520-0485(1976)006<0249:POSSTA>2.0.CO;2.

dos Santos, A. F., S. R. Freitas, J. G. Z. de Mattos, H. F. de Campos Velho, M. A. Gan, E. F. P. da Luz, and G. A. Grell, 2013: Using the firefly optimization method to weight an ensemble of rainfall forecasts from the Brazilian developments on the Regional Atmospheric Modeling System (BRAMS). *Adv. Geosci.*, **35**, 123–136, doi:10.5194/adgeo-35-123-2013.

Efron, B., and G. Gong, 1983: A leisurely look at the bootstrap, the jackknife, and cross-validation. *Amer. Stat.*, **37**, 36–48.

Garfinkel, C. I., M. M. Hurwitz, D. W. Waugh, and A. H. Butler, 2013: Are the teleconnections of central Pacific and eastern Pacific El Niño distinct in boreal wintertime? *Climate Dyn.*, **41**, 1835–1852, doi:10.1007/s00382-012-1570-2.

Huffman, G. J., R. F. Adler, M. Morrissey, D. T. Bolvin, S. Curtis, R. Joyce, B. McGavock, and J. Susskind, 2001: Global precipitation at one-degree daily resolution from multisatellite observations. *J. Hydrometeor.*, **2**, 36–50, doi:10.1175/1525-7541(2001)002<0036:GPAODD>2.0.CO;2.

Kao, H.-Y., and J. Y. Yu, 2009: Contrasting eastern-Pacific and central-Pacific types of ENSO. *J. Climate*, **22**, 615–632, doi:10.1175/2008JCLI2309.1.

Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran, 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, **285**, 1548–1550, doi:10.1126/science.285.5433.1548.

——, ——, Z. Zhang, T. LaRow, D. Bachiochi, E. Williford, S. Gadgil, and S. Surendran, 2000: Multimodel ensemble forecasts for weather and seasonal climate. *J. Climate*, **13**, 4196–4216, doi:10.1175/1520-0442(2000)013<4196:MEFFWA>2.0.CO;2.

——, L. Stefanova, A. Chakraborty, T. S. V. Vijaya Kumar, S. Cocke, D. Bachiochi, and B. Mackey, 2002: Seasonal forecasts of precipitation anomalies for the North American and Asian monsoons. *J. Meteor. Soc. Japan*, **80**, 1415–1426, doi:10.2151/jmsj.80.1415.

Kumar, A., A. G. Barnston, and M. P. Hoerling, 2001: Seasonal predictions, probabilistic verifications, and ensemble size. *J. Climate*, **14**, 1671–1676, doi:10.1175/1520-0442(2001)014<1671:SPPVAE>2.0.CO;2.

Lau, K.-M., K.-M. Kim, and S. S. P. Shen, 2002: Potential predictability of seasonal precipitation over the United States from canonical ensemble correlation predictions. *Geophys. Res. Lett.*, **29**, 1097, doi:10.1029/2001GL014263.

Mantua, N. J., S. R. Hare, Y. Zhang, J. M. Wallace, and R. C. Francis, 1997: A Pacific interdecadal climate oscillation with impacts on salmon production. *Bull. Amer. Meteor. Soc.*, **78**, 1069–1079, doi:10.1175/1520-0477(1997)078<1069:APICOW>2.0.CO;2.

Michaelsen, J., 1987: Cross-validation in statistical climate forecast models. *J. Climate Appl. Meteor.*, **26**, 1589–1600, doi:10.1175/1520-0450(1987)026<1589:CVISCF>2.0.CO;2.

Mo, K. C., 2003: Ensemble canonical correlation prediction of surface temperature over the United States. *J. Climate*, **16**, 1665–1683, doi:10.1175/1520-0442(2003)016<1665:ECCPOS>2.0.CO;2.

——, and D. P. Lettenmaier, 2014: Hydrologic prediction over the conterminous United States using the National Multi-Model Ensemble. *J. Hydrometeor.*, **15**, 1457–1472, doi:10.1175/JHM-D-13-0197.1.

Palmer, T. N., and Coauthors, 2004: Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER). *Bull. Amer. Meteor. Soc.*, **85**, 853–872, doi:10.1175/BAMS-85-6-853.

Reynolds, R. W., N. A. Rayner, T. M. Smith, D. C. Stokes, and W. Wang, 2002: An improved in situ and satellite SST analysis. *J. Climate*, **15**, 1609–1625, doi:10.1175/1520-0442(2002)015<1609:AIISAS>2.0.CO;2.

Shen, S. S. P., W. K.-M. Lau, K.-M. Kim, and G. Li, 2001: A canonical ensemble correlation prediction model for seasonal precipitation anomaly. Tech. Memo NASA/TM-2001-209989, 53 pp. [Available online at http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20010102849.pdf.]

Smith, T. M., and R. E. Livezey, 1999: GCM systematic error correction and specification of the seasonal mean Pacific/North America region atmosphere from global SSTs. *J. Climate*, **12**, 273–288, doi:10.1175/1520-0442-12.1.273.

Stensrud, D. J., J.-W. Bao, and T. T. Warner, 2000: Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Wea. Rev.*, **128**, 2077–2107, doi:10.1175/1520-0493(2000)128<2077:UICAMP>2.0.CO;2.

Vitart, F., 2014: Evolution of ECMWF sub-seasonal forecast skill scores. *Quart. J. Roy. Meteor. Soc.*, **140**, 1889–1899, doi:10.1002/qj.2256.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2008: Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quart. J. Roy. Meteor. Soc.*, **134**, 241–260, doi:10.1002/qj.210.