Predictability of Week 3-4 Average Temperature and Precipitation over the Contiguous United States

Timothy DelSole*, Laurie Trenary,

George Mason University, Fairfax, Virginia and

Center for Ocean-Land-Atmosphere Studies, Fairfax, Virginia

Michael K. Tippett

Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY, and Center of Excellence for Climate Change Research, Department of Meteorology, King Abdulaziz University, Jeddah, Saudi Arabia

Kathleen Pegion

George Mason University, Fairfax, Virginia and

Center for Ocean-Land-Atmosphere Studies, Fairfax, Virginia

Corresponding author address: 112 Research Hall, Mail Stop 2B3, George Mason University, 4400 University Drive, Fairfax, VA 22030 USA

E-mail: tdelsole@gmu.edu

ABSTRACT

This paper demonstrates that an operational forecast model can skillfully predict week 3-4 averages of temperature and precipitation over the contiguous U.S. This skill is demonstrated at the grid point level (about $1^\circ \times 1^\circ$), and by decomposing temperature and precipitation anomalies in terms of an orthogonal set of patterns that can be ordered by a measure of length scale, and then showing that many of the resulting components are predictable and can be predicted in observations with statistically significant skill. The statistical significance of predictability and skill are assessed using a permutation test that accounts for serial correlation. Skill is detected based on correlation measures but not based on mean square error measures, indicating that an amplitude correction is necessary for skill. The statistical characteristics of predictability are further clarified by finding linear combinations of components that maximize predictability. The forecast model analyzed here is version 2 of the Climate Forecast System (CFSv2) and the variables considered are temperature and precipitation over the contiguous U. S. during January and July. A 4-day lagged ensemble, comprising 16 ensemble members, is used. The most predictable component of winter temperature and precipitation are related to ENSO, and other predictable components of winter precipitation are shown to be related to the Madden-Julian Oscillation. These results establish a scientific basis for making week 3-4 weather and climate predictions.

1. Introduction

Operational weather forecasts are skillful out to 7-10 days (Simmons and Hollingsworth 2002), and operational seasonal forecasts are skillful out to 3-8 months (depending on season and model, Barnston et al. 2012), but there is relatively limited evidence that forecasts are skillful in the intermediate $3-4$ week range (Newman et al. 2003; Pegion and Sardeshmukh 2011; Wang et al. 2013). If skillful forecasts in the $3-4$ week range existed, they would have significant social and economic value because many management decisions in agriculture, food security, water resources, and disaster risk are made on this time scale. However, most studies that claim predictability in the $3-4$ week range identify this skill in the tropics (Li and Robertson 2015), in upper level quantities like geopotential height fields (Pegion and Sardeshmukh 2011), or in certain global climate indices (Wang et al. 2013), whereas the skill of midlatitude land surface quantities like 2m-temperature or precipitation tend to be negligible (Li and Robertson 2015). Johnson et al. (2013) develop an empirical model for predicting North American 2m-temperature out to four weeks based on a linear trend and statistical relations with the Madden-Julian Oscillation (MJO) and the El Niño-Southern Oscillation (ENSO), and find that this empirical model has skill in certain regions and phases of the MJO. This paper will show that an operational forecast model makes skillful predictions of week 3-4 average temperature and precipitation over the contiguous United States (CONUS).

Predictability of temperature and precipitation depends very much on the spatial and temporal scale under consideration. Beyond weather time scales (e.g., $7-10$ days), it is widely accepted that only large-scale spatial structures are predictable. Accordingly, we propose a novel approach to investigating subseasonal predictability using a set of spatial patterns that can be ordered by length scale. We will show that week 3-4 averages of time series corresponding to many of these spatial patterns can be skillfully predicted by a state-of-the-art prediction model. In addition, we

find linear combinations of these time series that maximize predictability, and show that many of these predictable components can be predicted with skill.

2. Data

The computations performed in this study are strongly constrained by the availability of forecasts, hence it is helpful to discuss data issues first. We analyze retrospective forecasts, called "hindcasts," from version 2 of the Climate Forecast System (CFSv2; Saha et al. 2014). The CFSv2 is a coupled atmosphere-ocean-land-ice model and is initialized based on analysis products for the atmosphere, ocean, land, and sea ice. The hindcasts under investigation were initialized at 0000, 0600, 1200, 1800 UTC of each day over the 12-year period from January 1999 to December 2010. Although these hindcasts were integrated out to 45 days, only the 2-week mean of weeks 3-4 were considered. Only one hindcast per initialization time is available, so a lagged-ensemble approach is employed whereby an average of forecasts initialized at different times but verifying at the same time were used. In general, skill increases with the size of the lagged ensemble until it saturates around 4 days (as shown in sec. 4). Accordingly, we consider hindcasts based on a 4-day lagged ensemble, which contains 16 members, derived from four hindcasts per day. To be clear, the 4-day lagged ensemble is computed from hindcasts initialized *at or before* time t and that verify between times $t+15$ through $t+28$ days (inclusive). We consider hindcasts of temperature and precipitation over CONUS initialized only in January and July, i.e., boreal winter and summer.

For verification, the 2-week mean temperature is compared to estimates from the NCEP/NCAR reanalysis (Kistler et al. 2001). Similarly, hindcasts of daily precipitation were verified relative to the Climate Prediction Center (CPC) Unified Gauge-Based Analysis (Chen et al. 2008).

Climatologies of daily temperature and precipitation are quite noisy and require significant smoothing. No significant dependence of hindcast climatology on lead time was detected, so

the model climatology for each calendar day was estimated by averaging all hindcasts verifying on the same day and over all lead times. In addition, the daily climatology was fit to a second order polynomial over the 76-day period starting from the first of each month. Various checks and visual comparisons were made to ensure that the estimated climatologies were reasonable.

MJO indices are computed from CFSv2 hindcasts in the manner of Trenary et al. (2016). Specifically, the familiar RMM1 and RMM2 indices of Wheeler and Hendon (2004) were derived from an EOF analysis of observations and then the resulting EOF patterns were projected on model variables. In contrast to the standard approach, a 120-day running mean was not subtracted from the indices, hence our MJO indices includes interannual variability.

3. Methods

This section describes our methods for (1) defining an orthogonal set of large-scale patterns, (2) quantifying predictability and skill, and (3) finding patterns that maximize predictability and skill.

a. Eigenvectors of the Laplace Operator

We project temperature and precipitation fields onto the eigenvectors of the Laplacian operator over CONUS. Laplacian eigenvectors provide a convenient orthogonal basis set that can be ordered by a measure of length scale. Special cases of Laplacian eigenvectors include Fourier series and spherical harmonics, which are used routinely to decompose time series by time scale and spatial structures by length scale, respectively. Eigenvectors of the Laplace operator over CONUS were obtained using a Green's function method described in DelSole and Tippett (2015), which should be consulted for details (codes are available upon request). The resulting spatial patterns are orthogonal with respect to an area-weighted inner product and ordered such that the first corresponds to a spatially uniform pattern over the domain (i.e., the largest spatial scale that fits in the

domain), and subsequent patterns correspond to dipoles, tripoles, quadrupoles, etc., of decreasing length scale. These vectors depend only on the geometry of the domain and therefore are data-independent, in contrast to Empirical Orthogonal Functions (EOFs). Thus, a single set of spatial patterns are used to analyze different variables and seasons.

Laplacian eigenvectors 2-10 over CONUS are shown in fig. 1. The first eigenvector is not shown because it equals a constant over the whole domain. The second and third eigenvectors measure the east-west and north-south gradients, respectively. The next two eigenvectors correspond to a tripole and quadrupole, and so on. The percent variance of observed 2-week means explained by the first 20 Laplacian eigenvectors is shown in fig. 2; similar percentages are found in the model (not shown). As expected, the explained variance tends to decrease with decreasing spatial scale.

b. Measure of Predictability

Predictability refers to the degree to which a variable *in a model* is predictable by that model. As such, predictability is an inherent property of a model that can be measured independently of observations. The standard approach to measuring predictability is to consider an ensemble of predictions initialized at equally likely states of the system. Although the CFSv2 reforecast data set does not have multiple ensemble members for the same initial condition day (i.e., a "burst" ensemble), an ensemble can be approximated by grouping hindcasts initialized six hours apart and that verify on the same day. The resulting ensemble often is called a *lagged ensemble* (Hoffman and Kalnay 1983). Let $f(t,t-\tau)$ denote the forecast *anomaly* initialized at time $t-\tau$ and verifying at time t, where τ is the lead time; time is measured in units of days. If E is the ensemble size, then the mean of the lagged ensemble is defined as

$$\tilde{f}(t,t-\tau) = \frac{1}{E}\sum_{e=0}^{E-1} f(t,t-\tau-e/4), \tag{1}$$

where $e/4$ arises because hindcasts were initialized 6 hours apart (i.e., 1/4 of a day apart).

If a variable is not predictable, then the ensemble members would be independent and the expected variance of the ensemble mean σ_S^2 would be $1/E$ times the expected variance of the climatological distribution σ_N^2. The standard test for this hypothesis is Analysis of Variance (ANOVA, Rowell 1998). To test the null hypothesis of no predictability, ANOVA uses the statistic

$$F = E\frac{\hat{\sigma}_S^2}{\hat{\sigma}_N^2}, \tag{2}$$

where $\hat{\sigma}_S^2$ is an estimate of the variance of ensemble means (i.e., "signal"),

$$\hat{\sigma}_S^2 = \frac{1}{T-1}\sum_{t=1}^{T}\left(\tilde{f}(t,t-\tau)\right)^2, \tag{3}$$

and $\hat{\sigma}_N^2$ is an estimate of the variance about the ensemble means (i.e., "noise"),

$$\hat{\sigma}_N^2 = \frac{1}{T(E-1)}\sum_{t=1}^{T}\sum_{e=0}^{E-1}\left(f(t,t-\tau-e/4)-\tilde{f}(t,t-\tau)\right)^2. \tag{4}$$

If the noise perturbations are independent and identically distributed Gaussian random variables, then F follows an F distribution with $T-1$ and $T(E-1)$ degrees of freedom, which can be used to test significance. Unfortunately, the independence assumption is unrealistic for forecasts initialized a few days apart because large-scale fields tend to be serially correlated on daily time scales. Therefore, the standard hypothesis test is not appropriate for subseasonal forecasts. We propose a block permutation test for deciding predictability. Specifically, under the null hypothesis of no predictability, the forecasts would be exchangeable in the sense that each value of F obtained from a permutation of (independent) samples is equally likely. Accordingly, we construct a permuted ensemble by drawing forecasts from random years. Importantly, the entire sequence of forecasts within a year are drawn, ensuring that the serial correlation across consecutive days is preserved. This sampling is tantamount to randomly permuting (or "shuffling") the years assigned to the forecasts. The statistic F is computed for the permuted ensemble, and this procedure is repeated many

times (i.e., 10,000 times). The rank of the F obtained from the unpermuted ensemble is evaluated relative to the values of F for the permuted ensembles. Under the hypothesis of exchangeability, the rank is uniformly distributed. The actual lagged ensemble is said to be predictable if the observed value of F exceeds the 95% percentile of the F values obtained from permuted samples.

c. Measure of Skill

Skill refers to the degree to which a forecast predicts the *observed* variable. Two standard measures of skill are mean square error (MSE) and correlation ρ. Significance tests for skill based on mean square error have been discussed by DelSole and Tippett (2014), while those based on correlation are standard. Unfortunately, these tests are not appropriate for forecasts initialized at daily intervals because of the serial correlation mentioned above. We again apply a permutation method in which the year labels for the observations are randomly permuted. By selecting the entire sequence of observations within a year, the serial correlation between observations on daily time scales is preserved. After shuffling the year labels for the observations, the correlation coefficient between forecasts and shuffled observations can be computed. This procedure is repeated many times (i.e., 10,000 times) to build up an empirical distribution for the correlation under the null hypothesis of independence. The 95th percentile of the resulting samples then defines the 5% significance threshold value for the correlation coefficient.

d. Predictable Component Analysis

In some cases, none of the time series for the Laplacian eigenvectors can be predicted with skill. However, this result does not prove that there is no skill, because it is possible that some linear combination of eigenvectors can be predicted with skill. To test this possibility, we find the linear combination of eigenvectors that maximize the predictability measure F in (2). This procedure

is formally equivalent to Predictable Component Analysis (see DelSole and Tippett 2007, for a review). We briefly review this procedure to clarify its application in our particular situation. Let the weights of the linear combination be $q_1, q_2, \ldots, q_M$ such that the quantity being forecasted is

$$f(t, t-\tau) = \sum_{m=1}^{M} q_m F_m(t, t-\tau), \tag{5}$$

where $F_m(t, t-\tau)$ is the forecast anomaly for the m'th Laplacian eigenvector. If the weights are collected into the vector $\mathbf{q}$, then the predictability error measure (2) can be written equivalently as

$$F = \frac{\mathbf{q}^T \boldsymbol{\Sigma}_S \mathbf{q}}{\mathbf{q}^T \boldsymbol{\Sigma}_N \mathbf{q}}, \tag{6}$$

where $\boldsymbol{\Sigma}_S$ and $\boldsymbol{\Sigma}_N$ are covariance matrices for the ensemble mean and residuals about the ensemble mean, respectively, defined as

$$(\boldsymbol{\Sigma}_S)_{ij} = \frac{E}{T-1} \sum_{t=1}^{T} \tilde{f}_i(t, t-\tau) \tilde{f}_j(t, t-\tau), \tag{7}$$

and

$$(\boldsymbol{\Sigma}_N)_{ij} = \frac{1}{T(E-1)} \sum_{t=1}^{T} \sum_{e=0}^{E-1} \left(f_i(t, t-\tau-e/4) - \tilde{f}_i(t, t-\tau) \right) \left(f_j(t, t-\tau-e/4) - \tilde{f}_j(t, t-\tau) \right). \tag{8}$$

To find an extremum, we differentiate (6) with respect to $\mathbf{q}$:

$$\frac{\partial F}{\partial \mathbf{q}} = \frac{2\boldsymbol{\Sigma}_S \mathbf{q}}{\mathbf{q}^T \boldsymbol{\Sigma}_N \mathbf{q}} - 2 \frac{\mathbf{q}^T \boldsymbol{\Sigma}_S \mathbf{q}}{(\mathbf{q}^T \boldsymbol{\Sigma}_N \mathbf{q})^2} \boldsymbol{\Sigma}_N \mathbf{q} \tag{9}$$

$$= \frac{2}{\mathbf{q}^T \boldsymbol{\Sigma}_N \mathbf{q}} \left(\boldsymbol{\Sigma}_S \mathbf{q} - \lambda \boldsymbol{\Sigma}_N \mathbf{q} \right), \tag{10}$$

where λ is the value of F for the linear combination defined by the weights $\mathbf{q}$. If $\boldsymbol{\Sigma}_N$ is positive definite, which is typically true when the number of Laplacian eigenvectors is much smaller than the sample size, then the derivative vanishes when $\mathbf{q}$ satisfies the generalized eigenvalue problem

$$\boldsymbol{\Sigma}_S \mathbf{q} = \lambda \boldsymbol{\Sigma}_N \mathbf{q}. \tag{11}$$

It can be proven that if the eigenvalues (and corresponding eigenvectors) are ordered in descending order, then the first eigenvector maximizes F, the second maximizes F subject to being uncorrelated with the first eigenvector (in a sense defined shortly), and so on. Moreover, the eigenvalues give the corresponding maximized F-values. These solutions define the predictable components, the first of which will be called the "most predictable component." Each eigenvector can be substituted in (5) to define the time series associated with that component. Because covariance matrices are symmetric, the resulting time series for different components are uncorrelated. The spatial structure of the predictable component is obtained from regression. The regression coefficient between the predictable component time series (5) and the m'th Laplacian eigenvector is

$$(\mathbf{p})_m = \langle f(t,t-\tau)F_m(t,t-\tau)\rangle = (\mathbf{\Sigma}_N\mathbf{q})_m\,. \tag{12}$$

The Laplacian eigenvectors are then summed using weights specified in the vector $\mathbf{p}$. Note that a regression coefficient can be computed for the m'th Laplacian eigenvector even if that vector was not included in the optimization procedure discussed above (e.g., when $m > \mathscr{M}$). We use $\mathscr{M} = 20$ Laplacian eigenvectors to construct the spatial pattern. This choice effectively imposes a prescribed level of spatial smoothing for the regression pattern.

Note that the above procedure yields a complete set of predictable components for each lead time τ. This lead time dependence is sensible because predictability is characterized by different patterns at different time scales. An alternative approach is to characterize predictability over all time scales, which can be done by maximizing a measure of predictability integrated over all lead times. This approach is called Average Predictability Time (APT, DelSole and Tippett 2009) analysis. APT analysis is not used here because we want to demonstrate the existence of predictability specifically for the week $3-4$ forecasts. Although APT analysis can find predictable components on subseasonal time scales, testing the hypothesis of predictability on subseasonal

time scales is not straightforward because the integral includes the short weather lead times that are predictable. By applying predictable component analysis for only one lead time, subseasonal predictability can be tested in isolation from predictability on other time scales.

The sampling distribution of the maximized F-values (i.e., the eigenvalues) under the null hypothesis of no predictability can be estimated using a permutation technique similar to that described above, in which the label for years assigned to forecasts are randomly permuted. The only extra step is that instead of drawing a single variable, an entire M-dimensional vector is drawn, corresponding to the amplitudes of the M Laplacian eigenvectors for the relevant forecast. Again, an essential element of the technique is to draw the entire sequence of forecasts within a year for the M eigenvectors, which preserves the serial correlation on daily time scales. After generating a mock ensemble forecast data set comprising T time steps and E ensemble members, the covariance matrices are computed and the generalized eigenvalue problem (11) is solved. This process is repeated many times (i.e., 10,000 times) to build up an empirical distribution for the eigenvalues.

4. Results

The correlation skill of 4-day lagged ensembles of temperature and precipitation of week $3-4$ hindcasts over CONUS during January and July is shown in fig. 3. Statistically insignificant values at the 5% level (according to the permutation test) are masked out. The figure shows that winter temperature and precipitation and summer temperature are skillfully predicted by the CFSv2 over a third to a half of the area of CONUS. Summer precipitation shows effectively no skill (e.g., the number of positive and negative correlations are approximately equal).

Our goal is to diagnose the predictability and skill shown in fig. 3 in terms of large-scale spatial structures. The predictability and skill of individual Laplacian eigenvectors of January temperature as a function of ensemble size is shown in fig. 4. Qualitatively similar results are obtained for other

variables and time periods. Not surprisingly, predictability decreases with ensemble size because each additional member is initialized further from the target and therefore contains more noise. The signal-to-noise ratio decreases by a factor of 2-3 from a 12-hour to a 4-day lagged ensemble. In contrast, the skill tends to increase with ensemble size, provided the skill is sufficiently large.

The predictability of week $3-4$ temperature and precipitation CFSv2 hindcasts projected onto individual Laplacian eigenvectors is shown in fig. 5. Predictability is quantified by the signal-to-noise (S/N) ratio F/E, where F is defined in (2). We use $E = 2$, which is equivalent to analyzing differences between hindcasts initialized six hours apart, as done in weather prediction studies (Simmons and Hollingsworth 2002). The figure reveals that several spatial structures of winter temperature and precipitation and summer temperature are predictable. Summer precipitation also is predictable, but for fewer spatial structures. In general, temperature is more predictable than precipitation, and winter is more predictable than summer. Note, however, that precipitation is more predictable than temperature for certain components during winter (e.g., 8th and 9th components)

Although the above results demonstrate week $3-4$ predictability, this result does not necessarily imply that the associated hindcasts are skillful (i.e., that the hindcasts can predict observed anomalies with skill). In most cases, mean square error shows no significant skill. Accordingly, we consider skill based on correlation, which is invariant to linear transformations of the forecast and thus does not penalize biases or errors in forecast amplitude. The skills of the hindcasts based on a 4-day lagged ensemble are shown in fig. 6. The figure shows that many spatial structures of winter temperature and precipitation and summer temperature can be predicted with skill by $3-4$ week hindcasts. The fact that skill exists for correlation but not for mean square error suggests that an amplitude correction is necessary for skill. Only one spatial structure (i.e., the 19th) of summer precipitation has skill exceeding the relevant significance level, but it is unlikely that it would be significant after the multiple comparisons required to identify it are taken into account. Thus, we

conclude that large-scale week $3-4$ winter temperature and precipitation and summer temperature can be predicted with skill, but find little evidence that large-scale, summer precipitation can be predicted with skill at weeks $3-4$.

Although no individual Laplacian eigenvector has significant skill for summer precipitation, this result does not necessarily imply that summer precipitation cannot be predicted with skill. In particular, it is possible that some linear combination of eigenvectors can be predicted with skill. To test this possibility, we apply Predictable Component Analysis to find linear combinations of Laplacian eigenvectors that maximize predictability. A critical step in this procedure is selecting the number of eigenvectors. This step is tantamount to a model selection problem and is one of the most challenging problems in statistics (Fukunaga 1990; Hastie et al. 2003; Taylor and Tibshirani 2015). Fortunately, we have found that our results are not sensitive to the precise number of eigenvectors in the range $9-20$ eigenvectors (we did not look beyond 20). To further validate our results, we have partitioned the years into two parts, a training sample in which the most predictable components are identified, and a verification sample onto which the predictable components are projected and used as an independent test of predictability. We find that the predictability and skill in the verification sample tends to saturate after about nine eigenvectors and remains nearly the same (or even grows) by 20 eigenvectors (not shown). The time series and associated regression pattern corresponding to individual predictable components are virtually independent of the number of eigenvectors greater then five or so. For subsequent calculations, we use the same number of eigenvectors, namely nine, for all months and variables.

The maximized signal-to-noise ratios for CFSv2 week $3-4$ hindcasts are shown in fig. 7. As above, we use ensemble size $E=2$ corresponding to differences between hindcasts initialized six hours apart. The shaded area shows the 95% confidence intervals for no predictability. The results suggest that all components are predictable (because all results lie outside the shaded confidence

region). However, precipitation components near the trailing end tend to be only marginally significant. There is a "kink" in the signal-to-noise spectra at one or two components, indicating predictability significantly greater than the background significance threshold.

The regression map between the most predictable component time series and relevant field is shown in fig. 8. The winter temperature and precipitation patterns are similar to the *observed* ENSO teleconnection patterns derived from monthly means (Yang and DelSole 2012), suggesting that CFSv2 week $3-4$ predictability arises from El Niño/La Niña events. The summer temperature pattern also bear some resemblance to model-ENSO teleconnection patterns (e.g., compare to fig. 7 of Wang et al. 2012), but the correspondence to the summer precipitation pattern is weak.

The skills of the predictable components are shown in fig. 9. The figure shows that the most predictable components have skill at week $3-4$ for winter temperature and precipitation and summer temperature. In contrast, the most predictable component of summer precipitation has no significant skill (it is too small to appear in the figure). About 2-3 predictable components of winter temperature and precipitation and summer temperature have skill. Confidence intervals for the correlation skills overlap (not shown), indicating that the correlations cannot be distinguished. It follows that the ranking according to skill cannot be determined based on the available data. Thus, the fact that the most predictable component is not the most skillful is not necessarily meaningful.

To gain insight into the nature of the predictability and skill, we show in fig. 10 time series of the most predictable components. These time series confirm that secular trends are small. In addition, for the components with the most skill, the time series exhibit relatively large jumps between years, but relatively small fluctuations within a year. This feature suggests that the predictability comes from predicting the overall mean during the month, rather than predicting variations within the month. To test this possibility, the forecasts within a month were decomposed into the sum of two terms, a monthly mean plus an anomaly relative to the monthly mean, then the correlation skill

of these two components were computed separately. The result, shown in fig. 11, shows that skill associated with the monthly mean often dominates. Moreover, skill in predicting the anomalies rarely exceeds 0.35, whereas skill in predicting monthly means frequently exceeds 0.35.

Given that predictability appears to be dominated by the monthly mean component, it is reasonable to explore relations with other variables by computing correlations between monthly mean quantities. The simultaneous squared correlation between each predictable component in CFSv2 and the NINO3.4 index is shown in fig. 12a. We call this measure R-squared because it corresponds to the coefficient of determination of a regression model for predicting the component based on NINO3.4. Because the NINO3.4 index is persistent on weekly time scales, its value in the model is very close to its initial value, which in turn is close to the observed value. Thus, these correlations measure the ENSO teleconnections *in the model*. We see that the most predictable component of winter temperature and precipitation in CFSv2 are highly correlated with ENSO.

In addition to ENSO, the Madden-Julian Oscillation (MJO) often is cited as a phenomenon that may give rise to subseasonal predictability (Vitart 2014). To explore this, we compute the coefficient of determination between the predictable component and the RMM1 and RMM2 indices defined in Wheeler and Hendon (2004). These indices were computed from daily CFSv2 fields, then averaged over week 3-4 hindcasts, and then averaged over the month so that a correlation could computed using only monthly values. The coefficient of determination R^2 is the correlation between the predictable component and the best linear combination of RMM1 and RMM2, and measures the fraction of variance of the predictable component that can be predicted from the MJO indices. The R^2 values computed from monthly mean quantities are shown in fig. 12b and reveal that the most predictable component of winter temperature and precipitation are significantly correlated with MJO activity.

It is well known that ENSO and MJO activity tend to be correlated. This correlation confounds the interpretation of pair-wise correlations. To clarify the relations further, we quantify the degree of relation after one of the indices has been regressed out. A convenient measure of the degree of relation between X and Y, after Z has been removed, is

$$R^2_{YX|Z} = \frac{SSE(Y|X) - SSE(Y|XZ)}{SSE(Y|X)}, \tag{13}$$

where $SSE(Y|X)$ is the sum square error of a regression prediction of Y based on X, and $SSE(Y|XZ)$ is the sum square error of a regression prediction of Y based on X and Z. The constant term is understood to be included in all regression models. The quantity $R^2_{YX|Z}$ lies between 0 and 1 and can be interpreted as the fraction of variance of Y explained by Z after the linear relation with X has been removed from all variables. In the case of ENSO after MJO has been removed (fig. 12c), only the leading predictable component of winter precipitation shows a significant relation with ENSO. In contrast, the leading component of winter temperature has a significant correlation with ENSO (see fig. 12a), but not after the MJO has been removed (which has an R-square of about 0.4, just below the significance threshold; see fig. 12c). This result does not necessarily mean the leading component of winter temperature is unrelated to ENSO, but rather that a correlation could exist but that the sample size (i.e., 12 years) may be too small to detect it. In the case of MJO after ENSO has been removed, shown in fig. 12d, the 3rd and 4th predictable components of winter precipitation shows a significant relation with MJO.

For completeness, we note a similar analysis was performed using the North Atlantic Oscillation (NAO) index and MJO indices. We find that the correlations with the predictable components are marginally significant, but that these correlations become insignificant when MJO has been regressed out (not shown).

5. Conclusion

This paper shows that an operational forecast model skillfully predicts week $3-4$ temperature and precipitation over the contiguous U.S. This skill can be identified at the grid point level (about $1^\circ \times 1^\circ$) and by projecting data onto an orthogonal set of large-scale CONUS patterns (as derived from the eigenvectors of the Laplace operator). An important aspect of this identification is a permutation significance test that accounts for serial correlation on daily time scales. Skill is detected based on correlation measures but not based on mean square error measures, indicating that an amplitude correction is necessary for skill. Our results differ from those of Li and Robertson (2015) perhaps because we analyzed weeks $3-4$ together rather than separately, and only one month at a time was analyzed.

Winter temperature and precipitation tend to have more predictability than their summer counterparts, with summer precipitation having the weakest predictability of all quantities considered in this paper. In addition, the most predictable components were identified by finding linear combinations of Laplacian eigenvectors that maximize signal-to-noise ratio. The results of this maximization procedure clarify the spatial structure of the predictable variability. The most predictable component during winter effectively represents the model's ENSO teleconnection pattern. Some predictable components of winter precipitation are associated with MJO activity. The skill of the predictable components is dominated by the skill in predicting the mean value during a month, rather than from predicting anomalies relative to the monthly mean. By explicitly identifying patterns in an operational forecast model that are predictable on subseasonal time scales and demonstrating that these patterns can be predicted with skill in observations, the above results provide a scientific basis for week $3-4$ predictions.

Acknowledgments. We thank two reviewers and the editor Joseph Barsugli for helpful comments that lead to improved clarity in the final paper. This research was supported primarily by the National Oceanic and Atmospheric Administration, under the Climate Test Bed program (NA10OAR4310264), and the MAPP program (NA14OAR4310184). Additional support was provided by the National Science Foundation (AGS-1338427), National Aeronautics and Space Administration (NNX14AM19G), the National Oceanic and Atmospheric Administration (NA14OAR4310160). The views expressed herein are those of the authors and do not necessarily reflect the views of these agencies.

References

Barnston, A., M. K. Tippett, M. L. L'Heureux, S. Li, and D. G. DeWitt, 2012: Skill of real-time seasonal ENSO model predictions during 2002–11: Is our capability increasing? *Bull. Am. Meteor. Soc.*, **93**, ES48–ES50.

Chen, M., W. Shi, P. Xie, V. B. S. Silva, V. E. Kousky, R. Wayne Higgins, and J. E. Janowiak, 2008: Assessing objective techniques for gauge-based analyses of global daily precipitation. *Journal of Geophysical Research: Atmospheres*, **113 (D4)**, n/a–n/a, doi:10.1029/2007JD009132, URL http://dx.doi.org/10.1029/2007JD009132.

DelSole, T., and M. K. Tippett, 2007: Predictability: Recent insights from information theory. *Rev. Geophys.*, **45**, RG4002, doi:10.1029/2006RG000202.

DelSole, T., and M. K. Tippett, 2009: Average predictability time: Part II: Seamless diagnosis of predictability on multiple time scales. *J. Atmos. Sci.*, **66**, 1188–1204.

DelSole, T., and M. K. Tippett, 2014: Comparing forecast skill. *Mon. Wea. Rev.*, **142**, 4658–4678.

DelSole, T., and M. K. Tippett, 2015: Laplacian eigenfunctions for climate analysis. *J. Climate*, **28**, 7420–7436.

Fukunaga, K., 1990: *An introduction to statistical pattern recognition*. 2nd ed., Academic Press, 591 pp.

Hastie, T., R. Tibshirani, and J. H. Friedman, 2003: *Elements of Statistical Learning*. corrected ed., Springer, 552 pp.

Hoffman, R. N., and E. Kalnay, 1983: Lagged average forecasting, an alternative to monte carlo forecasting. *Tellus A*, **35A (2)**, 100–118, doi:10.1111/j.1600-0870.1983.tb00189.x, URL http://dx.doi.org/10.1111/j.1600-0870.1983.tb00189.x.

Johnson, N. C., D. C. Collins, S. B. Feldstein, M. L. L'Heureux, and E. E. Riddle, 2013: Skillful wintertime north american temperature forecasts out to 4 weeks based on the state of enso and the mjo. *Weather and Forecasting*, **29 (1)**, 23–38, doi:10.1175/WAF-D-13-00102.1, URL http://dx.doi.org/10.1175/WAF-D-13-00102.1.

Kistler, R., and Coauthors, 2001: The NCEP–NCAR 50–year Reanalysis: Monthly means CD–ROM and documentation. *Bull. Am. Meteor. Soc.*, **82**, 247–267.

Li, S., and A. W. Robertson, 2015: Evaluation of submonthly precipitation forecast skill from global ensemble prediction systems. *Monthly Weather Review*, **143 (7)**, 2871–2889, doi:10.1175/MWR-D-14-00277.1, URL http://dx.doi.org/10.1175/MWR-D-14-00277.1.

Newman, M., P. D. Sardeshmukh, C. R. Winkler, and J. S. Whitaker, 2003: A study of subseasonal predictability. *Mon. Wea. Rev.*, **131 (8)**, 1715–1732.

Pegion, K., and P. D. Sardeshmukh, 2011: Prospects for improving subseasonal predictions. *Mon. Wea. Rev.*, **139 (11)**, 3648–3666.

Rowell, D. P., 1998: Assessing potential seasonal predictability with an ensemble of multidecadal GCM simulations. *J. Climate*, **11**, 109–120.

Saha, S., and Coauthors, 2014: The NCEP Climate Forecast System Version 2. *J. Climate*, **27**, 2185–2208.

Simmons, A. J., and A. Hollingsworth, 2002: Some aspects of the improvement in skill of numerical weather prediction. *Q. J. R. Meteorol. Soc.*, **128**, 647–677.

Taylor, J., and R. J. Tibshirani, 2015: Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, **112 (25)**, 7629–7634.

Trenary, L., T. DelSole, M. K. Tippett, and K. Pegion, 2016: A new method for determining the optimal lagged ensemble. *J. Adv. Model. Earth Syst.*, **submitted**.

Vitart, F., 2014: Evolution of ECMWF sub-seasonal forecast skill scores. *Quarterly Journal of the Royal Meteorological Society*, **140 (683)**, 1889–1899, doi:10.1002/qj.2256, URL http://dx.doi.org/10.1002/qj.2256.

Wang, H., A. Kumar, W. Wang, and B. Jha, 2012: U.s. summer precipitation and temperature patterns following the peak phase of el niño. *Journal of Climate*, **25 (20)**, 7204–7215, doi:10.1175/JCLI-D-11-00660.1, URL http://dx.doi.org/10.1175/JCLI-D-11-00660.1.

Wang, W., M.-P. Hung, S. J. Weaver, A. Kumar, and X. Fu, 2013: MJO prediction in the NCEP Climate Forecast System version 2. *Climate Dynamics*, 1–12.

Wheeler, M. C., and H. Hendon, 2004: An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Mon. Wea. Rev.*, **132**, 1917–1932.

Yang, X., and T. DelSole, 2012: Systematic comparison of ENSO teleconnection patterns between models and observations. *J. Climate*, **25**, 425–446.

LIST OF FIGURES

year; a time series beginning on the first of the month (indicated by a dot) is disconnected from time series of the previous year. The title of each panel indicates the month and variable of the predictable component. The correlation coefficient between the observed and hindcast time series is indicated in the title of each panel. 32

Fig. 11. Skill of predictable components for week $3-4$ CFSv2 hindcasts of temperature and precipitation over CONUS. The week 3-4 forecasts within a month were decomposed into the sum of two terms, a monthly mean and an anomaly relative to the monthly mean, then the correlation skill of the two components were computed separately. The skill of predicting monthly means is shown on the x-axis, and the skill of predicting anomalies is shown on the y-axis. The shaded box indicates areas in which both skills are below 0.35 (an approximate 5% significance threshold). The number label indicates the order of the predictable component, and the different colors denote different months and variables, as indicated in the legend key. 33

Fig. 12. R-squared values of the degree of association between predictable components of CFSv2 with (a) ENSO and (b) MJO activity. ENSO is measured by the observed NINO3.4 index while the MJO is measured by the RMM1 and RMM2 indices of Wheeler and Hendon (2004) as derived from the CFSv2. The bottom two panels show the "partial" R-squared values, which measures the association with ENSO after the MJO has been removed (c), and the association with MJO after ENSO has been removed (d). The horizontal dashed line shows the 5% significance threshold based 12 samples, i.e., the number of years in the period 1999-2010. 34

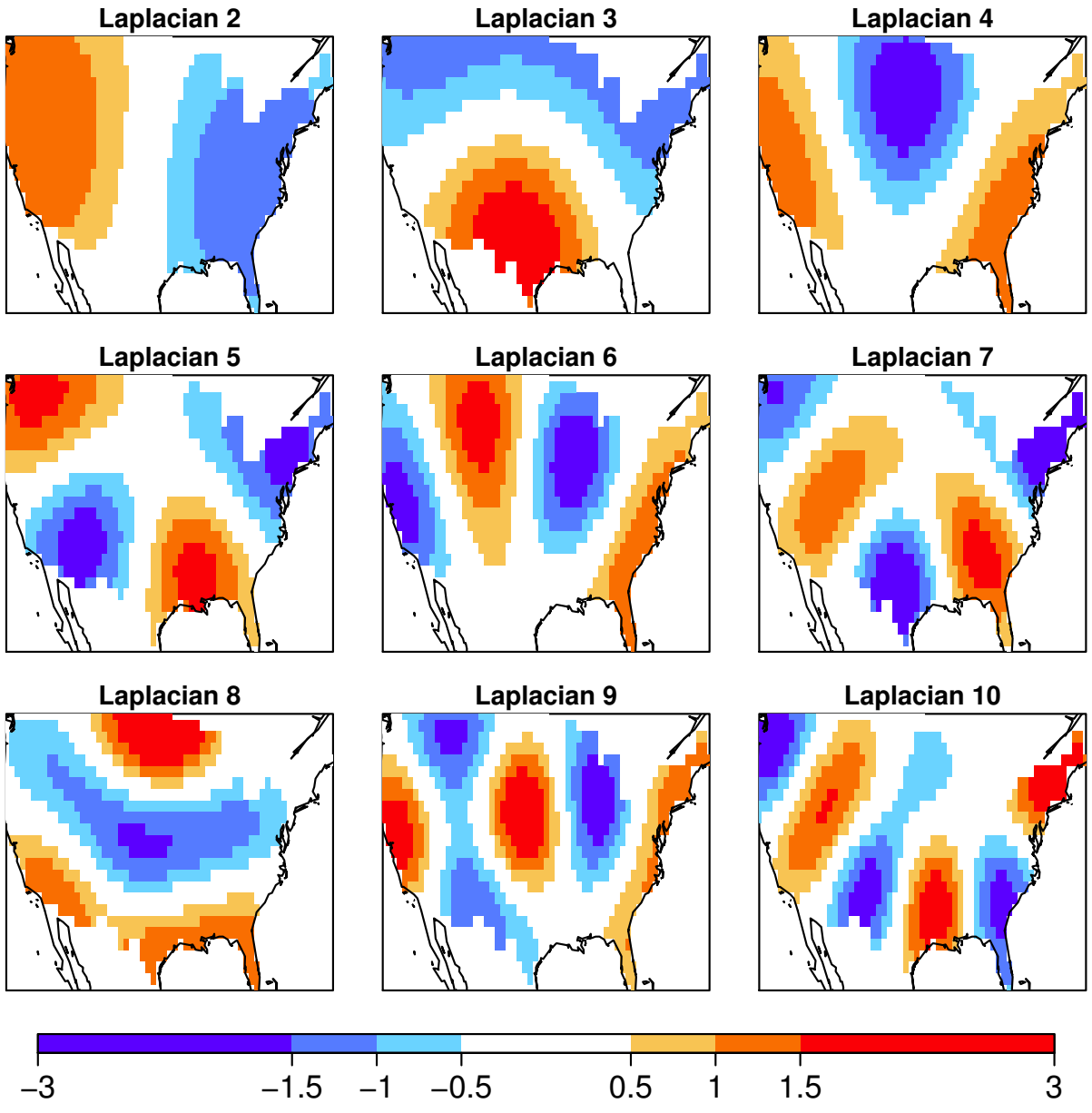


FIG. 1. Laplacian eigenvectors 2-10 over the contiguous United States.

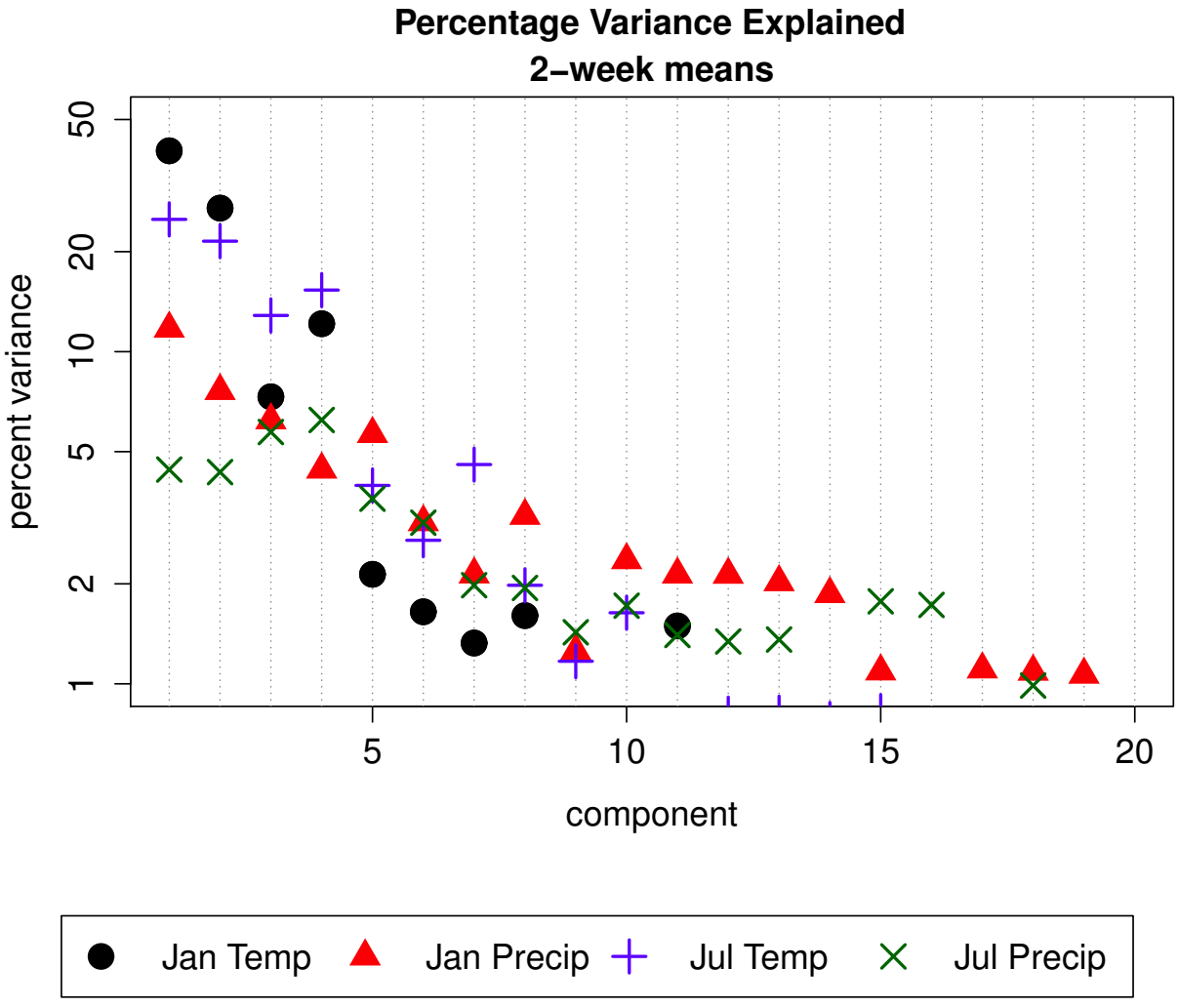


FIG. 2. Fraction of variance of observed 2-week means explained by individual Laplacian eigenfunctions 1-20.

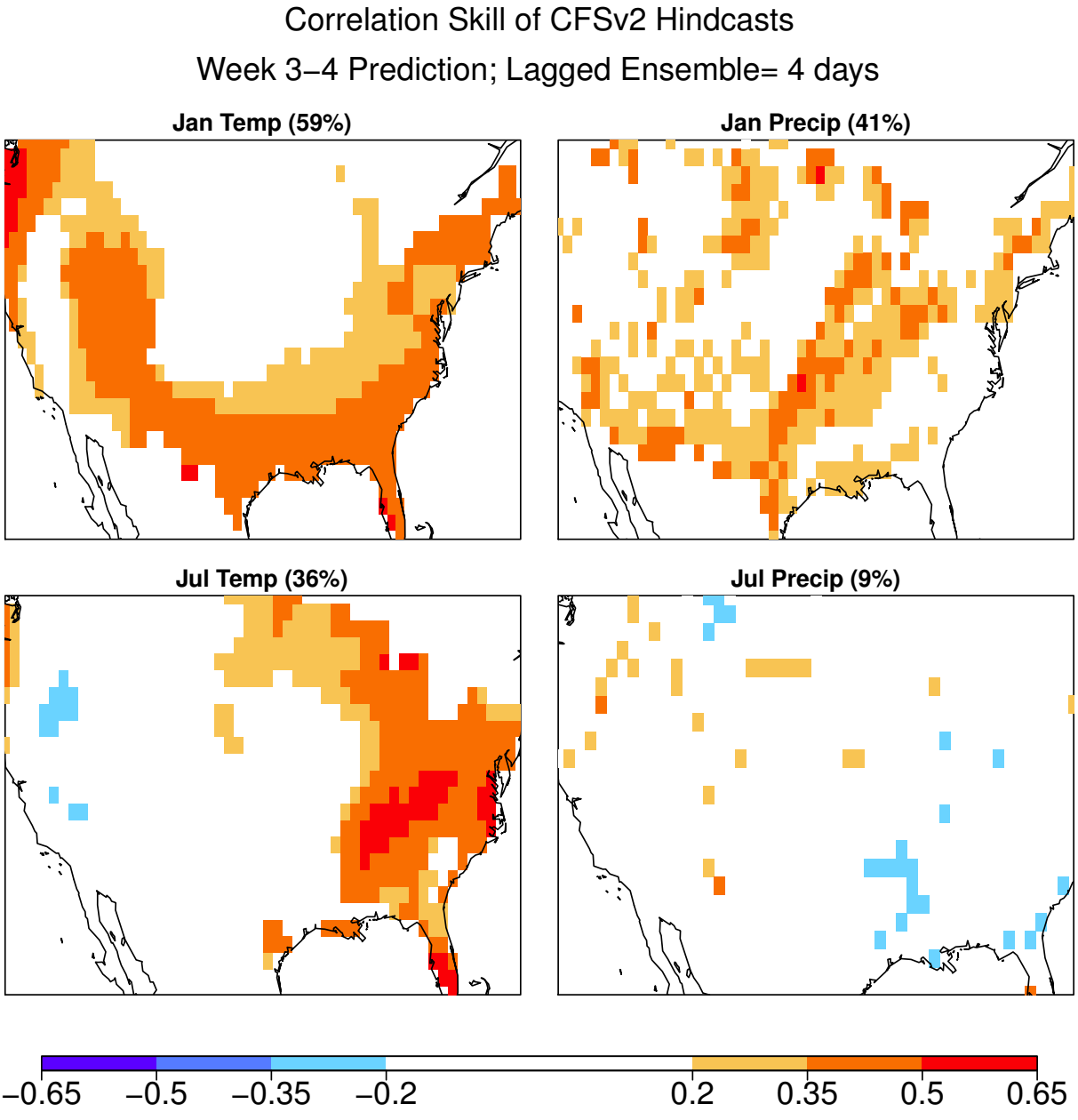


FIG. 3. Correlation skill of week $3-4$ temperature and precipitation CFSv2 hindcasts over CONUS during January and July, 1999-2010 (12 years). The hindcasts are based on a 4-day lagged ensemble (comprising 16 members drawn from 4x daily hindcasts). Values that are statistically insignificant at the 5% level (according to the permutation test) are masked out. The percentage area with significant correlation skill (positive and negative) is indicated in the title of each panel.

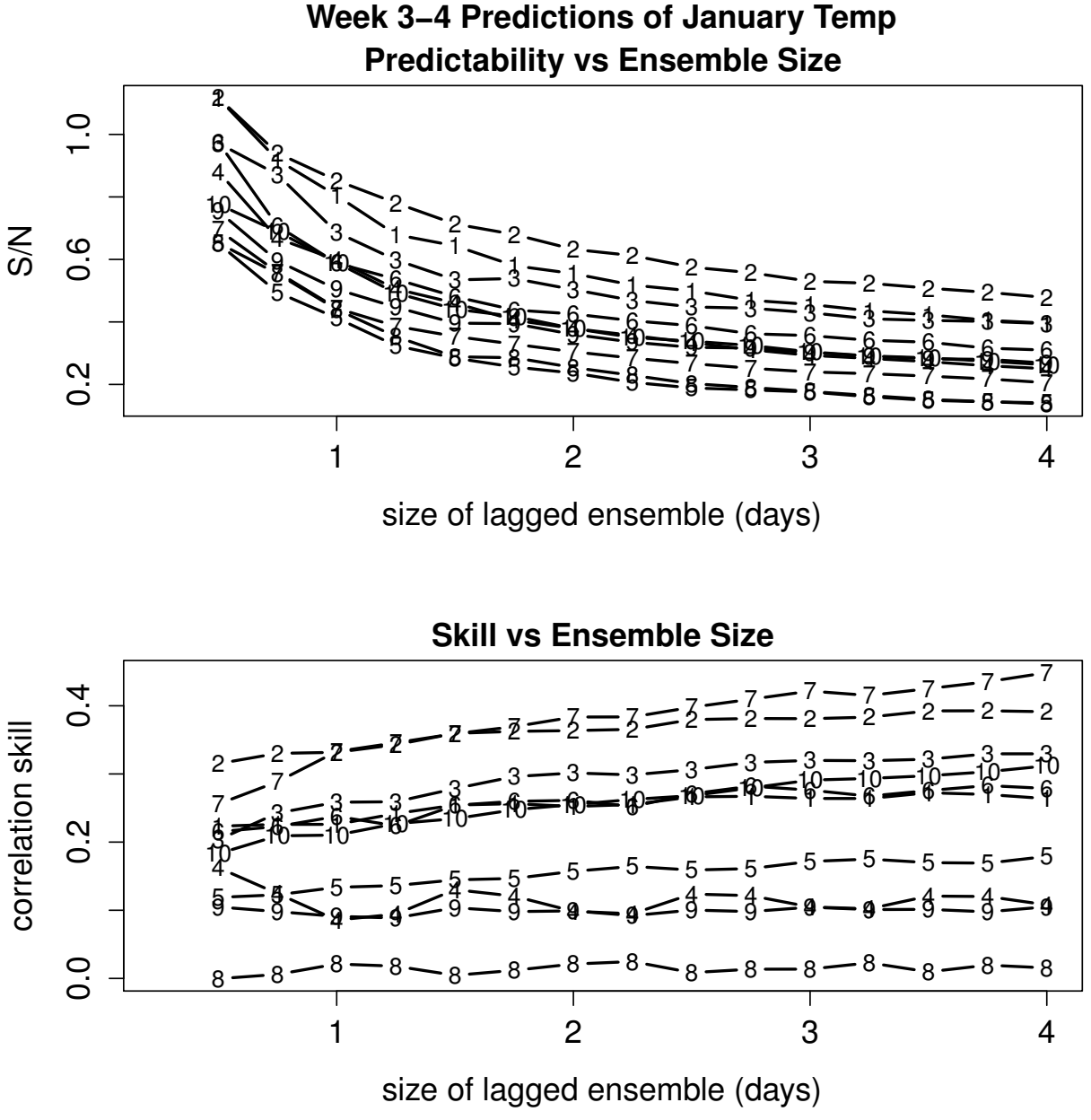


FIG. 4. Predictability (top) and skill (bottom) of week $3-4$ CFSv2 hindcasts of January temperature for individual eigenfunctions as a function of ensemble size (measured in days spanned by the lagged ensemble). The numbered labels indicate the Laplacian eigenfunction.

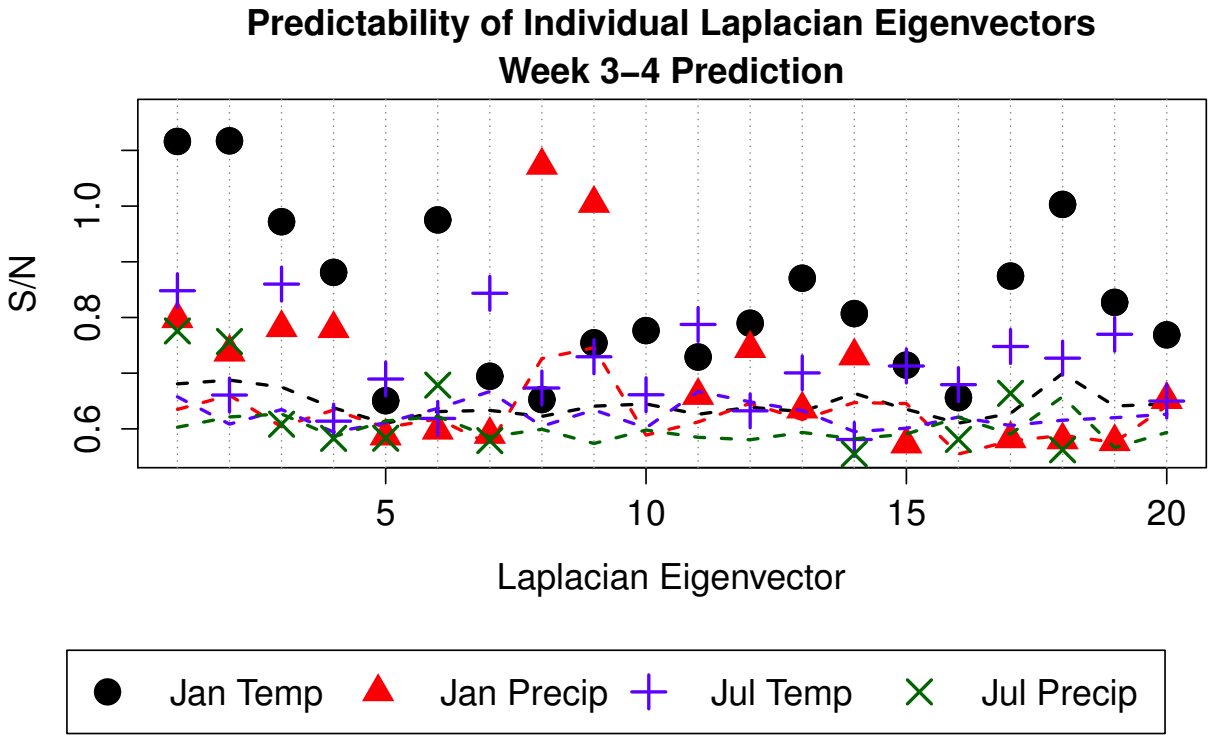
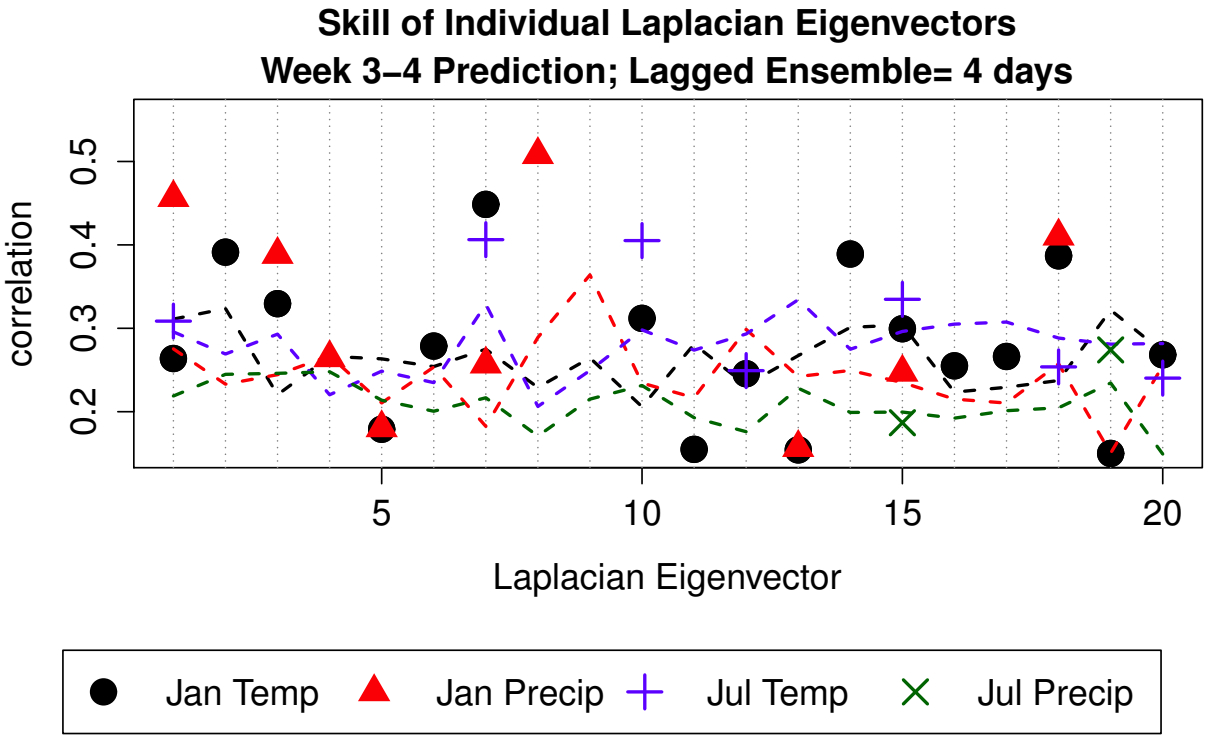


FIG. 5. Predictability (as measured by the signal-to-noise ratio) of week $3-4$ temperature and precipitation hindcasts over the contiguous U.S. during January and July from the CFSv2 for individual Laplacian eigenvectors (the first ten of which are shown in fig. 1). Different symbols correspond to different variables and months, as indicated in the bottom legend. The dashed lines show the 5% significance threshold estimated from 10,000 permutation samples, using the color corresponding to the relevant variable and month (e.g., the black dashed shows the significance thresholds for the black dots corresponding to January temperature). S/N ratios below the smallest significance threshold are not shown.



505 FIG. 6. Correlation skill of 4-day lagged ensemble hindcasts of week $3-4$ temperature and precipitation over
506 CONUS from the CFSv2 for individual Laplacian eigenvectors (the first ten of which are shown in fig. 1). The
507 format of the figure is the same as fig. 5.

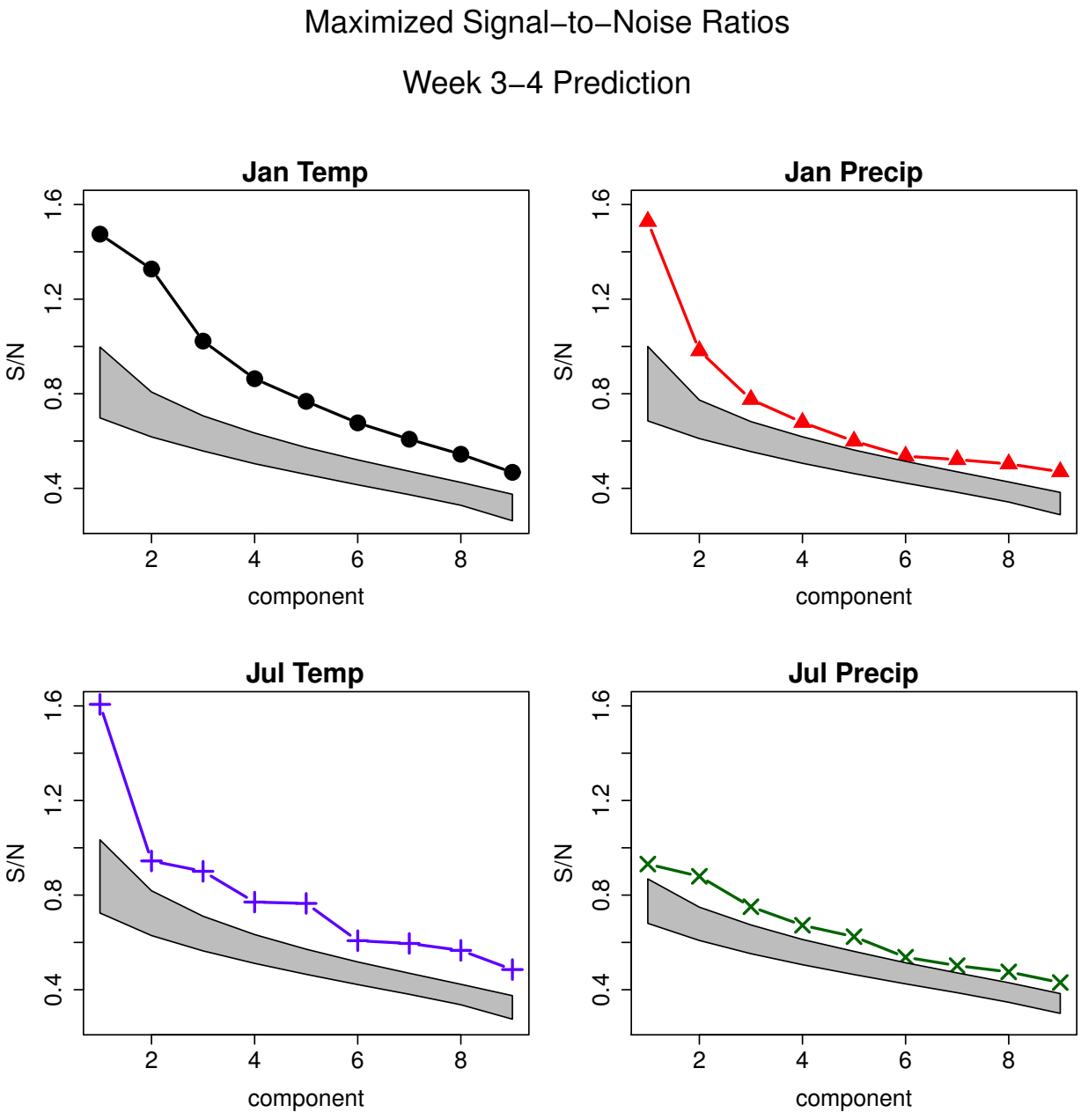


FIG. 7. Maximized signal-to-noise ratios of CFSv2 weeks $3-4$ hindcasts of temperature and precipitation over CONUS. The maximization is performed using the first nine Laplacian eigenfunctions over CONUS, which are shown in fig. 1. The shaded region shows the 95% confidence interval for no predictability estimated by permutation methods.

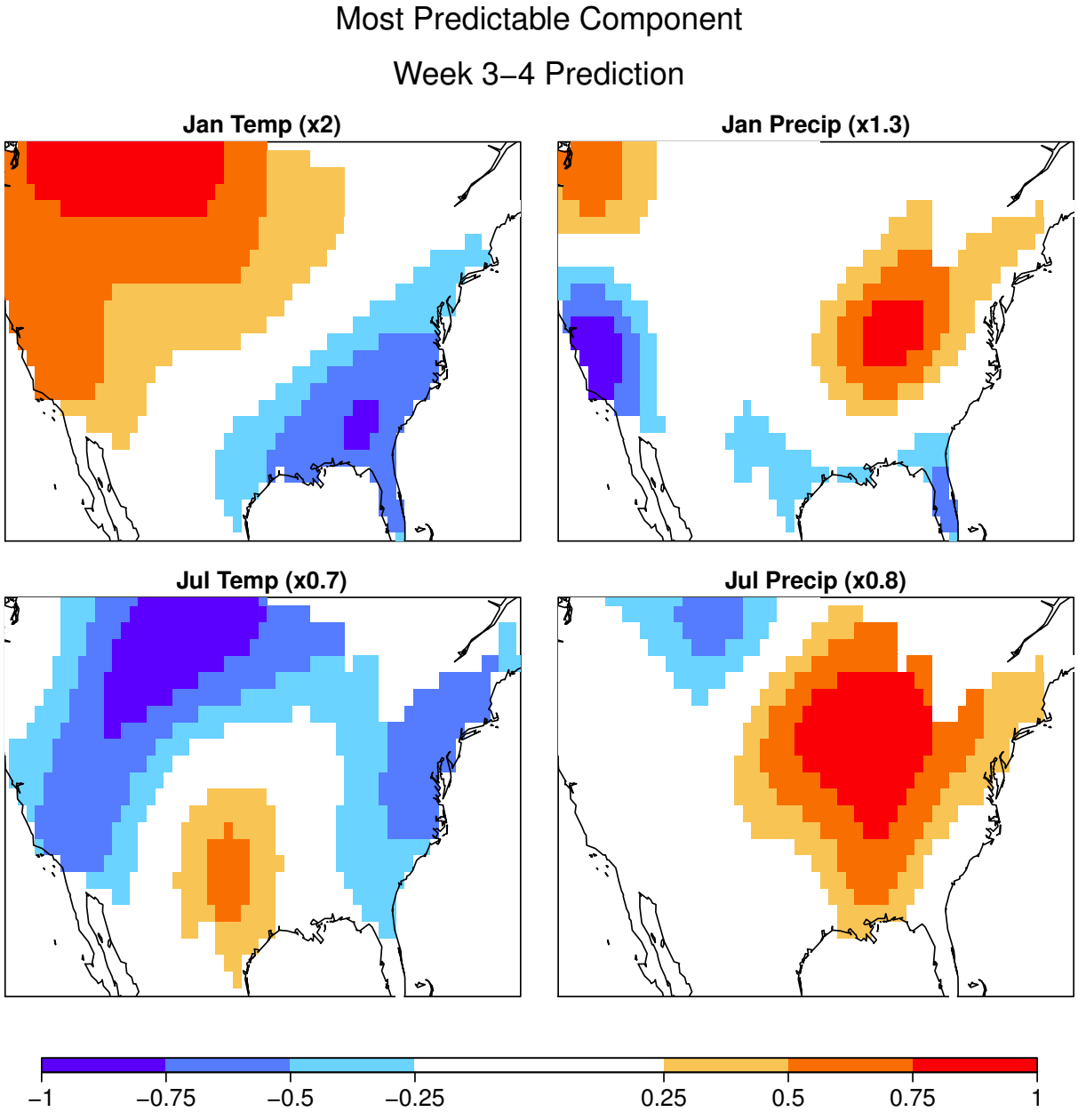


FIG. 8. Regression coefficients between the most predictable component time series and the associated variable. The regression map is derived by regressing time series onto the first 20 Laplacian eigenvectors. The choice of 20 imposes an implicit level of spatial smoothing. The pattern is normalized to lie between -1 and 1, and the multiplicative factor to obtain degrees Kelvin or mm/day, for temperature and precipitation, respectively is indicated in the title above each panel.

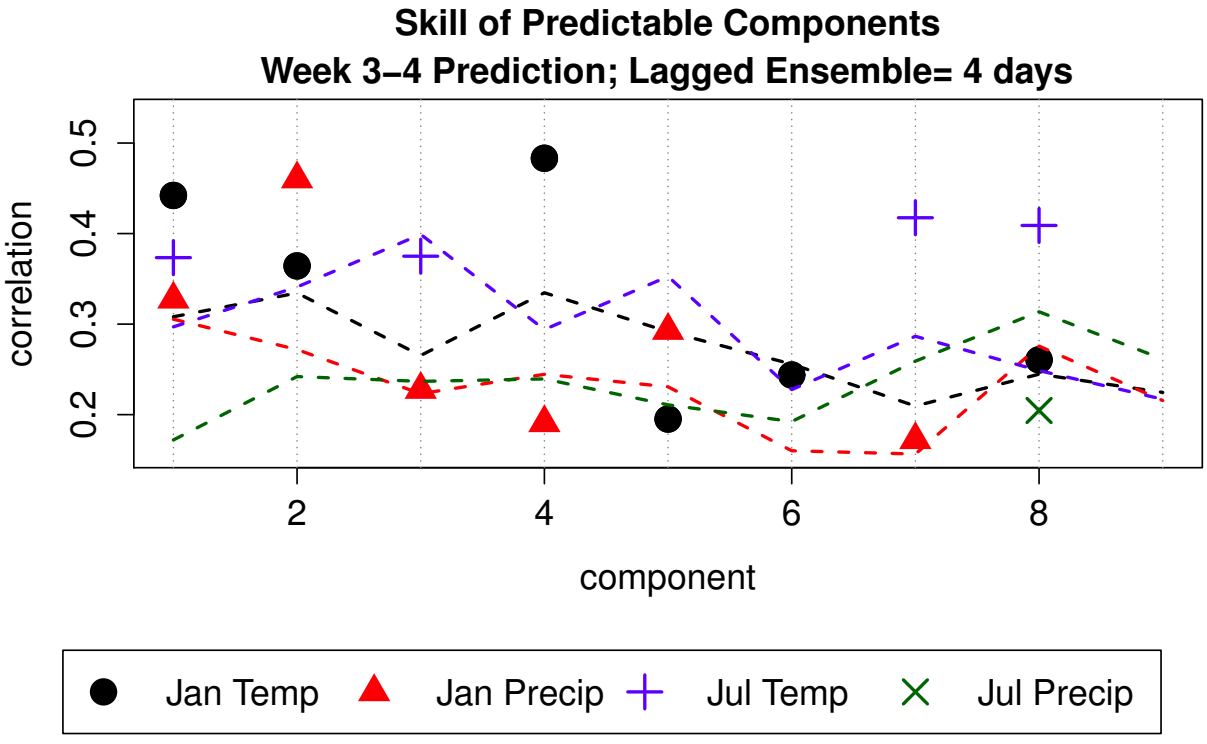
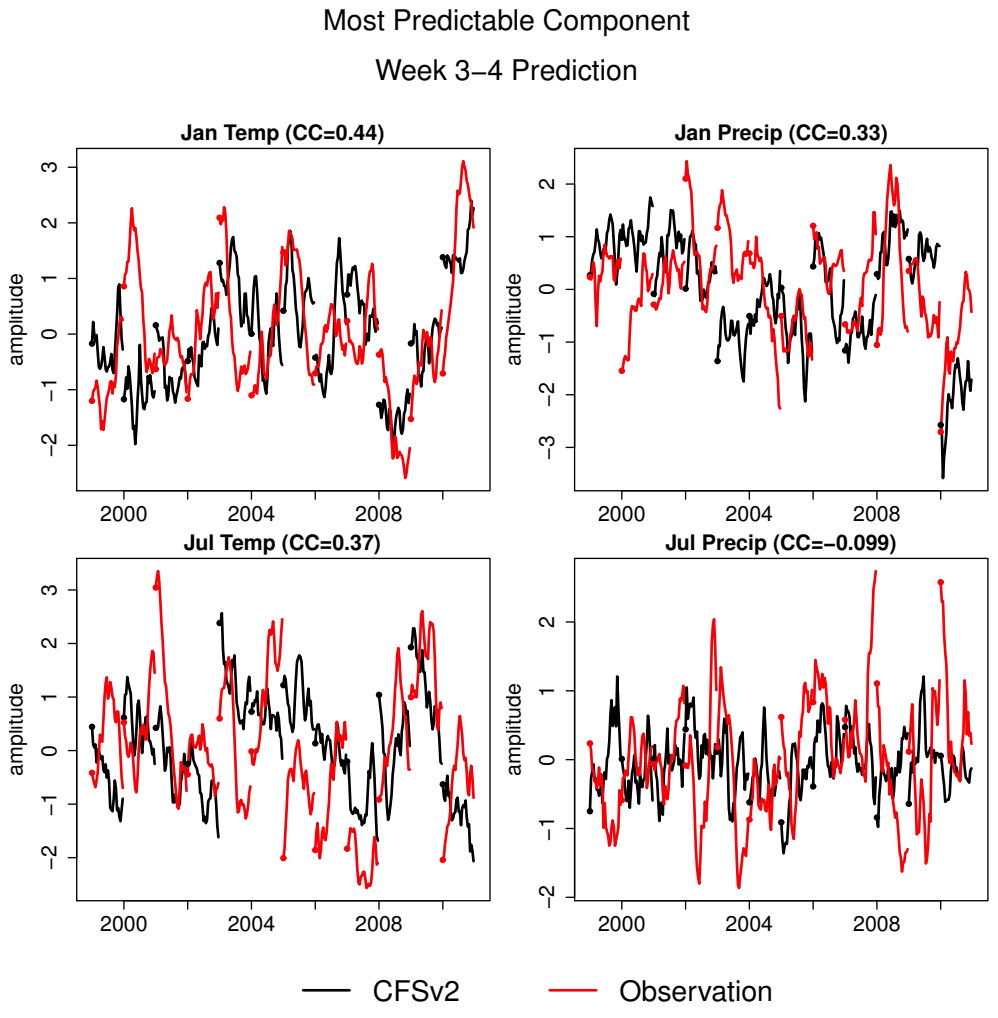


FIG. 9. The correlation skill between the predictable components (i.e., components that maximize the signal-to-noise ratio) of week $3-4$ temperature and precipitation hindcasts from CFSv2 for over CONUS. The format of the figure is the same as fig. 5.



520 FIG. 10. Time series of the most predictable components of week $3-4$ CFSv2 hindcasts of temperature (left)
521 and precipitation (right) over CONUS. Each time series shows a 2-week mean of a variable: for observations
522 (red), this time series corresponds to a 2-week running mean and serves as verification; for hindcasts (black),
523 each 2-week mean is computed separately by averaging leads 15-28 days of each hindcast initialized on each day
524 of the month. Forecasts initialized on consecutive days in a given month are plotted as a single time series for
525 each year; a time series beginning on the first of the month (indicated by a dot) is disconnected from time series
526 of the previous year. The title of each panel indicates the month and variable of the predictable component. The
527 correlation coefficient between the observed and hindcast time series is indicated in the title of each panel.

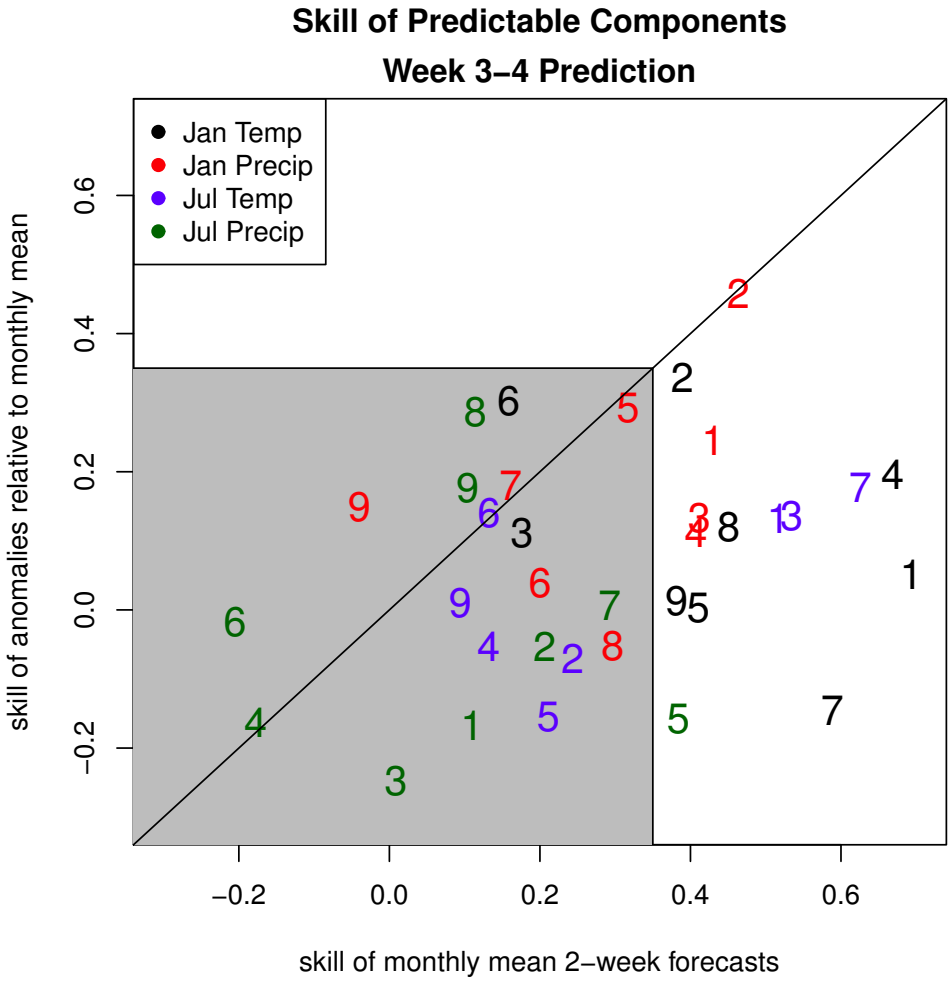
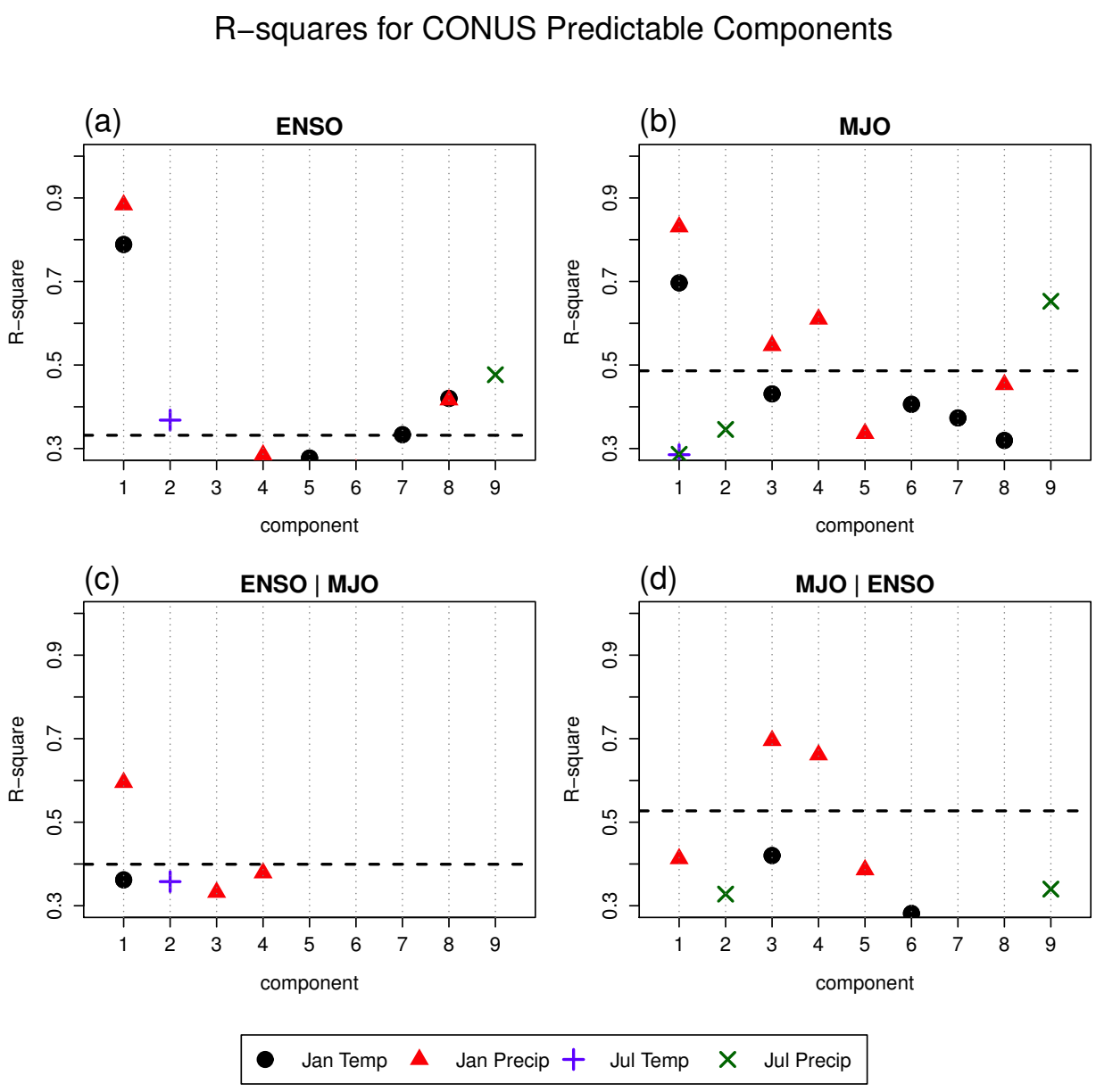


FIG. 11. Skill of predictable components for week $3-4$ CFSv2 hindcasts of temperature and precipitation over CONUS. The week 3-4 forecasts within a month were decomposed into the sum of two terms, a monthly mean and an anomaly relative to the monthly mean, then the correlation skill of the two components were computed separately. The skill of predicting monthly means is shown on the x-axis, and the skill of predicting anomalies is shown on the y-axis. The shaded box indicates areas in which both skills are below 0.35 (an approximate 5% significance threshold). The number label indicates the order of the predictable component, and the different colors denote different months and variables, as indicated in the legend key.



535 FIG. 12. R-squared values of the degree of association between predictable components of CFSv2 with (a)
536 ENSO and (b) MJO activity. ENSO is measured by the observed NINO3.4 index while the MJO is measured
537 by the RMM1 and RMM2 indices of Wheeler and Hendon (2004) as derived from the CFSv2. The bottom two
538 panels show the "partial" R-squared values, which measures the association with ENSO after the MJO has been
539 removed (c), and the association with MJO after ENSO has been removed (d). The horizontal dashed line shows
540 the 5% significance threshold based 12 samples, i.e., the number of years in the period 1999-2010.