Statistical Seasonal Prediction Based on Regularized Regression

TIMOTHY DELSOLE

Department of Atmospheric, Ocean, and Earth Sciences, and Center for Ocean–Land–Atmosphere Studies, George Mason University, Fairfax, Virginia

ARINDAM BANERJEE

Department of Computer Science and Engineering, University of Minnesota, Twin Cities, Minnesota

(Manuscript received 27 March 2016, in final form 24 August 2016)

ABSTRACT

This paper proposes a regularized regression procedure for finding a predictive relation between one variable and a field of other variables. The procedure estimates a linear prediction model under the constraint that the regression coefficients have smooth spatial structure. The smoothness constraint is imposed using a novel approach based on the eigenvectors of the Laplace operator over the domain, which results in a constrained optimization problem equivalent to either ridge regression or least absolute shrinkage and selection operator (LASSO) regression, which can be solved by standard numerical software. In addition, this paper explores an unconventional procedure whereby regression models are estimated from dynamical model output and then verified against observations—the reverse of the traditional order. The methodology is illustrated by constructing statistical prediction models of summer Texas-area temperature based on concurrent Pacific sea surface temperature (SST). None of the regularized regression models have statistically significant skill when estimated from observations. In contrast, when estimated from dynamical model output, the regression models have skill with respect to dynamical model data because of the substantially larger sample size available from dynamical model output. In addition, the regression models estimated from dynamical model data can predict observed anomalies with significant skill, even though no observations were used directly to estimate the regression models. The results indicate that dynamical models had no significant skill because they could not accurately predict the SST itself, not because they could not capture realistic SST teleconnections.

1. Introduction

It is well established that tropical sea surface temperatures (SSTs) influence midlatitude weather on seasonal time scales (Ropelewski and Halpert 1987; Trenberth et al. 1998; Straus et al. 2003; Shukla and Kinter 2006). This influence arises from the tendency of atmospheric convection to intensify over anomalously warm SSTs in the tropics and thereby excite perturbations in the atmosphere that propagate around the world (Opsteegh and Van den Dool 1980; Horel and Wallace 1981; Hoskins and Karoly 1981; Simmons et al. 1983). The spatial structure of the midlatitude response to tropical SST perturbations is a robust property in atmospheric general circulation models (Geisler et al. 1985; Yang and DelSole 2012). Seasonal weather could be predicted to some extent if its relation with SST perturbations could be identified. A common, but problematic, approach to identifying such relations is to compute a regression map between the quantity being predicted and SSTs and then select certain SST "hot spots" as indices that can serve as predictors. The problem with this approach is that when hundreds or thousands of regression coefficients are computed, as done to construct a regression map, there is a high probability that many values will exceed standard significance thresholds even in the absence of a predictive relation (DelSole and Shukla 2009). This procedure is an example of data fishing (or screening) and is widely discredited (Caldwell et al. 2014; Lo et al. 2015; Taylor and Tibshirani 2015).

A critical step in statistical seasonal prediction is identifying predictive relations between a variable and a field of other variables, where the number of field variables is much larger than the sample size. This problem

Corresponding author e-mail: Timothy DelSole, tdelsole@gmu.edu

DOI: 10.1175/JCLI-D-16-0249.1

occurs in many other fields, including economics, data classification, pattern recognition, and machine learning. It is therefore natural to investigate whether methods that have been applied successfully in other fields might be helpful in seasonal prediction. The fundamental idea in all such methods is to impose constraints on the unknown parameters, where the constraints are derived from physical reasoning or other prior knowledge that is independent of the data. The purpose of this paper is to develop prediction models based on the hypothesis that seasonal midlatitude weather is linearly related to SST, where the linear relation is defined by weighting coefficients that have smooth spatial structure. The smoothness hypothesis is motivated mostly by practical considerations: small-scale structure in the coefficients would be difficult to reproduce across dynamical models and difficult to observe, so it is prudent to filter them out. The smoothness constraint is imposed using a novel approach based on the eigenvectors of the Laplace operator in the ocean basin. Specifically, these eigenvectors can be ordered by a measure of spatial scale; hence, the smoothness constraint can be imposed by requiring that the coefficients for small-scale eigenvectors be ''small'' or zero. The resulting constrained minimization problem reduces to regularized regression, which can be solved by standard numerical packages.

In addition to proposing new prediction models, we also utilize dynamical models in an unconventional way. The conventional approach to constructing empirical prediction models (Barnston 1994) is to estimate the empirical model from observations, validate the model using cross validation, and then compare the model to dynamical models to understand the underlying physics. Instead, we reverse the order of this procedure. Specifically, we estimate empirical models from dynamical model output and then validate the empirical model using observations. Obviously, this approach, which has been suggested previously (Quan et al. 2006), will work only if the dynamical model from which the empirical model is estimated is realistic. To the extent that the dynamical model is adequate, a major advantage of this approach is that the sample size available from dynamical model output often is several times larger than that of observations, owing to the multiple ensemble members and start dates (or to availability of long control runs). The larger sample size opens opportunities for a richer variety of statistical or data-mining methodologies. This approach also provides an alternative way to compare dynamical models: empirical models derived from different dynamical models may have very different skills, which may clarify model errors or inadequacies.

The methodologies for constructing regularized regression models are described in the next section. In addition to the new methods, we also consider principal component regression, in which a variable is predicted based on a linear combination of a small number of leading principal components (Barnston and Smith 1996). To illustrate the methodologies, we apply them to predict summer Texas-area temperature based on simultaneous and antecedent Pacific SSTs. The datasets for this analysis are described in section 3. The result of applying the above methods to observations and dynamical model hindcasts is described in section 4. We show that none of the regression models estimated from observations have significant prediction skill. Next, regression models are estimated from dynamical model output without direct use of any observational data. In this case, the regression models derived from most dynamical models have skill with respect to dynamical model output, presumably because of the substantially larger sample size obtained by pooling multiple ensemble members and initial start months. These regression models are then applied to observational data to make predictions. We find that some regularized regression models produce skillful predictions of observational data, even though they were estimated from dynamical model output and even though the regularization procedure cannot produce skillful models from (shorter) observational data. We further examine the skill of multimodel regressions, the sensitivity of skill to sample size, and the skill based on antecedent SSTs. Finally, we show that North American Multimodel Ensemble (NMME) hindcasts themselves had no skill at predicting Texas-area temperature. The regression model results are used to argue that the lack of skill is not because the dynamical models do not capture realistic relations with SST but rather because the NMME models could not accurately predict the relevant SST patterns. We conclude with a summary and discussion of our results.

2. Methodology

We consider the problem of predicting a variable y given a spatial field X. Let y_n be the predictand at the nth time step and $X_{n,s}$ be the predictor at the nth time step and sth spatial location, where $n = 1, \ldots, N$ and $s = 1, \ldots, S$. We assume the two variables are related as follows:

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \epsilon, \tag{1}$$

where $\mathbf{w}$ is an S-dimensional vector of coefficients and ϵ is an unpredictable, random term. Ordinary least squares (OLS) estimates the coefficients $\mathbf{w}$ by minimizing the sum square residuals:

$$F_0(\mathbf{w}) = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2, \tag{2}$$

where $\|\cdot\|$ denotes the L_2 norm (i.e., $\|\mathbf{z}\|^2 = \mathbf{z}^T\mathbf{z}$ for any vector $\mathbf{z}$). For the type of problems considered in this paper, the number of predictors S far exceeds the sample size N, in which case the minimization problem is grossly underdetermined [i.e., coefficients $\mathbf{w}$ can be found such that the linear model fits the data perfectly, in the sense that the residuals in (2) vanish]. In this paper, we consider methods for estimating $\mathbf{w}$ based on minimizing the cost function:

$$F_\lambda(\mathbf{w}) = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \lambda R(\mathbf{w}), \tag{3}$$

where $R(\mathbf{w})$ is a regularization function that imposes a penalty on the complexity of $\mathbf{w}$, typically by constraining a vector norm, and λ is a parameter the controls the strength of the regularization.

A common regularization in seasonal prediction is principal component regression (PCR), in which the predictors are represented by a small number of principal components of the predictor data (Barnston and Smith 1996). In this case, the regularization function is ''sharp'' in the sense that the weights for the leading T principal components are unconstrained while the weights for the other principal components are constrained to be zero. The solution is most conveniently obtained by replacing the $N \times S$ predictor matrix $\mathbf{X}$ in (3) by the $N \times T$ principal component matrix $\tilde{\mathbf{X}}$ and then applying OLS to obtain the T coefficients $\mathbf{w}$.

The most common regularizations in regression are the L_1 and L_2 norms. The L_1 norm is the sum of the absolute values of the weights:

$$R_1(\mathbf{w}) = \sum_{s=1}^{S} |w_s|. \tag{4}$$

Minimizing (3) based on the L_1 norm (4) is called least absolute shrinkage and selection operator (LASSO) (Tibshirani 1996, 2011). LASSO tends to set certain elements of $\mathbf{w}$ exactly to zero, which facilitates interpretation by indicating that the corresponding predictors can be discarded. In contrast, optimization based on the L_2 norm is called ridge regression and tends to shrink all elements of $\mathbf{w}$ toward zero:

$$R_2(\mathbf{w}) = \sum_{s=1}^{S} w_s^2. \tag{5}$$

There are strong theoretical justifications for both ridge and LASSO: both can be derived from an empirical Bayes theory (Efron and Morris 1971), and both are shrinkage estimators, which tend to have less expected total squared error than maximum likelihood estimates, for a suitable choice of λ (Van Houwelingen 2001). Ridge and LASSO regression can be solved using standard mathematical software packages (e.g., R and MATLAB).

While LASSO and ridge have strong statistical justifications, they have only weak physical justification: if the predictor X is SST, then LASSO effectively assumes that only a few SST grid points influence midlatitude weather, whereas ridge effectively assumes that most coefficients are ''small.'' Neither of these assumptions is compelling. We propose a new regularization constraint based on the hypothesis that large-scale SST structures provide the most robust predictive information for seasonal weather. This hypothesis is motivated mostly by practical considerations: small-scale SST structures are difficult to observe and not robust across climate models, so they should be filtered out. This principle can be expressed equivalently by saying that if the predictors are represented in a basis set ordered by spatial scale, then most of the amplitudes of the basis vectors are zero or close to zero. This formulation is exactly a LASSO or ridge regression.

A natural basis set for filtering out short spatial scales is the eigenvectors of the Laplace operator. On a global domain, Laplacian eigenvectors are the spherical harmonics and easily computable. Over an ocean basin, these eigenvectors are difficult to compute using standard boundary conditions. Recent advances in signal processing (Saito 2008) have led to efficient algorithms for computing these eigenvectors in arbitrary domains. These eigenvectors typically satisfy unconventional boundary conditions, but the precise boundary conditions are irrelevant if the vectors are used merely as a basis set. We compute the eigenvectors using the Green's function technique of DelSole and Tippett (2015). The resulting eigenfunctions are orthogonal with respect to an area-weighted norm and normalized to unit-area-weighted norm. The leading Laplacian eigenvectors in the North Pacific are shown in Fig. 1. The first eigenfunction, not shown, is merely a constant and corresponds to the mean over the North Pacific basin. The second and third eigenfunctions measure the east–west and north–south gradients across the Pacific, respectively. Subsequent eigenfunctions are characterized by tripoles, quadrupoles, etc. of decreasing length scale.

To make use of the Laplacian eigenvectors, the T ''gravest'' Laplacian eigenvectors are projected onto the SST data to produce T time series, which are collected

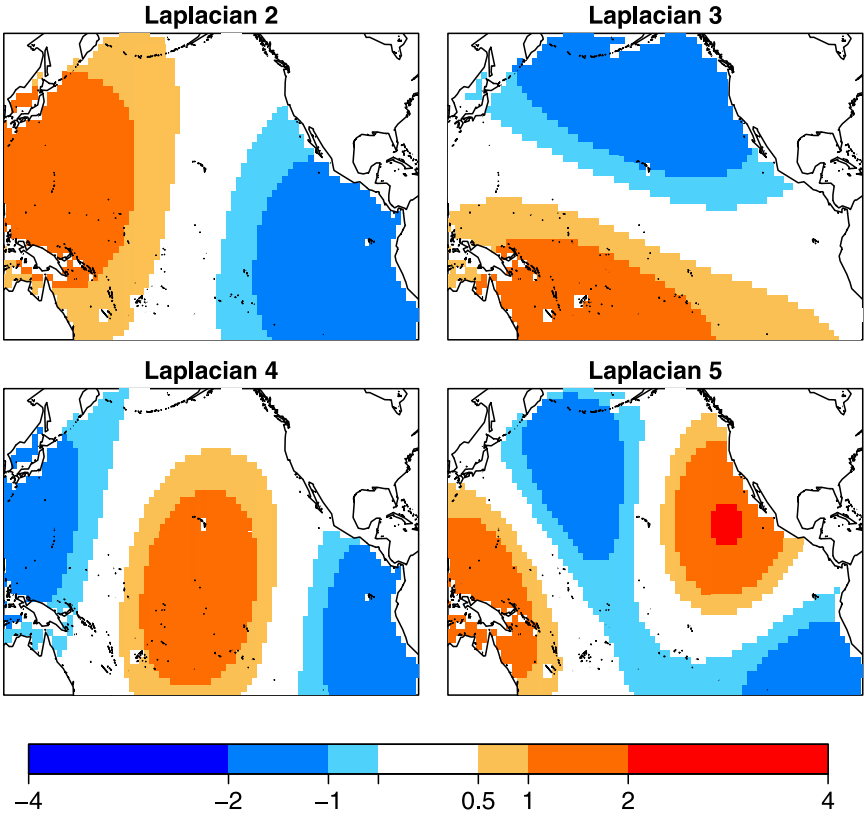

FIG. 1. Spatial patterns of Laplacian eigenvectors 2, 3, 4, and 5 in the North Pacific between 30°S and 60°N.

into the $N \times T$ predictor matrix $\tilde{\mathbf{X}}$. Then, the weights $\tilde{\mathbf{w}}$ for the Laplacian eigenvectors are determined by minimizing

$$F_\lambda(\tilde{\mathbf{w}}) = \|\mathbf{y} - \tilde{\mathbf{X}}\tilde{\mathbf{w}}\|^2 + \lambda R(\tilde{\mathbf{w}}), \tag{6}$$

where the regularization function $R(\tilde{\mathbf{w}})$ is either the L_1 or L_2 norm, or a "sharp" regularization that sets all coefficients beyond a threshold to zero. The maximum number of Laplacian eigenvectors is set to 50; our results are not sensitive to the choice of the upper limit.

In general, the regression model (1) should include an intercept term, but the intercept term should not be included in the penalty functions of (4) or (5); otherwise the solution would depend on the (arbitrary) choice of origin. It can be shown that minimizing (3) with the intercept is equivalent to minimizing (3) without the intercept, provided all variables are centered. Accordingly, all predictors and predictands were centered before finding the ridge and LASSO solutions.

It should be recognized that the coefficients estimated from ridge or LASSO depend on the scale (i.e., standard deviation) of the predictors. In contrast, OLS is invariant to nonsingular linear transformation of the predictors and thus does not depend on the scale of the predictors. In regularized regression, it is customary to rescale the predictors to have identical variances, which effectively penalizes all Laplacian functions equally. However, to be consistent with the smoothness hypothesis, small-scale patterns should be penalized more strongly than large-scale patterns. We explored a wide variety of penalty functions that increase monotonically with wavenumber and found that the final predictions are not sensitive to the choice of penalty function. Accordingly, a convenient approach is to simply use the time series of the Laplacian eigenvectors with no rescaling. This approach effectively imposes a scale-dependent penalty function because, as is well known in geophysical fluid dynamics, large-scale patterns tend to have larger variance than small-scale patterns (Charney 1971; Nastrom and Gage 1985). Consequently, the standard deviation of the time series decreases with wavenumber, and it can be shown that this corresponds to a penalty function that increases with wavenumber if the time series were normalized to the same variance.

A key question in regularized regression is how to select the regularization parameter λ. A common

TABLE 1. List of NMME models and relevant details (Kirtman et al. 2014). (FLOR is Forecast-Oriented Low Ocean Resolution; RSMAS is Rosenstiel School of Marine and Atmospheric Science. Additional acronym expansions are available online at http://www.ametsoc.org/PubsAcronymList.)

Full model name	Shortened model name	Year of first forecast	Status
NCEP CFSv1	CFSv1	2011	Retired
NCEP CFSv2	CFSv2	2011	Active
CMC1 CanCM3	CanCM3	2011	Active
CMC2 CanCM4	CanCM4	2011	Active
GFDL CM2.1-aer04	CM2.1-aer04	2011	Active
GFDL CM2.5-FLOR-A06	FLOR-A	2014	Active
GFDL CM2.5-FLOR-B01	FLOR-B	2014	Active
IRI ECHAM4.5-Anomaly	IRI-A	2011	Retired
IRI ECHAM4.5-Direct	IRI-D	2011	Retired
NASA GMAO-062012	NASA	2011	Active
COLA-RSMAS CCSM3	CCSM3	2011	Active
COLA-RSMAS CCSM4	CCSM4	2014	Active

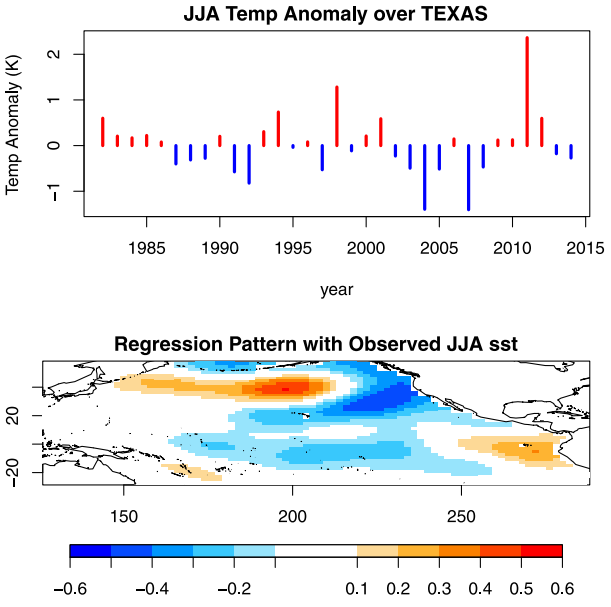

FIG. 2. (top) The observed June–August Texas-area temperature anomaly and (bottom) the regression coefficients with concurrent Pacific SSTs. Anomaly refers to departures from the 1982–2014 mean. The units of the regression coefficients are SST in kelvin per Texas-area temperature in kelvin.

approach is based on cross validation, in which one sample is withheld and the remaining sample is used to estimate a regression model, after which the resulting regression model is used to predict the withheld sample. The procedure is repeated for each sample in turn until all samples have been used at least once for validation. In the case of PCR, we emphasize that the empirical orthogonal functions (EOFs) and centering are recomputed for each new training set in the cross-validation procedure. In practice, though, the results are essentially unchanged if the EOFs are computed once for the whole period. However, we found that leave-one-out cross validation yielded unrealistically high skill scores (e.g., ≥ 0.9) when estimating models from observations, for reasons that were difficult to ascertain. To explore this situation further, we applied tenfold cross validation and found that the regression models estimated from observations had no significant skill for any choice of regularization. These conflicting results demonstrate the danger in estimating regression models from observations of the size considered here (i.e., 33 yr or fewer). However, different cross-validation methods yielded similar results when applied to dynamical model output. The results presented in section 4 are based on tenfold cross validation, which is a generally recommended method in the statistics literature (Hastie et al. 2009, p. 243).

Many studies suggest selecting the model that minimizes the cross-validated mean square error (CVMSE). In contrast, Hastie et al. (2009) recommend selecting the "simplest" model within one standard deviation of the minimum CVMSE. In the context of minimizing (6), a "simpler" model is a model with larger regularization parameter λ. However, our data are correlated with each other because we pool ensemble members and hindcasts initialized one month apart. Therefore, the sample standard deviation probably overestimates the true standard deviation. Nevertheless, the standard deviation of the skill score will be shown in the results to follow as a reference.

3. Data

We first attempt to identify relations between SST and land variables in observational data. For the predictand, we use summer (June–August) temperature over the land region bounded by 94°–106°W and 26°–36°N, which is centered over Texas. This area is chosen because the drought or heat wave in that region during 2011 raised critical questions about the role of ocean temperatures and the extent to which such events can be predicted in the future (Hoerling et al. 2012). The present paper extends previous studies by examining the extent to which such events can be predicted on seasonal time scales by dynamical models and by regression models using SSTs as predictors. Observational estimates of summer land temperature are from the dataset of Fan and Van den Dool (2008), which is a combination of data from the Global Historical Climatology Network,

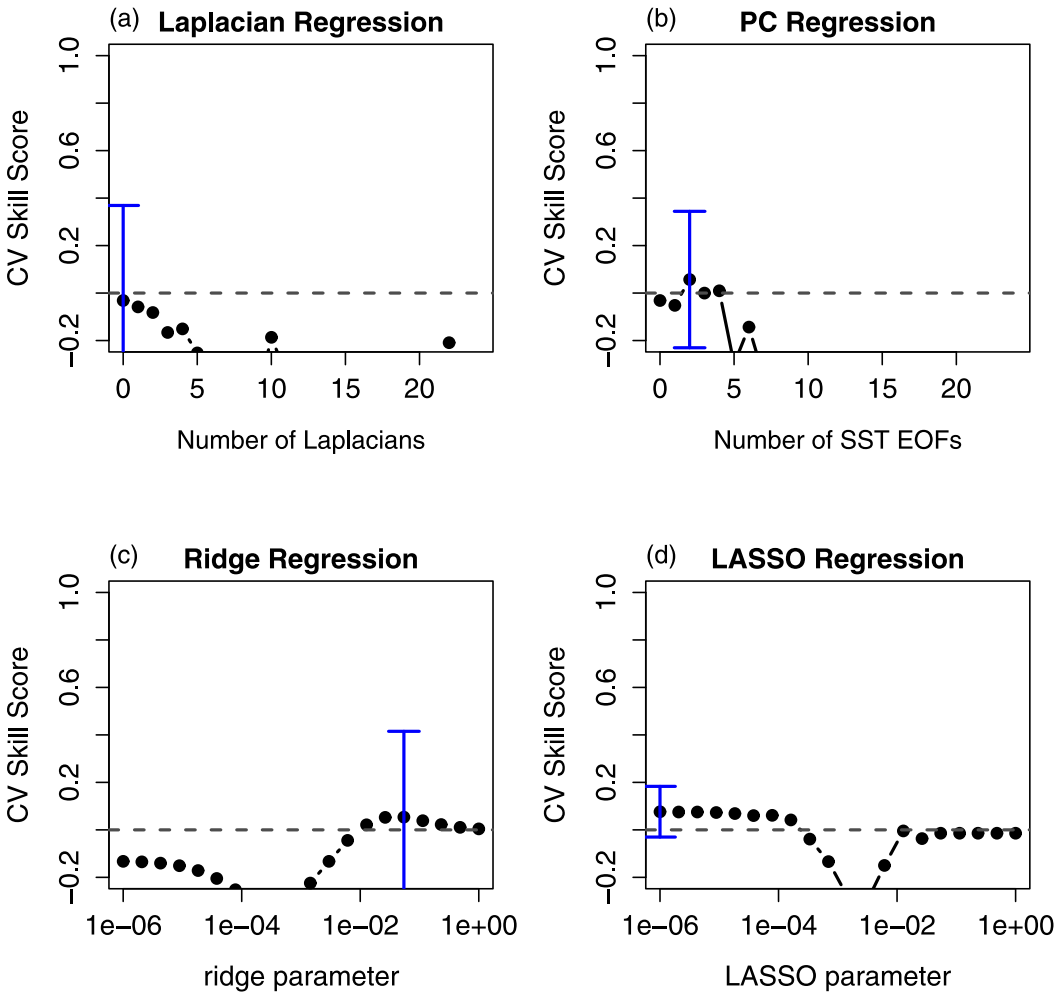

FIG. 3. CVSS for predicting Texas-area summer (JJA) temperature based on concurrent (JJA) Pacific SSTs between 30°S and 60°N using (a) a truncated set of Laplacians, (b) PCR, (c) ridge regression, and (d) LASSO. The error bar shows the standard error of the skill score at the maximum score.

version 2, and the Climate Anomaly Monitoring System.[1] For the predictors, we use the 3-month mean SST in the Pacific Ocean over 30°S–36°N derived from the NOAA Optimum Interpolation Sea Surface Temperature, version 2 (OISSTv2; Reynolds et al. 2002).[2]

We also derive SST–land relations from climate model simulations. Specifically, we use hindcasts and forecasts from the NMME (Kirtman et al. 2014) during the period 1982–2014. A list of models is given in Table 1. A hindcast refers to a dynamical model prediction of historical data in which the verification is available at initialization time. Each model generates an ensemble forecast in which multiple predictions are generated from slightly different initial conditions, each of which are plausible realizations of the state of the system given the available observations. The reason for choosing this dataset is that the associated dynamical models have been designed and validated specifically for seasonal prediction, so these models are likely to capture realistic seasonal relations between SST and land variables. Unfortunately, seasonal prediction datasets are relatively short (e.g., about 30 yr). To increase the sample size, we pool individual ensemble members initialized in the months preceding the June–August verification period. To be clear, we do not use ensemble averages, but rather we attempt to find the relation between summer land and SST in individual ensemble members. To avoid differences due to different ensemble sizes, we use an

[1] Downloaded from http://www.esrl.noaa.gov/psd/data/gridded/data.ghcncams.html.

[2] Downloaded from http://www.esrl.noaa.gov/psd/data/gridded/data.noaa.oisst.v2.html.

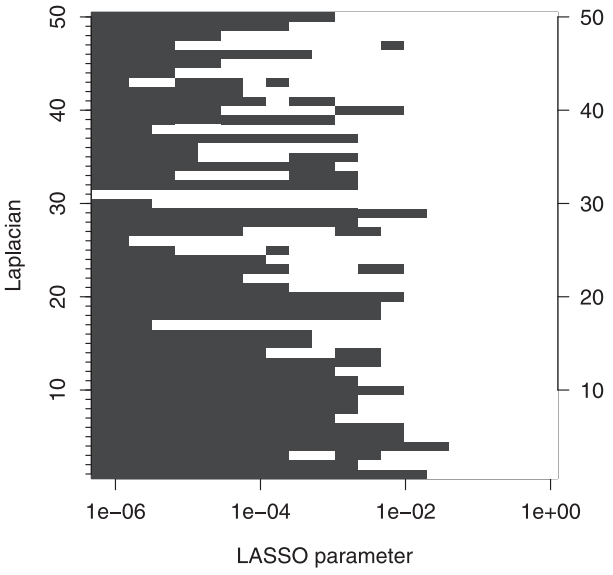

FIG. 4. Schematic indicating whether a given Laplacian eigenvector was selected (black) or not (white) by LASSO as a function of regularization parameter. LASSO is used to predict summer Texas-area temperature based on observed JJA Pacific SSTs between 30°S and 60°N.

equal number of ensemble members per model. Specifically, we use six members, which is the smallest ensemble size among the models in this dataset. For each model, then, there are 33 years, five initial months (January–May), and six ensemble members, giving a total of 990 samples per model. However, some models do not have the full 33 years; these models are designated as "retired" in Table 1.

4. Results

The variable we want to predict is summer (June–August) Texas-area temperature. We first consider predictions based on concurrent SST (i.e., June–August SST). Although predictions based on concurrent SST are not true predictions, they nevertheless are investigated extensively in seasonal prediction studies because they define teleconnection patterns and define an upper bound on predictability. Later (in section 4d) we consider time-lagged relations between SST and Texas-area temperature.

Both SST and Texas-area temperature exhibit a significant trend during the period under investigation. The trend presumably reflects the global warming signal and does not reflect a causal relation between land temperature and SST. If these trends are not removed, then all regression models have apparent skill even at large lags. To isolate predictability caused by SST anomalies, we remove the linear trend from SST and Texas-area temperature prior to analysis.

The observed anomaly time series, relative to the 1982–2014 mean, is shown in Fig. 2. The Texas heat wave of 2011 is evident. To gain insight into the Pacific SST pattern relevant to this prediction, it is customary to construct a regression map, which shows the least squares regression coefficient between Texas-area temperature T and local SST from the regression model:

$$\mathrm{SST}(s,n) = p(s)T(n) + \text{noise}, \tag{7}$$

where, as in section 2, s and n denote the spatial location and time step, respectively. The least squares solution is equivalent to estimating $p(s)$ independently and individually at each grid point. The resulting coefficients can be collected and displayed on a single map. The regression map $p(s)$ derived from observations, shown at the bottom of Fig. 2, is dominated by a hot spot in the north-central Pacific and negative values to the southeast. This regression pattern is quite similar to the actual SST anomaly that occurred in 2011 in association with the extreme Texas heat wave (Hoerling et al. 2012). Unfortunately, this pattern is not of direct use for prediction because the regression model (7) requires that the pattern be multiplied by Texas-area temperature, the variable we want to predict, which is not ordinarily available in a real prediction setting. Various studies have proposed procedures for deriving predictors from a regression map, but most of these procedures have been discredited (DelSole and Shukla 2009).

a. Predictions based on observations only

We use the methods discussed in section 2 to derive a prediction model for summer Texas-area temperature based on concurrent Pacific SST. The skill of the prediction model is measured by tenfold cross validation, as discussed in section 2. This procedure generates N predictions for N samples from which the mean square error (MSE) can be computed. The skill of the regression model is measured by the cross-validated skill score (CVSS):

$$\mathrm{CVSS} = 1 - \frac{\mathrm{MSE}}{\mathrm{var}(y)}, \tag{8}$$

where $\mathrm{var}(y)$ denotes the variance of summer mean Texas-area temperature. CVSS can be interpreted as a measure of the fraction of variance explained by the regression model.

The cross-validated skill of four regularized regression models are shown in Fig. 3. The error bar in each panel shows the standard error of the maximum score. None of the regression models show significant skill when estimated directly from observations. Results from PCR are essentially unchanged if the EOFs are

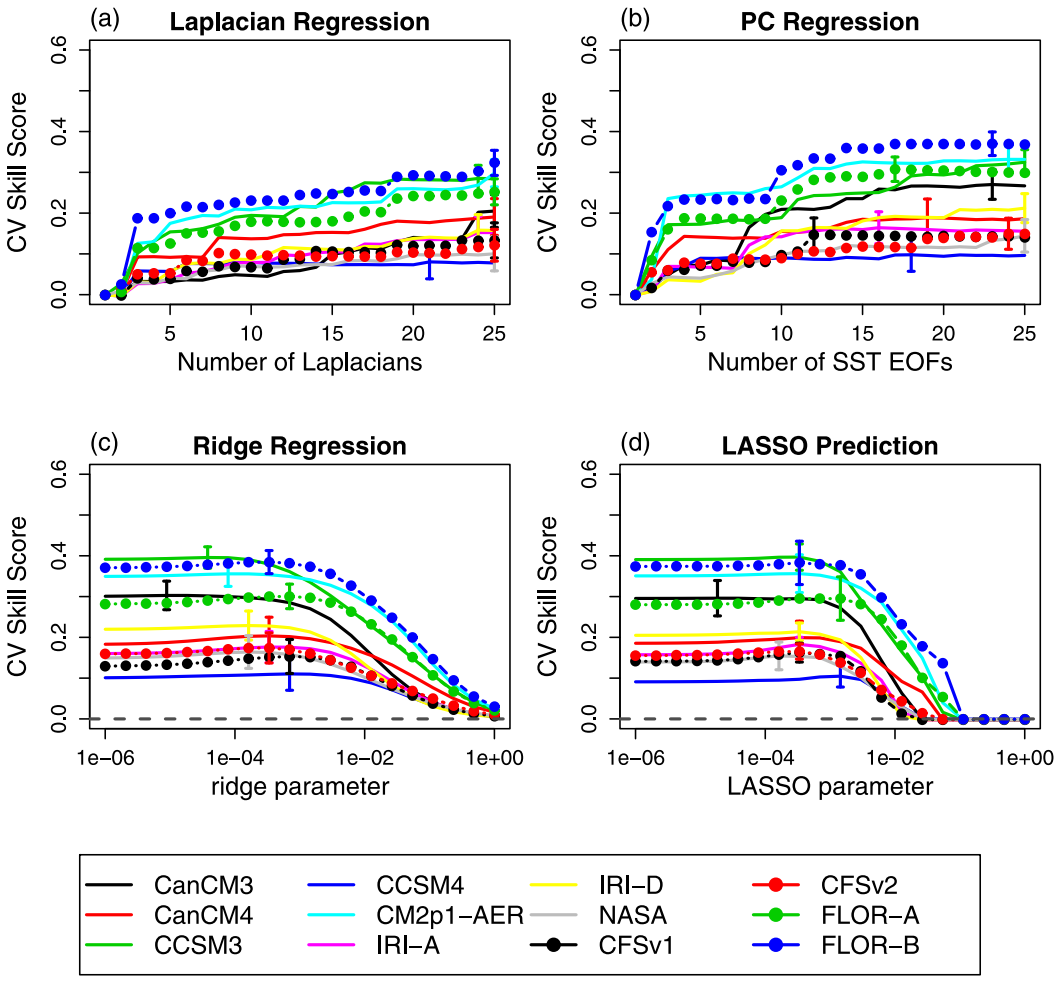

FIG. 5. CVSS for predicting Texas-area summer (JJA) temperature based on Pacific SSTs between 30°S and 60°N in the dynamical models using (a) a truncated set of Laplacians, (b) PCR, (c) ridge regression, and (d) LASSO. The error bars show the standard error of the skill score at the maximum score.

fixed during cross validation. In the remainder of the paper, the spatial patterns of the EOFs are computed from the entire dataset and held fixed during cross validation.

Recall that LASSO tends to set some coefficients exactly to zero. The distribution of nonzero coefficients as a function of regularization parameter is shown in Fig. 4. As expected, the number of nonzero coefficients decreases with increasing LASSO parameter. Also, the coefficients become slightly concentrated near the bottom of Fig. 4 as the regularization parameter increases, reflecting the filtering of small-scale spatial patterns with increasing regularization parameter.

b. *Regressions learned from dynamical models and tested in "model world"*

We now apply regularized regression to dynamical model output. Although the dynamical models were used originally to perform ensemble predictions, we disregard this fact and simply use the individual ensemble members as additional realizations of the system. This means that the initialization time is not relevant for prediction except insofar as it provides additional realizations. Accordingly, we pool the JJA Texas-area temperature and concurrent sea surface temperature from each ensemble member and for each start month January–May preceding the summer. The cross-validated skill scores for the various regression models are shown in Fig. 5. Many of the regression models have statistically significant skill for some choice of regularization parameter, in the sense that the cross-validated skill is positive and the standard error does not include zero. The fact that the skill is statistically significant (in contrast to observations) is attributed to the larger sample size as a result of pooling ensemble members and initial start months. Hastie et al. (2003)

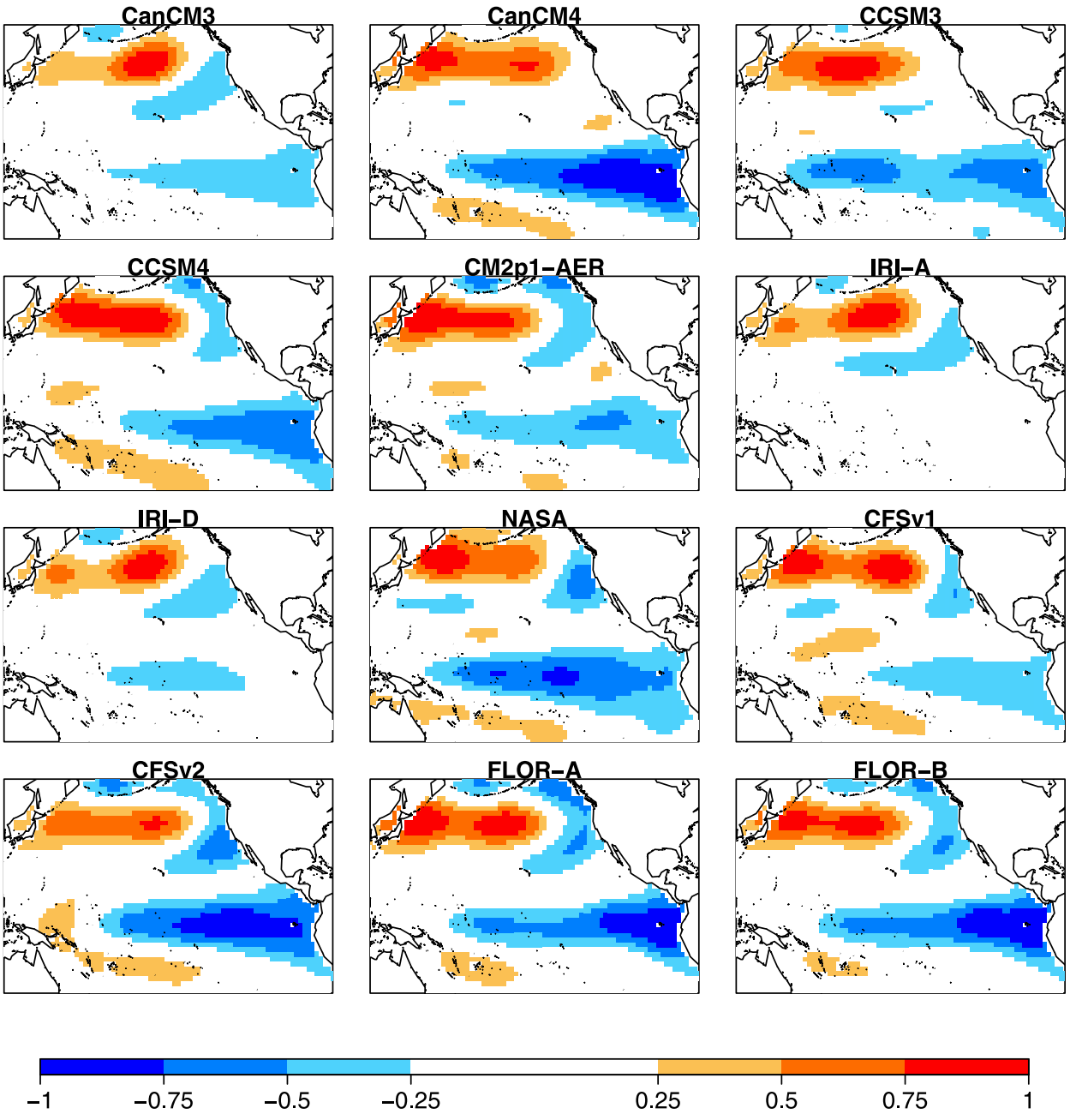

FIG. 6. Regression maps derived from NMME seasonal prediction models using LASSO. The regression pattern has been scaled so that the maximum absolute value is one. The regularization parameter is chosen to maximize the CVSS in each model separately.

suggest selecting the simplest model within one standard deviation of the best model. This criterion suggests selecting three Laplacian eigenvectors in Laplacian regression and 13 or fewer EOFs for PCR. It also suggests selecting a ridge parameter around 10^{-5} (when skill exists), and a LASSO parameter around 10^{-3}. Note that some models show skill even in the case of $\lambda \approx 0$, suggesting that the sample size is sufficiently large that regularization is not required (although it is noteworthy that the maximum number of Laplacian eigenvectors is 50, which is itself a kind of regularization).

To gain insight into the spatial structure associated with the predictive relations, we show the regression map between JJA SST and the Texas-area temperature predicted from those SSTs in the model. Equivalently, the regression map is defined in (7), except that the predictor $T(n)$ is replaced by the Texas temperature *predicted* by the regression model. For consistency, we use the

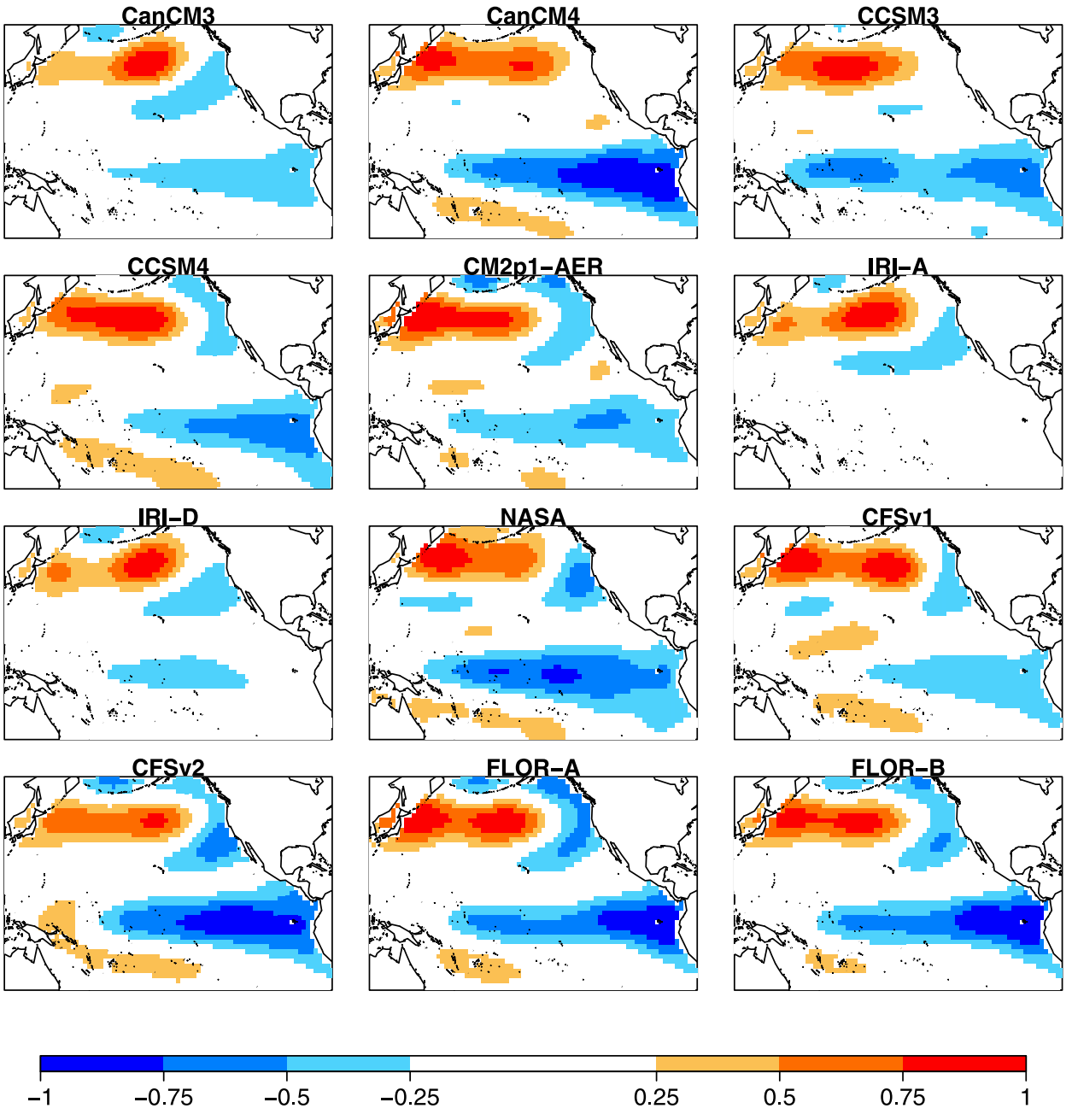

FIG. 7. As in Fig. 6, but using ridge regression.

regularization parameter that maximizes the CVSS in each model. The regression maps associated with LASSO and ridge are shown in Figs. 6 and 7, respectively. Most regression maps indicate that the predictive relation is dominated by a hot spot in the North Pacific and cooling to the southeast, consistent with the regression pattern derived from observations. We emphasize that these regression patterns were derived strictly from models with no input from observational data. The strong similarity between regression patterns derived independently from observations and distinct dynamical models gives us substantial confidence that SSTs and Texas-area temperature are indeed related in nature and that this relation is captured by dynamical models.

c. *Regressions learned from dynamical models and tested in observations*

We next test whether predictive relations derived from dynamical model output give skillful predictions of observations. Specifically, we derive a regression model

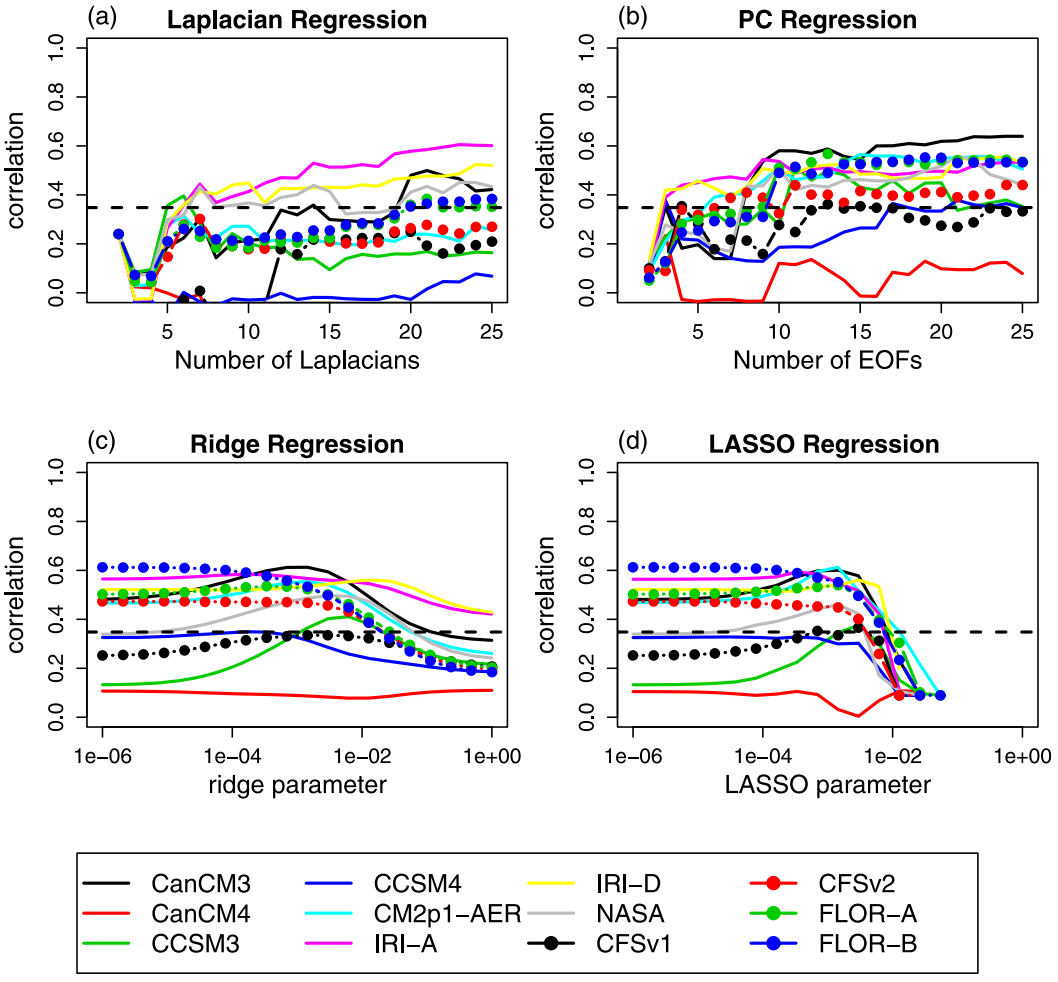

FIG. 8. Correlation skill of predicting summer Texas-area temperature using regression models derived from NMME dynamical forecast models, as a function of regularization parameter, using (a) a truncated set of Laplacians, (b) PCR, (c) ridge regression, and (d) LASSO. The horizontal dashed line is the 5% significance threshold of skill.

from each dynamical model and then use that regression model to predict observed JJA Texas-area temperature based on observed JJA Pacific SST. We emphasize that observations are not used to construct the regression models directly. Although observations were used to initialize the dynamical models, these observations do not constrain the land–SST relation in JJA because we use only hindcasts initialized in antecedent months. In this sense, JJA observations constitute independent verification data.

The correlation skill between predicted and observed Texas-area temperature for each model and regularization is shown in Fig. 8. Regression based on a truncated set of Laplacian eigenvectors (Fig. 8a) yields skillful models only when estimated from certain dynamical models and for truncations greater than five. However, we might have selected only three eigenvectors based on the training results shown in Fig. 5, which would not have yielded skill. In contrast, regularized regressions derived from dynamical models using PCR, ridge, and LASSO yield skillful predictions of observations for reasonable choices of regularization parameter, although for certain models the skill is insignificant, but still positive. The contrast between Figs. 3 and 8 is striking.

d. Prediction based on antecedent SST

Instead of predicting JJA Texas-area temperature using concurrent SSTs, we now discuss predictions using antecedent (MAM) SSTs. Repeating the above analysis, the correlation skill of regression models estimated from dynamical model output, and then used to predict observations, is shown in Fig. 9. In contrast to the results shown in Fig. 8, only two regression models have

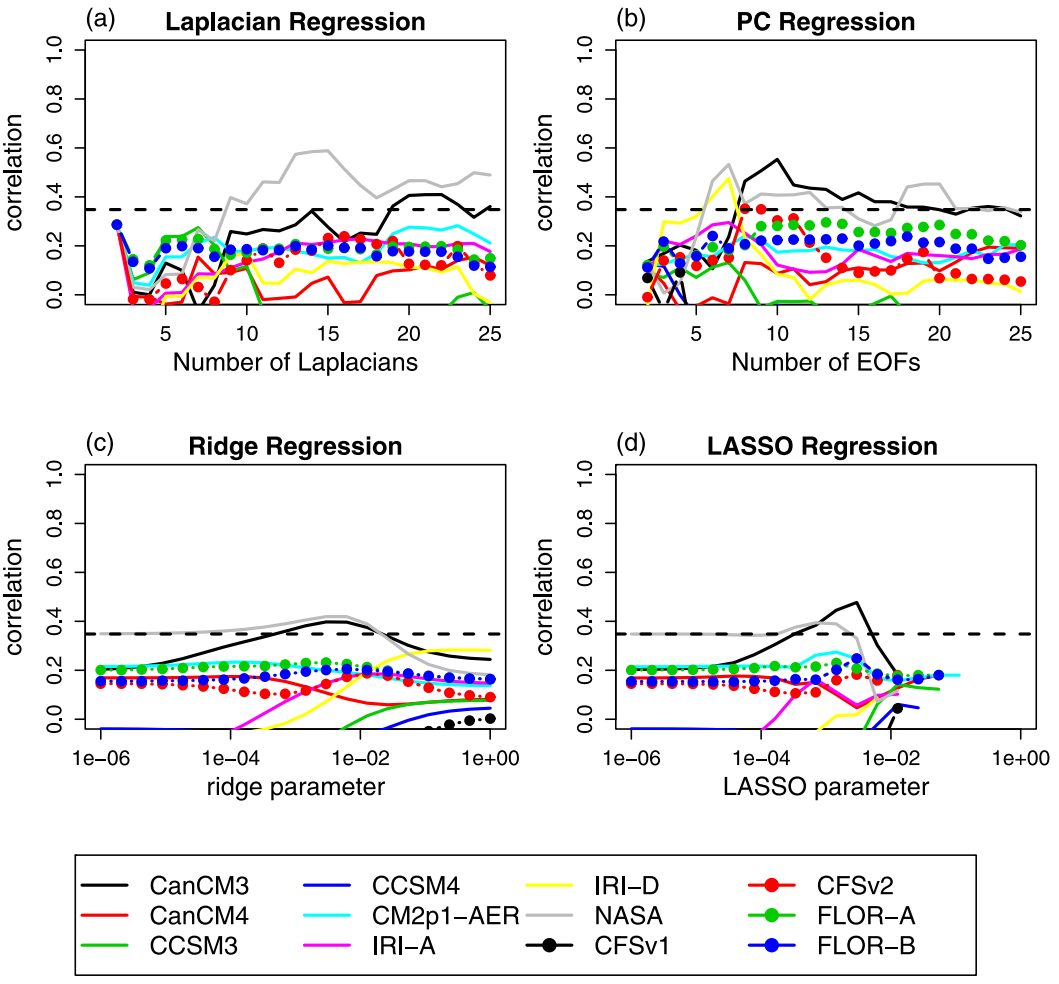

FIG. 9. As in Fig. 8, but using one-season (MAM) lagged SSTs.

statistically significant skill (viz., those estimated from NASA and CanCM3), and this skill occurs only for a narrow range of regularization parameters. However, these models do not have significant skill in their respective dynamical models (not shown), so it is not likely that they would have been chosen based on training data. This result suggests that while there exists a concurrent relation between SST and Texas-area temperature, there exists little to no "precursor" SST to Texas-area temperature, at least in dynamical models (Fig. 9), in observations (not shown), and from the multimodel regression (also not shown).

e. *Skill of NMME hindcasts*

We now use the above results to understand the skill of NMME hindcasts. For each model initialized in May, the correlation skill of Texas-area temperature is shown as the red triangles in Fig. 10. Figure 10 shows that none of the models have skill. Yet we showed above that regression models derived from dynamical model output produce skillful predictions based on concurrent SST. How can these results be reconciled? An obvious answer is that dynamical models were not able to predict the JJA SST accurately. In other words, predictability exists when the SSTs are known exactly, but the SSTs themselves could not be predicted with skill. To explore this hypothesis further, we show in Fig. 10 the correlation between model-predicted and regression-predicted JJA Texas-area temperature for May starts, where the regression prediction is based on concurrent SSTs in the model (black-filled circles). These correlations are fairly high and demonstrate that the regression model can recover much of the model hindcast when given the correct SSTs. If the model SSTs are then replaced by observed SSTs in the regression model, the resulting correlation skill, shown as the blue plus symbols in Fig. 10, typically are statistically significant and lie between the previous two correlations. These results

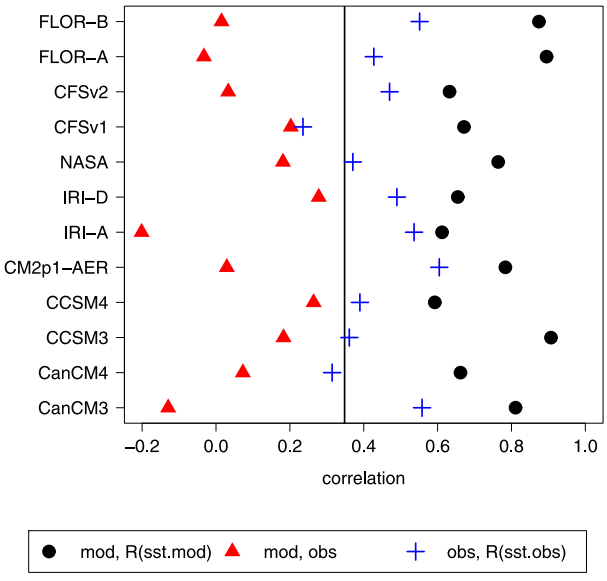

FIG. 10. Correlations of Texas-area temperature between model hindcast and regression based on model SST (black-filled circles), model hindcast and observed (red-filled triangles), and observed and regression based on observed SST (blue plus symbols). The regression is based on ridge regression using $\lambda = 10^{-5}$. Model hindcast quantities are based on a six-member mean initialized in May. The vertical line is the 5% significance threshold of skill.

suggest that the reason the NMME models had no significant skill in predicted Texas-area temperature is not because they could not capture the SST teleconnections but because the SST themselves could not be predicted accurately.

f. Multimodel regression

We next consider regressions based on a combination of dynamical models. Specifically, we estimate a regression model by pooling all 11 000 and more dynamical model hindcasts (as well as all initial start times and ensemble members) together. The resulting model is then used to predict observations. The resulting skill as a function regularization parameter is shown in Fig. 11. The multimodel regression model has significant skill and, moreover, tends to have skill near the top of the individual model skills (as can be seen by comparing Fig. 11 and Fig. 8).

g. Sensitivity to sample size

We now quantify the sensitivity of skill with respect to sample size. Figure 12 shows the correlation skill as a function of sample size of regression models for predicting observed Texas-area temperature, where the sample is drawn randomly from the 11 000 and more realizations of the multimodel hindcast dataset. For each sample size and for each choice of regularization parameter, 10 independent estimations are performed. The results indicate that negative correlations are common even for sample sizes around 100 and that statistically significant correlations become robust only for sample sizes greater than 500. For less than 50 samples (the real world of observations), the correlations tend to be insignificant. Truncating based on Laplacian eigenvectors fails to produce skillful regressions for large sample size, in contrast to PCR, ridge, or LASSO.

5. Summary and conclusions

A characteristic problem in statistical seasonal prediction is to find a predictive relation between one variable and a field of other variables (e.g., SST). To solve this problem, we proposed a regularized regression procedure in which a linear prediction model is estimated under a smoothness constraint on the coefficients. The smoothness constraint is motivated by the fact that small-scale SST structures are difficult to observe and not robust across models, so it is prudent to filter them out. This constraint was imposed by representing the predictor field by a sum of Laplacian eigenvectors and then constraining the corresponding weighting coefficients to be small or zero. Because the temporal variance of SST patterns tends to increase with spatial scale, constraining the coefficients of the Laplacian eigenvectors directly, without the standard rescaling, is tantamount to penalizing small-scale patterns more strongly than large-scale patterns. The resulting regression problem is equivalent to ridge or LASSO regression and can be solved by standard algorithms.

In addition, we explored an unconventional procedure in which an empirical prediction model is estimated from dynamical model output and then verified against observations—the reverse of the traditional order. Obviously, this approach works only if the underlying dynamical model is realistic. However, to the extent that the dynamical model is realistic, a major advantage of this approach is that the sample size available from dynamical models is orders of magnitude larger than of observational data, which allows for the estimation of significantly more robust empirical models. Also, large differences in the skill of the empirical models indicate large differences in the realism of the underlying dynamical models, which may provide clues for model development.

The above approach was used to predict summer Texas-area temperature based on concurrent Pacific SST during the period 1982–2014. For comparison, we also applied principal component regression, in which the regression model is based on a linear combination of

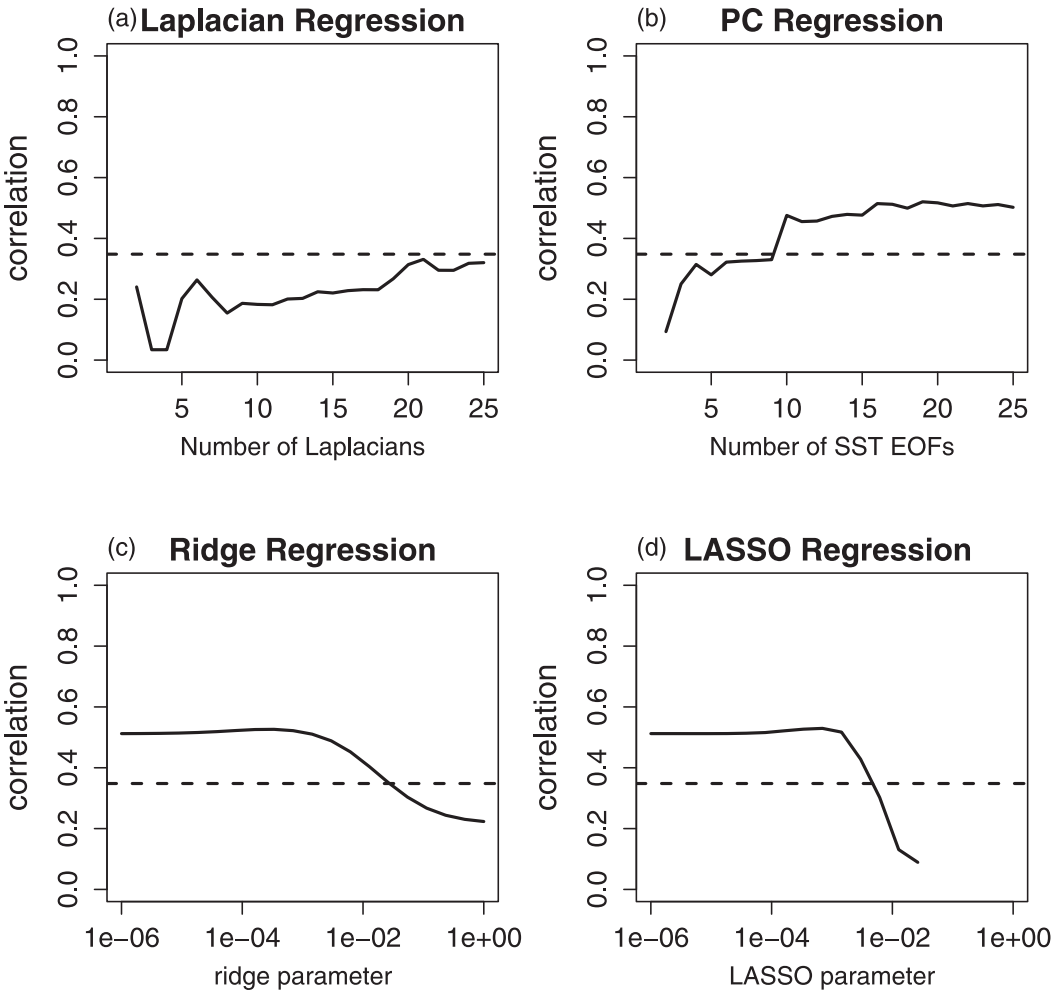

FIG. 11. As in Fig. 8, but for a multimodel regression model.

the leading EOFs of SST. An important consideration is that temperature during this period experiences significant trends and these trends can dominate the regression relations. Accordingly, all analyses were conducted for detrended data. When applied to observations, none of the regularized regression models had significant cross-validated skill. In contrast, when the regularized regression models were trained on dynamical model output, they were able to predict dynamical model output with significant skill. The higher skill is attributed to the larger sample size afforded by multiple ensemble members and start times. The associated regression patterns were similar across models, in the sense that all of them showed a ''hot spot'' in the North Pacific, but the cooling patterns in the tropical Pacific and eastern North Pacific were less robust. The regression models derived from dynamical models were then used to predict observations. We emphasize that the regression models derived from dynamical model output do not utilize any observational data. Although observations were used to initialize the dynamical models, these observations do not constrain the land–SST relation in JJA because we use only hindcasts initialized in antecedent months. In this sense, JJA observations constitute independent verification data. PCR, ridge, and LASSO models trained on dynamical model output gave skillful predictions of observational data for reasonable choices of regularization parameter. In addition, a multimodel regression model was estimated by pooling all model hindcasts together (over 11 000 samples) and shown to produce skillful predictions of observed Texas-area temperature, with skill near the top of any individual model skill.

Most regression models had insignificant skill in predicting observed anomalies with antecedent SSTs. Although a small number of models did have correlation skill exceeding the significance threshold, this occurred for a narrow range of regularization parameters that

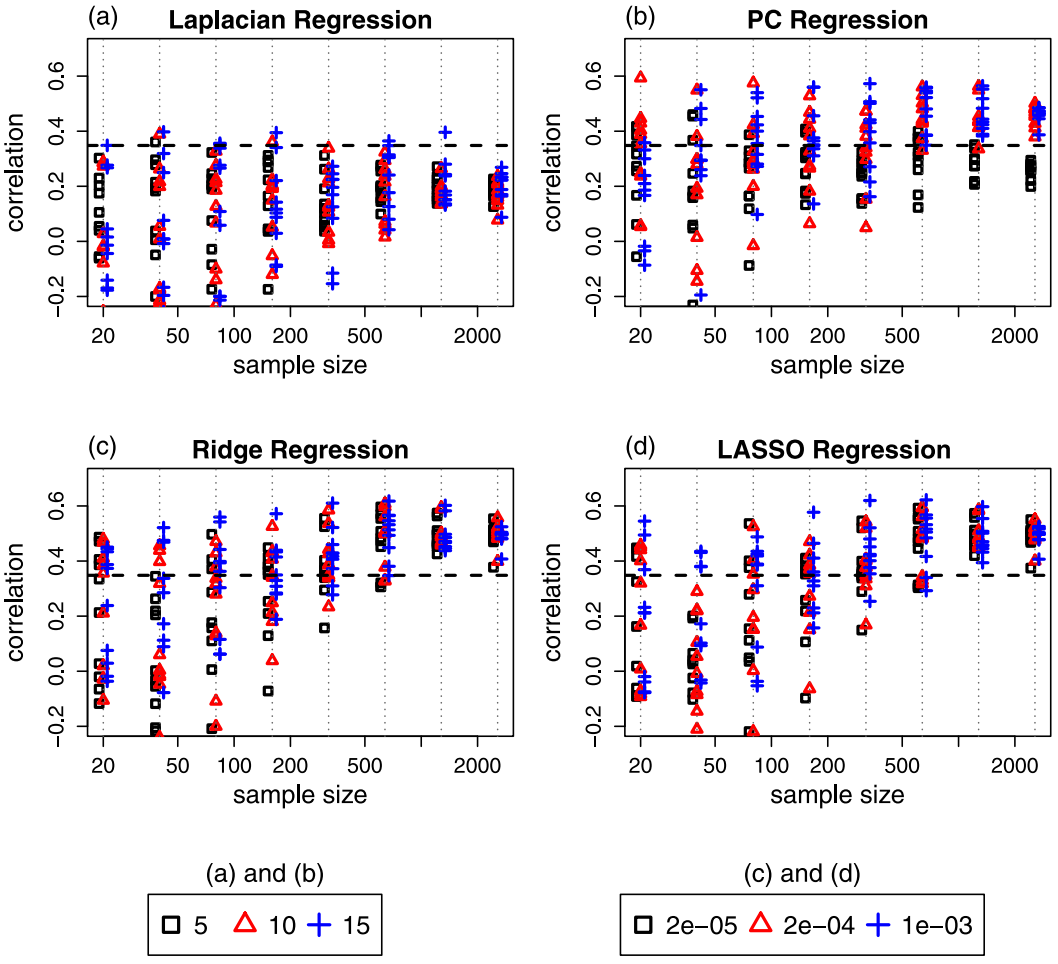

FIG. 12. Correlation skill of regression models for predicting observed JJA Texas-area temperature as a function of sample size, where the sample used to estimate the regression is drawn from dynamical model output. The different regressions are based on (a) a truncated set of Laplacians, (b) PCR, (c) ridge regression, and (d) LASSO. The different colors and symbols show different values of the regularization parameter, as indicated in the bottom legends, and the 10 points in each column show 10 different estimations from the full 11 000 and more sample multimodel hindcast dataset.

would not necessarily have been selected based on training data. The fact that skill apparently disappears when using antecedent SST suggests that summer Texas-area temperature is difficult to predict in advance. Consistent with this, none of the NMME hindcasts themselves had skill in predicting JJA Texas-area temperature based on May (or earlier) start months. How can the NMME models have no skill when regression models derived from them do? The key is to recognize that the regression models had skill only when using concurrent SSTs; they did not have skill when using antecedent SSTs. The fact that regression models derived from dynamical model output had skill suggests that the reason the NMME models had no skill is not because they did not capture realistic relations with SST but because they could not accurately predict the relevant SST pattern.

We also explored the sensitivity of skill to the sample size used to estimate the regression model. For sample sizes less than 50, few of the regression models had statistically significant skill. The regression models tended to have positive skill only for sample sizes greater than 100 and statistically significant skill for sample sizes greater than 500. These results imply that estimating robust, skillful regression models from 50 years of observational data is difficult.

This study demonstrates that regularized regression models trained on dynamical model output can yield empirical prediction models with significant skill, even though the same methods applied to observations do not

yield skillful empirical prediction models. This study also confirms that dynamical models can capture the SST regression pattern related to Texas-area temperature. It is especially noteworthy that a regression model derived strictly from dynamical model output can produce skillful predictions in observational data. Such consistency not only demonstrates the validity of the regression methodology but enhances confidence in the existence of a predictive relation in both models and observations.

Acknowledgments. We thank Michael K. Tippett, Andrew Rhines, acting as reviewer, and another reviewer for insightful comments that led to substantial improvements in methodology and clarifications in the final paper. T.D. was supported by the National Science Foundation (AGS-1338427 and CCF-1451945), the National Aeronautics and Space Administration (NNX14AM19G), and the National Oceanic and Atmospheric Administration (NA14OAR4310160). A.B. was supported by the National Science Foundation (CCF-1451945 and IIS-1447566). The views expressed herein are those of the authors and do not necessarily reflect the views of these agencies.

REFERENCES

Barnston, A. G., 1994: Linear statistical short-term climate predictive skill in the Northern Hemisphere. *J. Climate*, **7**, 1513–1564, doi:10.1175/1520-0442(1994)007<1513:LSSTCP>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.

Caldwell, P. M., C. S. Bretherton, M. D. Zelinka, S. A. Klein, B. D. Santer, and B. M. Sanderson, 2014: Statistical significance of climate sensitivity predictors obtained by data mining. *Geophys. Res. Lett.*, **41**, 1803–1808, doi:10.1002/2014GL059205.

Charney, J. G., 1971: Geostrophic turbulence. *J. Atmos. Sci.*, **28**, 1087–1095, doi:10.1175/1520-0469(1971)028<1087:GT>2.0.CO;2.

DelSole, T., and J. Shukla, 2009: Artificial skill due to predictor screening. *J. Climate*, **22**, 331–345, doi:10.1175/2008JCLI2414.1.

——, and M. K. Tippett, 2015: Laplacian eigenfunctions for climate analysis. *J. Climate*, **28**, 7420–7436, doi:10.1175/JCLI-D-15-0049.1.

Efron, B., and C. Morris, 1971: Limiting the risk of Bayes and empirical Bayes estimators—Part I: The Bayes case. *J. Amer. Stat. Assoc.*, **66**, 807–815.

Fan, Y., and H. Van den Dool, 2008: A global monthly land surface air temperature analysis for 1948–present. *J. Geophys. Res.*, **113**, D01103, doi:10.1029/2007JD008470.

Geisler, J. E., M. L. Blackmon, G. T. Bates, and S. Muñoz, 1985: Sensitivity of January climate response to the magnitude and position of equatorial Pacific sea surface temperature anomalies. *J. Atmos. Sci.*, **42**, 1037–1049, doi:10.1175/1520-0469(1985)042⟨1037:SOJCRT⟩2.0.CO;2.

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

——, ——, and ——, 2009: *Elements of Statistical Learning*. 2nd ed. Springer, 732 pp.

Hoerling, M., and Coauthors, 2012: Anatomy of an extreme event. *J. Climate*, **26**, 2811–2832, doi:10.1175/JCLI-D-12-00270.1.

Horel, J. D., and J. M. Wallace, 1981: Planetary-scale atmospheric phenomena associated with the Southern Oscillation. *Mon. Wea. Rev.*, **109**, 813–829, doi:10.1175/1520-0493(1981)109<0813:PSAPAW>2.0.CO;2.

Hoskins, B. J., and D. J. Karoly, 1981: The steady linear response of a spherical atmosphere to thermal and orographic forcing. *J. Atmos. Sci.*, **38**, 1179–1196, doi:10.1175/1520-0469(1981)038<1179:TSLROA>2.0.CO;2.

Kirtman, B. P., and Coauthors, 2014: The North American Multimodel Ensemble (NMME): Phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bull. Amer. Meteor. Soc.*, **95**, 585–601, doi:10.1175/BAMS-D-12-00050.1.

Lo, A., H. Chernoff, T. Zheng, and S.-H. Lo, 2015: Why significant variables aren't automatically good predictors. *Proc. Natl. Acad. Sci. USA*, **112**, 13 892–13 897, doi:10.1073/pnas.1518285112.

Nastrom, G. D., and K. S. Gage, 1985: A climatology of atmospheric wavenumber spectra of wind and temperature observed by commercial aircraft. *J. Atmos. Sci.*, **42**, 950–960, doi:10.1175/1520-0469(1985)042<0950:ACOAWS>2.0.CO;2.

Opsteegh, J. D., and H. M. Van den Dool, 1980: Seasonal differences in the stationary response of a linearized primitive equation model: Prospects for long-range weather forecasting? *J. Atmos. Sci.*, **37**, 2169–2185, doi:10.1175/1520-0469(1980)037⟨2169:SDITSR⟩2.0.CO;2.

Quan, X., M. Hoerling, J. Whitaker, G. Bates, and T. Xu, 2006: Diagnosing sources of U.S. seasonal forecast skill. *J. Climate*, **19**, 3279–3293, doi:10.1175/JCLI3789.1.

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

Ropelewski, C., and M. Halpert, 1987: Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation. *Mon. Wea. Rev.*, **115**, 1606–1626, doi:10.1175/1520-0493(1987)115<1606:GARSPP>2.0.CO;2.

Saito, N., 2008: Data analysis and representation on a general domain using eigenfunctions of Laplacian. *Appl. Comput. Harmon. Anal.*, **25**, 68–97, doi:10.1016/j.acha.2007.09.005.

Shukla, J., and J. L. Kinter, 2006: Predictability of seasonal climate variations: A pedagogical view. *Predictability of Weather and Climate*, T. N. Palmer and R. Hagedorn, Eds., Cambridge University Press, 306–341.

Simmons, A. J., J. M. Wallace, and G. W. Branstator, 1983: Barotropic wave propagation and instability, and atmospheric teleconnection patterns. *J. Atmos. Sci.*, **40**, 1363–1392, doi:10.1175/1520-0469(1983)040⟨1363:BWPAIA⟩2.0.CO;2.

Straus, D., J. Shukla, D. Paolino, S. Schubert, M. Suarez, P. Pegion, and A. Kumar, 2003: Predictability of seasonal mean atmospheric circulation during autumn, winter, and spring. *J. Climate*, **16**, 3629–3649, doi:10.1175/1520-0442(2003)016<3629:POTSMA>2.0.CO;2.

Taylor, J., and R. J. Tibshirani, 2015: Statistical learning and selective inference. *Proc. Natl. Acad. Sci. USA*, **112**, 7629–7634, doi:10.1073/pnas.1507583112.

Tibshirani, R., 1996: Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc.*, **58B**, 267–288, doi:10.1111/j.1467-9868.2011.00771.x.

——, 2011: Regression shrinkage and selection via the lasso: A retrospective. *J. Roy. Stat. Soc.*, **73B**, 273–282, doi:10.1111/j.1467-9868.2011.00771.x.

Trenberth, K. E., G. W. Branstator, D. Karoly, A. Kumar, N.-C. Lau, and C. Ropelewski, 1998: Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperatures. *J. Geophys. Res.*, **103**, 14 291–14 324, doi:10.1029/97JC01444.

Van Houwelingen, J. C., 2001: Shrinkage and penalized likelihood as methods to improve predictive accuracy. *Stat. Neerl.*, **55**, 17–34, doi:10.1111/1467-9574.00154.

Yang, X., and T. DelSole, 2012: Systematic comparison of ENSO teleconnection patterns between models and observations. *J. Climate*, **25**, 425–446, doi:10.1175/JCLI-D-11-00175.1.