Specification of Wintertime North American Surface Temperature

Timothy DelSole[*] and J. Shukla

George Mason University, Fairfax, VA and

Center for Ocean-Land-Atmosphere Studies, Calverton, MD

September 19, 2005

[*] Corresponding Author Address: Timothy DelSole, Center for Ocean-Land-Atmosphere Studies, 4041 Powder Mill Rd., Suite 302, Calverton, MD 20705-3106. Email address: delsole@cola.iges.org

Abstract

The extent to which wintertime North American surface temperature can be specified based on simultaneous sea surface temperature (SST) is quantified for the period 1982-1998. The term “specification” indicates that the predictor and predictands are not lagged in time, as would be the case for true prediction. Four state of the art, general circulation models (GCMs) and linear empirical models with predictors derived from observations and dynamical models are considered. Predictors are derived from model hindcasts using principal component analysis (PCA), canonical correlation analysis (CCA), and discriminant analysis. The last technique has appeared in the climate literature but its use in the present context appears new. A distinguishing feature of this paper is that several methods and models are compared in a common framework.

The specification skill of GCMs for the period 1982-1998 is statistically significant in the northwestern region near Washington, British Columbia, and central Canada, with some local correlations exceeding 0.6. The specification skill of GCMs is comparable to, or better than, the skill of the best empirical model, for the particular 17-year period examined.

No single specification strategy was found to improve the model hindcast skill in all cases. Predictors derived from discriminant analysis generally lead to larger skill than predictors based on PCA or CCA. The signal-to-noise ratio varies greatly among models and appears to be, if anything, inversely related to the specification skill when discriminants are used as predictors. Predictors based on 500hPa geopotential height can lead to specification skill at least as good as predictors based on land surface temperature. Evidence is presented for the existence of at least two distinct dynamically predictable components of land surface temperature arising from two distinct “flavors” of SST anomalies associated with El Nino and La Nina.

1 Introduction

This paper presents an assessment of the degree to which wintertime North American surface temperature can be specified based on observations and state-of-the-art atmospheric general circulation models. The analogous question for statistical forecasts has been addressed comprehensively by Barnston and Smith (1996; BS hereafter), who use canonical correlation analysis (CCA) to predict seasonal mean surface temperature based on sea surface temperature (SST). Although BS examine global surface temperature and precipitation over all seasons and four lead times, the present paper focuses particularly on wintertime North American surface temperature predicted with simultaneous SST. Following BS, this is called a *specification* problem rather than a *prediction* problem, since the SST and land surface temperature are not lagged in time. BS show that skill degrades with lead time, hence the specification skill is an upper bound on the prediction skill. BS find that, for these particular forecasts, the specification skill is modest except for a large region in the west-central part of the continent in which the skill is negligible (see their fig. 3). The largest skill was found in the southeast United States and western Canada, where correlation skill exceeds 0.6. Based on the results of CCA, BS suggest that the skill arises from ENSO variations in the SST predictor, which are associated with a roughly north-south dipole structure and an east-west dipole structure in land surface temperature, and an interdecadal trend of tropical SST warming. These results are consistent with previous studies by Barnston (1994) and Ropelewski and Halpert (1986).

In contrast to the statistical prediction problem, the extent to which wintertime North American surface temperature can be specified based on dynamical model forecasts remains an open question. Part of the reason for this is that dynamical models have significant biases and systematic errors which require some form of correction to render the associated predictions useful. Unfortunately, no universal method for correcting forecasts exists. Perhaps the most common correction is to subtract the forecast climatology from each field to produce unbiased forecasts. Corrections based on local linear regression, such as Model Output Statistics (MOS) (Wilks 1995), are used routinely at operational forecast centers to issue local forecasts based on dynamical model output. Rajagopalan et al. (2002) recently proposed a Bayesian methodology for combining the "prior" climatological distribution with an ensemble of dynamical model forecasts to produce a single, corrected, probabilistic forecast. This methodology, which is used at the International Research Institute to produce tier-two predictions of global precipitation and temperature, yields Rank Probability Skill Scores of 20% or less in the south and eastern part of the U.S., and negligible everywhere else in the U.S (Barnston et al. 2003). Feddersen et al. (1999) proposed an alternative correction method based on Canonical Correlation Analysis (CCA) or Singular Value Decomposition (SVD), in which large scale fields produced by the model essentially are replaced by other large scale fields derived from observations. Feddersen et al. (1999) found that while the method improved the temperature and precipitation forecasts in many parts of the world, it failed to improve the ECHAM4 forecasts for wintertime North American temperature, ostensibly because the leading CCA and SVD patterns do not agree in time when cross validated.

Anderson et al. (1999) performed a simple bias correction to two atmospheric GCMs forced by observed SSTs and found that the corresponding forecasts of 700hPa geopotential height produced anomaly correlations around 0.35, whereas CCA forecasts produced higher correlations around 0.5. To the extent that 700hPa height correlates with surface temperature, this result implies that empirical models give better forecasts than bias-corrected dynamical models of wintertime surface temperature in North America.

A shortcoming of these and other predictability studies is that each study adopts a distinctive correction technique without comparing that technique to others. Consequently, it is difficult to gain a sense of how the different correction techniques compare with each other, and how well statistically corrected dynamical models perform relative to empirical prediction schemes. This paper attempts to shed light on these issues. Rather than attempt to reproduce the exact forecast scheme in the above cited papers, the details of which may have been tailored to specific data sets or dynamical models, we have adopted our own straightforward versions of many of these techniques, and introduced some new ones. The primary data set in this study is a set of hindcasts by four state-of-the-art general circulation models. These GCMs were run with specified SST and hence do not constitute genuine forecasts. The skill of these hindcasts can be interpreted as an upper bound on the forecast skill of the same models run with *specified* SST. In general, the predictability of hindcasts can be attributed to the imposed SST or initial conditions. However, since each model is initialized at least one month prior to the verification period, and since the decorrelation time of atmospheric disturbances is at most two weeks (DelSole 2001), it is reasonable to assume that the initial condition has little affect on the predictability of the hindcasts. Therefore, we ignore the influence of the initial condition and assume all predictability can be attributed to the SST.

Statistical prediction and statistical correction are mathematically equivalent problems. The difference lies in the choice of predictors: statistical prediction uses predictors from direct observations, whereas statistical correction uses predictors from dynamical model forecasts (which in turn depend on observations through initial conditions). The main arguments for drawing predictors from dynamical model forecasts is that the models may capture important nonlinear dynamics which are too complex to be incorporated in statistical models, and they allow application of Monte Carlo techniques to construct more complete probabilistic information. However, drawing predictors from dynamical model forecasts and observations requires a method for isolating useful predictors in high dimensional fields. How should this be done?

Recently, DelSole (2004, 2005) proposed a framework for quantifying predictability theory based on information theory. In this framework, the critical quantity is not the distribution of the forecast, as used in virtually all other predictability studies, but the conditional distribution of the verification given the forecasts. DelSole (2005) further showed that this conditional distribution depends on the forecast sample only through "potential predictable components," which are components in the forecast with non-vanishing mutual information. If the forecast and initial conditions (and boundary conditions) are joint normally distributed, then, in principle, the potential predictable components can be obtained directly from CCA. DelSole (2005) showed that in this case there is no loss of generality in considering only the ensemble mean fields. If the system is non-Gaussian, then CCA can find some of these components by virtue of their nonzero correlation with the boundary conditions or initial conditions. Schneider and Griffies (1999) proposed a procedure, called predictable component analysis, which identifies components that minimize the ratio of the forecast error variance to the climatological variance. In a perfect model scenario, the forecast error variance can be identified with the "noise" of the forecast ensemble, where noise is defined as the deviation of a forecast member from the ensemble mean.

In this case, predictable component analysis is equivalent to a statistical procedure known as discriminant analysis for finding components that optimally discriminate between the ensemble mean and deviations from the ensemble mean, in the sense that these components maximize the signal to noise ratio. To avoid confusion, we will call this technique signal-to-noise discriminant analysis. In this paper, we propose using discriminant analysis to identify predictable components in the forecast, and then using the amplitude of these components as predictors of land surface temperature. In summary, this paper explores four methods for identifying useful predictors from high dimensional forecast fields:

1) Principal component analysis of the ensemble mean model forecast field.

2) CCA between ensemble mean model forecast field and verification.

3) CCA between ensemble mean model forecast field and imposed SST.

4) Signal-to-noise discriminant analysis of the model forecast field.

Note that the role of CCA differs in the second and third methods. In the second, CCA extracts the structures in the forecast that are most correlated with the verification, whereas in the third CCA extracts the model response to SST anomalies. Importantly, only the second method includes information from the verification.

Evaluating the “best” set of predictors requires comparing hundreds of forecasts and hence requires an objective method for measuring skill and for selecting models. Following Barnston (1994), we adopt the correlation coefficient to measure the skill at each grid point, and the spatially averaged correlation as a measure of global skill. We also introduce a new measure of skill called the localized mutual information, which is a metric with attractive properties suggested from information theory. Both measures give essentially the same conclusions. In regards to model selection, we simply choose the model that maximizes the cross-validation skill. It should be noted that this selection criteria is biased– that is, the selected model usually achieves less skill in independent data than in the cross-validated data (DelSole and Shukla 2002).

A description of the empirical model, skill metrics, data, and dynamical models used in this study are given in the next two sections. This is followed by a discussion of the skill of dynamical model forecasts. Section 5 discusses three methods for extracting predictors from observations and model forecasts, namely principal component analysis, canonical correlation analysis, and signal-to-noise discriminant analysis. Section 6 discusses the results of "classical" CCA forecasts, in which EOFs of land surface temperature are the predictand and EOFs of model fields are the predictors. Section 7 discusses the results of specifying local land surface temperature based on the different predictors discussed above. The paper ends with a summary of the major conclusions.

2 Dynamical Models and Data

2.1 General Comments About the Forecast Data

The dynamical model forecasts used in this study are ensemble hindcasts by four state-of-the-art, general circulation models (GCMs) which were compiled as part of the Dynamical Seasonal Prediction (DSP) experiment (see special issue of *Quarterly Journal of the Royal Meteorological Society*, July 2000). The GCMs are from the Center for Ocean-Land-Atmosphere Studies (COLA), the National Aeronautics and Space Administration (NASA), and the National Centers for Environmental Prediction (NCEP).

The geographic domain of all analyses was restricted to North America between 20°N - 60°N, and 140W-60W. The northern latitudinal limit was chosen to avoid sea-ice complexities. All fields were averaged in time over January-February-March (JFM). This averaging was done because analysis of individual months revealed little or no predictability. The mean at each grid point over all years was subtracted from all fields.

2.2 COLA V1.11

The V1.11 COLA AGCM is a global primitive equation model with 18 sigma levels and R40 spectral truncation. The dynamical core and physical parameterizations are modified versions of the operational medium range forecast model at NCEP. The parameterization of deep convection is the relaxed Arawaka-Schubert (RAS) scheme of Moorthi and Suarez (1992). The land-surface model is the simplified Simple Biosphere model (SSiB) of Xue et al. (1991). For further details, consult Shukla et al. (2000).

2.3 NASA

The NASA Seasonal to Interannual Prediction Project AGCM (NSIPP-1) is a global primitive equation model with 34 sigma levels and 2° x 2.5° horizontal resolution. The dynamical core is the fourth-order grid point model of Suarez and Takacs (1995). Deep convection is parameterized by the RAS scheme of Moorthi and Suarez (1992). The land surface model is the MOSAIC LSM of Koster and Suarez (1992). For further details, consult Bacmeister and Suarez (2002).

2.4 NCEP

The NCEP Seasonal Forecast Model (SFM) is a global primitive equation model with 28 sigma levels and T63 spectral truncation. Deep convection is parameterized by the RAS scheme of Moorthi and Suarez (1992). The land surface model is the two-layer LSM of Pan and Mahrt (1987). For further details, consult Kanamitsu et al. (2002).

2.5 COLA v2.2 and COLAv2.2-50

The V2.2 COLA AGCM is a global primitive equation model with 18 sigma levels and T63 spectral truncation. The dynamical core is that of the National Center for Atmospheric Research (NCAR) Community Climate Model version 3 (CCM3) described in Kiehl et al. (1998). The prognostic variables are treated spectrally, except for water vapor which is advected using the semi-Lagrangian technique. The parameterization of deep convection is the RAS scheme of Moorthi and Suarez (1992). The land surface model is the simplified Simple Biosphere model (SSiB) of Xue et al., with the revised soil and vegetation parameters of Dirmeyer and Zeng (1999). COLAv2.2 denotes hindcast data for the period 1982-1998, and COLAv2.2-50 denotes hindcast data for 1950-1999.

2.6 Initial conditions

The atmospheric state of each GCM was initialized in November based on the NCEP/NCAR Reanalysis for the years 1982–1998 (except for the COLAv2.2-50 runs, which covered the years 1950-1999). However, the starting date and ensemble size differs from model to model. COLA V1.11 and COLA V2.2 were initialized at 0000 UTC for the last five days in November (Nov. 26, 27, 28, 29, 30). Five additional initial conditions were constructed by adding to each initial condition the difference between the analyses at 12 hours before and after the date in question. This procedure results in a total of 10 initial conditions with which to produce a 10-member ensemble for each COLA model. NSIPP-1 was initialized at 0000 UTC

November 13-17, and at 1200 UTC 13-16 November, giving a total of 9 initial conditions with which to produce a 9-member ensemble. The NCEP SFM was initialized at 0000 UTC and 1200 UTC November 1-5, giving a total of 10 initial conditions with which to produce a 10-member ensemble. The extent to which differences in initial conditions influenced the skill of the specifications is difficult to ascertain based solely on this data set. However, the results of secs. 4, 6, and 7 reveal no systematic pattern between the skill of the specifications and the lead time, suggesting that differences in the initial conditions were not a significant factor in explaining the differences in the specifications.

The land surface initial condition for COLA V1.11, COLA V2.2, and NCEP SFM were obtained from a climatology compatible with the respective land surface models. The land surface initial condition for the NCIPP-1 model was obtained from a single, arbitrary December state from a previous experiment based on the NCIPP-1 model.

2.7 SST

The SST in the COLA V1.11, NCEP, and NSIPP-1 models were imposed at each time step based on a time-interpolated version of the optimal interpolation SST (OISST) data of Reynolds (see Reynolds and Smith 1994). The SST in the COLA V2.2 model was taken from the HadISST data of Rayner et al. 2003.

2.8 TCAS and Z500

The primary predictors in this study are land surface temperature (TCAS), 500hPa geopotential height (Z500), and SST. The land surface temperature is identified as the "canopy temperature" in the COLA and NASA models, and "ground temperature" in the NCEP model.

2.9 Observational Data

The SST data set used in all CCA calculations is the HadISST data set of Rayner et al. (2003), which is a monthly mean sea surface temperature field of on a 1° × 1° global grid. We consider SSTs only between 40°S and 40°N. The data set used for verifying the land surface temperature is the monthly surface temperature anomalies of Jones and Moberg (2003) on a 5˚ x 5˚ global grid. Grid boxes for which the land surface temperature data were missing for all three consecutive months during the time period under analysis were discarded. The verification grid is indicated in fig. 1, for instance, by the shaded grid cells.

3 Empirical Models and Skill Metrics

This section reviews the empirical models and skill metrics used in this study.

3.1 The Empirical Linear Prediction Model

The variable we want to specify, $y(t)$, is called the predictand, and the variables on which the specification is based, $x_1, x_2, \ldots, x_{K-1}$, are called the predictors. The specification equation is assumed to be of the linear form

$$y(t) = a_1 x_1(t) + a_2 x_2(t) + \ldots + a_{K-1} x_{K-1}(t) + a_K + \zeta(t), \qquad \mathbf{(1)}$$

where the K regression parameters $a_1, \ldots, a_K$ are to be determined from data, and ζ represents random error. Here t is discrete with N distinct values. The parameters $a_1, \ldots, a_K$ that minimize the mean squared errors ζ^2 are obtained by the method of least squares.

3.2 Skill Metrics

The primary measure of skill at a grid point used in this study is the correlation between forecast and observations. This metric is not affected by linear transformations of the forecast which are often employed to correct for systematic biases or height differences between model topography and observations. The 5% significance level of the correlation coefficient for 17 independent, normally distributed years is about 0.39.

In addition to the above local measure, a global measure of predictability and skill is needed. Mean square error is additive and hence can be aggregated to produce a single overall measure of skill. Thus, we define

$$MSE = [\,\overline{(f-v)^2}\,] \qquad \textbf{(2)}$$

where *f* denotes the forecast minus its 1982-1998 climatology, *v* denotes the verification minus its 1982-1998 climatology, square brackets denote a space average, and overbar denotes a time average. It is often more intuitive to normalize the mean square error such that it is unity for a perfect forecast ($f = v$) and vanishes for a climatological forecast ($f = 0$). We call this normalize MSE the "explained variance," EV, which is simply

$$EV = 1 - \frac{MSE}{[\overline{v^2}]}\,. \qquad \textbf{(3)}$$

The sampling distribution of *EV* depends on the spectrum of forecast variances among the grid points, so, in general, no a priori significance level can be stated for *EV*.

Unfortunately, neither *MSE* nor *EV* is invariant with respect to linear transformation of the data. This problem is acute over mountainous regions, where the model surface height is not necessarily accurate, and in regions with large biases. Both errors can be corrected to a large extent by linear regression methods. Barnston and Smith (1996) have used the spatially averaged correlation coefficient,

$$RHOM = [\, corr \;(f, v)\,] , \qquad \textbf{(4)}$$

as a global measure of forecast skill. This quantity is invariant with respect to linear transformations of the data. The sampling distribution of *RHOM* was derived by selecting pairs of uncorrelated Gaussian variables 17 times, corresponding to 17 years, averaging over 48 independent grid points, and computing the resulting correlation coefficient for 1000 trials. The simulated values were found to be less than 0.07 in 990 cases, so the value 0.07 has been adopted as the 1% significance level for RHOM.

The spatially averaged correlation does not seem to have a natural interpretation in the context of predictability. Recently, DelSole (2004, 2005) proposed a framework for quantifying predictability based on information theory in which a lower bound estimate of the true (but unmeasurable) predictability is given by the mutual information between *f* and *v*. This quantity can be interpreted as a measure of the dependence between two variables, and, equivalently, as the reduction of uncertainty in *v* when *f* becomes known. If the two variables *f* and *v* are bivariate normal, then the mutual information is

$$I(v\,;f) = -\frac{1}{2}\log\left(1-\rho^2\right) , \qquad \textbf{(5)}$$

where ρ is the correlation coefficient between the two variables. In the multivariate case, mutual information is given by

$$\mathrm{MI} = -\frac{1}{2}\log\left(1-\rho_1^{c2}\right)\left(1-\rho_2^{c2}\right)\ldots\left(1-\rho_M^{c2}\right), \qquad \textbf{(6)}$$

where $\rho_1^c, \rho_2^c, \ldots, \rho_M^c$ are *canonical correlations* between $\boldsymbol{f}$ and $\boldsymbol{v}$ (DelSole 2004). In practice, this quantity is undefined when the dimension of the system exceeds the number of independent samples. To overcome this problem, we propose a new metric, called *localized mutual information (LMI)*, which is the spatially averaged mutual information at each local region, modified to penalize against negative correlations:

$$\mathrm{LMI} = -\frac{\delta}{2A} \log\left(1-\rho_1|\rho_1|\right)\left(1-\rho_2|\rho_2|\right)\ldots\left(1-\rho_M|\rho_M|\right), \qquad \textbf{(7)}$$

where $\rho_1, \rho_2, \ldots, \rho_M$ are the *local* correlations between f and v at each grid point, δ is the area at each grid point (assumed constant), and A is the total area. *LMI* is proportional to *MI* in the case in which each field f and v have no autocorrelations in space. Although *LMI* can be dominated by a single large correlation, this situation did not occur in the present study. Note that for small, positive correlations,

$$\mathrm{LMI} \approx \frac{\delta}{2A}\left(\rho_1^2 + \rho_2^2 + \ldots + \rho_M^2\right). \qquad \textbf{(8)}$$

Since the squared correlation coefficient equals the fraction of explained variance in linear regression theory (DelSole and Chang 2003), (8) implies that, for small positive correlations, *LMI* is the spatial average of the fractional explained variance.

The sampling distribution of *LMI* can be obtained analytically from the known distribution of the sample correlation coefficient. However, it is much easier to compute the sampling distribution by Monte Carlo methods as discussed above. For a set of independent 17 years and 48 grid points, the estimated 1% significance level for *LMI* was found to be 0.03.

3.3 Anomaly Pattern Correlation

To connect with previous studies, we also measure the hindcast skill of individual years using the pattern correlation coefficient. For reasons that are discussed in the appendix, we consider only the "uncentered" correlation coefficient, defined as

$$ACC = \frac{[fv]}{\sqrt{[f^2][v^2]}} . \qquad \textbf{(9)}$$

3.4 Cross-Validation

Using the same data to train and select an empirical model leads to artificial skill, by which we mean that the value of skill is inflated relative to the skill the model would have if it were used to predict independent data. To avoid artificial skill to some extent, we adopt a cross-validation procedure (see Michaelsen (1987) and Barnston and Ropelewski (1992)). In this procedure, each year of an N-year data set is set aside in turn and a regression model is trained based on the remaining N–1 year data set. All aspects of the training– including the computation of the mean, EOFs, predictors (e.g., canonical variates), and regression coefficients– are based on the N - 1 years independent of the specification year. For each year held out, the empirical forecast model is used to "predict" the predictand based on the predictors in the withheld year. Repeating this procedure for each year yields a pair of specification-verification fields that can be used to measure the skill of the forecast scheme.

3.5 Model Selection

In this paper, we computed the average cross validated skill of regression models for every combination of (1) the type of predictors, (2) the number of predictors, and (3) the number of principal components used to construct predictors. Then, we chose the specific combination that minimized the mean square, cross validated forecast error. These computations reveal that the cross validated error depends sensitively on these parameters, implying that some objective selection criterion is absolutely essential for this type of study. It should be recognized that the level of skill found by this selection criterion is not likely to be maintained in independent data, because this criterion is itself subject to artificial skill (DelSole and Shukla 2002). In an effort to overcome this problem, we explored a variety of selection criteria, including Akaike's Information Criteria (AIC) (Burnham and Anderson 2002), statistical significance of the predictors (as determined by methods described in section 5), sensitivity of the predictor patterns to leave-one-out cross validation, etc. We found AIC to work very well in the local specification experiments described in sec. 7, in the sense that it often chose a model whose MSE differs from the optimal MSE by less than 5%. But AIC did not perform well in the large-scale specification experiments described in sec. 6. Despite its problems, the criterion of choosing the model with the minimum mean square, cross validated error has the virtue that it is easy to apply to univariate and multivariate regression models, and avoids uncertainties arising from untested or unreliable selection criteria.

4 Skill of the Dynamical Models

The skill of TCAS hindcasts by the dynamical models interpolated onto the observation grid is summarized in fig. 1 and table 2. All models show statistically significant skill in the northwest U.S. over Washington and California, with some locations having a correlation exceeding 0.6. The models also suggest skill in central and east Canada. The value of EV for two models is negative, indicating that the uncorrected hindcasts are systematically worse than a prediction based on climatology. In contrast, the mutual information of the hindcasts all lie in the range 0.11-0.20, which are well outside the 99% probability interval found earlier by simulation methods as discussed in sec. 3, and hence are unlikely to have arisen by chance.

The anomaly pattern correlation between observed and hindcast TCAS during 1982-1998 is shown in fig. 2. The El Nino years 1983, 1992, and 1998, and during the La Nina year 1989, were the most skillful hindcasts for all models, consistent with previous studies (QJRMS special issue, July 2000). The 1987 anomaly was predicted well by three models even though the SST indicated only a weak El Nino. The anomalies in 1984, 1990, and 1991 were poorly predicted by all models.

The signal to noise ratio of JFM land surface temperature for each dynamical model is shown in fig. 3 (in the notation of sec. 5, the signal to noise ratio at the k'th grid point is $(\mathbf{S}_{\langle \mathbf{f} \rangle})_{kk}$ / $(\mathbf{S}_{\mathbf{n}})_{kk}$). The maximum signal to noise ratios in all models tend to lie on a northwest-southeast patch in central Canada. All models suggest low signal to noise ratios in the west-central U.S, precisely where BS found no skill for their statistical specification. Regions of large signal-to-noise ratios are not necessarily associated with large specification skill (compare figs. 1 and 3). The COLAV1.11 signal to noise ratios are much larger than those of other models. The large ratios randomly distributed along coasts are probably an artifact of the SST specification. The noise variance of all four models (not shown) are fairly consistent with each other, so the above differences in signal to noise ratios can be attributed primarily to differences in signals.

5 Predictors

We test three methods for finding predictors, namely principal component analysis, canonical correlation analysis, and signal-to-noise discriminant analysis. These methods are reviewed below. Our notation is as follows. The *M*-member anomaly forecast data set at a fixed lead time is $\boldsymbol{f}_1(t), \boldsymbol{f}_2(t), \ldots, \boldsymbol{f}_M(t)$, where the subscript refers to the ensemble member and t refers to the year. The ensemble mean is denoted by $\langle\rangle$ and defined as

$$\langle \boldsymbol{f} \rangle = \frac{1}{M} \sum_{i=1}^{M} \boldsymbol{f_i}(t) \, . \qquad \textbf{(10)}$$

The sample covariance matrix of the ensemble mean forecast is defined as

$$\boldsymbol{S}_{\langle \boldsymbol{f} \rangle} = \overline{\left(\langle \boldsymbol{f} \rangle - \overline{\langle \boldsymbol{f} \rangle} \right) \left(\langle \boldsymbol{f} \rangle - \overline{\langle \boldsymbol{f} \rangle} \right)^T} \qquad \textbf{(11)}$$

where the bar ($^{-}$) denotes a time average over all years, and superscript T indicates a matrix transpose. The spread of the forecast ensemble about the ensemble mean is measured by the noise covariance matrix, defined as

$$\boldsymbol{S_n} = \left\langle \overline{\left(\boldsymbol{f} - \langle \boldsymbol{f} \rangle - \overline{\boldsymbol{f} - \langle \boldsymbol{f} \rangle} \right) \left(\boldsymbol{f} - \langle \boldsymbol{f} \rangle - \overline{\boldsymbol{f} - \langle \boldsymbol{f} \rangle} \right)^T} \right\rangle . \qquad \textbf{(12)}$$

The anomaly verification data is denoted $\boldsymbol{v}(t)$ and its covariance matrix is

$$\boldsymbol{S_v} = \overline{\left(\boldsymbol{v} - \bar{\boldsymbol{v}} \right) \left(\boldsymbol{v} - \bar{\boldsymbol{v}} \right)^T} \, . \qquad \textbf{(13)}$$

Finally, the cross-covariance matrix between $\boldsymbol{v}$ and $\langle \boldsymbol{f} \rangle$ is denoted

$$\boldsymbol{S}_{\boldsymbol{v} \langle \boldsymbol{f} \rangle} = \overline{\left(\boldsymbol{v} - \bar{\boldsymbol{v}} \right) \left(\langle \boldsymbol{f} \rangle - \overline{\langle \boldsymbol{f} \rangle} \right)^T} \, . \qquad \textbf{(14)}$$

5.1 Principal Component Analysis (PCA)

Principal component analysis (PCA) decomposes a multivariate time series into an ordered set of orthogonal, uncorrelated components such that the first K components capture the maximum possible variance out of all possible sets of K vectors. The spatial patterns associated with principal components, called empirical orthogonal functions (EOFs), are the eigenvectors of the sample covariance matrix. Thus, the EOFs of the ensemble mean forecast are the eigenvectors of $\mathbf{S}_{\langle \mathbf{f} \rangle}$, and the EOFs of the verification are the eigenvectors of $\mathbf{S}_{\mathbf{v}}$. Projecting the (normalized) EOFs on the original data set yields time series called the principal components (PCs), which can be interpreted as the amplitude of each EOF. PCA is discussed in von Storch and Zwiers 1999, for example, to which we refer the reader for further details.

Principal component analysis was performed on the JFM averaged anomaly fields. For each year being specified, the EOFs were computed from data strictly from the remaining years; that is, the EOFs were recomputed during cross-validation.

The leading EOFs and PCs of the ensemble mean TCAS of all four models are shown in fig. 4. It is immediately apparent that the NCEP EOF is unlike the other EOFs, in the sense that its maximum loadings occur over the Great Lakes whereas those of the other EOFs occur well to the northwest of the Great Lakes. Furthermore, the PCs of the other three models have relatively large amplitudes during the major ENSO years 1983, 1989, and 1998. In contrast, the leading PC of the NCEP model has the same sign for 1989 and 1998, even though the NINO3 values for those two years are of opposite sign.

5.2 Canonical Correlation Analysis (CCA)

Canonical correlation analysis (CCA) is a procedure for finding a linear combination of variables in one data set, and a second linear combination of variables in a second data set, such that the correlation between the resulting combinations is maximized. If the two data sets in question are the verification and ensemble mean forecast, then the first step in CCA is to solve the following eigenvalue problems

$$\begin{aligned} \boldsymbol{S}_{v\langle f\rangle}\, \boldsymbol{S}_{\langle f\rangle}^{-1}\, \boldsymbol{S}_{\langle f\rangle v}\, \boldsymbol{x} &= \lambda\, \boldsymbol{S}_{v}\, \boldsymbol{x} \\ \boldsymbol{S}_{\langle f\rangle v}\, \boldsymbol{S}_{v}^{-1}\, \boldsymbol{S}_{v\langle f\rangle}\, \boldsymbol{y} &= \lambda\, \boldsymbol{S}_{\langle f\rangle}\, \boldsymbol{y}, \end{aligned} \qquad \textbf{(15)}$$

where the eigenvectors $\boldsymbol{x}$ and $\boldsymbol{y}$ contain the weighting coefficients for the desired linear combination. The time series produced by a linear combination is called a canonical variate. For each canonical variate, there corresponds a canonical pattern. The canonical patterns $\boldsymbol{p}_x$ and $\boldsymbol{p}_y$ associated with the eigenvectors $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, are

$$\boldsymbol{p}_x = \boldsymbol{S}_v\, \boldsymbol{x} \qquad \boldsymbol{p}_y = \boldsymbol{S}_{\langle f\rangle}\, \boldsymbol{y}. \qquad \textbf{(16)}$$

The (suitably normalized) canonical variates can be interpreted as the amplitude of the canonical pattern at each point in time.

Importantly, CCA is invariant with respect to nonsingular linear transformations. Owing to this invariance, it can be shown that CCA is equivalent to an SVD of a "pre-whitened" cross-covariance matrix. By "pre-whitened," we mean that the data have been transformed such that their covariance matrix equals the identity matrix. An example of such a transformation, which is commonly done in CCA, is to project the data onto the leading EOFs and then to normalize the PCs to unit variance. The particular application by Feddersen et al. (1999) is not equivalent to this because they normalize each grid point, rather than each PC. Regression forecasts can be computed conveniently from the canonical patterns and variates. For details of this and other aspects of CCA, we refer the reader to Barnett and Preisendorfer (1987) and DelSole and Chang (2003).

In this work, CCA is performed only on the principal components of fields. Whenever a hindcast field is analyzed in CCA, only the ensemble mean is used. There is no loss of generality in using ensemble means if we consider only linear prediction methods (see DelSole 2005). Also, the number of PCs for the two variables is the same, so that the cross-covariance matrix (14) between the two variables is always square.

The leading TCAS pattern from CCA between TCAS and SST for each model is shown in fig. 5. The figure suggests that the linear response of each model to SST is characterized by a large-scale pattern oriented northwest-southeast over Canada, plus a smaller scale pattern of opposite sign to the south. The canonical pattern from the COLAv1.11 model has very poor cross validation properties and should not be given too much weight (the cross-validated canonical variates have nearly vanishing correlation). The leading canonical variate for the NCEP model has relatively weak amplitude during 1983, a strong El Nino, and relatively large amplitude during 1997, a weak El Nino. The other canonical variates have significant amplitude during the El Ninos of 1983 and 1998.

5.3 Signal-to-Noise Discriminant Analysis (S2NDA)

Signal-to-noise discriminant analysis (S2NDA) identifies the linear combination of forecast variables that best discriminates between fluctuations in the ensemble mean, called the signal, and fluctuations about the ensemble mean, called the noise. The optimal weighting coefficients $\boldsymbol{z}$ are the eigenvectors of the generalized eigenvalue problem

$$\boldsymbol{S}_{\langle f \rangle} \boldsymbol{z} = \lambda \boldsymbol{S}_n \boldsymbol{z} . \qquad \textbf{(17)}$$

The time series produced by different eigenvectors are mutually uncorrelated and ordered by decreasing discriminantory power, as measured by the ratio of signal variance to noise variance. To each eigenvector $\boldsymbol{z}$, there corresponds a discriminant pattern $\boldsymbol{d}$ given by

$$\boldsymbol{d} = \boldsymbol{S}_n \boldsymbol{z} . \qquad \textbf{(18)}$$

The (suitably normalized) disciminant time series gives the amplitude of the discriminant pattern at each point in time. This procedure is discussed in Venzke et al. (1999), to which the reader is referred for further details. Note that S2NDA does not involve observational data.

Theoretically, if the model is perfect, such that the noise variance equals the forecast *error* variance, then S2NDA is equivalent to predictable component analysis discussed by Schneider and Griffies (1999). If the variables are further assumed to be joint normally distributed, then S2NDA yields the same discriminant patterns as CCA between the ensemble mean forecast and observation (DelSole and Chang 2003; DelSole 2004, 2005). In this work, S2NDA is computed from the leading PCs of the “total” forecast members (no ensemble averaging), then the signal and noise components are computed directly from these PCs.

The leading discriminants of TCAS for each model are shown in fig. 6. The structures seen in fig. 6 are similar to those seen in fig. 5 and to some of those seen in fig. 4. All of the discriminant patterns have small loadings in the southwest and southeast U.S. Most discriminants have strongest amplitudes during strong ENSO years. Interestingly, the signal to noise ratios differ widely among the models, by as much a factor of 3. The figures also give the mutual information if the “noise” is identified with “error,” and the climatology is identified as the “signal plus noise,” given by

$$\mathrm{MI} = -\frac{1}{2}\log\left(\frac{noise}{signal + noise}\right) = +\frac{1}{2}\log\left(\frac{signal}{noise} + 1\right). \qquad \textbf{(19)}$$

6 Specification of Large Scale Temperature

In this section, we consider specification models in which both predictors and predictands are derived from CCA. More specifically, CCA is applied to the observed land surface temperature and a model hindcast variable to construct canonical variates, then, the amplitude of the model hindcast canonical variate is used as a predictor for the amplitude of the canonical pattern associated with the observed TCAS data. Because the predictands are derived from leading EOFs, which tend to be large scale, this procedure is essentially a specification of the large-scale temperature field. This will be called "classical CCA," since it is a standard method discussed in Barnett and Preisendorfer (1987) and DelSole and Chang (2003). Table 3 summarizes the results for CCA forecasts that minimize the cross validated, mean square error. Comparison between tables 2 and 3 show that (small sample) CCA does not consistently improve the skill of the original forecasts. However, in most cases the majority of the predictability is captured with 1-2 predictors and predictands. This suggest that the predictability of TCAS in the models depends on a few coherent patterns. The table shows that predictors based on model TCAS recover about as much skill as predictors based on model Z500. This result is consistent with the theory that the imposed SST drives a large-scale response in Z500 which in turn is correlated with the response in TCAS (which in turn is detected by CCA).

The local specification skill of the classical CCA forecasts, in which ensemble mean TCAS is a predictor, is shown in fig. 7. Relative to the skill of the raw model forecasts, the CCA forecast improves the local skill in many local regions, with a major exception being the

COLAv1.11 model. The failure of the latter model to produce good predictors appears to be related to the fact that the COLAv1.11 model has a very large signal that is not well correlated with observations, as mentioned in sec. 4. The negative correlations in the southwest U.S. are due to the fact that the forecast for those regions is close to climatology, which gives strong negative correlations under cross validation (Barnston and van den Dool 1993). The difference between the (uncentered) pattern correlations for the CCA forecast and the raw model hindcasts are shown in fig. 8. The figure shows that the classical CCA forecast improves the raw model hindcasts more often than it degrades, although in certain isolated years classical CCA leads to a substantial degradation of specification skill.

The last row of table 3 shows the skill of a classical CCA forecast of TCAS based on SST, in which the statistical model was trained using data during the period 1950-1999, but verified over the period 1982-1998. Comparison of the results in tables 2 and 3 reveal that the skill of the purely statistical forecast is comparable to that of the raw model hindcasts, and comparable to the skill of classical CCA corrected forecasts. This result shows that the current state-of-the-art dynamical models provide hindcasts of wintertime land surface temperature that are at least as skillful as statistical models. For reference, the local skill and yearly pattern correlations of this model are given in fig. 9.

Note that the best statistical model was a “one EOF” CCA forecast. This model is formally equivalent to predicting the amplitude of the leading EOF of observed TCAS, based solely on the leading PC of SST. In contrast, Barnston and Smith (1996; BS hereafter) reported that the best CCA forecast model was based on approximately 10 EOFs. This difference can be attributed to the difference in verification periods between our study and that in BS. In particular, when our CCA forecast is verified over the longer period 1950-1999, the cross-validated mean square error was minimized when 11 out of 13 canonical variates were used, more or less consistent with the finding of BS. Moreover, the structure of the skill map for these forecasts was reasonably consistent with the results of BS, and the magnitudes were comparable to that of BS, including in the southeast region where BS found correlations exceeding 0.6. The spatially averaged correlation was 0.45, which is slightly larger than BS's value of 0.38. It should be recognized that there are numerous differences between BS and this study, including that BS considered a slightly different region of North America (in particular, their domain extended to 80°N, which includes a substantial area of statistically significant skill which is not included in our analysis), used a different SST data set, namely that of Smith et al. (1996), used a different surface temperature data set, namely CAMS (Ropelewski et al. 1985), controlled for outliers, filled observation gaps by horizontal interpolation, and considered a different period, namely 1950-1992. The degree of similarity between our results and BS's results, despite the differences in analysis, suggests that our statistical forecast model has captured essentially the same predictability as reported in BS.

Figure 10 illustrates the skill of virtually all forecast models considered in this paper (some of which will be introduced in the next section). As can be seen, classical CCA forecasts based on NCEP model variables can give superior hindcasts overall. The spatial distribution of local correlation skill (fig. 7) reveals significant skill over a large area, with correlations exceeding 0.7 in many places. Although these specifications are based on the first two canonical variates, a forecast based on just the leading canonical variate explains *negative* 11% in the case of TCAS, and *negative* 23% in the case of Z500, strongly suggesting that all of the skill comes from the second canonical variate of each variable. The second canonical pattern pair and associated time series for TCAS are shown in fig. 11. We see that the observed temperature pattern is dominated by a dipole oriented primarily north-south, whereas the modeled pattern has a dipole oriented northwest-southeast. The lack of significant peaks at 1983, 1989, and 1998 in the variates suggests that this structure responds to a different "flavor" of ENSO than those measured by the NINO1, NINO2, or NINO3 indices, a point to which we will return.

The distinctiveness of the NCEP model, as compared to the other three models, becomes even more apparent when one compares the correlation of the model PCs with the SST PCs. In particular, the correlation between the leading PC of SST, and the leading PCs of TCAS in the NASA, NCEP, COLAv1.11, COLAv2.2 models, are 0.85, 0.20, 0.70, 0.71, respectively. These results show that the NCEP model responds relatively weakly to the leading EOF of SST, compared to the other models. In contrast, the cross-validated correlation between the *second* PC of SST, and the above models, are 0.2, 0.5, 0.1, 0.1. To obtain the spatial structure of the SST which gives rise to the NCEP-TCAS pattern shown in fig. 11, we computed the covariance between the associated time series (bottom panel of fig. 11) and local SST. The result is shown in fig. 12. We see that the strongest SST anomalies are found almost entirely in the NINO3.4 region. This structure turns out to be nearly the same as the *second* leading canonical pattern between NCEP TCAS and SST. It turns out that the first canonical pattern between NCEP TCAS and SST has dominant loadings in the NINO1, NINO2, NINO3 regions. However, forecasts based on this leading pattern lead to very poor forecasts (i.e., skill worse than climatology). These results suggest that the NCEP model does have some response to SST anomalies in NINO1-3 regions, but that this response is not representative of that which occurs in the true system. Rather, most of the skill in the NCEP model appears to arise from SST anomalies in the NINO3.4 region.

Barsugli and Sardeshmukh (2002) computed the sensitivity of seasonal anomalies over North America to localized SST anomalies in a particular version of the NCEP model, and found that the SST region of maximum sensitivity was mostly in the NINO4 region, which overlaps with the NINO3.4 region. It is noteworthy that we reached a similar conclusion with a similar model strictly from the results of the hindcast experiments. The present study adds to this conclusion by showing that it is model dependent, since significant responses to SST anomalies in the NINO3.4 region could not be detected in the other models. Moreover, our study not only confirms this sensitivity, but shows that this sensitivity is representative of the true system (as evidenced by the fact that a specification model based on the associated components in the NCEP model give nearly the highest skill of any model, as illustrated in fig. 10).

7 Specification of Small-Scale Temperature

In the previous section we examined hindcasts of large scale patterns defined by EOFs or linear combinations of EOFs. In this and the remaining sections, we consider the specification of *local* regions. More precisely, we consider specification of individual grid points of the surface temperature data set. Note that CCA is used in different ways in the different types of forecasts. In the "classical" CCA prediction, which was used in the previous section and has been discussed by Barnett and Preisendorfer (1987) and DelSole and Chang (2003), the forecaster first applies CCA to two data sets, one of which *must* be the field to be predicted. In such cases, the amplitude of one canonical variable is used to forecast the amplitude of the other. In contrast, the local prediction introduced here also applies CCA to two data sets, but neither of the data sets have to be the variable to be predicted. Rather, CCA produces time series, called canonical variates, which gives the amplitude of the corresponding canonical pattern at each point in time. These canonical variates are used as predictors in a linear regression prediction, as discussed in sec. 2.1. Thus, the predictors correspond to large-scale canonical patterns as before, but the predictands are local surface temperature quantities.

A list of all predictors derivable from CCA, signal-to-noise discriminants, and EOFs is given in table 4. The specification skill of local surface temperature using these predictors are given in tables 5-8 and illustrated in fig. 10. The following conclusions can be ascertained from these results:

1 When SST is a predictor, the best forecast (in a cross-validated mean square sense) is obtained using only the first PC of SST, and little advantage is gained by specifying small-scale TCAS over large scale TCAS (last rows of tables 3 and 5).

2 Specification skill of predictors based on the leading PC of ensemble mean hindcasts is usually less than that of the raw model hindcasts (c.f. tables 2 and 5).

3 When CCA is applied to observed TCAS and an ensemble mean model forecast, the resulting canonical variates lead to local predictions that are nearly indistinguishable from classical CCA forecasts (compare tables 3 and 6).

4 When CCA is applied to ensemble mean hindcast and SST, one of the two variates usually leads to greater skill than that of the raw model hindcasts (c.f. tables 2 and 7), but it is not clear which should be used a priori.

5 When the canonical variate involving SST is used as a predictor, the skill is at least as good as simply using the leading PC of SST (c.f. fig. 10).

6 Predictors based on Z500 can lead to specification skill at least as good as, if not better than, predictors based on TCAS (c.f. fig 10).

7 Predictors based on signal-to-noise discriminants lead to greater specification skill than that of the raw hindcasts in about half the cases (c.f. fig. 10). In contrast, classical CCA forecasts and predictors based on PCs of ensemble mean model forecast lead to improved specification skill in about a quarter of the cases.

8 In most cases, the optimal number of predictors is one, but the optimal number of principal components used to extract predictors varies considerably from 1-10. Presumably, these conclusions depend on the length of the data set.

9 Discriminants from the NCEP model consistently had moderate signal-to-noise ratios (~ 3) but the largest specification skill (RHOM ~ 0.3). Discriminants from the COLAv1.11 model consistently had the largest signal-to-noise ratios (>4) but the smallest specification skill (RHOM < 0.25). In several cases two models have virtually the same skill in specifying land surface temperature, but very different signal to noise ratio (over a factor of two difference). These results demonstrate that the signal to noise ratio, and hence the degree of potential predictability in the model, is not necessarily related to the skill with which the discriminants can be used to specify land surface temperature.

10 Surprisingly, the use of longer training sets (from 17 years to 50 years) rarely improved the specification skill of the regression model for the verification period 1982-1998 (compare COLAv2.2 to COLAv2.2-50 and HADSST to HADSST-50 in tables 3-8). Only predictors based on discriminant analysis consistently led to improved skill for longer data sets. The single most skillful specification model of this study used

predictors derived from the TCAS discriminants of COLAv2.2-50, with a spatial averaged correlation coefficient of 0.44 for the period 1982-1998.

11 Even if a model has a strong response to SST, as indicated by strong canonical correlations between model variable and SST, the skill based on the canonical variates may be worse than climatology.

12 No useful specification model could be derived from COLAv1.11's TCAS.

The fact that the TCAS variable from COLAv1.11 model consistently gives poor specification skill deserves comment. These poor forecasts are especially surprising given that the leading PCs of the COLAv1.11 are similar to the leading PCs of other models. For instance, the correlation coefficient between the leading PC of the COLAv1.11 model and COLAv2.2 model is 0.7, yet the cross validated specification skill based on these predictors is -0.10 and 0.21, respectively. Such a large difference in specification skill for two highly correlated predictors suggest a coding error, but this possibility has been checked extensively and nothing has been found to suggest an error. Moreover, the result is not mathematically impossible in the sense that at each grid point the correlations are consistent with the requirement of positive definiteness of correlation matrices. The problem appears to be that the COLAv1.11 model has a strong signal in TCAS that is not well correlated to the observed TCAS variability.

8 Summary and Discussion

The major results of this paper, which of course are confined to the methods and data used in this paper, are as follows:

1. This paper quantified the specification skill of wintertime North American surface temperature by four state-of-the-art, general circulation models, integrated with observed

SST. The spatially averaged point-by-point correlation skill was 0.22 - 0.32, depending on the model, although the local skill exceeded 0.6 in some regions (over Washington, British Columbia, and North Dakota, depending on the model). The models were most skillful during strong ENSO years.

2. The best empirical model (as measured by RHOM, EV, or LMI) of wintertime North American surface temperature for the period 1982-1998 was a "classical" CCA forecast using *one* PC of SST as a predictor, and trained on the 1950-1999 data. This forecast, which had a spatially averaged correlation skill of 0.27, is equivalent to predicting the amplitude of the leading EOF of observed TCAS, using the leading PC of SST as a predictor. The leading PC of SST is a common measure of ENSO variability, thus specifications based on this predictor indicates the degree of predictability due to ENSO.

3. It follows from the above results that the specification skill of state-of-the-art general circulation models is comparable to, or sometimes better than, the skill of the best empirical models, for the particular 17-year period examined.

4. The effectiveness of different specification strategies was found to be model dependent. No single strategy improved the model forecast skill in all cases. It should be recognized that differences between different strategies could very well arise from sampling error. A notable feature of this paper is that these strategies were evaluated within a common framework with common data sets, allowing direct comparison between different specification strategies.

5. Predictors based on 500hPa geopotential height can lead to specification skill at least as good as, if not better than, predictors based on TCAS.

6. The optimal number of predictors of local land surface temperature is often one, and rarely exceeds four. Moreover, the specification skill of these predictors is comparable to that of the raw model hindcasts. This result suggests that much of the predictability in the dynamical models can be captured by just a few predictable components.

7. The single best forecast of land surface temperature, derived strictly from the 1982-1998 data, was a “classical” CCA forecast using the ensemble mean NCEP TCAS as a predictor. Interestingly, the skill of this forecast does not appear to arise from the same ENSO influences as the other models, since the canonical variate that dominated the skill had relatively small amplitudes during the years 1983, 1989, and 1998.

8. This paper proposed the use of signal-to-noise discriminants as predictors of land surface temperature. In many cases specifications based on discriminants led to greater skill than specifications based on principal components or canonical variates. In fact, the single best forecast for the period 1982-1998 was obtained from discriminant analysis of the 50-year COLAv2.2 integrations, whose spatially averaged correlation skill for the period 1982-1998 was 0.44.

9. The signal to noise ratio varied widely among the models, by a factor of 3 in some cases, and, if anything, an inverse relation was found between signal to noise ratio and the specification skill of predictors derived from discriminant analysis.

None of the statistical correction methods investigated in this paper guarantees an improvement in skill in every case in every model. It is not unreasonable to suppose that the effectiveness of statistical correction methods depends on the model and require case by case experiments. It is perhaps worth noting that a comparison based on mean square error would make the dynamical models look worse owing to their excessive variances.

As noted in the paper, the model selection criterion was applied only to the average skill rather than to the skill at each grid point independently. Although applying the selection criterion at each grid point undoubtedly would improve the forecast skill, such an endeavor would have increased the data management requirements of this project, which already were considerable (more than 1,000,000 forecasts were constructed for this paper). Note that while the *number* of predictors was constrained to be the same at each point, individual regression coefficients vary from point to point. Hence, the selected forecast at different grid points can respond in different ways to the same SST forcing.

The above results appear to contradict certain previous studies and thus deserve further comment. Feddersen et al. (1999), Kang et al. (2004), and others have applied statistical corrections to model hindcasts similar to those considered here and seem to report more success than found here. However, of these studies, only Feddersen et al. (1999) applied statistical correction to wintertime land surface temperature hindcasts, and in that case they found that the corrections did not lead to improvement. The enhancement in skill reported in the above papers appears to be concentrated primarily in the tropics, where the signal to noise ratio is much larger than in midlatitudes. It would be interesting to perform a study similar to ours for the tropics.

Anderson et al. (1999) found that classical CCA forecasts had larger anomaly pattern correlations than bias corrected dynamical model hindcasts, leading them to conclude that statistical models "produce considerably better simulations" than dynamical models. In contrast, we find the opposite: the skill of dynamical models is comparable to, or better than, empirical models. There are numerous differences between the two studies which could explain this discrepancy, including the use of different, more recent dynamical models, different forecast variables (TCAS vs. 700hPa height), and different data sets for applying CCA. Another possibility, however, is that Anderson et al.'s conclusion may be based on a questionable measure of skill. In particular, Anderson et al. (1999) measure skill by the spatial anomaly correlation over a "PNA region" within 20°N-80°N and 180°-60°W, of which 63% is covered by ocean. The dynamical models in that study were all integrated with observed SST as a lower boundary condition. It seems plausible to us that the 700hPa height over the oceans are determined primarily by the SST directly underneath in both the models and observations. However, the models presumably contain systematic errors, whereas CCA could probably lift the correct SST-700hPa height relation from observations. This reasoning suggests that this particular measure of skill may possibly favor the statistical models compared to the dynamical models simply because the ocean dominates the verification area.

Our results suggest that three of the dynamical models respond predominantly to one "flavor" of ENSO, characterized by maximum anomalies in the NINO3 region, while the fourth model responds predominantly to a different "flavor" of ENSO, characterized by maximum anomalies in the NINO3.4 region (both flavors also have significant amplitudes in midlatitudes, but the precise structure depends on the method used to extract the pattern). Perhaps some combination of the two patterns will lead to a better specification model. The fact that classical CCA forecasts over the 50-year period optimize skill when about 11 predictors are used, a result supported by Barnston and Smith (1996), suggests that many more "flavors" of predictable patterns remain to be identified and simulated. It is hoped that the evidence presented here of dynamical predictability arising from distinct SST patterns can guide modelers in the development of improved dynamical prediction models.

9. Appendix: Centered vs. Uncentered Pattern Correlations

In this appendix we briefly discuss the merits of using centered versus uncentered anomaly pattern correlations.

In model comparison studies, it is routine practice to consider only the “anomaly pattern,” which is the field after the sample mean over all years has been subtracted. By definition, the time mean anomaly vanishes, but the spatial mean anomaly may not. When computing the pattern correlation between two anomaly fields, an investigator must decide whether to subtract the spatial average of the anomaly. If the spatial average is *not* subtracted, then we call this an *uncentered* anomaly pattern correlation, and compute it according to (9). If the spatial average is subtracted, then we call this a *centered* anomaly pattern correlation, and compute it according to

$$C\_ACC = \frac{[\,(f-[f])\,(v-[v])\,]}{\sqrt{[\,(f-[f])^2\,]\,[\,(v-[v])^2\,]}}\,. \tag{20}$$

The centered anomaly pattern correlation is simply the standard correlation coefficient in classical statistics. Many studies do not state clearly whether the centered or uncentered pattern correlation coefficient is used. Both versions are convenient skill scores since they both lie between -1 and +1, give +1 for a perfect forecast, and vanish for a climatological forecast (i.e., for $f = 0$). There seems to be a tendency to prefer the centered version since it corresponds to the version used in classical statistics for which various sampling distributions are known. However, these sampling distributions require knowledge of the number of degrees of freedom in the spatial field, which is rarely known. The distinction between these two metrics is not important for global scale domains in which the anomaly field contain compensating positive and negative values, which might explain why many studies neglect to note the distinction.

An instructive example in which the uncentered and centered correlations give different results is shown in fig. 13, which shows the anomaly pattern for the observed TCAS and the ensemble mean TCAS hindcasted by three models for the year 1986 (by "anomaly pattern," we mean the field minus the average field over 1982-1998). We draw attention to the fact that the anomaly for the COLAv2.2 model is mostly negative whereas the observed anomaly is mostly positive. Most forecasters would conclude from these figures that the skill of the COLAv2.2 hindcast for this year was poor. Yet, the centered ACC is positive for this case because the patterns share some similarity after the spatial average of each field is subtracted. In contrast, the uncentered ACC is strongly negative, indicating a poor forecast. In any real forecast scenario, a forecaster would examine the anomaly pattern of a forecast, not the anomaly minus its spatial average. Thus, the uncentered ACC appears to conform better to a forecasters subjective ranking of a forecast than the centered ACC. The other anomaly patterns shown in the figure also have very different centered and uncentered ACCs. In each case, we believe the uncentered ACC corresponds more closely to a forecasters subjective ranking of the hindcasts than the centered ACC. The uncentered ACC can be interpreted as a measure of how well the hindcasts reproduce the sign *and* anomaly structure of the verification. Therefore, we have preferred the uncentered ACC in this study to quantify the skill of hindcasts as a function of year.

Acknowledgements

We gratefully acknowledge the ensemble hindcast data set made possible by the Dynamical Seasonal Prediction (DSP) project. The particular data sets used in this study were provided by Siegfried Shubert for the NASA model and Arun Kumar for the NCEP model. This research was supported by the National Science Foundation (ATM0332910), National Aeronautics and Space Administration (NNG04GG46G), and the National Oceanographic and Atmospheric Administration (NA04OAR4310034).

References

Anderson, J., H. van den Dool, A. Barnston, W. Chen, W. Stern, and J. Ploshay, 1999: Present-day capabilities of numerical and statistical models for atmospheric extratropical seasonal simulation and prediction. *Bull. Amer. Meteor. Soc.*, **80**, 1349–1362.

Bacmeister, J. T., and M. J. Suarez, 2002: Wind stress simulations and the equatorial momentum budget in an AGCM. *J. Atmos. Sci.*, **59**, 3051–3073.

Barnett, T. P., and R. Preisendorfer, 1987: Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis. *Mon. Wea. Rev.*, **115**, 1825-1850.

Barnston, A. G., 1994: Linear statistical short-term climate predictive skill in the Northern Hemisphere. *J. Climate.*, **7**, 1513-1564.

Barnston, A. G., and C. F. Ropelewski, 1992: Prediction of ENSO episodes using canonical correlation analysis. *J. Climate*, **5**, 1316–1345.

Barnston, A. G., and H. M van den Dool, 1993: A Degeneracy in Cross-Validated Skill in Regression-based Forecasts. *J. Climate,* 6, 963–977.

Barnston, A. G., and T. M. Smith, 1996: Specification and prediction of global surface temperature and precipitation from global SST using CCA. *J. Climate*, **9**, 2660–2696.

Barnston, A. G., S. J. Mason, L. Goddard, D. G. DeWitt, S. E. Zebiak, 2003: Multimodel ensembling in seasonal climate forecasting at IRI, *Bull. Amer. Meteor. Soc.*, **85**, 1019–1037.

Barsugli, J. J., and Sardeshmukh P. D., 2002: Global atmospheric sensitivity to tropical SST anomalies throughout the Indo-Pacific basin. *J. Climate*, **15**, 3427–3442.

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

Cover, T. M., and J. A. Thomas, 1991: *Elements of Information Theory*. Wiley, 576 pp.

DelSole, T., 2001: Optimally Persistent Patterns in Time-Varying Fields. *J. Atmos. Sci*., **58**, 1341-1356.

DelSole, T., 2004: Predictability and Information Theory. Part I: Measures of Predictability. *J. Atmos Sci.*, **61**, 2425-2440.

DelSole, T., 2005: Predictability and Information Theory. Part II: Imperfect Forecasts. *J. Atmos Sci*., **62**, 3368-3381.

DelSole, T., and J. Shukla, 2002: Linear Prediction of Indian Monsoon Rainfall. *J. Climate*, **15**, 3645-3658.

DelSole, T., and P. Chang, 2003: Predictable component analysis, canonical correlation analysis, and autoregressive models. *J. Atmos. Sci*., **60**, 409-416.

Dirmeyer, P. A., and F. J. Zeng, 1999: An update to the distribution and treatment of vegetation and soil properties in SSiB. *COLA Tech. Rep*. 78, 25 pp.[Available from the Center for Ocean–Land–Atmosphere Studies, 4041 Powder Mill Road, Suite 302, Calverton, MD 20705.].

Feddersen, H., A. Navarra, and M. N. Ward, 1999: Reduction of model systematic error by statistical correction for dynamical seasonal prediction. *J. Climate*., **12**, 1974–1989.

Jones, P. D., and A. Moberg, 2003: Hemispheric and large-scale surface air temperature variations: An extensive revision and an update to 2001. *J. Climate*., **16**, 206–223.

Kanamitsu, M. Coauthors, 2002: NCEP dynamical seasonal forecast system 2000. *Bull. Amer. Meteor. Soc*., **83**, 1019–1037.

Kang, I.-S., J. Y. Lee, and C.-K. Park, 2004: Potential Predictability of Summer Mean Precipitation in a Dynamical Seasonal Prediction System with Systematic Error Correction. *J. Climate*., **17**, 834-844.

Kiehl, J. T., J. J. Hack, G. Bonan, B. A. Boville, D. L. Williamson, and P. J. Rasch, 1998: The National Center for Atmospheric Research Community Climate Model: CCM3. *J. Climate*, **11**, 1131–1149.

Krishnamurti, T.N., C.M Kishtawal, Z. Zhang, T. LaRow, D. Bachiochi, E. Williford, 2000: Multimodel ensemble forecasts for weather and seasonal climate. *J. Climate*, **13**, 4196-4216.

Koster, R. D., and M. J. Suarez, 1992: A comparative analysis of two land surface heterogeneity representations. *J. Climate*, **5**, 1379–1390.

Michaelsen, J., 1987: Cross-validation in statistical climate forecast models. *J. Appl. Meteor*., **26**, 1589–1600.

Moorthi, S., and M. J. Suarez, 1992: Relaxed Arakawa–Schubert: A parameterization of moist convection for general circulation models. *Mon. Wea. Rev*., **120**, 978–1002.

Pan, H.-L., and L. Mahrt, 1987: Interaction between soil hydrology and boundary layer developments. *Bound.-Layer Meteor*., **38**, 185–202.

Rajagopalan, B., U. Lall, and S. E. Zebiak, 2002: Categorical climate forecasts through regularization and optimal combination of multiple GCM ensembles. *Mon. Wea. Rev*., **130**, 1792–1811.

Rayner, N.A., Parker D.E., Horton E.B., Folland C.K., Alexander L.V., Rowell D.P., Kent E.C., and Kaplan A., 2003: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res*., **108**, 4407, doi:10.1029/2002JD002670.

Reynolds, R. W., and T. M. Smith, 1994: Improved global sea surface temperature analyses using optimal interpolation. *J. Climate*., **7**, 929–948.

Ropelewski, C. F., Janowiak J. E., and Halpert M. F., 1985: The analysis and display of real time surface climate data. *Mon. Wea. Rev*., **113**, 1101–1107.

Ropelewski, C. F., and Halpert M. S., 1986: North American precipitation and temperature patterns associated with the El Niño-Southern Oscillation (ENSO). *Mon. Wea. Rev*., **114**, 2352-2362.

Schneider, T. and S. M. Griffies, 1999: A conceptual framework for predictability studies. *J. Climate*, **12**, 3133-3155.

Smith, T.M., Reynolds R.W., Livezey R.E., and Stokes D.C., 1996: Reconstruction of historical sea surface temperatures using empirical orthogonal functions. *J. Climate*., **9**, 1403-1420.

Shukla, J., Paolino D. A., Straus D. M., DeWitt D., Fennessy M., Kinter J. L., Marx L., and Mo R., 2000: Dynamical seasonal predictions with the COLA atmospheric model. *Quart. J. Roy. Meteor. Soc*., **126**, 2265-2291.

Suarez, M. J., and L. L. Takacs, 1995: Documentation of the Aires/GEOS dynamical core version 2. *NASA Tech. Memo* 104606, Vol. 10, 56 pp.

Venzke, S., M. R. Allen, R. T. Sutton, and D. P. Rowell, 1999: The atmospheric response over the North Atlantic to decadal changes in sea surface temperature. *J. Climate*, **12**, 2562–2584.

von Storch, H., and F. W. Zwiers, 1999: *Statistical Analysis in Climate Research*. Vol. 12, Cambridge University Press, 484 pp.

Wallace, J. M., and D. S. Gutzler, 1981: Teleconnections in the geopotential height field during the Northern Hemisphere winter. *Mon. Wea. Rev*., **109**, 785–812.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*. International Geophysics Series, Vol. 59, Academic Press, 464 pp.

Xue, Y., Sellers P. J., Kinter J. L., and Shukla J., 1991: A simplified biosphere model for global climate studies. *J. Climate*., **4**, 345-364.

10. Table Captions

Table 1: Description of contents in subsequent tables.

Table 2: Skill statistics of the ensemble mean dynamical model hindcasts of TCAS for the period 1982-1998. See table 1 for explanation of the different columns.

Table 3: The cross-validated skill of "classical" CCA forecasts of JFM surface temperature, with predictors derived from JFM ensemble mean dynamical hindcasts, verified over the period 1982-1998. By "classical," we mean the predictors and predictands are selected from the leading principal components of both fields, and the statistical forecast is computed by methods that are standard in CCA. The table shows results only for forecasts that minimized the cross validated mean square error for the given model and variable. See table 1 for explanation of the different columns. The last 3 rows of the table show the CCA forecast based on training data in the period 1950-1999, but verified in cross validated sense in the period 1982-1998.

Table 4: Predictors of *local* JFM surface temperature examined in this paper, and the tables which summarize the skill statistics for each predictor. The brackets $\langle\rangle$ indicate ensemble average.

Table 5: Skill statistics of linear regression forecasts of *local* JFM surface temperature, with predictors based on the leading principal components of the model variable indicated in the table, for the period 1982-1998. The table shows only forecasts that minimized the mean square error for the given model and variable. The last two rows of the table show the skill of a purely statistical local prediction derived from the leading principal components of SST, trained over the periods 1950-1999 (last row) and 1982-1998 (second to last row) (both forecasts are cross validated over the same period 1982-1998). "COLAv2.2-50" indicates that the 50-year period

1950-1999 was used to obtain the principal components and train the regression forecast derived from COLAv2.2, but the skill is cross validated over the same period as all the others, namely 1982-1998. See table 1 for explanation of the different columns.

Table 6: Skill statistics of linear regression forecasts of *local* JFM surface temperature with predictors based on canonical variates, for the period 1982-1998. In all cases tabulated, the observed land surface temperature is the predictand used in CCA. The table lists the predictors used in CCA, which also is the canonical variate used for linear regression forecasts. The table shows only forecasts that minimized the mean square error for the given model variable. The last row of the table shows the skill of a purely statistical local forecast derived from the canonical variates of SST, trained over the period 1950-1999 but verified over the period 1982-1998

Table 7: Skill statistics of linear regression forecasts of *local* JFM surface temperature with predictors based on canonical variates, for the period 1982-1998. In all cases tabulated, SST is the predictor used in CCA, and the model variable indicated in the table is the predictand used in CCA. The canonical variate used as a predictor for linear regression is stated in the "predictor variable" column. The table shows only forecasts that minimized the mean square cross validated error in the period 1982-1998 for the given model variable.

Table 8: Results of linear regression forecasts of *local* JFM surface temperature with predictors derived from signal-to-noise discriminants of the dynamical models indicated in the table for the period 1982-1998. The last two rows show the results for discriminants estimated from the period 1950-1999, but whose forecasts were verified over the period 1982-1998. The table shows only forecasts that minimized the mean square error for the given model variable.

11. Figure Captions

Figure 1: Correlation between observed and model forecasted JFM land surface temperature for the period 1982-1998. Each model forecast grid was interpolated onto the observation grid. The 5% significance level of the correlation coefficient for this data set is 0.39. Areas with no shading indicate areas with insufficient data.

Figure 2: The uncentered anomaly pattern correlation between observed and model forecasted JFM land surface temperature corresponding to the raw hindcasts displayed in fig. 1.

Figure 3: Local signal-to-noise ratio of JFM land surface temperature of four dynamical model hindcasts for the period 1982-1998. Signal is defined as the ensemble mean hindcast, and noise as deviations about the ensemble mean. Large signal-to-noise ratios near coastal boundaries are an artifact of the prescribed SST boundary conditions.

Figure 4: Leading empirical orthogonal functions (EOFs) and principal components (PCs) of ensemble mean, JFM, land surface temperature of four dynamical model hindcasts for the period 1982-1998. The percent of variance explained by each model, relative to the total variance of the ensemble mean, is indicated above each EOF. The PCs have been normalized to unit variance, and the EOFs have been normalized such that the PC times the EOF, summed over all components, exactly reproduces the original data.

Figure 5: Leading canonical pattern between JFM, ensemble mean, land surface temperature in each model, and the simultaneous JFM mean SST, for the period 1982-1998. The number of principal components in each CCA is indicated above each pattern. The associated SST patterns are nearly identical to each other (all pair-wise SST anomaly correlation coefficients exceed 0.98). The canonical variates have been normalized to unit variance, and the

canonical pattern has been normalized such that the variate times the pattern, summed over all components, exactly reproduces the original data.

Figure 6: Signal-to-noise discriminants of JFM land surface temperature hindcasts for the period 1982-1998. These structures optimize the signal-to-noise ratio in each model. The value of the signal-to-noise ratio and associated mutual information (see text for explanation) are indicated above the variates. The variates have been normalized to unit variance, and the discriminants have been normalized such that the variate times the discriminant, summed over all components, exactly reproduces the original data.

Figure 7: Correlation between observed and CCA-corrected model hindcasts of JFM land surface temperature for the period 1982-1998. The 5% significance level of the correlation is 0.39. Areas with no shading indicate areas with insufficient data.

Figure 8: Difference between the (uncentered) anomaly pattern correlation of "CCA-corrected" model forecasts of JFM land surface temperature, and the raw model forecasts. Positive values indicate that the CCA-corrected model forecasts has more skill than the uncorrected forecasts. By "CCA-corrected," we mean that CCA has been applied to the observations and ensemble mean hindcasts, and then a classical CCA forecast for the observed temperature has been constructed using model hindcast as a predictor.

Figure 9: Specification skill of a "classical" CCA forecast of JFM land surface temperature, using JFM SST as predictors. The top figure shows the cross validated correlation between observed and forecasted JFM surface temperature for the period 1982-1998, and the bottom shows the (uncentered) anomaly correlation of the forecast. The CCA forecast was trained using data in the period 1950-1999. The overall specification skill metrics are: EV =

0.14, LMI = 0.14, MSE = 2.95, RHOM = 0.27.

Figure 10: Spatially averaged correlation coefficient (RHOM) between locally observed TCAS and the locally forecasted value by the models indicated on the abscissa for the period 1982-1998. The listed models should be self evident from the discussion in secs. 6-7 and table 4. The dashed line gives the spatially averaged correlation coefficient for a linear specification of TCAS based on the first PC of SST (RHOM = 0.17).

Figure 11: The *second* leading canonical pair of patterns between observed TCAS and ensemble mean forecasted TCAS for the NCEP model. CCA was applied to the leading 6 principal components of each field. The bottom figure shows the amplitude of each canonical pattern.

Figure 12: Covariance between local SST and the time series shown above, which is the time series displayed in fig. 11 for the second leading canonical variate between the NCEP hindcasted TCAS and observed land surface temperature. The top figure can be interpreted as an estimate of the SST pattern that is maximally correlated with the specified time series.

Figure 13: Anomaly pattern of the observed JFM land surface temperature, and of the ensemble mean TCAS hindcast of three dynamical models, for the year 1986. The anomaly pattern is computed with respect to the 1982-1998 average. The explained variance ("EV"), centered anomaly pattern correlation ("C_ACC"), and uncentered anomaly pattern correlation ("UC_ACC") of each field, relative to the observed anomaly, is shown at the bottom left in each figure. Note that the centered and uncentered correlations differ subtantially, and that the uncentered version is more consistent with how a forecaster would subjective rank the hindcasts.

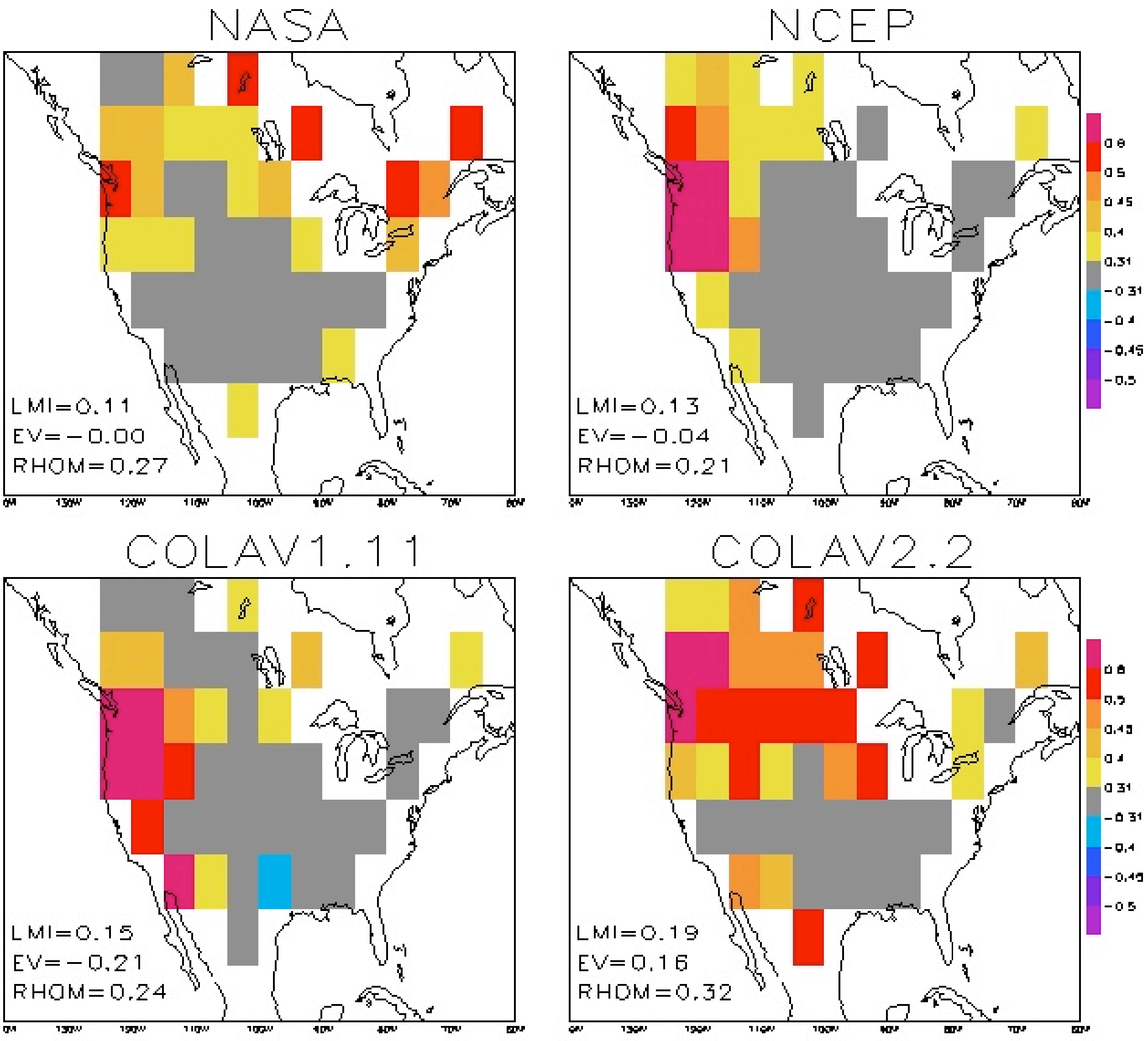

Figure 1: Correlation between observed and model forecasted JFM land surface temperature for the period 1982-1998. Each model forecast grid was interpolated onto the observation grid. The 5% significance level of the correlation coefficient for this data set is 0.39. Areas with no shading indicate areas with insufficient data.

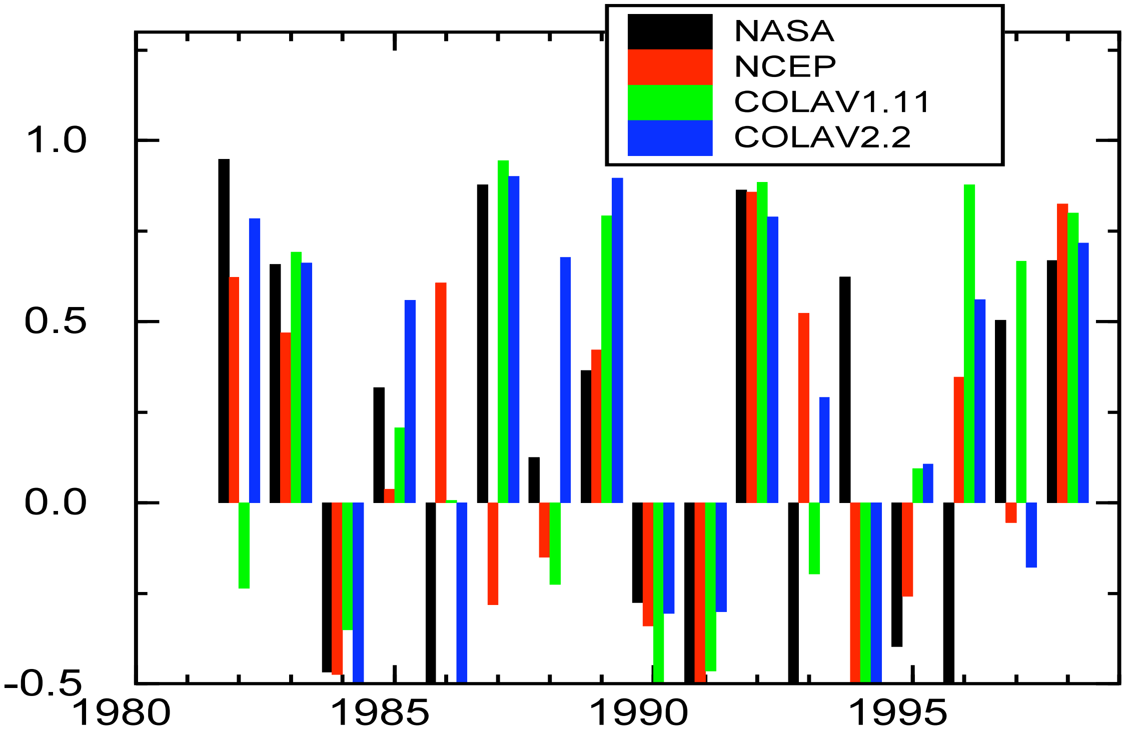

Figure 2: The uncentered anomaly pattern correlation between observed and model forecasted JFM land surface temperature corresponding to the raw hindcasts displayed in fig. 1.

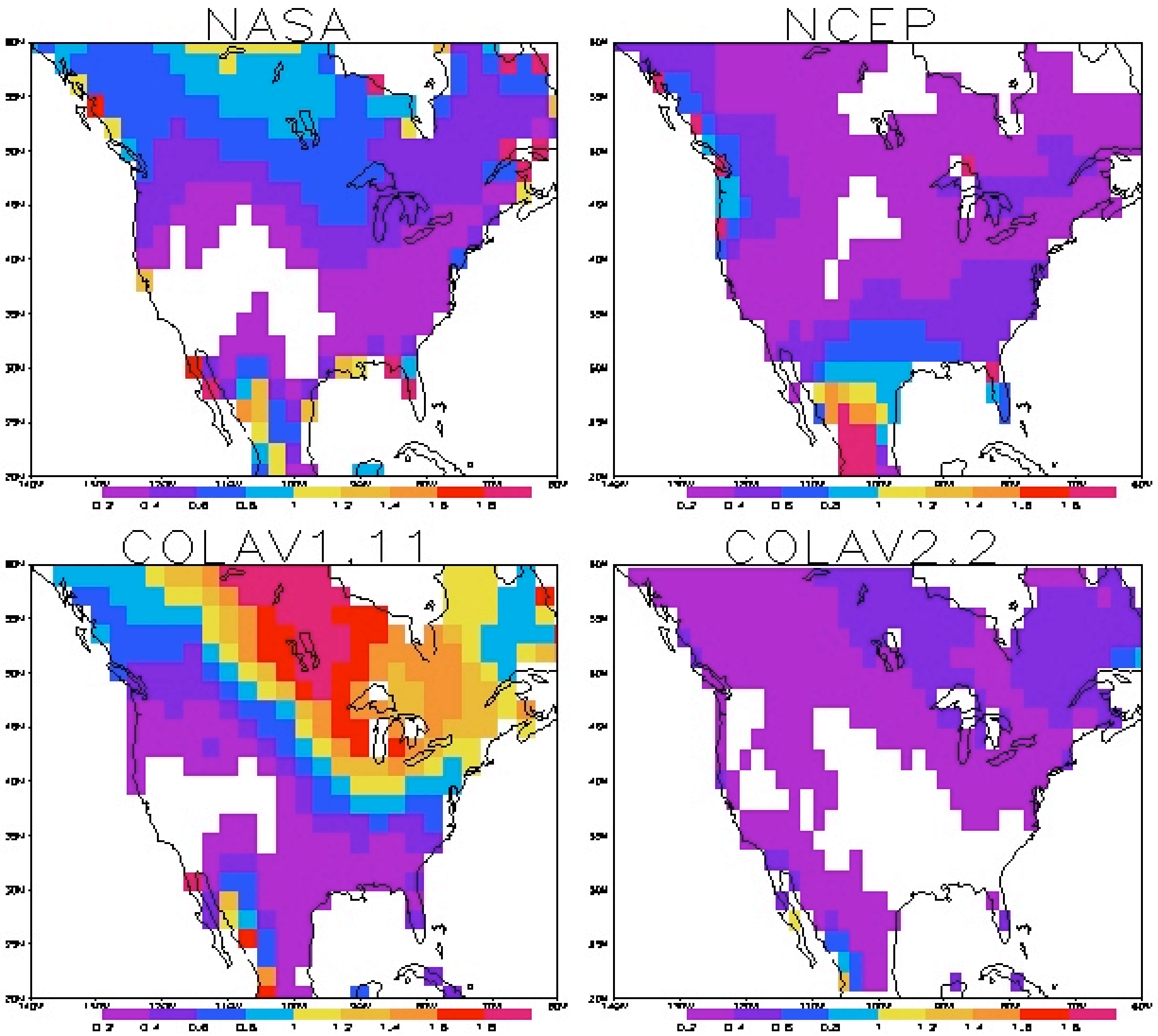

Figure 3: Local signal-to-noise ratio of JFM land surface temperature of four dynamical model hindcasts for the period 1982-1998. Signal is defined as the ensemble mean hindcast, and noise as deviations about the ensemble mean. Large signal-to-noise ratios near coastal boundaries are an artifact of the prescribed SST boundary conditions.

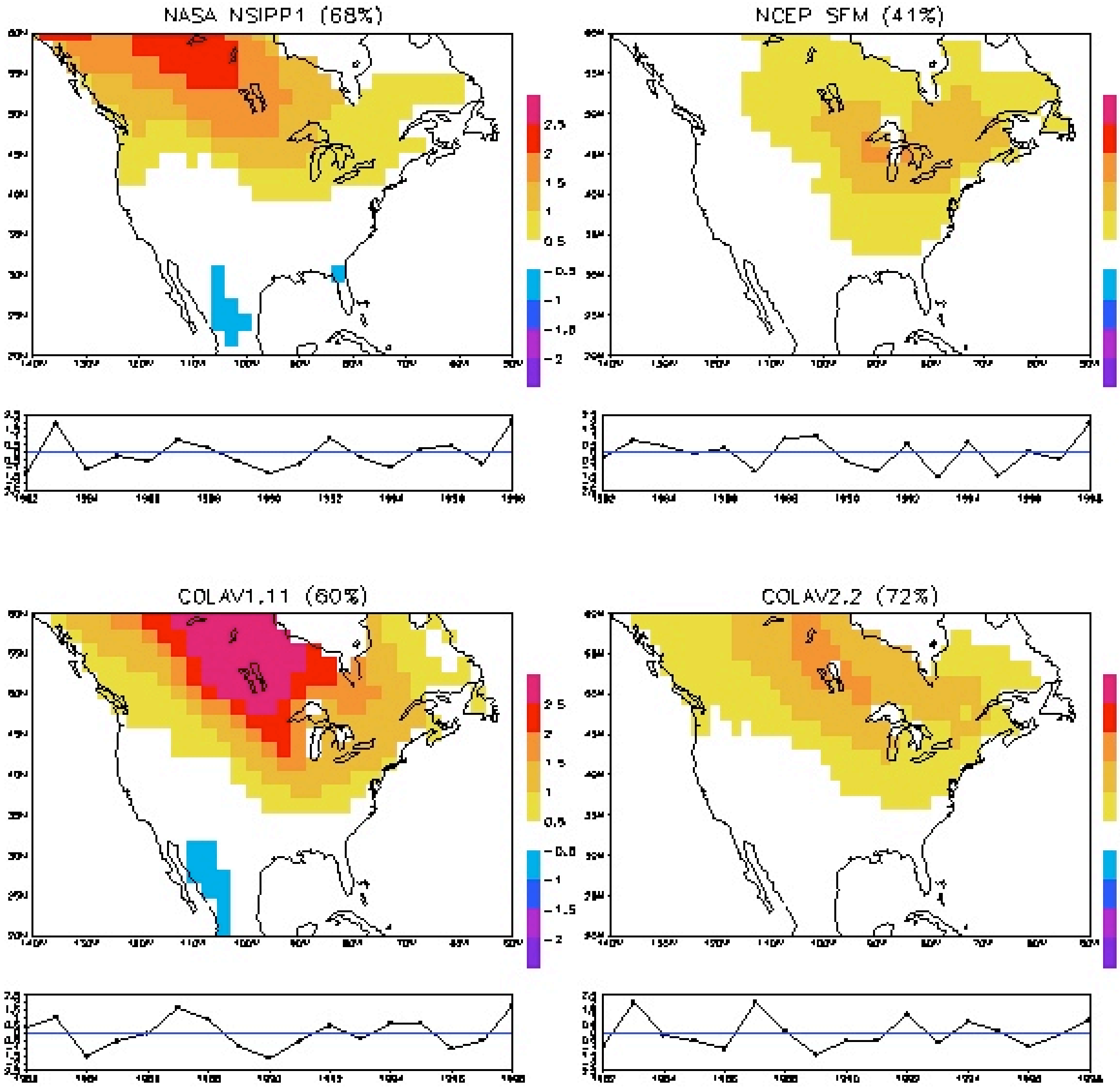

Figure 4: Leading empirical orthogonal functions (EOFs) and principal components (PCs) of ensemble mean, JFM, land surface temperature of four dynamical model hindcasts for the period 1982-1998. The percent of variance explained by each model, relative to the total variance of the ensemble mean, is indicated above each EOF. The PCs have been normalized to unit variance, and the EOFs have been normalized such that the PC times the EOF, summed over all components, exactly reproduces the original data.

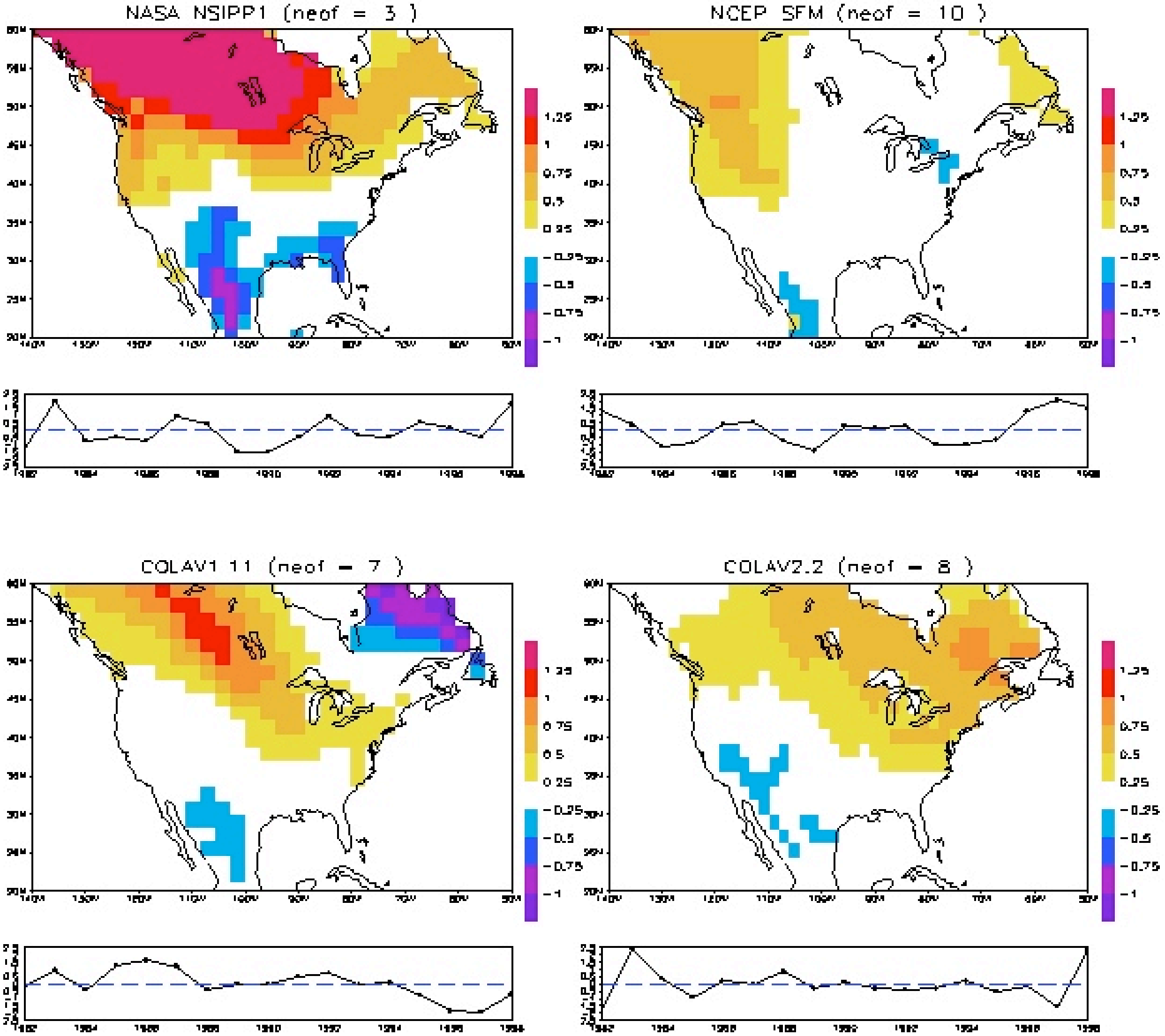

Figure 5: Leading canonical pattern between JFM, ensemble mean, land surface temperature in each model, and the simultaneous JFM mean SST, for the period 1982-1998. The number of principal components in each CCA is indicated above each pattern. The associated SST patterns are nearly identical to each other (all pair-wise SST anomaly correlation coefficients exceed 0.98). The canonical variates have been normalized to unit variance, and the canonical pattern has been normalized such that the variate times the pattern, summed over all components, exactly reproduces the original data.

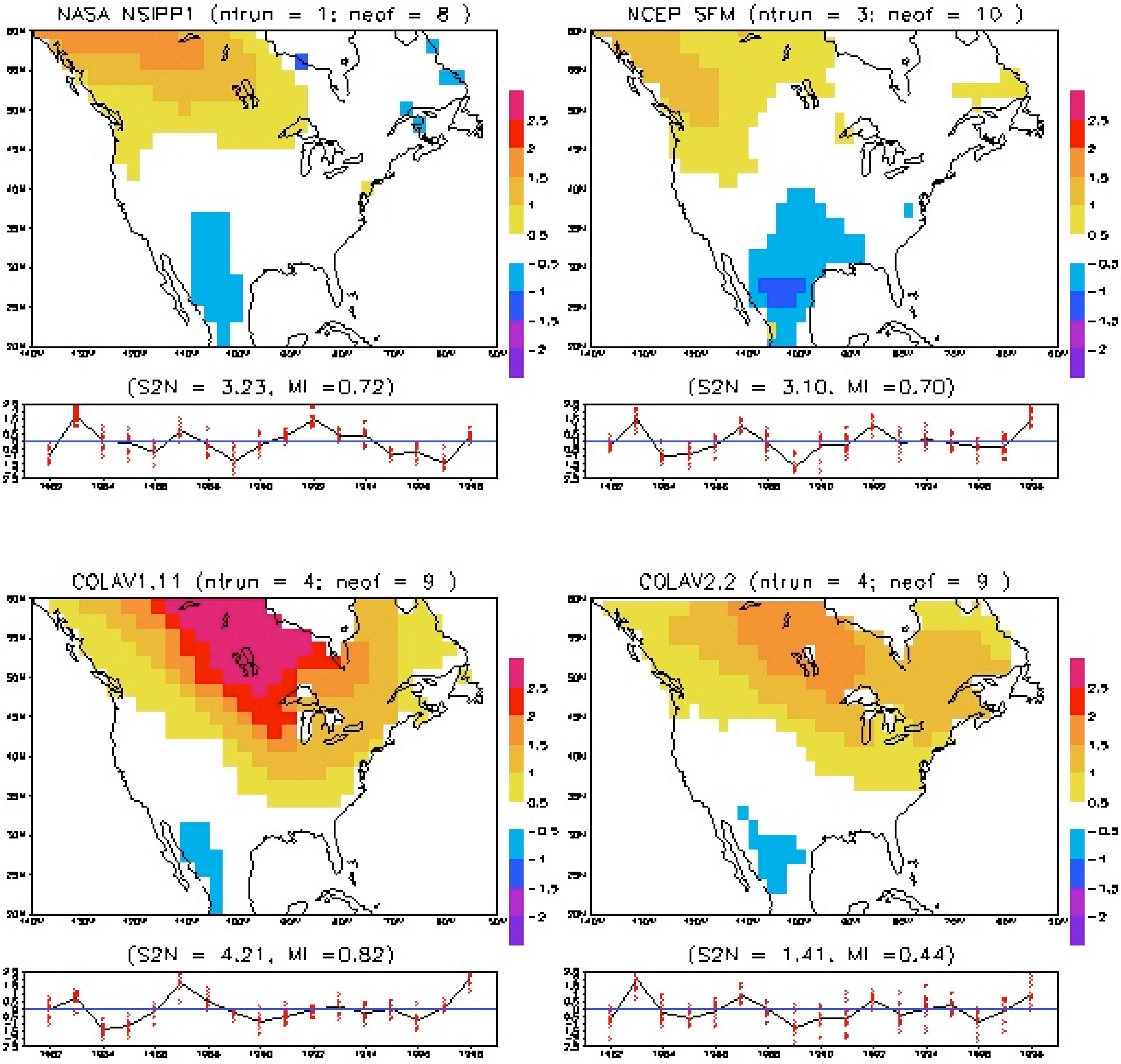
NASA NSIPP1 (ntrun = 1; neof = 8)
(S2N = 3.23, MI =0.72)
NCEP SFM (ntrun = 3; neof = 10)
(S2N = 3.10, MI =0.70)
COLAV1.11 (ntrun = 4; neof = 9)
(S2N = 4.21, MI =0.82)
COLAV2.2 (ntrun = 4; neof = 9)
(S2N = 1.41, MI =0.44)
2.5
2
1.5
1
0.5
-0.5
-1
-1.5
-2

Figure 6: Signal-to-noise discriminants of JFM land surface temperature hindcasts for the period 1982-1998. These structures optimize the signal-to-noise ratio in each model. The value of the signal-to-noise ratio and associated mutual information (see text for explanation) are indicated above the variates. The variates have been normalized to unit variance, and the discriminants have been normalized such that the variate times the discriminant, summed over all components, exactly reproduces the original data.

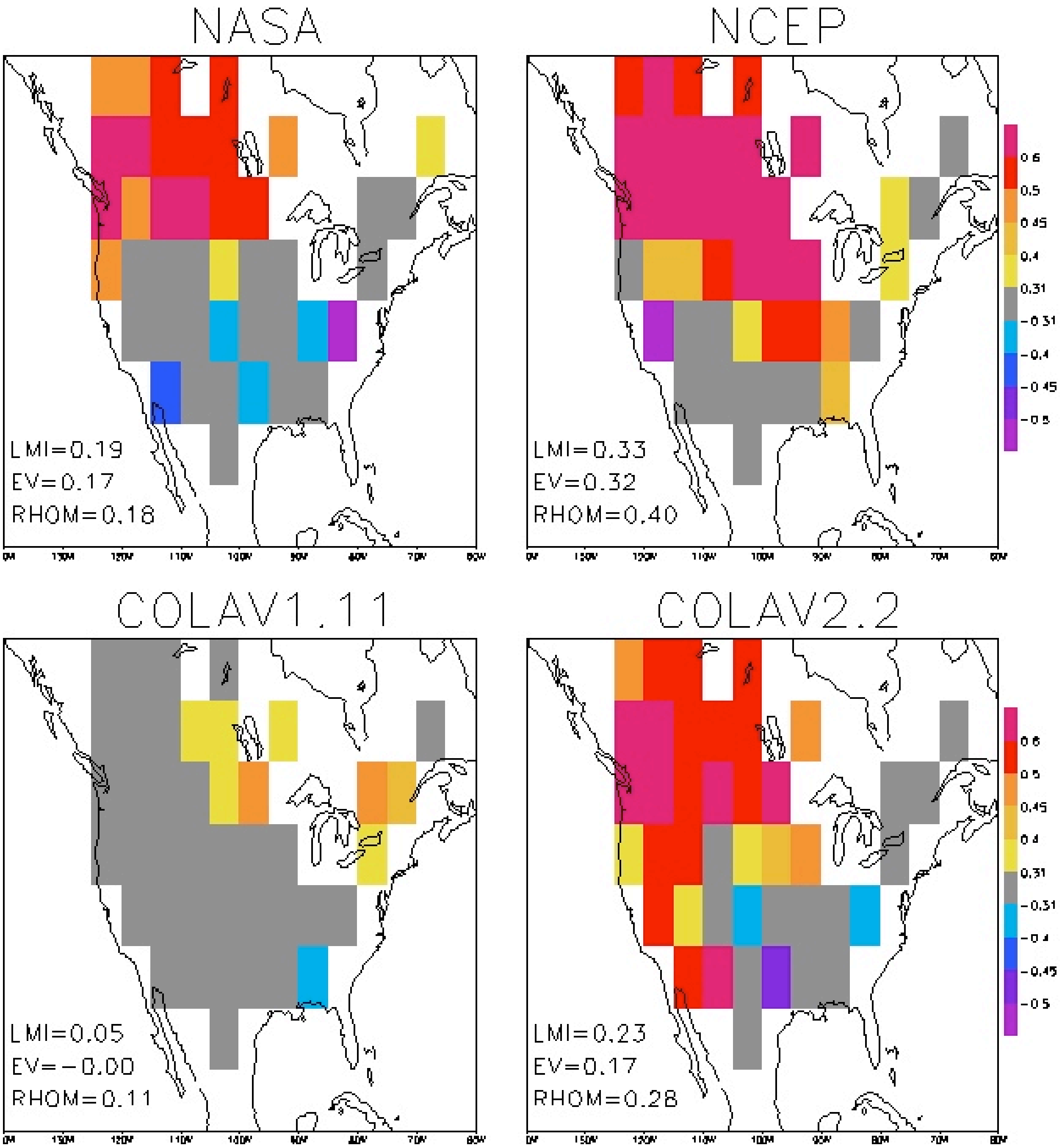

Figure 7: Correlation between observed and CCA-corrected model hindcasts of JFM land surface temperature for the period 1982-1998. The 5% significance level of the correlation is 0.39. Areas with no shading indicate areas with insufficient data.

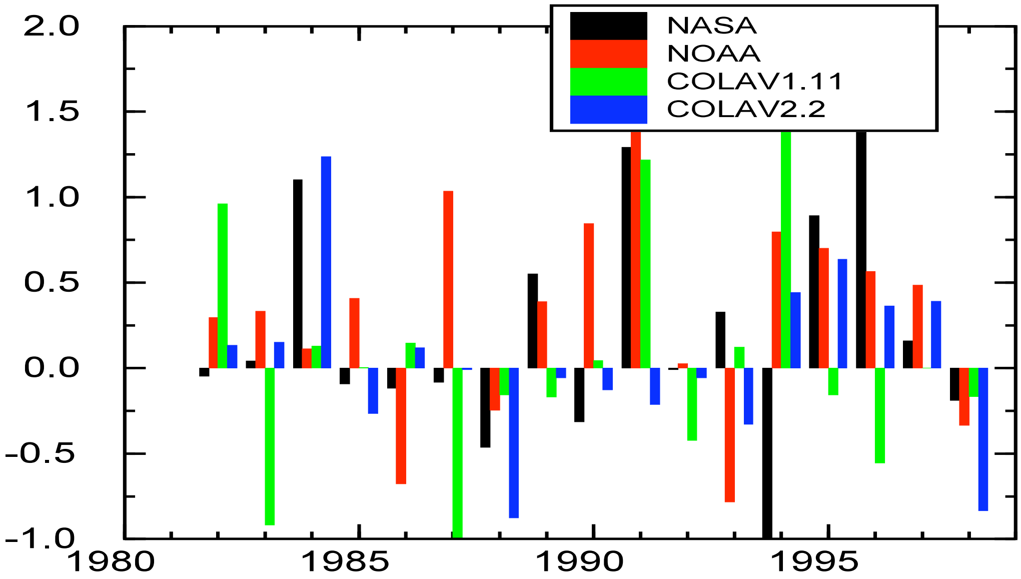

Figure 8: Difference between the (uncentered) anomaly pattern correlation of “CCA-corrected” model forecasts of JFM land surface temperature, and the raw model forecasts. Positive values indicate that the CCA-corrected model forecasts has more skill than the uncorrected forecasts. By “CCA-corrected,” we mean that CCA has been applied to the observations and ensemble mean hindcasts, and then a classical CCA forecast for the observed temperature has been constructed using model hindcast as a predictor.

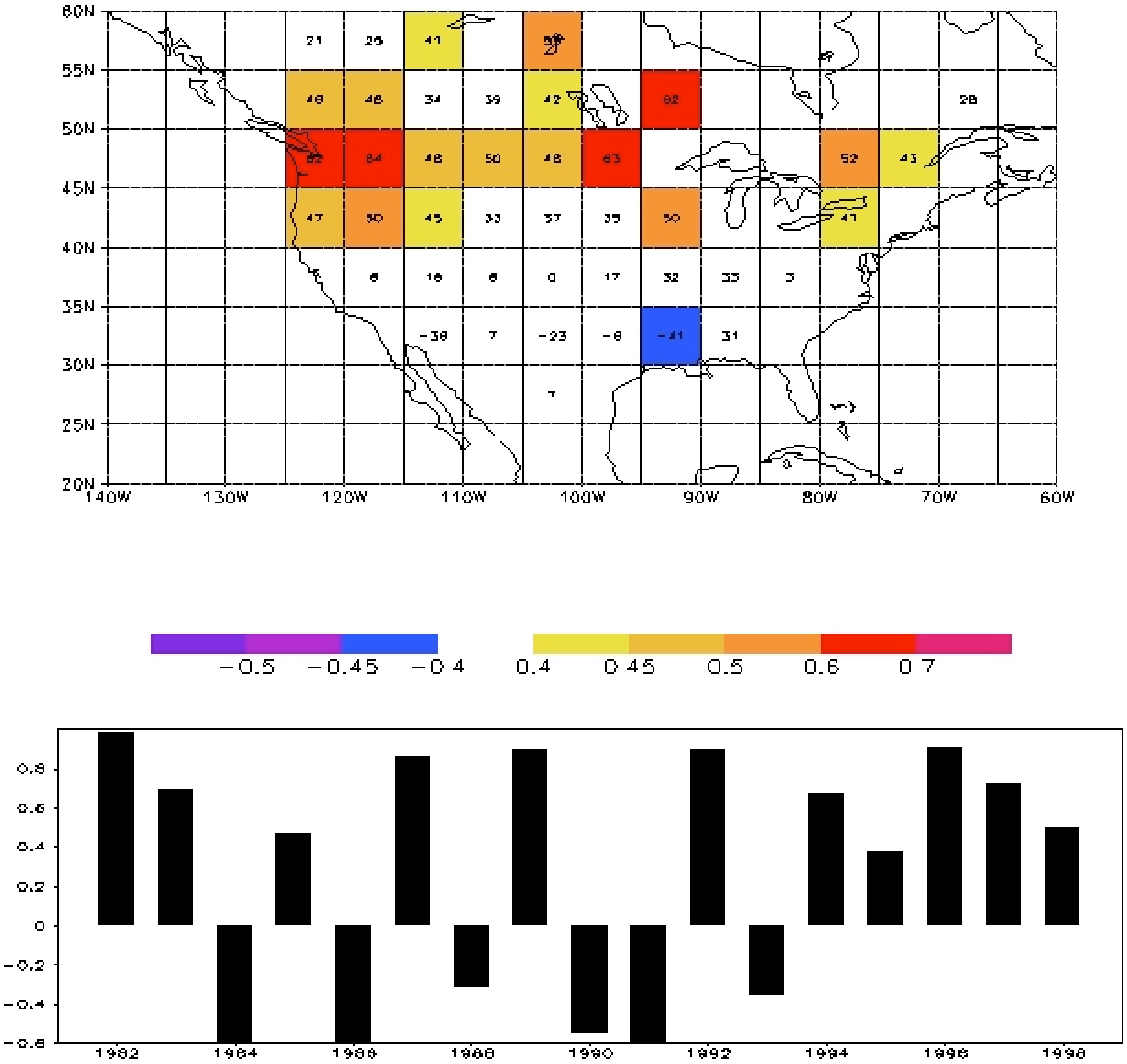

Figure 9: Specification skill of a "classical" CCA forecast of JFM land surface temperature, using JFM SST as predictors. The top figure shows the cross validated correlation between observed and forecasted JFM surface temperature for the period 1982-1998, and the bottom shows the (uncentered) anomaly correlation of the forecast. The CCA forecast was trained using data in the period 1950-1999. The overall specification skill metrics are: EV = 0.14, LMI = 0.14, MSE = 2.95, RHOM = 0.27.

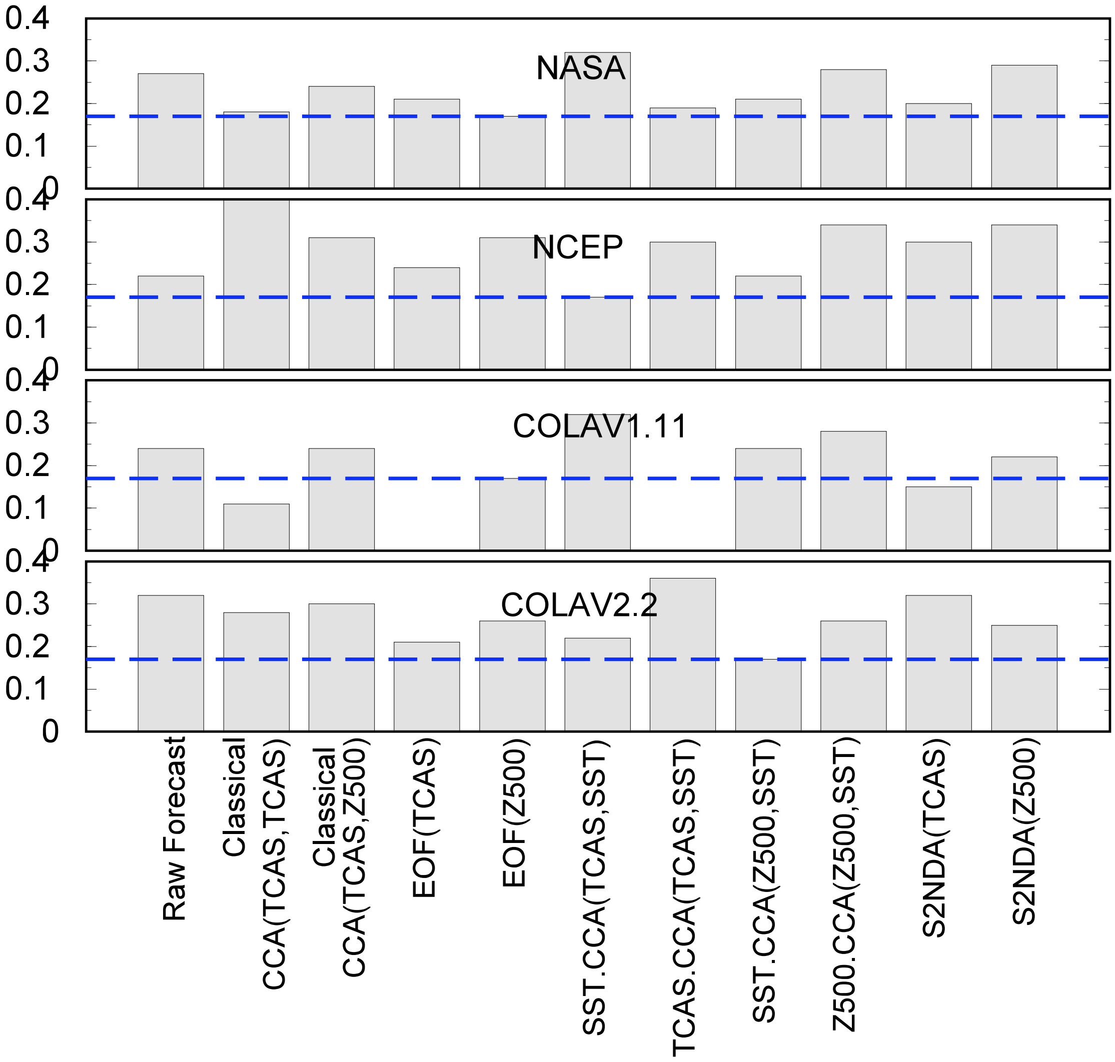

Figure 10: Spatially averaged correlation coefficient (RHOM) between locally observed TCAS and the locally forecasted value by the models indicated on the abscissa for the period 1982-1998. The listed models should be self evident from the discussion in secs. 6-7 and table 4. The dashed line gives the spatially averaged correlation coefficient for a linear specification of TCAS based on the first PC of SST (RHOM = 0.17).

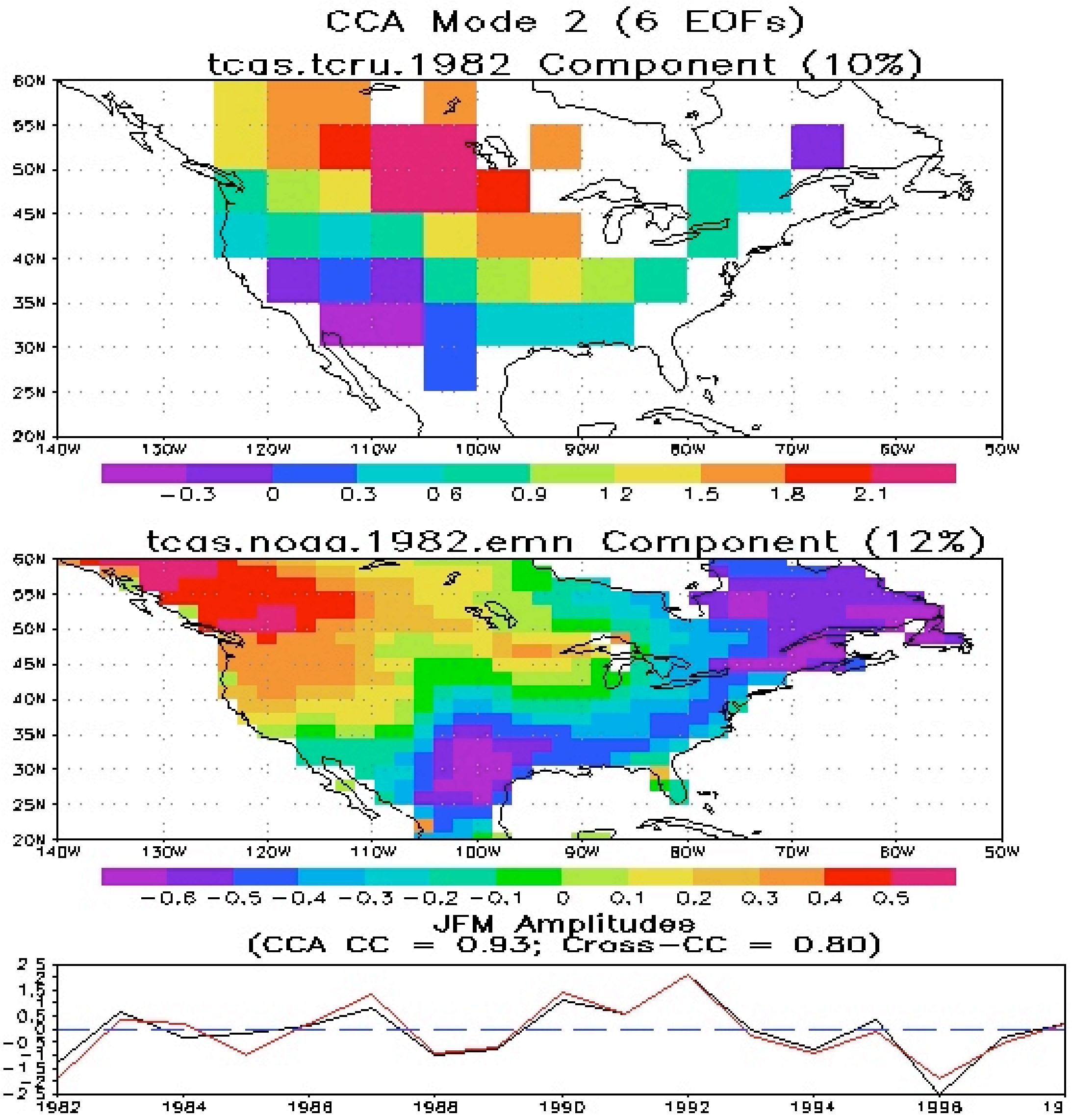

Figure 11: The *second* leading canonical pair of patterns between observed TCAS and ensemble mean forecasted TCAS for the NCEP model. CCA was applied to the leading 6 principal components of each field. The bottom figure shows the amplitude of each canonical pattern.

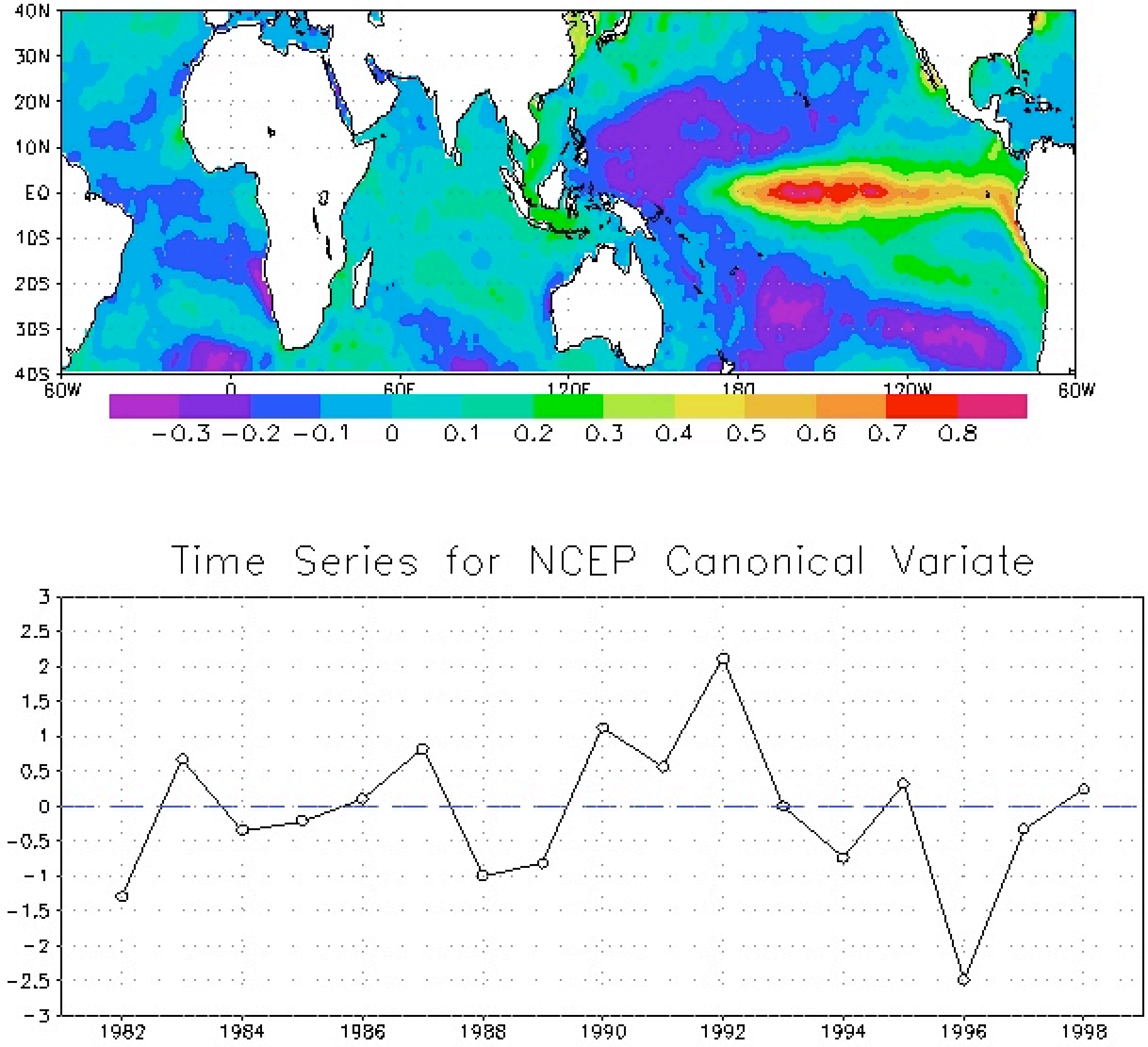

Figure 12: Covariance between local SST and the time series shown above, which is the time series displayed in fig. 11 for the second leading canonical variate between the NCEP hindcasted TCAS and observed land surface temperature. The top figure can be interpreted as an estimate of the SST pattern that is maximally correlated with the specified time series.

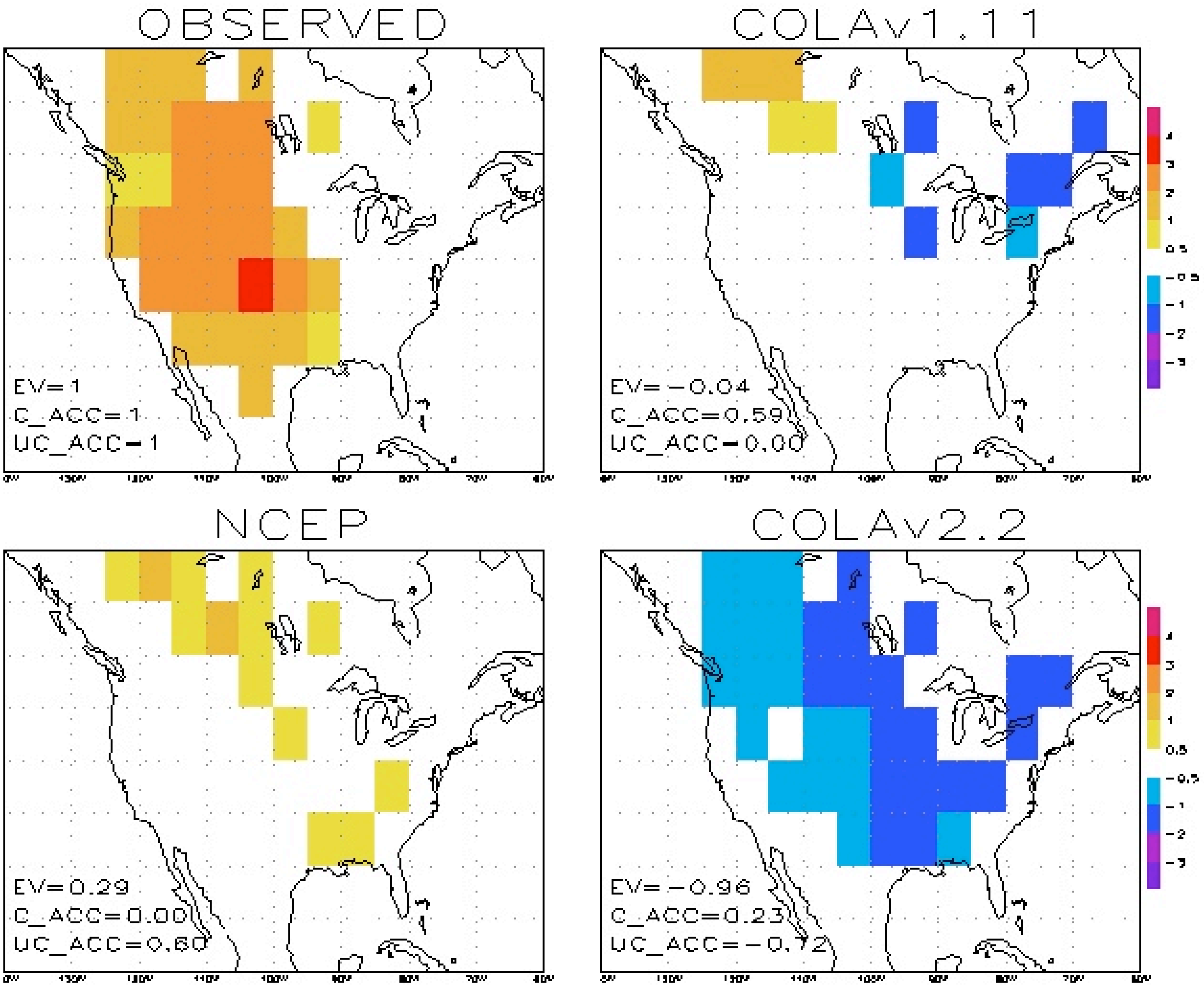

Figure 13: Anomaly pattern of the observed JFM land surface temperature, and of the ensemble mean TCAS hindcast of three dynamical models, for the year 1986. The anomaly pattern is computed with respect to the 1982-1998 average. The explained variance ("EV"), centered anomaly pattern correlation ("C_ACC"), and uncentered anomaly pattern correlation ("UC_ACC") of each field, relative to the observed anomaly, is shown at the bottom left in each figure. Note that the centered and uncentered correlations differ subtantially, and that the uncentered version is more consistent with how a forecaster would subjective rank the hindcasts.

12. Tables

Column Label	Description
Variable	variable used in CCA, EOF, or S2NDA.
predictor variable	indicates which of the two patterns from CCA (either SST or the model variable) was used as a predictor for surface temperature
npred	number of canonical variates used as predictors in linear regression
neof	number of EOFs used to perform CCA
EV	spatially averaged explained variance of cross validated linear regression forecasts
LMI	spatial average localized mutual information of cross validated linear regression forecasts
rhom	spatial average correlation coefficient of cross validated linear regression forecasts
mse	spatially averaged mean square error of cross validated linear regression forecasts

Table 1: Description of contents in subsequent tables.

Model	EV	LMI	rhom
NASA	0.00	0.11	0.27
NCEP	-0.04	0.13	0.21
COLAv1.11	-0.21	0.15	0.24
COLAv2.2	0.16	0.19	0.32

Table 2: Skill statistics of the ensemble mean dynamical model hindcasts of TCAS for the period 1982-1998. See table 1 for explanation of the different columns.

Variable	Model	npred	neof	EV	LMI	mse	rhom
TCAS	NASA	2	8	0.17	0.13	2.85	0.18
Z500	NASA	1	1	0.09	0.10	3.14	0.24
TCAS	NCEP	2	6	0.32	0.30	2.34	0.40
Z500	NCEP	2	5	0.17	0.18	2.85	0.31
TCAS	COLAv1.11	1	9	0.00	0.04	3.49	0.11
Z500	COLAv1.11	1	1	0.09	0.10	3.13	0.24
TCAS	COLAv2.2	3	6	0.17	0.20	2.85	0.28
Z500	COLAv2.2	1	1	0.20	0.18	2.77	0.30
TCAS	COLAv2.2-50	1	1	0.10	0.07	3.11	0.17
Z500	COLAv2.2-50	1	1	0.14	0.14	2.95	0.27
SST	**HADSST50**	**1**	**1**	**0.14**	**0.14**	**2.95**	**0.27**

Table 3: The cross-validated skill of "classical" CCA forecasts of JFM surface temperature, with predictors derived from JFM ensemble mean dynamical hindcasts, verified over the period 1982-1998. By "classical," we mean the predictors and predictands are selected from the leading principal components of both fields, and the statistical forecast is computed by methods that are standard in CCA. The table shows results only for forecasts that minimized the cross validated mean square error for the given model and variable. See table 1 for explanation of the different columns. The last 3 rows of the table show the CCA forecast based on training data in the period 1950-1999, but verified in cross validated sense in the period 1982-1998.

Variable	Analysis	Table
⟨TCAS⟩_MODEL	from EOF(⟨TCAS⟩_MODEL)	table 5
⟨Z500⟩_MODEL	from EOF(⟨Z500⟩_MODEL)	table 5
⟨TCAS⟩_MODEL	from CCA(TCAS_OBS, ⟨TCAS⟩_MODEL)	table 6
⟨Z500⟩_MODEL	from CCA(TCAS_OBS, ⟨Z500⟩_MODEL)	table 6
SST	from CCA(⟨TCAS⟩_MODEL, SST)	table 7
⟨TCAS⟩_MODEL	from CCA(⟨TCAS⟩_MODEL, SST)	table 7
SST	from CCA(⟨Z500⟩_MODEL, SST)	table 7
⟨Z500⟩_MODEL	from CCA(⟨Z500⟩_MODEL, SST)	table 7
TCAS_MODEL	from S2NDA(TCAS_MODEL)	table 8
Z500_MODEL	from S2NDA(Z500_MODEL)	table 8
SST	from EOF(SST)	table 5
SST	from CCA(TCAS_OBS, SST)	table 6

Table 4: Predictors of *local* JFM surface temperature examined in this paper, and the tables which summarize the skill statistics for each predictor. The brackets ⟨⟩ indicate ensemble average.

Variable	Model	neof	EV	LMI	mse	rhom
TCAS	NASA	4	0.07	0.10	3.21	0.21
Z500	NASA	1	0.08	0.07	3.16	0.17
TCAS	NCEP	4	0.18	0.17	2.81	0.24
Z500	NCEP	5	0.12	0.19	3.04	0.31
TCAS	COLAv1.11	1	-0.07	-0.04	3.73	-0.10
Z500	COLAv1.11	1	0.09	0.08	3.15	0.17
TCAS	COLAv2.2	1	0.14	0.12	2.98	0.21
Z500	COLAv2.2	1	0.19	0.16	2.80	0.26
TCAS	COLAv2.2-50	1	0.08	0.04	3.17	0.09
Z500	COLAv2.2-50	2	0.07	0.08	3.21	0.22
SST	**HADSST**	**1**	**0.12**	**0.08**	**3.04**	**0.17**
SST	**HADSST50**	**1**	**0.13**	**0.12**	**3.01**	**0.22**

Table 5: Skill statistics of linear regression forecasts of *local* JFM surface temperature, with predictors based on the leading principal components of the model variable indicated in the table, for the period 1982-1998. The table shows only forecasts that minimized the mean square error for the given model and variable. The last two rows of the table show the skill of a purely statistical local prediction derived from the leading principal components of SST, trained over the periods 1950-1999 (last row) and 1982-1998 (second to last row) (both forecasts are cross validated over the same period 1982-1998). “COLAv2.2-50" indicates that the 50-year period 1950-1999 was used to obtain the principal components and train the regression forecast derived from COLAv2.2, but the skill is cross validated over the same period as all the others, namely 1982-1998. See table 1 for explanation of the different columns.

Variable	Model	npred	neof	EV	LMI	mse	rhom
TCAS	NASA	2	8	0.17	0.13	2.85	0.18
Z500	NASA	1	1	0.08	0.07	3.16	0.17
TCAS	NCEP	2	6	0.32	0.30	2.34	0.40
Z500	NCEP	2	5	0.17	0.18	2.85	0.30
TCAS	COLAv1.11	1	1	-0.07	-0.03	3.73	-0.10
Z500	COLAv1.11	1	2	0.09	0.10	3.14	0.19
TCAS	COLAv2.2	3	7	0.16	0.21	2.90	0.29
Z500	COLAv2.2	1	1	0.19	0.16	2.80	0.26
TCAS	COLAv2.2-50	1	1	0.08	0.04	3.17	0.09
Z500	COLAv2.2-50	1	12	0.12	0.11	3.04	0.25
SST	**HADSST50**	**1**	**1**	**0.13**	**0.12**	**3.01**	**0.22**

Table 6: Skill statistics of linear regression forecasts of *local* JFM surface temperature with predictors based on canonical variates, for the period 1982-1998. In all cases tabulated, the observed land surface temperature is the predictand used in CCA. The table lists the predictors used in CCA, which also is the canonical variate used for linear regression forecasts. The table shows only forecasts that minimized the mean square error for the given model variable. The last row of the table shows the skill of a purely statistical local forecast derived from the canonical variates of SST, trained over the period 1950-1999 but verified over the period 1982-1998

Model		predictor						
Variable	Model	variable	npred	neof	EV	LMI	mse	rhom
TCAS	NASA	SST	2	9	0.14	0.16	2.95	0.32
TCAS	NASA	TCAS	2	3	0.11	0.10	3.07	0.19
Z500	NASA	SST	1	4	0.13	0.11	3.00	0.21
Z500	NASA	Z500	4	6	0.09	0.18	3.14	0.28
TCAS	NCEP	SST	1	3	0.15	0.10	2.94	0.17
TCAS	NCEP	TCAS	8	10	0.19	0.22	2.78	0.30
Z500	NCEP	SST	1	3	0.13	0.12	2.98	0.22
Z500	NCEP	Z500	5	6	0.27	0.31	2.53	0.34
TCAS	COLAv1.11	SST	3	9	0.18	0.16	2.83	0.32
TCAS	COLAv1.11	TCAS	1	7	0.00	0.00	3.50	0.00
Z500	COLAv1.11	SST	1	8	0.15	0.14	2.93	0.24
Z500	COLAv1.11	Z500	3	10	0.15	0.19	2.92	0.28
TCAS	COLAv2.2	SST	1	3	0.12	0.11	3.03	0.22
TCAS	COLAv2.2	TCAS	4	8	0.29	0.29	2.46	0.36
Z500	COLAv2.2	SST	1	1	0.12	0.08	3.04	0.17
Z500	COLAv2.2	Z500	1	1	0.19	0.16	2.80	0.26
TCAS	COLAv2.2-50	SST	2	13	0.17	0.20	2.87	0.36
TCAS	COLAv2.2-50	TCAS	1	10	0.17	0.16	2.86	0.23
Z500	COLAv2.2-50	SST	3	15	0.22	0.25	2.68	0.39
Z500	COLAv2.2-50	Z500	2	7	0.11	0.10	3.06	0.23

Table 7: Skill statistics of linear regression forecasts of *local* JFM surface temperature with predictors based on canonical variates, for the period 1982-1998. In all cases tabulated, SST is the predictor used in CCA, and the model variable indicated in the table is the predictand used in CCA. The canonical variate used as a predictor for linear regression is stated in the "predictor variable" column. The table shows only forecasts that minimized the mean square cross validated error in the period 1982-1998 for the given model variable.

Model							
Variable	Model	npred	neof	EV	LMI	mse	rhom
TCAS	NASA	1	8	0.17	0.12	2.86	0.20
Z500	NASA	1	10	0.20	0.20	2.76	0.29
TCAS	NCEP	3	10	0.21	0.20	2.72	0.30
Z500	NCEP	3	4	0.21	0.25	2.72	0.34
TCAS	COLAv1.11	4	9	0.06	0.07	3.24	0.15
Z500	COLAv1.11	1	10	0.15	0.13	2.94	0.22
TCAS	COLAv2.2	4	9	0.18	0.24	2.84	0.32
Z500	COLAv2.2	1	7	0.19	0.17	2.78	0.25
TCAS	COLAv2.2-50	4	14	0.24	0.34	2.60	0.44
Z500	COLAv2.2-50	1	1	0.18	0.18	2.83	0.27

Table 8: Results of linear regression forecasts of *local* JFM surface temperature with predictors derived from signal-to-noise discriminants of the dynamical models indicated in the table for the period 1982-1998. The last two rows show the results for discriminants estimated from the period 1950-1999, but whose forecasts were verified over the period 1982-1998. The table shows only forecasts that minimized the mean square error for the given model variable.