Can Optimal Projection Improve Dynamical Model Forecasts?

LIWEI JIA
Center for Ocean–Land–Atmosphere Studies, Calverton, Maryland

TIMOTHY DELSOLE
George Mason University, Fairfax, Virginia, and Center for Ocean–Land–Atmosphere Studies, Calverton, Maryland

MICHAEL K. TIPPETT
International Research Institute for Climate and Society, Columbia University, Palisades, New York, and Center of Excellence for Climate Change Research, Department of Meteorology, King Abdulaziz University, Jeddah, Saudi Arabia

(Manuscript received 15 April 2013, in final form 13 December 2013)

ABSTRACT

An optimal projection for improving the skill of dynamical model forecasts is proposed. The proposed method uses statistical optimization techniques to identify the most skillful or most predictable patterns, and then projects forecasts onto these patterns. Applying the method to seasonal mean 2-m temperature from the Ensemble-Based Predictions of Climate Changes and Their Impacts (ENSEMBLES) multimodel hindcast dataset reveals that the method improves skill only in South America and Africa, suggesting that the benefit of optimal projection is limited to certain regions, but can be substantial. Further investigation reveals that the improvement in skill comes not from optimal projection itself, but from the EOF prefitering that is done to reduce the dimension of the optimization space. Thus, much of the improvement attributable to optimal projection can be achieved by suitable EOF filtering. Interestingly, models are found to generate patterns that project only weakly on observational datasets but are strongly correlated between models. An important by-product of the method is a concise summary of the skillful or predictable structures in a given forecast. For the ENSEMBLES dataset, the method convincingly demonstrates that most of the seasonal prediction skill over continents comes from two components, ENSO and the global warming trend. In addition, the method can be used to determine whether a pattern exists that is well predicted by one model but not by another model (complementary skill).

1. Introduction

Many regression techniques for improving the skill of dynamical forecasts are applied pointwise or to smoothed fields. For instance, Hamill et al. (2004) applied a logistic regression technique to improve medium-range probabilistic forecasts at each observation location. Krishnamurti et al. (1999, 2000) showed improved forecast skill by applying multiple regression to combine forecasts from multiple dynamical models into a single forecast for each grid. More advanced techniques such as ridge regression also have been applied to multimodel forecasts in a pointwise sense (Peña and Van den Dool 2008; DelSole 2007). DelSole et al. (2013) proposed a new technique, called scale-selective ridge regression, that ensures that the regression coefficients applied to forecasts have no small-scale structure.

An alternative approach to improving skill is to apply regression techniques to entire fields or patterns. DelSole and Shukla (2006) and Tippett (2006) proposed projections based on patterns that maximized predictability in dynamical models. In this paper, “skill” refers to how well a model predicts the real world, while “predictability” refers to how well a model predicts itself. Both studies found that such filtering lead to limited, though positive, improvements in forecast skill. However, only North America and Africa were studied, and DelSole
and Shukla (2006) found that the effectiveness of this approach was model dependent and in many cases equivalent to predicting just the empirical orthogonal functions of observations. The purpose of this paper is to explore this projection method more systematically and globally, and to consider alternative approaches to optimal projection.

Ideally, pattern-based forecast correction would transform a given forecast field into another field that has higher skill. If the transformation is linear, then the transformation operator can be estimated by multivariate regression methods. Unfortunately, the available observational record is too short to estimate such operators. In this work, we identify patterns that are optimal in some sense, and then project the patterns onto verification data and predict just their amplitudes. This method, which we call the projection method, is tantamount to filtering out unpredictable or unskillful patterns. One might attempt to rescale the patterns based on their skill, but this rescaling requires estimating additional parameters from data and raises serious problems with validating and comparing forecasts. Therefore, we predict just the amplitudes of these patterns without rescaling.

In the context of forecasting, two definitions of “optimal” naturally suggest themselves. First, optimal could mean “most predictable,” in the sense that the components being predicted are the best predicted components within the model itself, without regard to whether these components can be predicted in the real world. These components can be identified from predictable component analysis (PrCA) (Déqué 1988; Renwick and Wallace 1995; Schneider and Griffies 1999; DelSole and Tippett 2007). An important advantage of this approach is that it is not limited by the observational dataset: the number and accuracy of the components can be improved merely by increasing the number of ensemble members and the number of initial conditions in a set of numerical predictability experiments. However, a disadvantage of PrCA is that the obtained components may not be predictable in the real world. Second, optimal could mean “most skillful,” in the sense that the components that are retained have the maximum skill relative to observations. These components can be identified by modifying PrCA to maximize skill, as will be shown in this paper. The disadvantages of this second approach are that it is severely constrained by the available observational record, and the observations used to identify skillful components become unavailable for independent validation. Both definitions of “optimal” are considered in this paper, but only the results from optimizing skill are shown because of the higher prediction skill than optimizing predictability.

Although our proposed projection method does not involve linear regression directly, it does implicitly. In the (unrealistic) case in which data are not limiting, the ideal linear patterns would be obtained from applying canonical correlation analysis to the verification and forecast. The resulting patterns are related to linear regression in that they diagonalize the multivariate linear regression that maps forecast into verification field (DelSole and Chang 2003; DelSole and Tippett 2008). If the patterns are based on the most predictable patterns of an ensemble forecast, then the resulting patterns diagonalize the linear regression operator for predicting the ensemble mean given an ensemble member (DelSole and Tippett 2007). In these cases, one might characterize our patterns as “regression informed.” Projection methods can be generalized to handle combinations of model forecasts, but this generalization is not unique, is considerably more complex, and requires even larger samples, and so is deferred to a future study.

The methods for identifying predictable and skillful components are discussed in section 2. The dataset on which these methods are applied is described in section 3. Our results are discussed in section 4 and summarized and discussed in section 5.

2. Methodology

To identify spatial patterns that are skillful or predictable, we seek a linear combination of variables that maximize skill or predictability. This linear approach does not assume that the underlying dynamics are linear. Rather, the linear combination will eventually be used to define a basis set for representing variability, similar to the way principal components or spherical harmonics are used to represent variability in nonlinear systems. It is possible that nonlinear combinations might yield better skill or predictability, but it is difficult to identify such combinations without a priori model for constraining the nonlinear combinations. Let \( Y_{ns} \) denote the variable to be predicted for the \( nth \) sample (identified with time) at the \( sth \) spatial grid cell. The climatological mean has been subtracted from all variables, so \( Y_{ns} \) denotes the anomaly with respect to the climatological mean. Let \( N \) and \( S \) denote the number of samples and spatial grid cells, respectively. We want to study the skill or predictability of a linear combination of variables. Let the weights of the linear combination be an \( S \)-dimensional vector \( \mathbf{q} \), such that the resulting index is a time series given by

\[
 r_n = \sum_s Y_{ns} q_s .
\]  

Note that the time mean of \( r_n \) vanishes because the time mean of \( Y_{ns} \) has been removed. The prediction of \( Y_{ns} \),
denoted $\hat{Y}_{ns}$, is based on the ensemble mean forecast, while the corresponding index is denoted $\hat{r}_n$. The skill or predictability of this index is

$$\lambda = 1 - \frac{\sigma_E^2}{\sigma_C^2} \tag{2}$$

where $\sigma_E^2$ and $\sigma_C^2$ depend on whether skill or predictability is being measured. For predictability, these variances are defined as

$$\sigma_E^2 = \frac{1}{NE} \sum_{n=1}^{E} \sum_{e=1}^{N} (r_{ne} - \hat{r}_n)^2 \quad \text{and}$$

$$\sigma_C^2 = \frac{1}{N} \sum_{n=1}^{N} (\hat{r}_n - \bar{r})^2, \tag{3}$$

where $r_{ne}$ is the $e$th ensemble member at the $n$th verification time, $\bar{r}$ is the average over all ensemble members and verification times, and $E$ is the total number of ensemble members. In this case, $\sigma_E^2$ is an estimate of the intraensemble spread, often called “noise,” and $\sigma_C^2$ is an estimate of the climatological variance of the ensemble forecast. It can be shown that

$$\sigma_C^2 - \sigma_E^2 = \frac{1}{N} \sum_{n=1}^{N} (\hat{r}_n - \bar{r})^2, \tag{5}$$

where the right-hand side is an estimate of the variance of the ensemble means, often called “signal.” Thus, $\lambda$ can be interpreted as the signal-to-total ratio of predictability. On the other hand, for skill, the variances are defined as

$$\sigma_E^2 = \frac{1}{N} \sum_{n=1}^{N} (r_n - \hat{r}_n)^2 \quad \text{and}$$

$$\sigma_C^2 = \frac{1}{N} \sum_{n=1}^{N} (r_n - \bar{r})^2, \tag{6}$$

where $r_n$ is the verification at the $n$th verification time, and $\bar{r}$ is the average over all verification times. In this case, $\sigma_E^2$ is the mean square error of the forecast and $\sigma_C^2$ is the observed climatological variance, and $\lambda$ is equivalent to the squared error skill score (SESS). In both cases, our measure effectively compares the error of a dynamical model forecast to the error of a forecast based on the climatological mean. Because the denominator is the same for all forecasts, differences in our measure effectively imply differences in mean square error. Since only differences in skill are considered in this paper, the denominator merely sets the scale of the measure and does not alter the ranking of the models based on mean square error.

The parameter $\lambda$ can be written explicitly in terms of weights $\mathbf{q}$ as

$$\lambda = 1 - \frac{\mathbf{q}^T \Sigma_E \mathbf{q}}{\mathbf{q}^T \Sigma_C \mathbf{q}}, \tag{8}$$

where the superscript $T$ denotes the transpose operation, $\Sigma_E$ is the error covariance matrix, and $\Sigma_C$ is the climatological covariance matrix. Again, $\Sigma_E$ and $\Sigma_C$ depend on whether skill or predictability is being measured. For predictability,

$$(\Sigma_E)_{ss'} = \frac{1}{NE} \sum_{n=1}^{E} \sum_{e=1}^{N} (\hat{Y}_{ns} - \bar{Y}_{es})(\hat{Y}_{ns'} - \bar{Y}_{es'}) \quad \text{and}$$

$$(\Sigma_C)_{ss'} = \frac{1}{N} \sum_{n=1}^{N} (\bar{Y}_{es} - \bar{Y}_{s})(\bar{Y}_{es'} - \bar{Y}_{s}), \tag{9}$$

where $\bar{Y}_{es}$ is the ensemble forecast of the $e$th ensemble member for the $n$th verification time at the $s$th grid cell, and $\bar{Y}_{s}$ is the average over all ensemble members and verification time at the $s$th grid. For skill,

$$(\Sigma_E)_{ss'} = \frac{1}{N} \sum_{n=1}^{N} (\bar{Y}_{ns} - \bar{Y}_{ns})(\bar{Y}_{ns'} - \bar{Y}_{ns'}) \quad \text{and}$$

$$(\Sigma_C)_{ss'} = \frac{1}{N} \sum_{n=1}^{N} (\bar{Y}_{ns} - \bar{Y}_{s})(\bar{Y}_{ns'} - \bar{Y}_{s}), \tag{11}$$

where $\bar{Y}_{ns}$ denotes the observed value for the $n$th verification time at the $s$th grid, and $\bar{Y}_{s}$ is the average over all verification time at the $s$th grid. The weights $\mathbf{q}$ that maximize $\lambda$ are found by differentiating $\lambda$ with respect to $\mathbf{q}$ and setting the result to zero. This calculation leads to the generalized eigenvalue problem

$$(\Sigma_C - \Sigma_E) \mathbf{q} = \lambda \Sigma_C \mathbf{q}. \tag{13}$$

The properties of generalized eigenvalue problems have been discussed extensively in Schneider and Griffies (1999) and DelSole and Tippett (2007) and hence will only be summarized here. For $S$-dimensional matrices, there will exist $S$ eigenvectors. For each eigenvector there exists a corresponding eigenvalue $\lambda$. It is convenient to order the eigenvalues in descending order, in which case the first eigenvector $\mathbf{q}_1$ gives the weights that maximize $\lambda$, the second eigenvector $\mathbf{q}_2$ gives the weights that maximize $\lambda$ subject to the constraint...
and so on. The constraint (14) is equivalent to the constraint that the indices (1) derived from two eigenvectors (with distinct eigenvalues) are uncorrelated. It is convenient to normalize the indices to have unit variance, in which case \( q^T \Sigma_c q = 1 \). The regression pattern \( p \) associated with the index \( (1) \) can be shown to be
\[
p = \Sigma_c q.
\]

For typical global hindcast dataset, the number of grid points exceeds the number of samples, so the above covariance matrices are singular and the eigenvalue problem (13) cannot be solved. A standard approach is to project the data onto the leading principal components of the predictand, and then to maximize \( \lambda \) only in the subspace spanned by the leading principal components. To choose the number of principal components, we tested the sensitivity of \( \lambda \) to the number of principal components for different domains (as those shown in Fig. 1). Overall, \( \lambda \) is not sensitive when 10 or more principal components are selected. Thus, we chose 10 principal components from observations in this study. The principal components are derived in each domain separately and explain around 70% of total variance for the globe and global oceans, and more than 90% for continental domains, the tropical Atlantic (20°S–20°N), and the tropical Pacific (20°S–20°N).

Another issue is that overfitting leads to serious biases when the dimension of the subspace is not a small fraction of the sample size. Despite this bias, subjective judgment of the optimized ratios turns out to be adequate for this paper, for reasons that will become clear.

Having identified the predictable or skillful components as above, we then project the forecasts onto the leading components. This projection onto a reduced subspace is tantamount to filtering out unpredictable or unskillful components.

3. Data

The multimodel hindcast dataset used in this study is from the Ensemble-Based Predictions of Climate Changes and Their Impacts (ENSEMBLES) project (http://www.ensembles-eu.org/). This dataset consists of 7-month hindcasts by five state-of-the-art coupled atmosphere–ocean general circulation models (AOGCMs) from the Met Office, Météo France, European Centre for Medium-Range Weather Forecasts (ECMWF), Leibniz Institute of Marine Sciences at Kiel University, and Euro-Mediterranean Centre for Climate Change in Bologna. Each model produced a nine-member ensemble hindcast. These were averaged to construct an ensemble mean forecast for each model. Considering only ensemble mean forecasts is justified when only linear combinations of forecasts are considered (DelSole 2007). We emphasize that most of the analyses in this paper are based on ensemble mean forecasts. All models include major radiative forcings and were initialized using realistic estimates from observations. More details of the data can be found in Weisheimer et al. (2009).

We examined 2-m temperature hindcasts initialized on the first of November of each year during 1960–2005 to illustrate the proposed technique. We considered the
November–January (NDJ) mean of temperature. Hindcasts from all models were interpolated to a common 2.5° × 2.5° grid. The hindcasts of ensemble members from a specific model were centered by subtracting the grand mean of the model.

The data used to verify 2-m temperature hindcasts are the observation-based surface temperature data from the National Centers for Environmental Prediction (NCEP)–National center for Atmospheric Research (NCAR) reanalysis (Kistler et al. 2001), which was interpolated on the same grid as dynamical model hindcasts. The verification data were centered relative to the mean of 1960–2005.

4. Results

a. Comparison of skill and predictability among models

To gain insight into the model dependence of skill and predictability, we first determine whether a pattern exists that is well predicted by one model but not by another model, that is, whether there exists “complementary skill” (Kirtman and Min 2009). To this end, we determine the most skillful pattern of one model and then project this pattern onto the hindcasts by other models. The resulting skill values [i.e.,  \( \lambda \) in (2)] of the extracted component are shown in Fig. 2 for 10 domains, where the \( x \) axis indicates the reference model used to maximize skill and the far right \( x \)-axis value, denoted by m, shows the skill of individual models for the component that maximizes the skill of multimodel mean hindcast (i.e., average over ensemble mean hindcasts of all models). Each number in the figure indicates a specific model. The principal components used to represent the data in each domain were computed separately from observations using empirical orthogonal function (EOF) analysis. For global or oceanic domains, the skill values are largely consistent with each other, indicating that the models share similar skillful components. For continental domains, however, dramatic differences in skill are found. In South America, for instance, the pattern using model 1 [ECMWF Integrated Forecast System Model, version 31 release 1 (IFS31R1)] or 4 [Met Office Hadley Centre Global Environment Model, version 2 (HadGEM2)] as a reference model is predicted with positive skill in models 1 and 4 but is predicted with negative skill in all other models. This result implies the existence of a particular pattern that is predicted with skill by models 1 and 4 and predicted with no skill by other models, a result that may be of interest in model comparisons. Also, in Europe, the most skillful pattern of each model has skill that is separated from the skill of other models. Note that over North America model 2 provides a more skillful prediction of the most skillful component of model 5 than model 5 itself. Other less extreme examples can be seen in the figure. Such examples merely show that some models are superior to others in predicting certain patterns. In principle, these skill differences could be exploited in such a way as to maximize skill over specific continents. The best procedure for combining forecasts is an open question that is beyond the scope of the present study. However, note that the skill values for the most skillful component of the multimodel mean are close to each other and usually close to the maximum skill of any model. This result suggests that the multimodel mean captures most of the skill that is common among models. For this reason, we focus on the multimodel mean dataset in our analysis.

Similar calculations based on maximizing predictability reveal that, in all domains, predictability of the most predictable component is high and consistent among models (not shown). This result implies that the models have predictability concentrated in similar patterns. However, the fact that the squared error skill score of the most predictable component is relatively low and sensitive to models over continental domains (not shown) suggests that the source of the model’s predictability does not correspond to nature.

b. Filtering based on skillful components

We performed filtering based on skillful components and predictable components. The improvement in SESS resulting from filtering unskillful patterns was found to be nearly the same as that resulting from filtering unpredictable patterns (now shown), except in South America and Africa, where filtering unskillful patterns gives higher SESS. Accordingly, we present results of filtering unskillful patterns. The independent prediction error of the filtered prediction is estimated by cross validation techniques. Specifically, one year of data is withheld, the remaining years are used to maximize multimodel mean skill in the subspace spanned by the leading 10 principal components, and then the resulting components are used to predict the withheld year from each model. The procedure was repeated using a different withheld year in turn until all years had been withheld exactly once. The predictions from all models were averaged to generate a mean prediction. After constructing a mean prediction based on the most skillful patterns, we then computed the SESS at each grid cell and then averaged the SESS over the domain to compute a single SESS for each domain. Here, the total observations, rather than the 10 principal components of observations, were used to compute SESS.
The cross-validated SESS for NDJ mean 2-m temperature in 10 geographic domains is shown in Fig. 3. The horizontal axis shows the accumulated number of skillful components used to construct the prediction. For instance, the x-axis value 2 means that both the first and second most skillful patterns were pooled together to construct the prediction. The horizontal line in each panel gives the SESS of a simple average of all model

![Fig. 3](image-url)
hindcasts, which does not involve any filtering or maximization at all. In general, the SESS values are highest in the tropical Pacific and Atlantic, and lowest in Europe. The nearly zero SESS in Europe implies no skill can be identified, at least within the subspace spanned by the leading principal components used to represent the state. Except for South America and Africa, the SESS values resulting from removing unskillful patterns are smaller or close to the SESS of the simple multimodel mean. Thus, the benefit of removing unskillful patterns is evident in only two domains. The skill of the simple average sometimes exceeds the skill of the filtered forecast because, as we will show shortly, the EOF prefiltering in these cases reduces skill prior to skill maximization.

Note that a significance test is not essential for drawing the above conclusion. As discussed in section 2, the optimized skill pattern tends to have inflated skill
because of overfitting. Such inflation harms the skill under a proper cross validation procedure. Nevertheless, the skill of the projected hindcasts does not exceed the skill of the multimodel mean except over South America and Africa, suggesting little bias. It is worth emphasizing that the observation EOFs and climatology were recomputed for each year withheld in the cross-validation procedure, as is proper.

c. Spatiotemporal structure of skillful components

To understand the spatiotemporal structure of the skillful components, we show the global patterns and time series for the two most skillful components of the multimodel mean hindcasts in Fig. 4. The first pattern shows large amplitudes over central and eastern tropical Pacific and is characteristic of an El Niño–Southern Oscillation (ENSO) pattern. The time series for this component are highly correlated across the five models and observations. The second most skillful pattern corresponds to a warming trend, with most of the warming occurring in the Northern Hemisphere polar regions. These results provide a convincing demonstration that most of the NDJ temperature predictability is attributable to two components, namely ENSO and a global warming trend. Optimizing skill over regional domains (not shown) also yields ENSO and trend patterns, except in Europe, which has no significant skill pattern for NDJ temperature.

The fact that ENSO and the global warming trend dominate prediction skill on seasonal time scales is well established now, but the ability of our method to isolate these components cleanly is attractive. Also, the fact that our optimized method found no other components with comparable skill implies that no other major component of predictability of seasonal mean temperature exists in the leading EOFs of this multimodel hindcast dataset.

d. Filtering based on observation EOFs

As shown in Fig. 3, the SESS from using all 10 components tends to be close to the peak SESS in each domain. However, using all 10 components derived from 10 EOFs is equivalent to simply projecting the forecast onto 10 EOFs. That is, the skill maximization is unnecessary when all 10 skillful components are included in the prediction. This result implies that most of the improvement in skill resulting from filtering comes not from optimizing skill directly, but from filtering out EOFs!

To test the above conclusion, we simply project forecasts onto the leading EOFs derived from observations. No skill optimization procedure is performed. Projecting EOFs onto observations merely removes patterns from the forecast that have low variance in observations. The cross-validated SESS as a function of the accumulated number of observation EOFs used to represent the prediction is shown in Fig. 5. Except for South America and Africa, the SESS generally increases with increasing
number of EOFs. In the case of South America and Africa, the skill score reaches a peak for a small number of observation EOFs (e.g., <10). Comparison with the SESS for optimally filtered forecasts, shown in Fig. 3, reveals relatively little difference between the peak SESS based on skillful patterns and the SESS of prediction using only a few observation EOFs. These results strongly suggest that there is relatively little benefit from optimizing skill directly, and that most of the benefit from removing unskillful patterns comes from simply removing patterns that have low variance in observations. This conclusion is substantially clarified by the skill optimization calculations: the latter calculations eliminate the possibility that some linear combination of observed EOFs could perform better than just predicting the EOFs themselves.

Interestingly, the convergence of SESS to the multimodel mean is slower for global and ocean domains than for smaller domains, suggesting that global-scale EOFs are an inefficient basis set for representing skillful
variability, possibly attributable to a larger number of degrees of freedom.

Note that the cross-validated SESS from using all EOFs (i.e., the last point along the x axis) is smaller or close to the SESS of multimodel mean prediction (solid horizontal line) in all domains except in South America and Africa. In the latter regions, the lower SESS of the simple multimodel mean is caused by patterns generated by dynamical prediction models that do not project on observations. In general, model generated patterns that are orthogonal to observations can only add error to the prediction. The EOF filtering is able to improve skill by removing such patterns.

Since temperature variance can vary substantially between tropics and midlatitudes, and between ocean and land, it is possible that the EOFs used here, which maximize area-weighted variance, might overemphasize midlatitude land regions. To check this possibility, we repeated the above analyses, but applied EOF analysis to data in which each grid point is normalized by its own standard deviation. The results from normalized EOFs are close to the unnormalized results presented here, except that the SESS values are lower. Thus, we showed results only for the unnormalized EOFs.

e. Unrealistic patterns

So far we have explored filtering based on predictable or skillful patterns. An alternative approach is to filter based on "unrealistic" patterns that can be defined strictly from the verification dataset without reference to the hindcasts. For instance, model patterns that are orthogonal to reanalysis fields may be called "unrealistic with respect to reanalysis." Such patterns are a major source of error in South America (see section 4d). Identifying unrealistic patterns may be just as helpful as identifying predictable or skillful patterns. In this section, we briefly explore unrealistic model patterns.

The unrealistic patterns can be defined as

$$\mathbf{N} = \mathbf{X} - \mathbf{E} (\mathbf{E}^\top)^{-1} \mathbf{X}, \quad (16)$$

where $\mathbf{E}$ is the matrix consists of all EOFs from observation, $\mathbf{E}^\top$ is the corresponding pseudoinverse matrix of $\mathbf{E}$ [i.e., $(\mathbf{E}^\top)^{-1} \mathbf{E} = \mathbf{I}$], and $\mathbf{X}$ denotes the multimodel hindcasts. The leading EOF of the unrealistic patterns is shown in Fig. 6 for South America and Africa. The two patterns explain 27.2% and 15.9% of the total variance of $\mathbf{N}$. Different colors in the time series indicate different models. In Africa, the leading EOF of the unrealistic patterns shows largest loadings in central Africa, and the associated time series shows an increasing trend in all models. Surprisingly, the amplitude of this pattern is strongly correlated among models. In the case of South America, the time series fluctuate with nearly a 4-yr period, perhaps suggesting an ENSO influence. Again, the time series of this unrealistic pattern are strongly correlated among models.

It is unclear what meaning should be attached to model-generated patterns that are orthogonal to reanalysis. For instance, these same patterns might project significantly on other observation-based datasets, implying that the definition of "unrealistic" depends on dataset. To explore this issue further, we projected the leading EOF of (16) onto estimates of NDJ temperature derived from the Climate Research Unit (CRU) time series, version 3.1 (TS3.1), dataset (Mitchell and Jones 2005) (shown as the thick black curve in Fig. 6). For Africa, the resulting time series shows a much weaker trend than in the models, while for South America the time series is insignificantly correlated with model variability. These results demonstrate that dynamical models can agree on the behavior of a pattern, yet this behavior is not clearly reflected in either reanalysis or CRU data. Nevertheless, neither reanalysis nor CRU data are the truth, so the possibility exists that this unrealistic pattern may be realistic relative to truth.

5. Summary and discussion

This paper proposed an optimal projection technique for improving the skill of dynamical model forecasts. The method is optimal in the sense that it identifies patterns that are either the most predictable or the most skillful. Because the method is optimal, no other projection method can improve the skill or predictability of the forecasts as much as the proposed methods, at least in the limit of large sample size and within the subspace spanned by the leading EOFs used to represent the state.

The proposed method was applied to NDJ mean 2-m temperature from the ENSEMBLES multimodel hindcasts with November initial conditions. The method first was used to compare the skill of different models on a pattern-by-pattern basis. Specifically, the most skillful pattern of a particular model was determined, and then the skill of this pattern as predicted by other models was calculated. For global or oceanic domains, the models predicted each other’s most skillful patterns with comparable skill. For continental domains, however, significant differences were found, including cases in which a few models predicted a domain with positive skill while all other models had negative skill. In principle, such differences could be exploited to improve the skill of model forecasts by selectively filtering certain patterns and models from the multimodel ensemble. However, validating such a procedure with a 46-yr hindcast dataset...
presents formidable challenges that lie beyond the scope of this paper.

An interesting result is that maximizing skill of the multimodel mean dataset resulted in patterns that were skillfully predicted by all the models, with skill values close to the maximum skill of any individual model. Since this result was obtained within the context of optimal filtering, it strongly validates the use of the multimodel mean forecasts in seasonal prediction.

The cross-validated skill of the optimally filtered prediction, as a function of the number of skillful patterns included in the prediction, was found to give nearly the same skill as the original multimodel mean prediction in all domains except in South America and Africa. This result implies that projection techniques are likely to be beneficial only in a few locations, where models have relatively large errors. Since our technique is optimal, this pessimistic conclusion is fairly definitive, at least within the subspace spanned by the 10 leading EOFs of the observed seasonal mean temperature.

An important by-product of the above analysis is that it provides a concise summary of the skillful structures in a given forecast. For the ENSEMBLES dataset, the method convincingly demonstrates that the vast majority of the skill of seasonal prediction skill comes from only two components, ENSO and the global warming trend.

In the case of South America, optimal projection techniques substantially improved the skill. In fact, the multimodel mean prediction had negative skill in this domain, whereas the filtered prediction had positive skill over a wide range of filtering levels. Thus, although the benefits of optimal projection are limited, the benefits can be substantial in certain domains. However, the results revealed that the skill using the best choice of filtering was nearly the same as the skill based on using all 10 skillful components, which in turn is precisely equivalent to simply projecting the forecasts onto the leading 10 observed EOFs that involves no sophisticated skill optimization.

Thus, the improvement in skill attributable to projecting came not from selecting patterns with maximum skill, but merely from EOF prefiltering (i.e., filtering patterns that have low variance in observations). To investigate this issue further, the cross-validated skill as a function of the accumulated number of observed EOFs kept in the

![FIG. 6. The leading EOF pattern and associated time series of model hindcasts after reanalysis patterns have been removed, for NDJ mean 2-m temperature in (left) South America and (right) Africa. The South American and African patterns explain 27.2% and 15.9% of the residual variance after reanalysis patterns have been removed. The thick black curves in the bottom panels show the projection of the pattern onto the CRU TS3.1 dataset.](http://journals.ametsoc.org/doi/pdf/10.1175/JCLI-D-13-00232.1)
prediction was examined. As anticipated, the skill increased and approached the skill of the multimodel mean forecast with the increase of number of EOFs in all domains except in South America and Africa. In South America, however, the skill was substantially higher than the multimodel mean prediction for all EOF truncations, with maximum skill at two EOFs. The skill of two EOFs was nearly the same as the skill obtained from optimal projection, suggesting that the benefits of optimal projection are comparable to the benefits of simple EOF filtering. Again, since the EOF filtering is compared to an optimal method, this conclusion is fairly definitive for this hindcast dataset.

Technically, our results show that optimal projection can improve dynamical model forecasts, which answers the question stated in our title. However, this improvement occurs only over two continents, namely South America and Africa, and optimal projection actually degrades the skill over global or oceanic domains. Furthermore, the improvement is dominated not by the optimal projection itself, but by the EOF prefiltering that is performed prior to optimization. EOF filtering may be considered suboptimal filtering. Thus, based on these considerations, we would not recommend optimal projection methods for improving forecasts. Instead, our results suggest that merely filtering patterns with weak variances in observations gives most of the benefit of filtering over South America and Africa. Despite the negative nature of our conclusion, we emphasize that negative results that are clear and unambiguous constitute progress.

We argue that the proposed techniques could still prove valuable to a wide range of other forecast problems. For instance, results presented here show that some models predict certain spatial patterns substantially better than other models. Such information may be helpful in identifying aspects of models that lead to improved forecasts. Conversely, some models predict certain spatial patterns much worse than other models. Once such patterns are identified, it seems likely that monitoring the skill of such patterns during model development may lead to efficient improvement strategies. The proposed techniques could also be used to develop weighting strategies in multimodel predictions for models with different skill.

A significant source of the poor skill over South America is the fact that models generate a host of “unrealistic” patterns, defined as patterns that are orthogonal to the observational dataset. A surprising result to emerge from this analysis is that the leading EOFs of these unrealistic patterns are strongly correlated among models. An important caveat should be kept in mind when interpreting the above results: specifically, the datasets used in this study were constrained to match the observational dataset. In contrast, much larger datasets can be generated for other purposes, such as classical predictability experiments. Indeed, if a sufficiently large number of classical predictability experiments could be performed, the most predictable components could be determined in a much larger dimensional space, allowing more accurate determination of these patterns, which in turn may yield much greater benefits in skill than those found here.

Acknowledgments. This research was supported primarily by the National Oceanic and Atmospheric Administration, under the Climate Test Bed program (NA10OAR4310264). Additional support was provided by the National Science Foundation (ATM0332910, ATM0830062, ATM0830068), the National Aeronautics and Space Administration (NNG04GG46G, NNX09AN50G), and the National Oceanic and Atmospheric Administration (NA04OAR4310034, NA09OAR4310058, NA05OAR4311004, NA10OAR4310210, NA10OAR4310249). The views expressed herein are those of the authors and do not necessarily reflect the views of these agencies.

REFERENCES


