Model Fidelity versus Skill in Seasonal Forecasting

TIMOTHY DELSOLE AND JAGADISH SHUKLA

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

(Manuscript received 30 March 2009, in final form 10 March 2010)

ABSTRACT

The relation between skill and fidelity of seasonal mean hindcasts of surface temperature by seven coupled atmosphere–ocean models is investigated. By definition, fidelity measures the agreement between model and observational climatological distributions, and skill measures the agreement between hindcasts and their corresponding verifications. While a relation between skill and fidelity seems intuitively plausible, it has not been checked systematically, nor is it mandated mathematically. New measures of skill and fidelity based on information theory are proposed. Specifically, fidelity is measured by the area average relative entropy between the climatological distributions of the model and observations, and skill is measured by the area averaged mutual information between forecast and verification. The fidelity measure is found to be dominated by the term measuring mean bias; that is, the discrepancy in climatological means is much larger than the discrepancy in climatological variances. Moreover, the mean bias is negatively correlated with skill at most initial months, lead times, and regions examined. Thus, models that more closely replicate the observed climatological mean tend to have better skill.

1. Introduction

The purpose of this paper is to test whether the skill of dynamical model hindcasts is related to the fidelity with which the model simulates the climatology. Here, fidelity refers to the degree to which the *climatology* of the hindcasts matches the observed climatology, and skill refers to the degree to which individual hindcasts match individual verifications. In general, models with poor climatologies are expected to have poor skill. For instance, storm track models suggest that the characteristics of synoptic fluctuations about the mean flow depend on the mean flow itself (Chang et al. 2002; DelSole 2004b), suggesting that a more accurate mean flow would translate into more accurate synoptic forecasts. Similarly, idealized models of ENSO consistently show that ENSO variability depends on the structure of the mean thermocline (Kirtman and Schopf 1998; Fedorov et al. 2003), suggesting that coupled atmosphere–ocean models that simulate the mean thermocline more accurately also will predict ENSO more accurately, all other things being equal. In extreme cases, models that generate incorrect circulation fields cannot be expected to have any skill.

Nevertheless, it should be recognized that skill and fidelity need not be related. For instance, the climatological mean of a linear model can be controlled independently of the variability about the mean (e.g., by adjusting the mean forcing); hence, the quality of linear forecasts about the climatological mean can be independent of the quality of the climatological mean itself. Another counterexample is provided by DelSole et al. (2008) and Yang et al. (2008), who added empirical forcing terms to atmospheric general circulation models to improve the climatological mean, but found no consistent improvement in skill.

In practice, dynamical model forecasts drift from the observed climate, leading to a difference from the observation. To mitigate the effects of drift, forecasters routinely subtract the climatological mean from the forecast and verification separately and then compare the corresponding anomalies. Numerous studies have confirmed that anomalies forecasted by dynamical models have skill even on seasonal time scales [see the July 2000 issue of the *Quarterly Journal of the Royal Meteorological Society* and Palmer and Hagedorn (2006)]. In most prediction studies, however, the specific climatological mean that is subtracted from the forecast often is discarded without commenting on how close it is to the observed climatological mean. This practice makes it difficult to ascertain whether a relation exists between skill and fidelity.

Corresponding author address: Timothy DelSole, 4041 Powder Mill Rd., Calverton, MD 20705.
E-mail: delsole@cola.iges.org

DOI: 10.1175/2010JCLI3164.1

The existence of a skill–fidelity relation in dynamical models would have significant implications for model development and climate change projections. For instance, if skill and fidelity were related, then one approach to improving the skill of a dynamical model is to improve its climatology [although DelSole et al. (2008) and Yang et al. (2008) show that simple empirical corrections do not generally improve skill]. Within the context of climate change projections, the question arises as to whether some projections are more trustworthy than others. Shukla et al. (2006) found that models that more accurately simulated the climatological distribution of surface temperature of the past 100 yr also tended to produce higher values of global warming for a doubling of CO_2 concentration. This result suggests that the projected warming due to increasing greenhouse gases is likely to be closer to the highest projected estimates—*assuming that models with greater fidelity also have greater skill*. Unfortunately, the last assumption cannot be checked because verifications for climate change projections are not available. However, such a relation can be tested for shorter lead-time hindcasts, for which verifications are available.

Since fidelity is a measure of how well the climatology of forecasts replicates the observed climatology, quantifying fidelity requires a measure of the difference between two *distributions*. Moreover, we seek a measure that allows different variables with different units to be taken into account. Also, we want a single measure that measures the fidelity of spatially varying variables as a whole. Similarly, skill is a measure of how well the forecast and verification covary, but it is not clear how to combine measures of skill at different locations and variables to give a measure of skill as a whole.

This paper proposes new measures of fidelity and skill that have several attractive properties. Specifically, we propose measuring fidelity based on the spatially averaged *relative entropy* between the climatological distributions of the model and observations, and measuring skill based on the spatially averaged *mutual information* between the forecast and corresponding verification. These measures are central to information theory (Cover and Thomas 1991) and will be discussed more fully in sections 2 and 3. Second, we evaluate and test the significance of these measures for a set of seasonal hindcast derived from the Development of a European Multimodel Ensemble System for Seasonal to Interannual Predictions (DEMETER) project, as discussed in section 5. We find that skill tends to increase with fidelity. We conclude with a summary and discussion of our results.

2. Measures of skill and fidelity

Skill refers to the degree to which forecasts and verifications are related to each other. In information theory, the natural measure of the dependence between variables is mutual information (Cover and Thomas 1991), defined as

$$M(\mathbf{f};\, \mathbf{v}) = \int p(\mathbf{f},\, \mathbf{v}) \log\left[\frac{p(\mathbf{f},\, \mathbf{v})}{p_f(\mathbf{f})p_v(\mathbf{v})}\right] d\mathbf{f}\, d\mathbf{v}, \tag{1}$$

where $p(\mathbf{f}, \mathbf{v})$ is the joint distribution between forecast and verification, and $p_f(\mathbf{f})$ and $p_v(\mathbf{v})$ are the respective marginal distributions. The above integral is interpreted as a multivariate integral over the support of $\mathbf{f}$ and $\mathbf{v}$. If the forecast is independent of the verification, and hence the forecast has no skill, then by definition

$$p(\mathbf{f},\, \mathbf{v}) = p_f(\mathbf{f})p_v(\mathbf{v}). \tag{2}$$

Substitution of (2) into (1) gives $M(\mathbf{f}; \mathbf{v}) = 0$, which shows that mutual information vanishes if the forecast and verification are independent. It turns out that mutual information vanishes if *and only if* (2) is true. Thus, mutual information provides a fundamental measure of skill, in the sense that it vanishes if and only if the forecast and verification are independent.

Fidelity refers to the degree to which the distribution of all forecasts matches the observed climatological distribution. As such, fidelity depends only on the *marginal* distributions while skill depends on the *joint* distributions; that is, fidelity is measured independently of skill. In information theory, an often used measure of the difference between two distributions is the relative entropy (Cover and Thomas 1991), defined as

$$R = \int p_v(\mathbf{x}) \log\left[\frac{p_v(\mathbf{x})}{p_f(\mathbf{x})}\right] d\mathbf{x}. \tag{3}$$

If the distribution of all forecasts equals the observed climatological distribution, then $p_v(\mathbf{x}) = p_f(\mathbf{x})$ and $R = 0$, showing that relative entropy vanishes if fidelity is perfect. It turns out that relative entropy vanishes if *and only if* the forecast and verification have the same climatological distribution. Because relative entropy vanishes for perfect fidelity, it is perhaps more accurate to say that relative entropy measures *discrepancy*.

Both mutual information and relative entropy have a number of mathematical properties that make their use attractive. First, both measures are invariant to invertible nonlinear transformations of the state. This invariance implies that variables with different units or natural variances can be included in a single state vector without scaling, since such scaling cannot affect the value of the measure. Second, if selected state variables are mutually independent, then mutual information and relative entropy are separately *additive*. For instance,

TABLE 1. Designation and source of the hindcasts from the DEMETER project used in this paper.

Designation	Center
CER	European Centre for Research and Advanced Training in Scientific Computation, France
ECM	ECMWF, International
ING	Istituto Nazionale de Geofisica e Vulcanologia, Italy
LOD	Laboratoire d'Océanographie Dynamique et de Climatologie, France
MET	Centre National de Recherches Météorologiques, Météo-France, France
MPI	Max-Planck Institut für Meteorologie, Germany
UKM	Met Office, United Kingdom

TABLE 2. The designation and boundaries of regions used to compute area-averaged skill. Only land points are included in the defined regions.

Designation		Boundaries
North America	NAM	20°–80°N, 168°–50°W
South America	SAM	60°–20°N, 110°–25°W
Europe	EUR	35°–80°N, 30°W–50°E
Asia	ASIA	0°–80°N, 50°W–180°
Africa	AFR	40°S–35°N, 20°W–50°E
Australia	AUS	50°–10°S, 110°W–180°
Tropics	TRP	23°S–23°N, 0°–360°
Globe	GLB	60°S–80°N, 0°–360°

the relative entropy of two independent variables equals the sum of the relative entropy of the individual variables. Third, relative entropy and mutual information have an immense number of applications in statistics, signal detection, stock market analysis, and communication theory. Thus, mutual information and relative entropy have a wide variety of other interpretations that may be useful beyond predictability itself.

Both mutual information and relative entropy are nonnegative and unbounded. These measures can be converted into "scores" between 0 and 1 using the following transformations:

$$\text{skill score} = 1 - e^{-2M} \qquad \text{fidelity score} = e^{-2R}. \tag{4}$$

Joe (1989) has shown how these transformations reduce to standard statistical quantities in certain cases. For instance, if the forecast and observations are bivariate, normally distributed, then the above skill score reduces to the squared correlation coefficient. In this paper, we consider M and R directly rather than the above score measures.

Unfortunately, the probability distributions required to evaluate the above quantities are not known and hence must be estimated from finite samples. Estimation of these quantities for multivariate data is difficult. A typical approach is to project variables onto a few leading principal components and then evaluate the relative entropy and mutual information in this reduced space (Shukla et al. 2006). Unfortunately, the results of this approach pertain to the particular principal components of the dataset, and can be sensitive to the chosen number of principal components. [Shukla et al. (2006) analyzed 100-yr datasets using the leading 15 components of two different "flavors" of principal components.] To obtain more robust and reproducible estimates, we propose calculating each measure individually and independently at each point, and then averaging each measure over a selected domain to obtain a single measure for the domain. This approach ignores the influence of spatial correlations, which is a compromise for estimating these quantities with small sample sizes and no prior information. It is sensible to average the relative entropy and mutual information, as opposed to some other function of these quantities, because these measures are additive for independent events. In general, the properties of relative entropy and mutual information noted above do not carry over to their spatially averaged counterparts. Nevertheless, reasons for preferring these measures are noted at the end of this section.

Second, we assume that the variables are normally distributed. In this case, both relative entropy and mutual information can be written in closed form in terms of the mean and variance of the variables (DelSole 2004a). When all of these assumptions and modifications are incorporated, our measure of discrepancy (or inverse fidelity) becomes

$$D = D_B + D_V, \tag{5}$$

$$D_B = \sum_{\text{grid}} \frac{(\mu_v - \mu_f)^2}{\sigma_v^2} \Delta A, \quad \text{and} \tag{6}$$

$$D_V = \sum_{\text{grid}} \left[\frac{\sigma_f^2}{\sigma_v^2} - \log\left(\frac{\sigma_f^2}{\sigma_v^2}\right) - 1 \right] \Delta A, \tag{7}$$

TABLE 3. The designation and boundaries of regions used to compute area-averaged skill. Only ocean points are included in the defined regions.

Designation		Boundaries
North Atlantic	NAT	30°–60°N, 90°W–0°
North Pacific	NPC	30°S–60°S, 130°–240°E
Indian	IND	10°S–20°N, 30°–120°E
Tropical Atlantic	TAS	20°S–30°N, 80°W–20°E
Niño-3	NIÑO-3	5°S–5°N, 210°–270°E
Niño-4	NIÑO-4	5°S–5°N, 160°–210°E
Niño-3.4	NINO34	5°S–5°N, 170°–150°W
Globe	GLB	60°S–80°N, 0°–360°

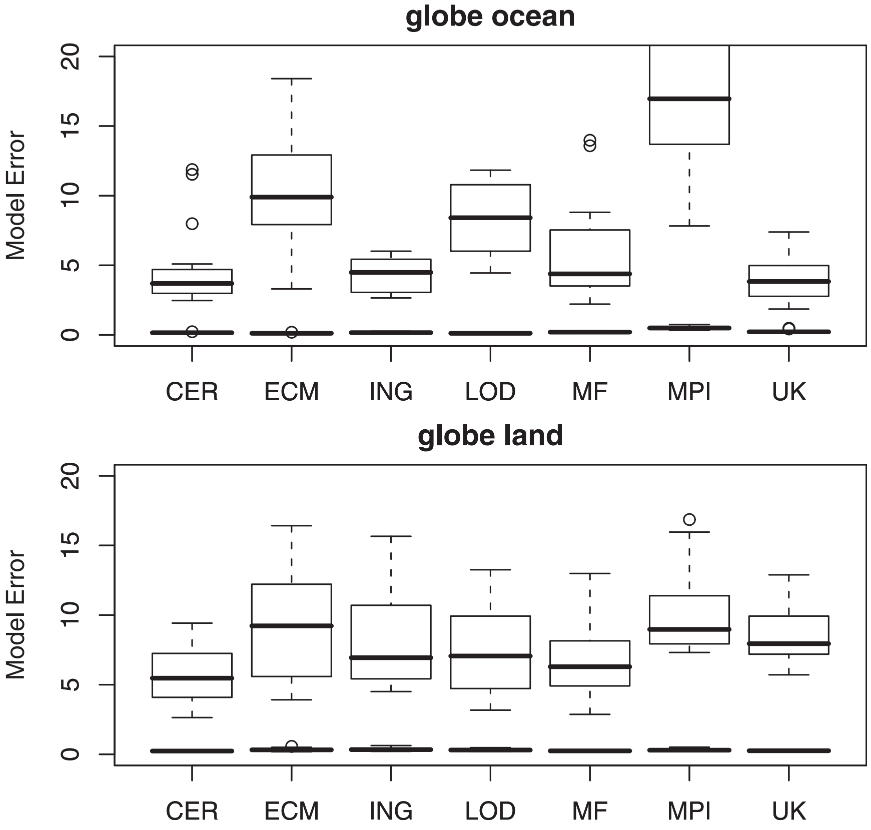

FIG. 1. Box-and-whisker plots for the area-averaged model error (as measured by spatially averaged relative entropy; described below) between the hindcast and observed climatology of 3-month-average T2 m over (top) ocean and (bottom) land as a function of model. In each panel, the top box-and-whisker plot indicates the normalized spatially averaged bias D, defined in (6) and the bottom box-and-whisker plot indicates spatially averaged differences in variance D_V defined in (7); both quantities are dimensionless. Only grid points between 60°S and 80°N are included in the area averages. The plot graphically depicts the range of values (16 in all, corresponding to the 4 initial months and 4 hindcast seasons) as follows: the top and bottom edges of the box indicate the top and bottom quartiles, the centerline in the box denotes the median, and the whiskers at the top and bottom extend to the most extreme data points, which are no more than 1.5 times the interquartile range from the box.

and our measure of skill S is

$$S = -\sum_{\text{grid}} \log(1-\rho^2)\Delta A, \qquad (8)$$

where μ_v, σ_v^2 are the mean and variance of the verification in each grid box, respectively; μ_f, σ_f^2 are the mean and variance of the forecast in each grid box, respectively; ρ is correlation between the ensemble mean forecast and verification in each grid box; and ΔA is the fractional area of a grid box (including cosine of latitude).

The skill measure (8) is invariant to the sign of the correlation ρ; a forecast that is negatively correlated with verification is deemed just as skillful as a forecast that is positively correlated. Whether this convention is appropriate depends on the study. To avoid counting negative correlations toward skill, one might average skill only over grid cells with positive correlations, but this approach would lead to biased skill estimates since the omitted cells were identified a posteriori. Another approach is to replace ρ^2 with $\rho|\rho|$, which penalizes against negative correlations, but then we lose contact with information theory. It turns out that both of these approaches give results similar to each other and similar to the original definition (at least in the present paper). Hence, we use (8) directly for our measure of skill.

Similarly, the term D_B in (6) is a measure of *forecast bias*, normalized by the observed standard deviation. This ratio is similar to that used in the familiar t test to test

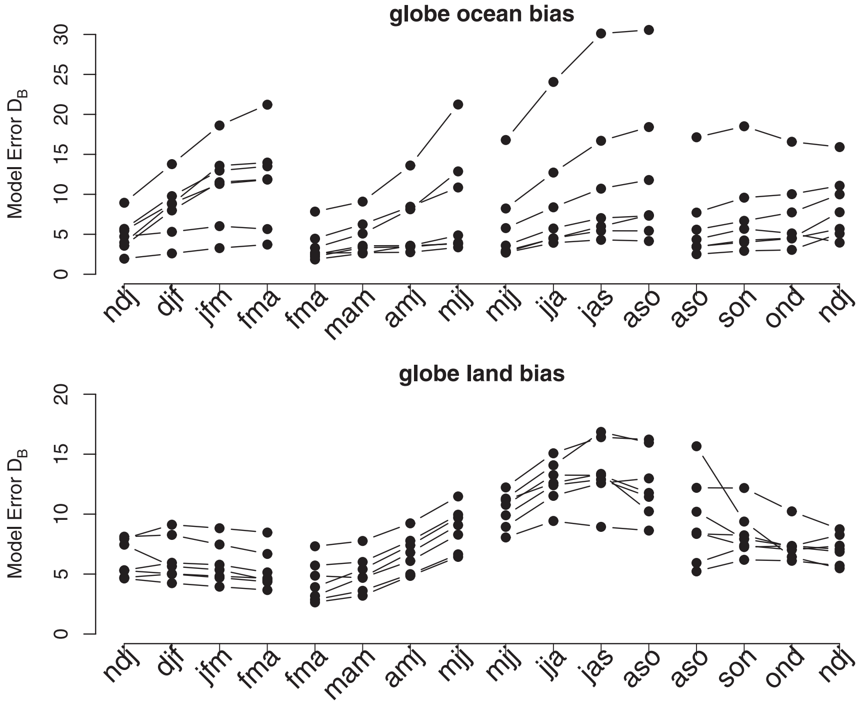

FIG. 2. The model error, as measured by the normalized bias D_B defined in (6) in 3-month-average T2 m over (top) ocean and (bottom) land as a function of the verifying 3-month average. Each curve corresponds to one of the 7 hindcast models, and begins and ends in the corresponding 3-month-average period of the 6-month hindcasts.

a difference in means. The term D_V in (7) is a measure of the difference in variances between the forecasts and verifications. For instance, D_V vanishes if $\sigma_f = \sigma_v$ and is positive otherwise. Thus, D_B measures how well the climatological means agree between forecasts and verifications, and D_V measures how well the climatological variability agrees between forecasts and verifications.

The correlation coefficient ρ, needed to evaluate skill in (8), is computed between the *ensemble mean hindcast* and verification, for each initial day, lead time, and grid box separately. The climatological mean μ_f and variance σ_f^2 of the hindcasts, needed to evaluate discrepancy in (5), is computed as the average over *all ensemble members* and all years, for a fixed initial day, lead time, and grid box. The climatological mean and variance of the verifications are computed similarly, except with only one realization. The reasons for using an ensemble mean forecast to measure skill but individual ensemble means to measure fidelity are discussed in DelSole (2005). Loosely speaking, the best single forecast is the ensemble mean, so it is appropriate to use the ensemble mean when estimating forecast skill. In contrast, fidelity measures the degree to which the model simulates the climatology. In a sense, fidelity is a property of a single realization of the forecast model, so it is appropriate to use individual ensemble members to measure fidelity. Ensemble means would be inappropriate for measuring fidelity because they would have too little variance compared to unaveraged realizations.

We further note that the above measures are closely related to more traditional measures. For instance, for small correlations our skill measure is approximately

$$S = -\sum_{\text{grid}} \log(1 - \rho^2)\Delta A \approx \sum_{\text{grid}} \rho^2 \Delta A, \tag{9}$$

which is simply the spatially averaged squared correlation coefficient. Since the squared correlation can be interpreted as the fraction of variance explained by the forecast, S can be interpreted as the average fractional variance explained by the forecast. Similarly, for weakly varying climatological variance σ_v^2, the bias term in the relative entropy is approximately

$$D_B \approx \frac{1}{\sigma_v^2} \sum_{\text{grid}} (\mu_v - \mu_f)^2 \Delta A, \tag{10}$$

which is proportional to the spatially averaged squared bias in the domain. Even though our measures reduce

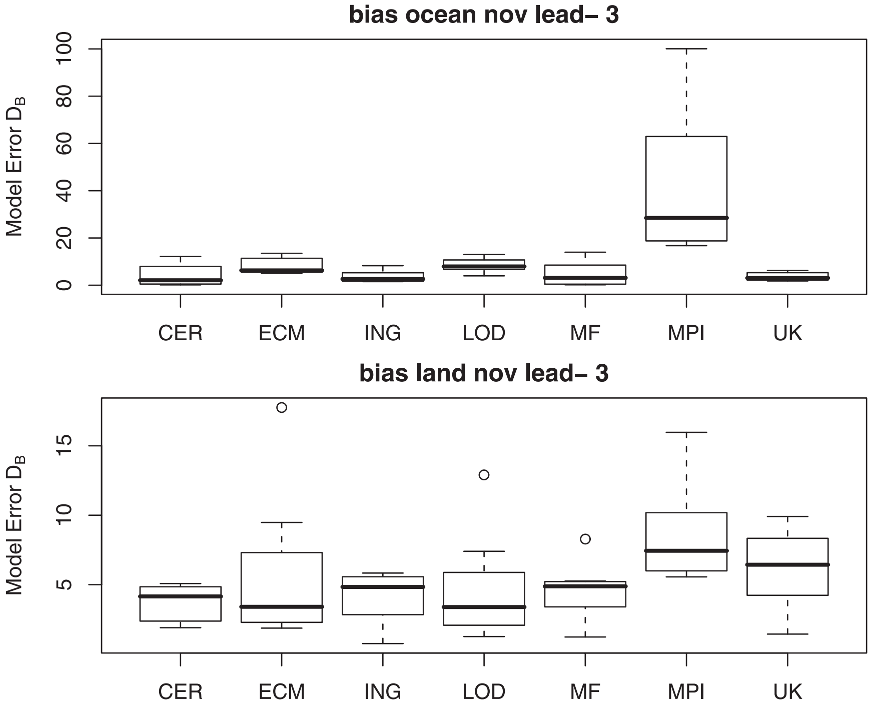

FIG. 3. Box-and-whisker plots of model error, as measured by the normalized bias D_B defined in (6) in 3-month-average T2 m over (top) ocean and (bottom) land for hindcasts initialized in November and verified in February–April (“lead-3”). Each box-and-whisker plot depicts the values of D_B in 8 different regions.

approximately to traditional measures, we argue that the proposed measures based on information theory nevertheless are preferable owing to 1) the measures can be applied to multiple physical variables with different units, 2) the measures are additive when variables are independent, 3) the measures generalize to non-Gaussian distributions, and 4) the measures generalize to multivariate distributions, at least in principle.

The fact that the bias term D_B normalizes the forecast variance by the climatological variance σ_v^2 deserves further comment. In general, spatially averaged squared biases can be highly misleading when the natural variances differ by orders of magnitude, as is the case for precipitation rate, or when variables have different units. On the other hand, this normalization becomes problematic for regions with very small variances, because then the average is dominated by a few points with very small variance estimates. An example of this is precipitation in desert regions. However, in this case the precipitation is highly non-Gaussian, so our metrics derived from the Gaussian assumption are inappropriate. A better approach would be to account for the non-Gaussian structure of precipitation directly and then to compute the corresponding skill and fidelity measures. We are currently exploring this possibility.

3. Statistical significance of skill and fidelity

The question arises as to whether the observed skill and fidelity are larger than would be expected “by random chance.” The appropriate significance test is difficult to derive since mutual information and relative entropy are averaged over a domain with unknown statistical characteristics. Accordingly, we adopt resampling techniques.

To estimate the sampling distribution of skill, under the null hypothesis that the forecast and verification are independent, we first select the model, initial month, and lead time, and then pair the actual verification of each year with a randomly selected ensemble mean forecast out of the hindcasts from that model, initial month, and lead time. This procedure yields a set of forecast–verification pairs of the same size as the original sample (i.e., 22 in this paper), from which the skill of each domain can be computed. This procedure is repeated 100 times for each model, initial month, and lead time.

To estimate the sampling distribution of the fidelity, under the null hypothesis that the forecast was drawn from the verification distribution, we first select a season, then randomly select *with replacement* the same number of verification fields in that season as years in the sample (i.e., 22 in this paper). From the resulting

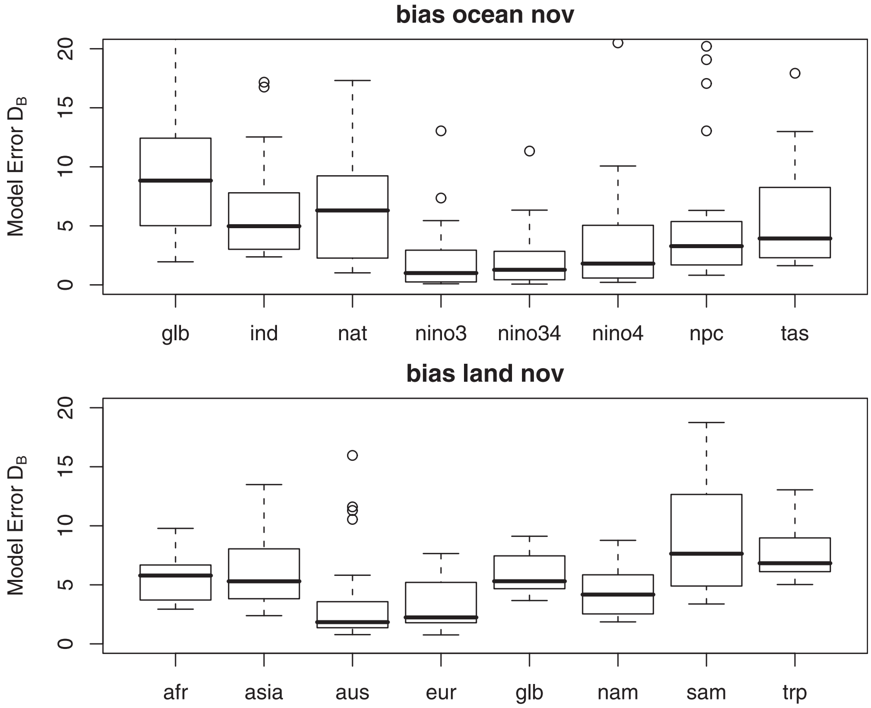

FIG. 4. As in Fig. 3, but for the average over all models. The maximum value in the top panel has been fixed at 20, which "clips" the extreme values produced by the MPI model. Each box-and-whiskers plot depicts the values of D_B for the 7 DEMETER models.

dataset, the mean and variance are calculated at each point, and then the spatially averaged relative entropy between the two resampled datasets is computed. This procedure is repeated 100 times for each season.

To determine whether the sampling distributions depend on the model, initial month, or lead time, the Kolmogorov–Smirnov (KS) test was performed on every combination of these parameters. For a fixed model and region, no dependence on initial month and lead time could be detected more often than would be theoretically expected (i.e., no more than 5% of the tests rejected the hypothesis of no difference at the 5% level). Consequently, samples from different initial months and lead times were pooled, increasing the number of samples from 100 to 1600. After pooling, the 95th percentiles were determined for each model and region.

If the observed skill exceeds the 95th percentile for that model and region, then the hypothesis that the model has no skill is rejected at the 5% significance level (the model "has skill"). A similar procedure was performed for spatially averaged relative entropy, including the bias and variance terms separately. For reference, the 5% threshold values for skill S range between 0.03 and 0.10, depending on model and region (with the global average producing the smallest threshold values). The 5% threshold values of bias D_B range between 0.08 and 0.15, and those for variance D_V range between 0.08 and 0.28.

4. Data

The variable studied in this paper is 2-m temperature (T2 m). The hindcasts used in this study are from the DEMETER project. This dataset consists of 6-month hindcasts by seven global coupled atmosphere–ocean models. The seven models are listed in Table 1. These hindcasts were initialized 4 times a year, namely on 1 February, 1 May, 1 August, and 1 November. For each initial day, an ensemble of nine hindcasts was produced by each model. Only hindcasts made for the years 1980–2001 are considered, since this period is the only one in which hindcasts from all seven models were available. Further details of the DEMETER hindcasts can be found in Palmer et al. (2004). The verification dataset is the National Centers for Environmental Prediction–Department of Energy Global Reanalysis 2 (Kanamitsu et al. 2002).

All datasets were interpolated onto a common 2.5° × 2.5° grid. Only 3-month averages are considered. To facilitate a description of the results, the skill and fidelity measures were averaged over selected regions of the globe. These regions follow those of Barnston and Smith

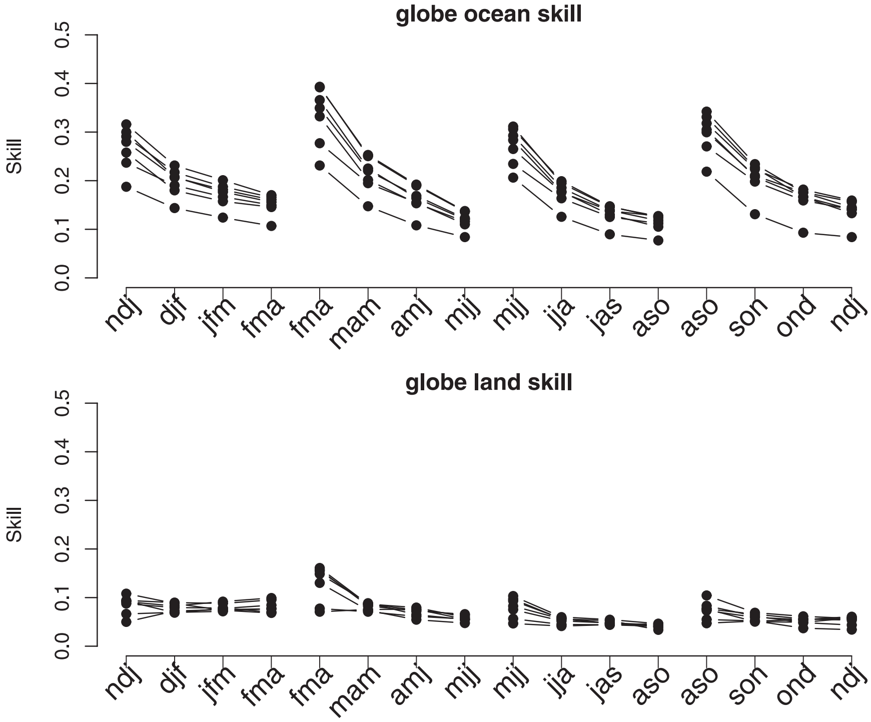

FIG. 5. The skill *S* defined in (8) in predicting 3-month-average T2 m over (top) ocean and (bottom) land as a function of the verifying 3-month average. Each curve corresponds to one of the seven hindcast models, and begins and ends in the corresponding 3-month-average period of the 6-month hindcasts. All skills displayed are statistically significant at the 5% level.

(1996) and DelSole and Shukla (2006), and are tabulated in Tables 2 and 3.

5. Results

The spatially averaged relative entropy between the climatology of the hindcasts and verifications was evaluated for all domains listed in Tables 2 and 3. The contribution due to bias was found to be much larger (often by an order of magnitude) than the contribution due to differences in variance, for all initial conditions, lead times, and domains examined. Virtually all (over 99%) biases are statistically significant. A typical result is shown in Fig. 1, which shows the range of the spatially averaged relative entropy for hindcasts verifying in the last 3 months of a 6-month hindcast. The top box-and-whiskers plot in Fig. 1 shows the term D_B, which quantifies the bias in the climatology, while the bottom box-and-whiskers plot shows the term D_V, which quantifies the differences in the variance. The bottom box-and-whisker plots can hardly be distinguished from straight lines owing to their small values relative to the bias terms.

Since the bias term D_B dominates the spatially averaged relative entropy, there is little loss of accuracy in ignoring the D_V term. Furthermore, the bias term D_B has a simple interpretation as a normalized measure of the difference in the climatological means. Finally, no obvious relation exists between skill *S* and the discrepancy in the variance D_V. Accordingly, we hereafter ignore the D_V term in fidelity and consider only the bias term D_B in fidelity.

The average bias D_B of each model over land and ocean is shown in Fig. 2 as a function of lead time and start date. The top panel in Fig. 2 reveals that the bias over the ocean generally increases monotonically with lead time, suggesting climate drift. However the bottom panel in Fig. 2 shows that the T2 m bias at the end of the integration period tends to be close to the bias at the beginning of the next hindcast. This result suggests that the hindcast bias effectively saturates over land in the first 3 months of integration.

As an example of the typical range of biases, we show in Fig. 3 the biases for hindcasts initialized in November and verified in February–April. The top panel in Fig. 3 reveals that the bias of the MPI model is much larger than that of the others over the ocean. This result holds for other initial months and lead times. The other models tend to have overlapping bias ranges between 2 and 10 units. Results for other initial months are similar, except that the magnitudes change, as indicated in Fig. 2.

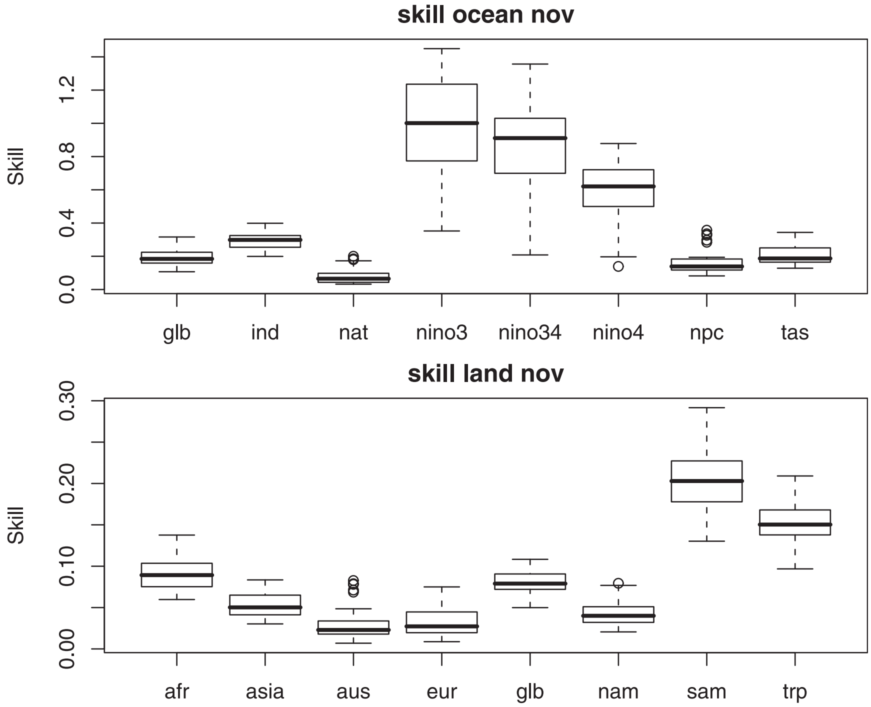

FIG. 6. Box-and-whisker plots of the skill S defined in (8) in 3-month-average T2 m over (top) ocean and (bottom) land for hindcasts initialized in November. The box-and-whisker plots depict the range of S for 7 DEMETER models and 4 verification seasons (for a total of 28). All values of S are included regardless of statistical significance.

The fact that the MPI model has larger biases relative to other models is probably due to the fact that the MPI model was initialized differently from the others. Specifically, the MPI model was initialized from coupled runs relaxed to observed sea surface temperatures, whereas most of the other models used initial conditions more closely related to the observations [e.g., the initial atmospheric and land states were taken directly from the 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis; Palmer et al. (2004)]. As such, the initial states for the MPI model are probably closer to the model's own climatology than to the observations, which may have advantages with respect to climate drift, but will contribute to a large bias by our measure. The fact that the MPI model is an ''outlier'' raises the possibility that it might have a disproportionate influence on the results. This turns out not to be true: our major conclusions remain the same whether or not the MPI is included in the pool of models.

A typical range of biases over different regions is shown in Fig. 4, for hindcasts initialized in November and verified in February–April. Figure 4 reveals that the median bias term D_B tends to be smallest in the equatorial Pacific and largest for the global average (note that the selected ocean regions do not cover the globe completely, so the global ocean bias is not a weighted average of individual regional biases). The bias in South American (''sam'') tends to be larger than that of other land areas, and this difference becomes more pronounced at other lead times.

Since the skill of the DEMETER hindcasts has been investigated extensively (Palmer et al. 2004; Hagedorn et al. 2005), we limit our discussion of skill to a few basic points. Of the predefined regions, 92% of the ocean regions and 54% of the land regions have hindcasts with statistically significant skill. As an illustrative example, the skill of predicting global average T2 m as a function of lead time is shown in Fig. 5. In this particular case, all relevant skill values are statistically significant at the 5% level. The skill over the ocean is generally larger than the skill over land. Although skill generally decreases with lead time, sometimes it increases with lead time. These increases tend to be small and short lived compared to the decreases. In addition, the magnitude of the increase tends to be within a standard deviation of the corresponding sampling variability of skill, suggesting that these increases are random fluctuations about the ''true skill'' due to sampling error.

The range of skill values over different geographic regions is shown in Fig. 6, where we can see that the skill depends significantly on geographic region, with tropical

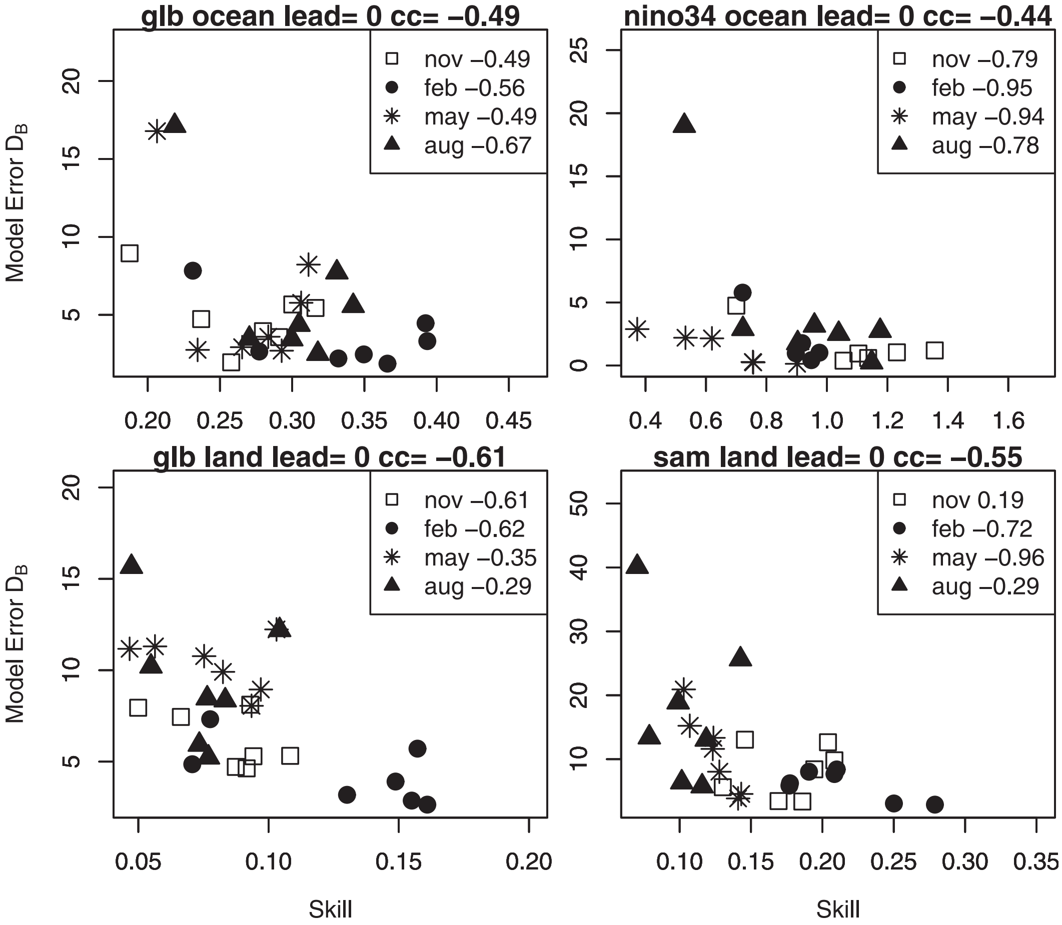

FIG. 7. Scatterplots of the skill (S) vs model error, as measured by the normalized bias D_B, of hindcast T2 m by 7 coupled atmosphere–ocean models from the DEMETER project, for 4 different initial months verifying in the first 3 months of integration. Shown are results for (top left) global ocean, (bottom left) global land, (top right) Niño-3.4 region, and (bottom right) sam (the regions are defined in Tables 2 and 3). The numbers in the legend give the correlation coefficients for the data having the indicated initial month. The correlation coefficient for all the points is indicated at the top of each panel.

regions generally exhibiting more skill than extratropical regions. The equatorial Pacific is hindcasted with the most skill and Australia with the least skill. South America also stands out as a region with consistently larger skill than other land areas.

Scatterplots of skill versus bias for selected regions and lead times are shown in Fig. 7. The correlation between skill and bias for each initial month is indicated in the legend, while the correlation over all initial months is indicated at the top of each panel. Figure 7 shows a clear tendency for the skill S to be inversely related to the bias D_B. Figure 7 also shows that the negative correlations are not caused by single outliers.

The skill–fidelity relation over South America deserves special consideration. Recall that this region was found to have the largest bias among the land regions (see Fig. 4), but also the largest skill (see Fig. 6). This positive relation seems inconsistent with our conclusion that skill and bias are inversely related. However, the results shown in Fig. 7 show that skill and bias are inversely related *within* South America. Thus, the inverse skill–bias relation holds in a relative sense rather than in an absolute sense.

The correlations between skill and bias for each region, initial day, and lead time a shown in Figs. 8 and 9. No correlation was computed if at least one model had statistically insignificant skill, where insignificant skill implies that the skill measure is dominated by sampling errors and thus will not have a relation with the bias. Figures 8 and 9 show a strong tendency for the correlations to be negative, especially on the global scale [counts of the positive (negative) correlations are 20 (49)

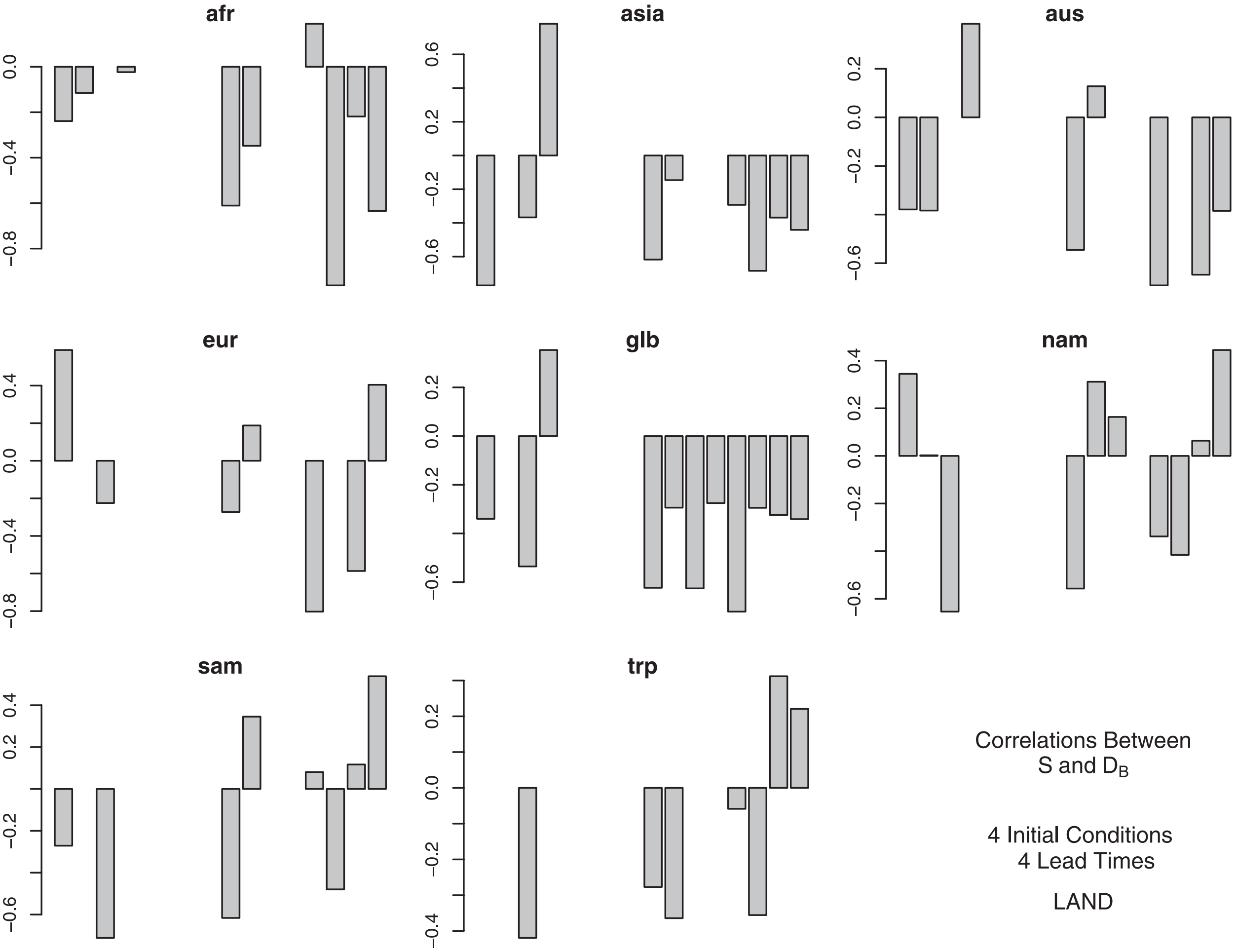

FIG. 8. The correlation coefficients between skill (S) and model error as measured by the normalized bias D_B of hindcast T2 m over land by 7 coupled atmosphere–ocean models analyzed separately for 8 regions (indicated at the top of each plot), 4 initial months, and 4 verifying 3-month periods. The results are organized such that the first 4 bars give the correlations for hindcasts starting in November, then the next 4 are for hindcasts starting in February followed by those starting in May and in August (giving a total of 16 correlations in each region). Only results in which all 7 DEMETER models have statistically significant skill for the given initial month and lead time are included.

for land and 13 (105) for ocean]. If no relation existed between skill and fidelity, then the sample correlation would be just as likely to be positive as negative.

6. Summary and discussion

This paper investigated the skill and fidelity of seasonal mean hindcasts of surface temperature by seven coupled atmosphere–ocean models. The skill and fidelity measures were based on spatially averaged mutual information and relative entropy, respectively.

Spatially averaged relative entropy was dominated by the term measuring the difference in climatological means. Over oceans, the bias term tended to increase monotonically with lead time, suggesting climate drift, whereas over land it was nearly as large at the beginning of the forecast as at the end of the previous forecast, suggesting saturation within the first 3 months. The bias of individual models ranged over similar values, except for the MPI model, which tended to have larger bias than the other models.

Of the predefined geographical regions, 92% of the ocean regions and 54% of the land regions were found to have hindcasts with statistically significant skill. Not surprisingly, the skill over the ocean tended to be larger than the skill over land. The skill varied substantially with region, with the tropics tending to have larger skill than the extratropics. The most skillful regions examined were the Niño-3 and Niño-3.4 regions, whereas the least skillful region was Australia. South America was found to have the most skill over land.

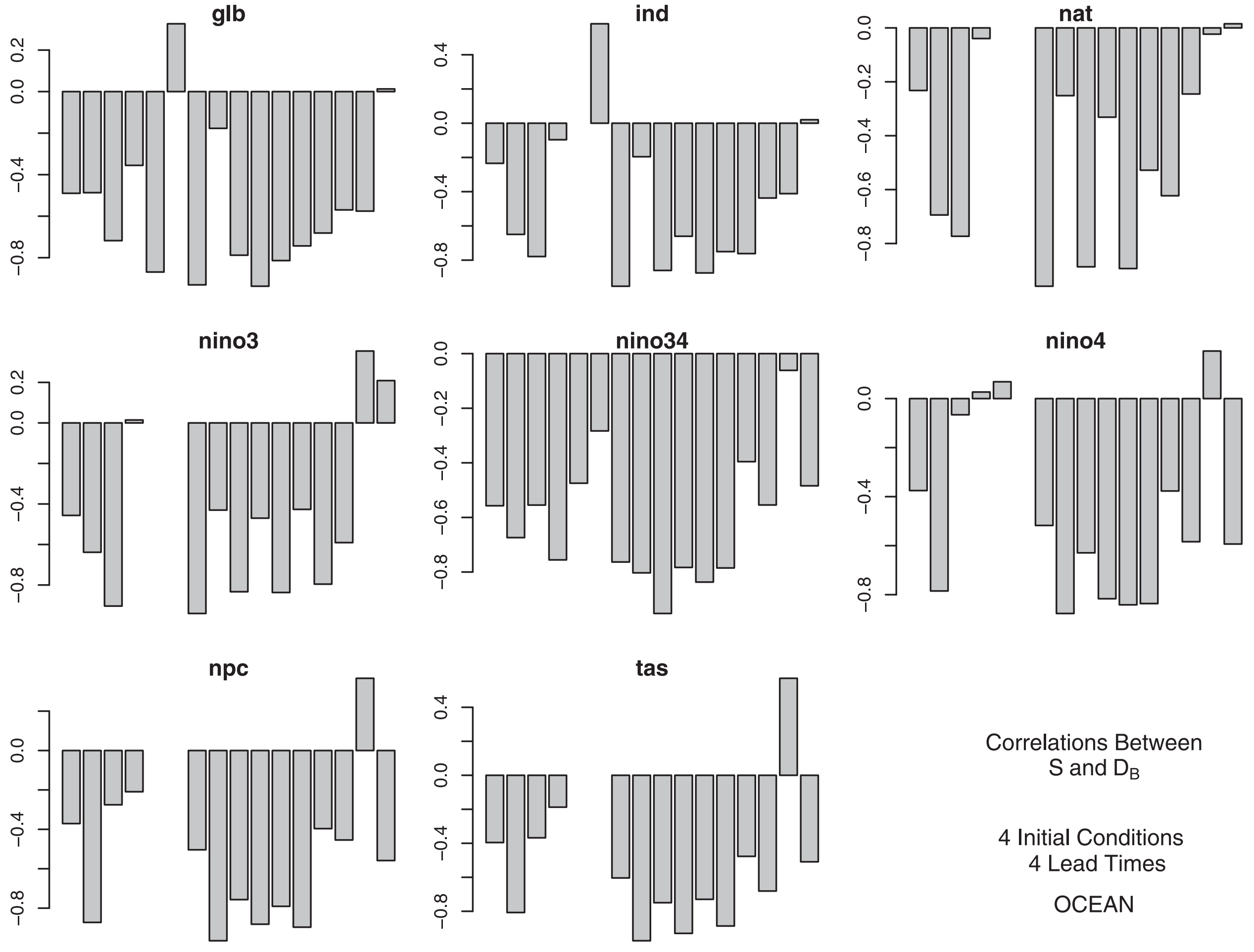

FIG. 9. As in Fig. 8, but for ocean regions.

The bias and skill of different models at the same initial month, lead time, and region tend to be negatively correlated. This relation holds whether the MPI model is included or not, and tends to be stronger for larger regions. We find that negative correlations between bias and skill outweigh positive correlations by over 2 to 1 over land regions, and over 8 to 1 over ocean regions. If no relation existed, then these ratios would be about 1 to 1. We conclude that skill and fidelity are positively related in this hindcast dataset: models that poorly simulate the observed climatological mean tend to have poor seasonal prediction skill, while models that more closely replicate the observed climatological mean tend to have better seasonal prediction skill.

We note that the bias–skill relation applies in a relative sense rather than an absolute sense. For example, over land, South America was found to have the largest bias as well as the largest skill. However, bias and skill are negatively correlated *within South America*. Also, bias and skill are not expected to be related over small domains. Indeed, a bias in one region ought to affect the skill in neighboring regions, even if the bias in the neighboring regions is not particularly large. In general, the bias–skill relation seems most physically justified over global domains rather than local domains.

Despite the clear relation between skill and bias demonstrated here, one should not jump to the conclusion that the skill of a model can be improved merely by improving its fidelity empirically. Recently, DelSole et al. (2008) and Yang et al. (2008) empirically corrected dynamical forecast models by subtracting the climatological mean tendency error at each time step. The empirically corrected model was found to have substantially less bias, but the hindcast skill was not consistently improved. Likewise, Pan et al. (2009, manuscript submitted to *Climate Dyn.*) found that applying a constant heat flux correction to a coupled model considerably reduced the bias in the tropical ocean surface temperature, but led to no improvement in the seasonal forecast skill. Thus, merely reducing the bias by empirical means is not sufficient to improve the skill.

The fact that skill and fidelity are related may be relevant to multimodel methods. For instance, Giorgi and Mearns (2002) and Tebaldi et al. (2004, 2005) propose methods of combining climate change projections in which the weight given to a projection is inversely related to the degree of bias (among other things). If no relation exists between skill and fidelity, then allowing a weight to depend on bias would be questionable. Demonstration of this relation therefore supports the approach. However, the precise relation between weights and fidelity in multimodel combinations is not obvious because individual weights cannot be interpreted as relative model ''reliability'' (Hasselmann 1979; Kharin and Zwiers 2002).

The skill–fidelity relation found in this paper lends credibility to the argument that models that better replicate the past climatology also produce more skillful forecasts. Shukla et al. (2006) found that the models with greater fidelity also tend to show stronger warming to the same change in greenhouse gas concentration. If the skill–fidelity relation found here for seasonal hindcasts holds for climate projections, then the results of Shukla et al. (2006) imply that the projected warming due to increasing greenhouse gas concentration is likely to be closer to the highest projected estimates among the current generation of climate models.

Acknowledgments. We thank two anonymous reviewers for constructive comments on an earlier version of this paper, especially for pointing out a bias in our original methodology related to missing data. This bias was circumvented in the present paper by using reanalysis data with no missing data. We also acknowledge insightful discussions with Michael Tippett. This research was supported by the National Science Foundation (ATM0332910, ATM0830062, and ATM0830068), the National Aeronautics and Space Administration (NNG04GG46G and NNX09AN50G), and the National Oceanic and Atmospheric Administration (NA04OAR4310034 and NA09OAR4310058). The views expressed herein are those of the authors and do not necessarily reflect the views of these agencies.

REFERENCES

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,** 2660–2697.

Chang, E. K. M., S. Lee, and K. L. Swanson, 2002: Storm track dynamics. *J. Climate,* **15,** 2163–2183.

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

DelSole, T., 2004a: Predictability and information theory. Part I: Measures of predictability. *J. Atmos. Sci.,* **61,** 2425–2440.

——, 2004b: Stochastic models of quasigeostrophic turbulence. *Surv. Geophys.,* **25,** 107–149.

——, 2005: Predictability and information theory. Part II: Imperfect forecasts. *J. Atmos. Sci.,* **62,** 3368–3381.

——, and J. Shukla, 2006: Specification of wintertime North America surface temperature. *J. Climate,* **19,** 2691–2716.

——, M. Zhao, P. Dirmeyer, and B. Kirtman, 2008: Empirical correction of a coupled land–atmosphere model. *Mon. Wea. Rev.,* **136,** 4063–4076.

Fedorov, A. V., S. Harper, S. Philander, B. Winter, and A. Wittenberg, 2003: How predictable is El Niño? *Bull. Amer. Meteor. Soc.,* **84,** 911–919.

Giorgi, F., and L. Mearns, 2002: Calculation of average uncertainty range and reliability of regional climate changes from AOGCM simulations via the ''reliability ensemble averaging'' (REA) method. *J. Climate,* **15,** 1141–1158.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting - I. Basic concept. *Tellus,* **57A,** 219–233, doi:10.1111/j.1600-0870.2005.00103.x.

Hasselmann, K. F., 1979: On the signal-to-noise problem in atmospheric response studies. *Meteorology of the Tropical Ocean,* D. B. Shaw, Ed., Royal Meteorological Society, 251–259.

Joe, H., 1989: Relative entropy measures of multivariate dependence. *J. Amer. Stat. Assoc.,* **84,** 157–164.

Kanamitsu, M., W. Ebisuzaki, J. Woollen, S.-K. Yang, J. Hnilo, M. Fiorino, and G. L. Potter, 2002: NCEP–DOE AMIP-II Reanalysis (R-2). *Bull. Amer. Meteor. Soc.,* **83,** 1631–1643.

Kharin, V. V., and F. W. Zwiers, 2002: Climate predictions with multimodel ensembles. *J. Climate,* **15,** 793–799.

Kirtman, B. P., and P. S. Schopf, 1998: Decadal variability in ENSO predictability and prediction. *J. Climate,* **11,** 2804–2822.

Palmer, T. N., and R. Hagedorn, Eds., 2006: *Predictability of Weather and Climate*. Cambridge University Press, 718 pp.

——, and Coauthors, 2004: Development of a European Multimodel Ensemble System for Seasonal to Inter-annual Prediction (DEMETER). *Bull. Amer. Meteor. Soc.,* **85,** 853–872.

Shukla, J., T. DelSole, M. Fennessy, J. Kinter, and D. Paolino, 2006: Climate model fidelity and projections of climate change. *Geophys. Res. Lett.,* **33,** L07702, doi:10.1029/2005GL025579.

Tebaldi, C., L. Mearns, D. Nychka, and R. Smith, 2004: Regional probabilities of precipitation change: A Bayesian analysis of multimodel simulations. *Geophys. Res. Lett.,* **31,** L24213, doi:10.1029/2004GL021276.

——, R. Smith, D. Nychka, and L. Mearns, 2005: Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles. *J. Climate,* **18,** 1524–1540.

Yang, X., T. DelSole, and H.-L. Pan, 2008: Empirical correction of the NCEP Global Forecast System. *Mon. Wea. Rev.,* **136,** 5224–5233.