A Simple Empirical Model for Decadal Climate Prediction

OLIVER KRUEGER

Institute for Coastal Research, Helmholtz-Zentrum Geesthacht, Geesthacht, Germany

JIN-SONG VON STORCH

Max Planck Institute for Meteorology, Hamburg, Germany

(Manuscript received 12 March 2010, in final form 17 October 2010)

ABSTRACT

Decadal climate prediction is a challenging aspect of climate research. It has been and will be tackled by various modeling groups. This study proposes a simple empirical forecasting system for the near-surface temperature that can be used as a benchmark for climate predictions obtained from atmosphere–ocean GCMs (AOGCMs). It is assumed that the temperature time series can be decomposed into components related to external forcing and internal variability. The considered external forcing consists of the atmospheric CO_2 concentration. Separation of the two components is achieved by using the Intergovernmental Panel on Climate Change Fourth Assessment Report (IPCC AR4) twentieth-century integrations. Temperature anomalies due to changing external forcing are described by a linear regression onto the forcing. The future evolution of the external forcing that is needed for predictions is approximated by a linear extrapolation of the forcing prior to the initial time. Temperature anomalies owing to the internal variability are described by an autoregressive model. An evaluation of hindcast experiments shows that the empirical model has a cross-validated correlation skill of 0.84 and a cross-validated rms error of 0.12 K in hindcasting global-mean temperature anomalies 10 years ahead.

1. Introduction

Decadal climate prediction is one of the grand challenges in climate research. To achieve this goal, a new generation of climate (earth system) models, in combination with advanced data assimilation techniques, has been and will be used in internationally coordinated efforts (Smith et al. 2007; Taylor et al. 2008). The results will contribute to the upcoming Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report. A summary of recent achievements in decadal climate prediction is given by Meehl et al. (2009).

When evaluating predictions produced by complex climate models, simple forecast schemes should be studied for at least two reasons. First, using simple schemes allows us to address the question of efficiency, that is, the extent to which a complex prediction model is more skillful than simple schemes. In this sense, a simple scheme serves as a benchmark for the evaluation of predictions produced by complex climate models. The idea of using simple forecast schemes as benchmarks is known and has been applied for both weather and seasonal forecasts (see, e.g., Livezey 1999; van Oldenborgh et al. 2005). Second, a simple scheme, when properly designed, becomes useful for quantifying different sources of prediction skills. This is a more difficult task in the framework of complex climate models.

It is generally believed that the skill of decadal predictions originates from the response to a changing external forcing and from low-frequency internal variability. The present paper aims at a simple benchmark model that allows a quantification of the prediction skill related to each source. The model is formulated for annual-mean near-surface air temperature anomalies. It is derived from observed and simulated temperature anomaly records over the common period from 1883 to 1999. The observed records are available from the 5° × 5° variance adjusted historical surface temperature dataset (HadCRUT3v; Brohan et al. 2006). The simulated records are available from the World Climate Research Programme's (WCRP's) Coupled Model Intercomparison Project phase 3 (CMIP3)

Corresponding author address: Oliver Krueger, Institute for Coastal Research, Helmholtz-Zentrum Geesthacht, Max-Planck-Str. 1, 21502 Geesthacht, Germany.
E-mail: oliver.krueger@hzg.de

DOI: 10.1175/2010JCLI3726.1

multimodel dataset. Both observed and simulated anomalies are defined relative to the mean of the period from 1961 to 1990. After having derived our simple model, predictions of temperature anomalies can be performed from given information about the external forcing and the temperature at the initial time, without performing any climate model simulation.

A benchmark for interannual to decadal prediction has recently been suggested by Laepple et al. (2008). Different from the method by Laepple et al., which describes the temperature responses to the external forcing through an ensemble mean of climate change simulations, our model explicitly relates the temperature response to the external forcing in terms of the model parameters.

Following the model description in sections 2 and 3, the hindcast skill of our simple model will be evaluated for temperature on both the global and local scale in section 4. A summary is given in section 5.

2. Assumptions and fitting procedures

The simple empirical model is based on three assumptions. First, the temperature time series x_t is assumed to be decomposable into two components,

$$x_t = x_t^f + x_t^i, \tag{1}$$

where x_t^f represents the response to the external forcing and x_t^i indicates the internal variability obtained without the changing external forcing. Second, the response is assumed to depend linearly on the CO_2 concentration C via

$$x_t^f = \alpha + \beta C_{t-1} + \epsilon_t^f. \tag{2}$$

The forcing–response relationship described in Eq. (2) is assumed to be valid for the coming decade, consistent with the view that the temperature response to the greenhouse gas forcing is insensitive to the detailed evolution of the forcing in the coming decades of the twenty-first century (Zwiers 2002). Term α indicates the temperature anomaly that would be obtained by reducing C to zero; β describes the response, delayed by one year, to the forcing C. From a physical point of view, a logarithmic dependence between temperature and CO_2 concentration might be expected. However, a linear approximation of a logarithmic forcing–response relationship is simple and thus serves the purpose of a benchmark. Furthermore, a response time in our model with a larger delay time, rather than a delay time of one year, is plausible because of the large inertia of climate components such as the ocean (Meehl et al. 2005). We will come back to this point in section 4.

Finally, it is assumed that the internal variability can be described by an autoregressive model of order 1 [AR(1)]:

$$x_t^i = \varphi + \phi x_{t-1}^i + \epsilon_t^i. \tag{3}$$

With these assumptions, the temperature anomalies are modeled by

$$x_t = \alpha + \beta C_{t-1} + \varphi + \phi x_{t-1}^i + \epsilon_t^f + \epsilon_t^i, \tag{4}$$

with α, β, φ, and ϕ being the model parameters and ϵ_t^f and ϵ_t^i being the error time series of the forcing–response regression and internal variability, respectively. Note that the temperature response to other greenhouse gases is largely described through the atmospheric CO_2 concentration, as they have a similar temporal evolution while the responses to other external forcings, such as solar irradiance, are covered by α and ϵ_t^f. Also note that φ represents an offset in internal variability that depends on the CO_2 forcing and the volcanic activity in the baseline period. By including φ, the hindcasts and the observed anomalies have, relative to the same baseline period, the same offset, ensuring that the rms errors (RMSEs) are independent from the baseline period. Furthermore, other components that are thought of to counteract temperature changes due to changes in the CO_2 concentration are not included explicitly. The effects of such components, for instance, sulfate loading, would hardly be statistically distinguishable from the influence of CO_2.

The model parameters are derived as follows: first, parameters α and β are estimated by fitting Eq. (2) to the past evolution of the CO_2 concentration (provided by the ENSEMBLES project available online at http://www.cnrm.meteo.fr/ensembles/public/results/results.html) into the ensemble mean of the IPCC Fourth Assessment Report (AR4) twentieth-century integrations (Meehl et al. 2007) that do not include volcanic forcing. The ensemble mean is considered to be an estimate of x_t^f, since the ensemble average leads to a cancellation of different internal variabilities simulated by various ensemble members. The ensemble mean might be a biased estimator of x_t^f or might not completely represent x_t^f. However, the ensemble mean is the best estimator we have today for x_t^f. Thus, we assume the influence of potential bias in the ensemble mean to be negligible.

After having estimated α and β, x_t^i is obtained from the observations x_t^{obs} by

$$x_t^i = x_t^{\text{obs}} - \hat{x}_t^f, \tag{5}$$

where $\hat{x}_t^f$ is calculated using the estimates of α and β and the known time evolution of C. The last parameters of the model, the AR(1)-coefficients φ and ϕ, are then estimated from the time series (5). Note that we use the nonvolcanic CMIP3 ensemble mean to estimate $\hat{x}_t^f$. Consequently, any response to volcanic activity is a component of x_t^i. This response cannot be described by an AR(1)-model.

The described procedures are repeated following a bootstrap scheme that is used later to cross-validate the hindcast skill: the observed and simulated temperature records cover the common period from 1883 to 1999. A training period, from which the coefficients α, β, φ, and ϕ are estimated, is this entire period minus a 10-yr gap. Coefficients α, β, φ, and ϕ are repeatedly derived 107 times from such training periods.

For the globally averaged temperature anomaly, the averaged values of the model parameters from the cross-validation are -3.4387 K, 0.01041 K (ppm CO_2)$^{-1}$, -0.0049 K, and 0.6099 for α, β, φ, and ϕ, respectively. Thus, if the CO_2 concentration was reduced to zero, the temperature would be about 3.4 K colder than the mean value derived from 1961 to 1990. The value of β indicates a temperature increase of about 0.1°C per 10 ppm increase of CO_2 concentration. To what extent the derived values are physically meaningful needs to be examined elsewhere.

The described fitting procedures are only optimal when the residuals of the regression and internal variability, ϵ_t^f and ϵ_t^i, are independent and normally distributed. By performing Kolmogorov–Smirnov tests, ϵ_t^f and ϵ_t^i are found to be both normally distributed. The independence, on the other hand, only holds for ϵ_t^i. Here ϵ_t^f is autocorrelated. The autocorrelation can be removed using a more sophisticated model, such as the generalized least squares (GLS) model (Cochrane and Orcutt 1949). Nevertheless, we chose to refrain from GLS models as we want to keep our prediction scheme as simple as possible. Furthermore, our results indicate that GLS models do not significantly improve the skill of the hindcast experiments (Krueger 2009).

Also note that it is possible to derive a model similar to Eq. (4) by directly regressing the full temperature to the forcing C and modeling the residual by an AR process, without involving climate model simulations at all. This approach is not considered for two reasons. First, model (4), which relies on Eq. (1) and identification of x_t^f through an ensemble of climate integrations, allows a more accurate specification of the sources of prediction skill. When regressing the temperature onto the forcing directly (without using the climate model ensemble), the temporal behavior of x that results from internal variability can modify the coefficients α and β. Similarly and to an unknown degree, coefficients φ and ϕ are connected to the temporal behavior of x that comes from the responses to the external forcing. Consequently, prediction skill resulting from the part described by coefficients α and β and that described by φ and ϕ cannot completely be attributed to the external forcing and internal variability. Second, by including additional data into the fitting procedure, overfitting is effectively reduced. As a consequence, model (4) derived from a climate model ensemble has reasonably higher skill than a model obtained from directly regressing temperature on C (not shown). Lean and Rind (2008, 2009), who performed such a direct regression, analyzed surface temperature records and found that anthropogenic forcing is mainly responsible for the atmospheric temperature increase. Since they used datasets that already distinguish between components of natural variability and of external forcing, they did not need to estimate x_t^f and x_t^i independently from each other. After estimating the future external forcing and natural variability, they used their model to predict the surface temperature for the following two decades.

3. The prediction scheme and hindcast experiments

When predicting the temperature anomaly at lead time t, x_t, using Eq. (4), one needs to know the evolution of the forcing C up to lead time $t - 1$. This study only considers predictions at lead times up to 10 yr. Within such a prediction interval, the CO_2 concentration changes little with time. Thus, the lagged value of C at lead time t can be approximated via a Taylor expansion in terms of the value of C at a time prior to the prediction, C_i:

$$C_{t-1} = C_i, \tag{6}$$

or in terms of both the value of C and that of the initial tendency at a time prior to the prediction, C_i and $\dot{C}_i$:

$$C_{t-1} = C_i + \dot{C}_i t. \tag{7}$$

The linear-varying-forcing approximation (7) is more accurate than the constant-forcing approximation (6) in describing the forcing at lead time t. Despite that, it remains to be examined whether the approximation (7) is also superior to the approximation (6) in predicting temperature changes. In the following, both possibilities will be considered.

Using Eq. (4) in accordance with neglecting the residuals ϵ_t^f and ϵ_t^i, a hindcast at lead time t is given by

$$x_t = \alpha + \beta C_{t-1} + \varphi \sum_{j=0}^{t-1} \phi^j + \phi^t x_0^i. \tag{8}$$

Here x_0^i indicates the value of the internal variability part x^i at the initial time. The subscript 0 denotes the initial time of a hindcast; x_0^i is to be obtained by subtracting the modeled response $x_0^f = \alpha + \beta C_{-1}$ from the observed temperature anomaly at the initial time. The subscript -1 indicates one time step before the initial time, and $\varphi \sum_{j=0}^{t-1} \phi^j + \phi^t x_0^i$ corresponds to the prediction of x^i at lead time t with $t > 0$.

To make a hindcast, C_{t-1} in Eq. (8) is approximated using either the constant-forcing approximation (6) or the linear-varying-forcing approximation (7). The values C_i and $\dot{C}_i$ in Eqs. (6) and (7) are chosen to be the values at one time step prior to the initial time ($i = -1$), rather than directly at the initial time. This choice ensures that for $t = 0$ the correct initial value of the internal variability x_0^i is obtained. Together with this choice, the hindcast at lead time t with $t = 1, \ldots, 10$ is given either by

$$x_t = \alpha + \beta C_{-1} + \varphi \sum_{j=0}^{t-1} \phi^j + \phi^t x_0^i \tag{9}$$

or

$$x_t = \alpha + \beta C_{-1} + \beta C_{-1} t + \varphi \sum_{j=0}^{t-1} \phi^j + \phi^t x_0^i. \tag{10}$$

The difference of C at two consecutive time steps before the initial time determines $\dot{C}_{-1}$.

The hindcast skill will be cross-validated following the bootstrap scheme described in section 2. Using Eqs. (9) and (10), hindcast experiments were carried out every year from 1883 to 1989. The length of each hindcast is 10 yr, and a total of $n = 107$ hindcasts is obtained. The 10-yr periods excluded from the estimation of parameters α, β, φ, and ϕ represent these hindcast periods. The cross-validated hindcast skill is measured by the correlation and the RMSE of the hindcasts. The correlation, calculated from hindcasts at lead time t and respective observations, measures the extent to which the hindcast is in phase with the observations. On the other hand, the RMSE quantifies the magnitude of hindcast errors. High hindcast skills should be related to large values of the correlation *and* small values of the RMSE. We also provide associated uncertainty estimates for both hindcast skills that result from the choice of hindcast periods via bootstrapping. The determined 95% confidence intervals spread from −0.008 to 0.010 K around derived values of the hindcast RMSE and from −0.228 to 0.097 around the values of the hindcast correlation.

For future predictions the averaged model parameters (averaged over 107 estimates) can be used. We found that the hindcast skill based on the averaged model parameters is almost identical to the cross-validated skill based on model parameters that are estimated from the whole data period excluding the hindcast period.

4. Results

Consider first the hindcasts for the globally averaged temperature anomaly. Table 1 displays the skill of hindcasts for lead times of 1 yr, 2 yr, and the average of years 3–5 and 6–10. Figure 1 shows the correlation skill and the RMSE as a function of lead time t. Relative to persistence (line with unfilled diamonds), the hindcasts of our model (10) (thick line) have higher correlation skill and a smaller RMSE. The correlation skill of our model stays larger than 0.79 and is about 0.87 for a lead time of 1 yr and 0.84 for a lead time of 10 yr. The RMSE increases from ~0.09 K in the first year to ~0.12 K in year 10. For comparison, persistence hindcasts have a correlation skill lower than about 0.7 and the RMSE is around 0.16 K for lead times longer than 7 yr.

Table 1. Cross-validated correlation skill and RMSE of hindcasts of globally averaged temperature anomalies obtained from prediction model (10) for lead times of 1 yr and 2 yr and the average of yr 3–5 and 6–10.

	Year			
	1	2	3–5	6–10
Correlation	0.871	0.817	0.841	0.882
RMSE (K)	0.095	0.113	0.102	0.093

A notable difference between the skill of our model and that of persistence lies in its dependence on lead time. The skill of persistence, as measured by both the correlation and the RMSE, decreases monotonically with lead time. On the contrary, the correlation skill of model (10) first decreases within the first 5 yr and then increases. This behavior is more clearly demonstrated in Fig. 2, which zooms in on high correlations and on high RMSEs around 0.12 K. Further, the RMSE of our model increases less strongly with t than those of persistence hindcasts. To understand this time dependence, the hindcast skill of the following cases is evaluated.

$$\text{Version (a):} \quad x_t = \alpha + \beta C_{-1} + \beta \dot{C}_{-1} t;$$

$$\text{Version (b):} \quad x_t = \varphi \sum_{j=0}^{t-1} \phi^j + \phi^t x_0^i.$$

The first case, version (a), concentrates on the skill originating from the response to the changing external forcing. For this case, the correlation skill increases with t (thin line in Fig. 1 left panel), and the RMSE hardly decreases with t (thin line in Fig. 1 right panel). Version (b) is different, as it focuses on the skill originating from the internal variability. The skill, as measured by both the correlation and the RMSE, is highest for the first years and diminishes with t (lines with filled dots in Fig. 1). Moreover, the skill in predicting the internal variability using an AR(1)-process is low. The correlation even becomes negative for a lead time greater than 3 yr. This result suggests that, while the internal variability is responsible for the high skill at short lead times, the source of skill in our model at long lead times is the response to

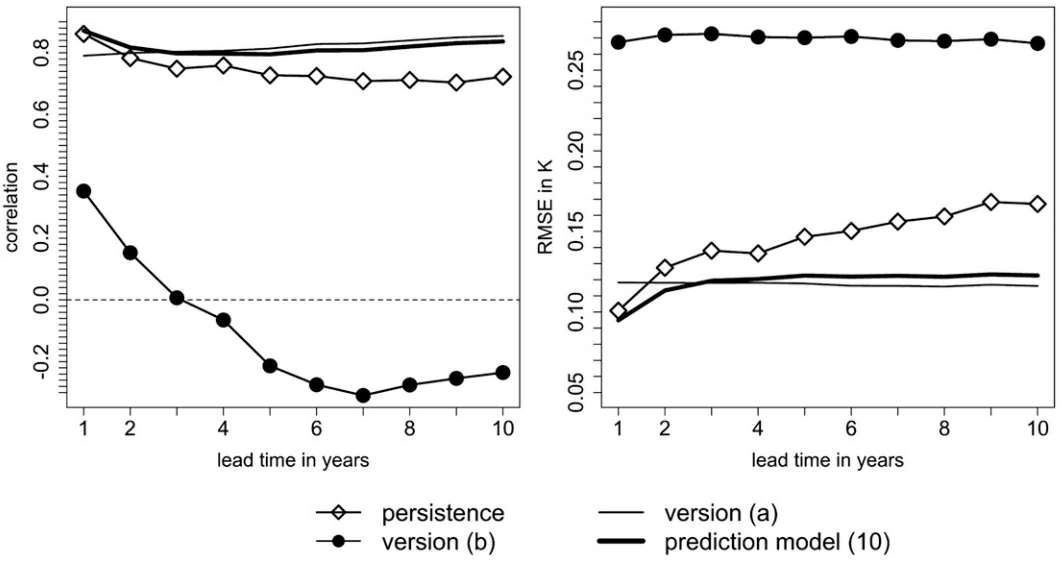

FIG. 1. Cross-validated (left) correlation skills and (right) RMSEs of globally averaged temperature anomaly predictions, as derived from the prediction model (10) (thick solid lines), version (a) that concentrates on the role of the external forcing (thin solid lines), version (b) that concentrates on the role of internal variability (lines with filled circles), and the persistence forecast (lines with diamonds).

the CO_2 forcing. If the prediction model captures the response to the changing CO_2 forcing and if the forcing changes persist, then the skill resulting from such responses will increase with lead time, as the responses amplify over time (von Storch 2008). In our model, the response to the CO_2 forcing is captured by the coefficient β, which implies a warming. It is this warming trend whose amplitude depends on the external forcing prior to the hindcast that makes the correlation skill increase with lead time and keeps the RMSE at about the same level. It should be noted that the importance of the external forcing stands out more clearly when the internal variability is only crudely described, as is the case when using an AR(1)-model.

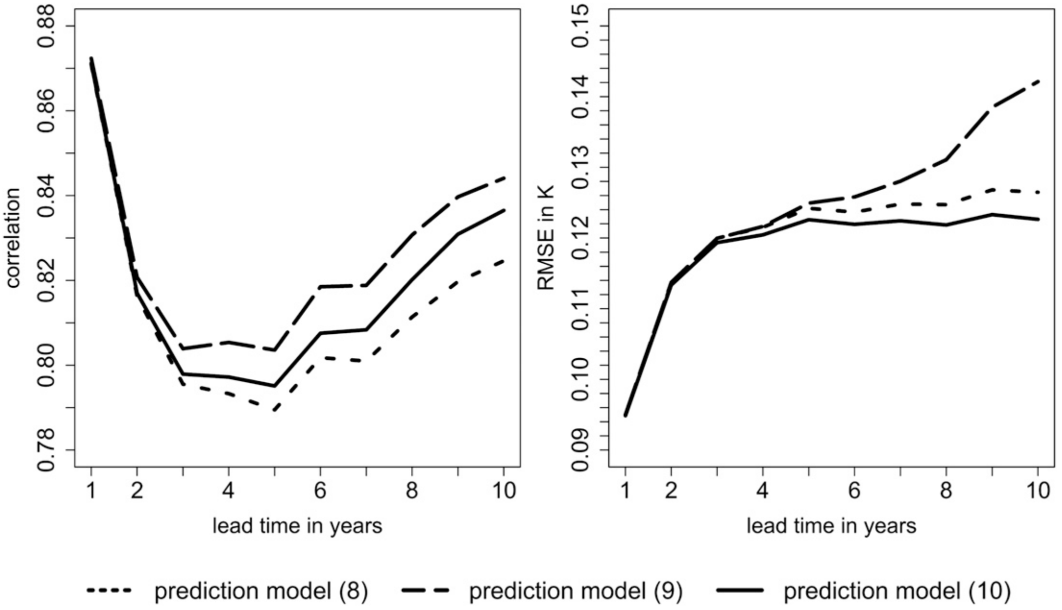

FIG. 2. Cross-validated (left) correlation skill and (right) RMSE of hindcasts of globally averaged temperature anomalies obtained from prediction models with three different forcing formulations. Model (8) does not include any forcing approximation, model (9) makes use of a constant-forcing approximation, and model (10) of a linear-varying approximation.

Additionally, the influence of the forcing formulation is examined by comparing the hindcast skill obtained by adopting actual values of the forcing without any approximation with that obtained by approximating the forcing via Eq. (6) or (7). The prediction models corresponding to different forcing formulations are given by Eqs. (8)–(10). We performed hindcasts according to these equations. The resulting skills are shown in Fig. 2. Adopting actual values of the forcing yields the lowest correlation. Making use of the constant-forcing approximation results in the highest correlation and highest RMSE, while the linear-varying-forcing approximation leads to the lowest RMSE. In terms of the RMSE, the linear-varying-forcing approximation gives the best results. The differences in the RMSE between the constant-forcing approximation, the linear-varying approximation, and adopting actual values of the CO_2 concentration are small in the first years and increase with lead time. However, the differences in the hindcast correlation of about 0.01–0.02 are small and not significant at the 5% level. These small differences are consistent with the analysis of Hawkins and Sutton (2009), that external forcing becomes important for longer lead times in decadal predictions but does not play an important role in the first decades.

Despite the smallness, the differences in the correlation appear to be systematic at large lead times. These differences can be caused by the assumption that the response to the external forcing occurs at a time lag of 1 yr, as formulated in Eqs. (2) and (8). We further analyzed the time lag of the response by modifying the forcing–response relationship described in Eq. (2): instead of a forcing–response relation with a 1-yr time lag, we considered relations with different time lags. By performing hindcasts with respective models, we found that the highest skill is obtained when assuming a time lag of about 7–10 yr between the forcing and the response, though the associated change in the skill is small (Krueger 2009). The result indicates that the systematic differences in Fig. 2 are partially caused by the less optimal assumption of a 1-yr time lag between the forcing and the response. However, we refrained from further optimizing our model since the improvement is expected to be small.

The above statement about the forcing–response relation cannot be directly applied to hindcast experiments using coupled atmosphere–ocean GCMs, as these models do not rely on any time-lag assumption and resolve the forcing–response relation automatically. Nevertheless, the above analysis provides a quantitative estimate of the skills due to different formulations of the external forcing. In particular, the use of different formulations of the external forcing does not lead to substantial changes in skill on the time scale of a decade.

For a comparison of the RMSE from our model to those obtained by Laepple et al. (2008) and Smith et al. (2007), we performed hindcasts over the same periods, namely 1930–2006 (Laepple et al. 2008) and 1982–2004 (Smith et al. 2007). Note that we used the averaged model parameters (see section 2) and the CO_2 concentration of the Special Report on Emissions Scenarios (SRES) A1B scenario (Nakicenovic et al. 2000), which has been used in the last IPCC report to obtain hindcasts for the years after 2000. For global-mean temperature, Smith et al. found an RMSE of about 0.07 K at a lead time of 1 yr and 0.13 K at a lead time of 9 yr, while we obtain 0.09 and 0.10 K, respectively, for the same period. Thus, we have lower skill at a lead time of 1 yr, but higher skill at a lead time of 9 yr. The result is consistent with the fact that much more effort has been made by Smith et al. to produce a more realistic simulation of the internal variability. It is also possible that their higher skill in the first years results from including the description of volcanic activity for the hindcast period. As we use the nonvolcanic CMIP3 ensemble mean, our model cannot explicitly predict the temperature response to volcanic activity. Laepple et al., on the other hand, obtained an RMSE of about 0.11 and 0.14 K for lead times of 1 and 9 yr, which is larger than the 0.10 and 0.12 K that we obtained with our model for the same period at lead times of 1 and 9 yr.

Consider now the skill in hindcasting local temperature anomalies. For this purpose, the prediction model (10) has been derived and fitted to the temperature anomalies in each of the 5° × 5° grid boxes. Figure 3 shows maps of the cross-validated hindcast correlation skill and hindcast RMSE of the prediction model (10) at lead times of 1 and 10 yr. In the case of the hindcast correlation, positive values indicate skill. Thus, only grid boxes with correlations significantly greater than 0 are shown, which has been determined via a one-sided *t* test at the 5% level with Fisher-Z transformed values. White areas also indicate grid boxes where the observed data are missing.

For a lead time of 1 yr, correlation skills larger than 0.6 are found mainly over the Atlantic and the northern part of the Indian Ocean (Fig. 3a). Over Northern Hemispheric lands, the skill is, for the most part, below 0.2–0.3 and becomes insignificant at some locations. These low skills likely result from the high internal variability in the respective regions, although boundary conditions are also important for regional skill (Lee et al. 2006). Predicting 10 years ahead (Fig. 3b) degrades the predictability at many grid points. Nevertheless, the skill in the northern part of the Indian Ocean and in the South Atlantic remains above 0.6.

The hindcast RMSE (Figs. 3c,d) is mostly consistent with the behavior of the correlation skill. It is small over some regions that are covered by high hindcast correlations; for example, the RMSE in the tropical Indian Ocean is about

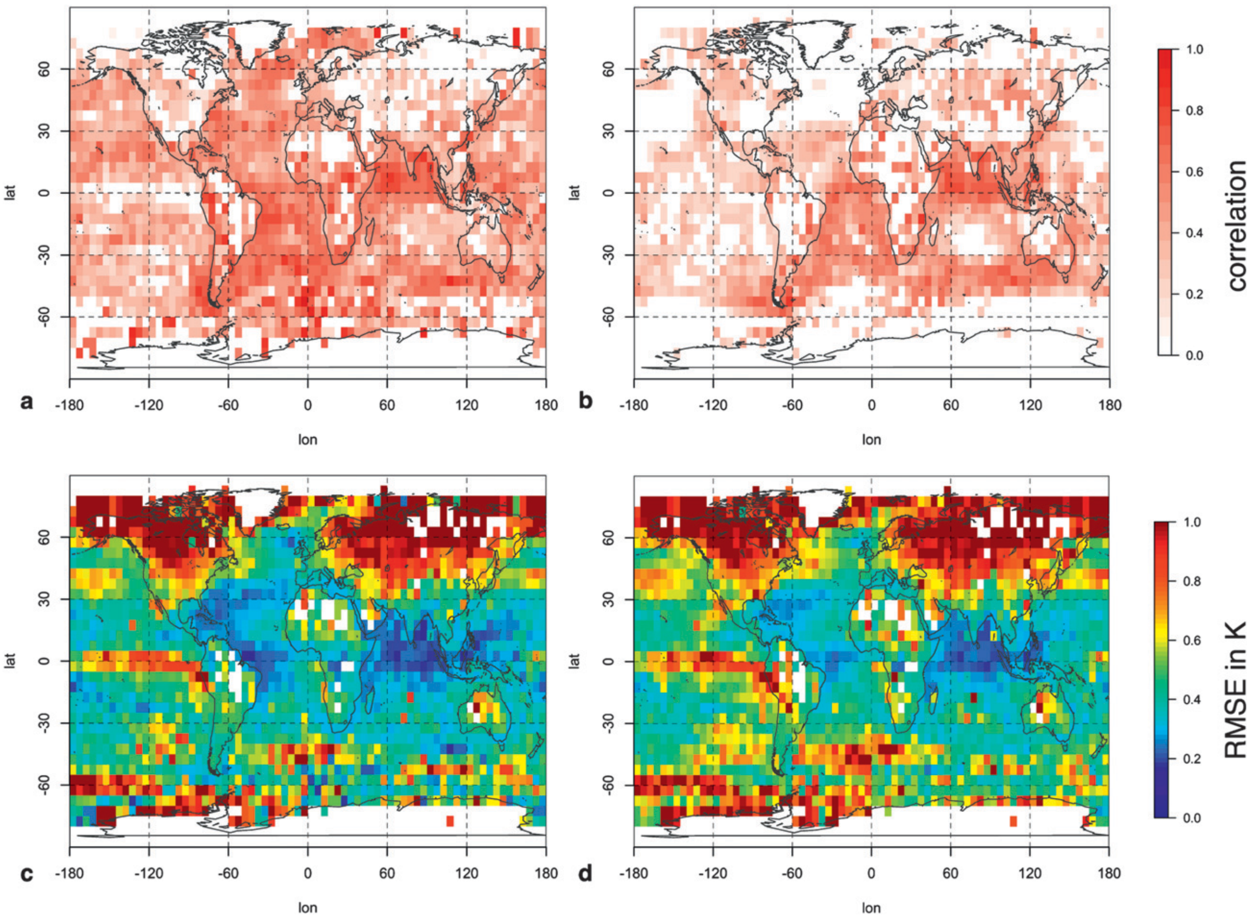

FIG. 3. Cross-validated hindcast skills for hindcasts obtained from model (10) fitted to temperatures in each of the 5° × 5° grid boxes. The hindcast correlation at (a) 1-yr and (b) 10-yr lead time, and the hindcast RMSE at (c) 1-yr and (d) 10-yr lead time. Only correlations significantly greater than 0 at the 5% level are shown.

0.2–0.4 K at a lead time of 1 (Fig. 3c) and 10 yr (Fig. 3d). The RMSE is often large in areas where the hindcast correlation is low or insignificant. The North Pacific, the western North Atlantic, and the ENSO region are such examples where the RMSE is at least 0.8 K at a lead time of 10 yr. In the North Atlantic, where the Atlantic multidecadal oscillation (AMO) (Trenberth and Shea 2006) dominates, and in the North Pacific, which is influenced by the Pacific interdecadal oscillation (Mantua et al. 1997), internal variability on decadal time scales is the main source of decadal predictability (Keenlyside et al. 2008; Latif et al. 2006). Our results are also in agreement with Hawkins and Sutton (2009), who found the internal variability to be the main source of uncertainty on the decadal time scale for regional predictions. Based on an AR(1)-process, our model cannot properly describe the internal variability regionally.

There are also regions with high hindcast correlations and high RMSE values. The correlation over Central Asia and the southern Indian Ocean is larger than 0.6 at a lead time of 10 yr, while the RMSE is larger than 1 K. Although temperature predictions over such areas are in phase with observed temperatures, the simple model is not able to forecast the magnitude of the temperature. The highest RMSE values are obtained over continental regions. The RMSE over Asia, Europe, North America, and parts of Africa is larger than 1 K at 10-yr lead time in many grid boxes, indicating that the simple model over- or underestimates the temperature in such regions. Since the internal variability of continental climate is stronger than that of oceanic climate, the model lacks the ability to describe areas influenced by continental climate. Note that the low skill in the North Atlantic can also be, at least partly, attributed to some of the CMIP3 models that reveal remarkable temperature biases there (van Oldenborgh et al. 2009). The high skills in the tropical Indian Ocean are also identified in other studies (Pohlmann et al. 2009). They mainly result from the correct simulation of the upward temperature trend in the twentieth century.

5. Summary

This study aims at a simple empirical model for decadal forecasts of surface temperature anomalies, which can be used to benchmark numerical climate models. The skill of the simple model originates from the temperature response to changes in the CO_2 concentration and from the internal variability. On the global scale, the response to CO_2 represents the main source of prediction skill and even results in a correlation skill increase with lead time to ~0.84 at year 10. Furthermore, the temperature response to CO_2 is able to keep the RMSE nearly unchanged with increasing lead time. At year 10, the RMSE is about 0.12 K. The role of the externally forced response may be somewhat overemphasized through the simple description of internal variability via an AR(1)-process. On local scales, the response to CO_2 becomes less important relative to the internal variability. A proper description of internal variability is crucial for skillful predictions. As our simple model uses a rudimentary description for the local internal variability, the skill, as measured by the hindcast correlation and the RMSE, is generally lower than that obtained for the global-mean temperature anomalies. This is particularly true over Northern Hemisphere lands and in the tropical Pacific and becomes more evident at long lead times.

Acknowledgments. We thank Daniela Matei and Wolfgang Mueller for helpful suggestions and constructive discussion. We are grateful to the anonymous reviewer and Thomas Laepple for providing very thoughtful comments.

We acknowledge the modeling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI), and the WCRP's Working Group on Coupled Modelling (WGCM) for their roles in making available the WCRP CMIP3 multimodel dataset. Support of this dataset is provided by the Office of Science, U.S. Department of Energy.

REFERENCES

Brohan, P., J. Kennedy, I. Harris, S. Tett, and P. Jones, 2006: Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *J. Geophys. Res.,* **111,** D12106, doi:10.1029/2005JD006548.

Cochrane, D., and G. Orcutt, 1949: Application of least squares regression to relationships containing auto-correlated error terms. *J. Amer. Stat. Assoc.,* **44,** 32–61.

Hawkins, E., and R. Sutton, 2009: The potential to narrow uncertainty in regional climate predictions. *Bull. Amer. Meteor. Soc.,* **90,** 1095–1107.

Keenlyside, N., M. Latif, J. Jungclaus, L. Kornblueh, and E. Roeckner, 2008: Advancing decadal-scale climate prediction in the North Atlantic sector. *Nature,* **453** (7191), 84–88.

Krueger, O., 2009: A simple empirical model for decadal climate prediction. Diplomarbeit, Meteorological Institute, University of Hamburg, 99 pp. [Available from Meteorological Institute, University of Hamburg, Bundesstr. 55, 20146 Hamburg, Germany.]

Laepple, T., S. Jewson, and K. Coughlin, 2008: Interannual temperature predictions using the CMIP3 multi-model ensemble mean. *Geophys. Res. Lett.,* **35,** L10701, doi:10.1029/2008GL033576.

Latif, M., M. Collins, H. Pohlmann, and N. Keenlyside, 2006: A review of predictability studies of Atlantic sector climate on decadal time scales. *J. Climate,* **19,** 5971–5987.

Lean, J. L., and D. H. Rind, 2008: How natural and anthropogenic influences alter global and regional surface temperatures: 1889 to 2006. *Geophys. Res. Lett.,* **35,** L18701, doi:10.1029/2008GL034864.

——, and ——, 2009: How will earth's surface temperature change in future decades. *Geophys. Res. Lett.,* **36,** L15708, doi:10.1029/2009GL038932.

Lee, T., F. Zwiers, X. Zhang, and M. Tsao, 2006: Evidence of decadal climate prediction skill resulting from changes in anthropogenic forcing. *J. Climate,* **19,** 5305–5318.

Livezey, R., 1999: The evaluation of forecasts. *Analysis of Climate Variability: Applications of Statistical Techniques,* H. von Storch and A. Navarra, Eds., Springer Verlag, 177–196.

Mantua, N., S. Hare, Y. Zhang, J. Wallace, and R. Francis, 1997: A Pacific interdecadal climate oscillation with impacts on salmon production. *Bull. Amer. Meteor. Soc.,* **78,** 1069–1079.

Meehl, G., W. Washington, W. Collins, J. Arblaster, A. Hu, L. Buja, W. Strand, and H. Teng, 2005: How much more global warming and sea level rise? *Science*, **307** (5716), 1769–1772.

——, C. Covey, T. Delworth, M. Latif, B. McAvaney, J. Mitchell, R. Stouffer, and K. Taylor, 2007: The WCRP CMIP3 multimodel dataset. *Bull. Amer. Meteor. Soc.,* **88,** 1383–1394.

——, and Coauthors, 2009: Decadal prediction: Can it be skillful? *Bull. Amer. Meteor. Soc.,* **90,** 1467–1485.

Nakicenovic, N., and Coauthors, 2000: *Special Report on Emissions Scenarios*. Cambridge University Press, 612 pp.

Pohlmann, H., J. Jungclaus, A. Köhl, D. Stammer, and J. Marotzke, 2009: Initializing decadal climate predictions with the GECCO oceanic synthesis: Effects on the North Atlantic. *J. Climate,* **22,** 3926–3938.

Smith, D., S. Cusack, A. Colman, C. Folland, G. Harris, and J. Murphy, 2007: Improved surface temperature prediction for the coming decade from a global climate model. *Science,* **317** (5839), 796–799.

Taylor, K., R. Stouffer, and G. Meehl, cited 2008: A summary of the CMIP5 experiment design. [Available online at http://www.clivar.org/organization/wgcm/references/Taylor_CMIP5.pdf.]

Trenberth, K. E., and D. J. Shea, 2006: Atlantic hurricanes and natural variability in 2005. *Geophys. Res. Lett.,* **33,** L12704, doi:10.1029/2006GL026894.

van Oldenborgh, G., M. Balmaseda, L. Ferranti, T. Stockdale, and D. Anderson, 2005: Evaluation of atmospheric fields from the ECMWF seasonal forecasts over a 15-yr period. *J. Climate,* **18,** 3250–3269.

——, S. Drijfhout, A. van Ulden, R. Haarsma, A. Sterl, C. Severijns, W. Hazeleger, and H. Dijkstra, 2009: Western Europe is warming much faster than expected. *Climate Past,* **5,** 1–12.

von Storch, J., 2008: Toward climate prediction: Interannual potential predictability due to an increase in CO_2 concentration as diagnosed from an ensemble of AOGCM integrations. *J. Climate,* **21,** 4607–4628.

Zwiers, F., 2002: The 20-year forecast. *Nature,* **416** (6882), 690–691.