NOTES AND CORRESPONDENCE

Climate Predictions with Multimodel Ensembles

VIATCHESLAV V. KHARIN AND FRANCIS W. ZWIERS

Canadian Centre for Climate Modelling and Analysis, Meteorological Service of Canada, Victoria, British Columbia, Canada

2 January 2001 and 2 November 2001

ABSTRACT

Several methods of combining individual forecasts from a group of climate models to produce an ensemble forecast are considered. These methods are applied to an ensemble of 500-hPa geopotential height forecasts derived from the Atmospheric Model Intercomparison Project (AMIP) integrations performed by 10 different modeling groups. Forecasts are verified against reanalyses from the European Centre for Medium-Range Weather Forecasts. Forecast skill is measured by means of error variance. In the Tropics, the simple ensemble mean produces the most skillful forecasts. In the extratropics, the regression-improved ensemble mean performs best. The ''superensemble'' forecast that is obtained by optimally weighting the individual ensemble members does not perform as well as either the simple ensemble mean or the regression-improved ensemble mean. The sample size evidently is too small to estimate reliably the relatively large number of optimal weights required for the superensemble approach.

1. Introduction

The skill of climate predictions is limited by internal atmospheric variability that is largely unpredictable beyond the deterministic predictability limit of about two weeks (e.g., Lorenz 1982). A standard approach that is used to reduce climate noise in model predictions is to average an ensemble of forecasts initiated from different initial conditions (e.g., Kharin et al. 2001, and references therein).

With the availability of climate predictions produced by several dynamical models, multimodel ensemble forecasting has drawn some attention recently. In particular, two recent studies (Krishnamurti et al. 1999, 2000) discuss a superensemble approach, in which a multimodel linear regression technique is used to improve deterministic forecasts locally.

Other studies have also used linear methods to produce better predictions by combining several independent forecasts. For example, Danard et al. (1968) and Thompson (1977) showed that the mean square error of forecasts constructed from a particular linear combination of two independent predictions is less than that of the individual predictions. Fraedrich and Smith (1989) discussed a linear regression method to combine two statistical forecast schemes and applied this method to long-range forecasting of the monthly mean tropical Pacific sea surface temperatures. In a recent paper, Derome et al. (2001) discussed a linear multivariate method for blending climate forecasts produced by two dynamical models.

Ensembles of predictions can be used for deterministic or probability forecasting. Doblas-Reyes et al. (2000) investigate the performance of multimodel climate predictions produced by three general circulation models and find that the multimodel approach offers a systematic improvement when using the ensemble to produce probabilistic forecasts. They find that the multimodel ensemble improves skill only marginally when verifying the ensemble mean, however. On the other hand, Krishnamurti et al. (2000) find an apparent systematic improvement in mean square error for a multimodel forecast over that of the individual model forecasts.

With the increasing popularity of multimodel ensemble forecasting, it is important to understand its virtues and limitations. In this paper, we consider deterministic forecasting only. The goal is to evaluate the performance of several forecasts constructed from a multimodel ensemble of forecasts, as measured by forecast error variance. We use monthly 500-hPa geopotential height data from several model simulations of the 1979–88 period produced for the Atmospheric Model Intercomparison Project (AMIP; Gates 1992). These integrations can be viewed as two-tier climate forecasts (Bengtsson et al. 1993) in which the lower boundary condition predic-

Corresponding author address: Francis W. Zwiers, Canadian Centre for Climate Modelling and Analysis, University of Victoria, P.O. Box 1700, Stn CSC, Victoria, BC V8W 2Y2, Canada.
E-mail: francis.zwiers@ec.gc.ca

tions are error free and in which initial conditions at the beginning of each forecast period are not specified.

The language that is used to describe ensemble-forecasting schemes can be confusing because similar words are used in the literature to describe schemes with several different configurations. The ensemble of forecasts may have been produced with a single dynamical model, or it may consist of a number of forecasts, each produced with a different model. Krishnamurti et al. call the latter a *superensemble.* However, this nomenclature appropriates a name that more aptly describes ensemble-forecasting schemes in which each of several models is used to produce an ensemble of forecasts. In this paper, we use the term ensemble to denote a collection of forecasts, each of which is produced with a different model.

The plan of the paper is the following. First we describe a general linear regression method that will serve as the basis for constructing several improved forecasts. Results are presented as a function of the size of the multimodel ensemble, and we conclude with a summary.

2. Multimodel linear regression

Let $\{X_i(t)\}$, $i = 1, \ldots, M$ denote an ensemble of forecasts produced by M models at a fixed location. An arbitrary linear combination of these forecasts is given by

$$F(t) = a_0 + \sum_{i=1}^{M} a_i X_i(t). \tag{1}$$

The $M + 1$ coefficients a_i, $i = 0, \ldots, M$ may be subject to some constraints, as is discussed below. Within the bounds of those constrains, coefficients may be chosen to minimize the mean square error,

$$\text{MSE} \equiv \overline{(F - Y)^2} = \text{Var}(F - Y) + B^2, \tag{2}$$

of the multimodel regression forecast F. Here Y represents the verifying observations, the overbar denotes time averaging, and $B = \overline{F} - \overline{Y}$ is the mean bias. Depending on the constraints, several versions of this general regression approach are possible, as will be outlined below. These versions will be identified variously by symbols C, U, or R depending upon the specific constraints and the extent to which the coefficients in Eq. (1) are adjustable.

With no constraints on the coefficients, Eqs. (1)–(2) describe the standard multimodel linear regression technique that is described by Krishnamurti et al. (2000). We denote this forecast by R_{all}, where the subscript all is used to indicate that all coefficients are adjustable and thus that all ensemble members may be weighted differently by the regression. The regression coefficients a_i are found by solving $M + 1$ linear equations

$$\sum_{i=1}^{M} a_i \,\text{Cov}(X_i, X_k) = \text{Cov}(X_k, Y), \qquad k = 1, \ldots, M,$$

$$a_0 + \sum_{i=1}^{M} a_i \overline{X}_i = \overline{Y},$$

where $\text{Cov}(X_i, X_k)$ and $\text{Cov}(X_k, Y)$ are the covariances and cross covariances between the model forecasts and the model forecasts and verifying observations Y, respectively.

Some studies (e.g., Krishnamurti et al. 2000) have interpreted the regression coefficients as indicators of the relative model ''reliability.'' However, this interpretation is not generally correct. To appreciate the difficulty of doing so, consider a simple example in which there are two model forecasts X_1 and X_2, one that systematically oversimulates and the other that undersimulates the amplitude of the predictable signal β in the observations Y. In particular, suppose

$$Y = \beta + \epsilon, \qquad X_1 = 0.5\beta + \epsilon_1, \quad \text{and}$$

$$X_2 = 1.5\beta + \epsilon_2.$$

Here ϵ, ϵ_1, and ϵ_2 represent the unpredictable internal variability that is present in the observations and in the corresponding model forecasts. For convenience, we assume that these are independent random variables with 0 time mean and the same variance σ_ϵ^2. The forecasts X_1 and X_2 have the same mean square error $\text{MSE} = 0.5^2\sigma_\beta^2 + 2\sigma_\epsilon^2$, where σ_β^2 is the variance of the predictable signal β. The regression coefficients are given by $a_1 = \sigma_\beta^2/(5\sigma_\beta^2 + 2\sigma_\epsilon^2)$ and $a_2 = 3a_1$. Thus, in this example and in general, equally reliable model forecasts may not necessarily be weighted equally when combined optimally. Hasselmann (1979) discusses this problem in greater generality and points out that the optimal coefficients weight the model forecasts so as to maximize the signal-to-noise ratio. Similar ideas apply to the optimal climate change signal detection problem (e.g., Hasselmann 1997; Zwiers 1999).

An attractive property of the multimodel linear regression forecast R_{all} is that it is superior to all other linear combinations of the individual forecasts when very large ''training'' datasets are available for estimating the regression coefficients. In practice, however, the coefficients depend on estimates of covariances $\text{Cov}(X_i, X_k)$ and $\text{Cov}(X_k, Y)$ that must be obtained from relatively short datasets. In these circumstances, estimating too many regression coefficients leads to overfitting that causes a degradation in skill (Davis 1976).

Overfitting results in optimistically biased estimates of forecast skill when skill is assessed with the same data that are used to train the regression model. This bias is avoided by assessing skill in verification data that are independent of the training data. Cross validation (Michaelson 1987) is used here, as described subsequently, to increase the amount of verification data.

Overfitting can be reduced by imposing constraints on the coefficients in the regression equation and thereby effectively reducing the number of free parameters that must be estimated from the data. One approach is to set some coefficients to 0, thereby reducing the number of models providing forecasts that enter into Eq.

(1). However, because the forecasts contain chaotic ''noise,'' reducing the multimodel ensemble size will likely increase the noise variance in the linear combination and thus may cause skill degradation. Other kinds of constraints that effectively reduce the number of coefficients to be estimated may also be imposed. For example, one might assume that the regression coefficients are constant throughout the year. This constraint reduces the number of coefficients that must be estimated per unit of data. It also implies a statistical model that does not take the cyclical behavior of climate data into account, however. Nonetheless, the improved forecast may benefit from relatively more robust coefficient estimates when only short training datasets are available.

Several forecast variants derived from Eq. (1) by imposing various constraints on the coefficients are examined below. Some, such as the climatological forecast, are trivial. Others take advantage of the multimodel ensemble. All are unbiased. The mean bias B in Eq. (2) often contributes substantially to MSE. However, the mean bias can be removed whenever historical model simulations and the corresponding observational records are available, as is the prerequisite for any statistical forecast improvement scheme.

We consider the following unbiased forecast variants.

- The trivial *climatological forecast* C, or zero-anomaly forecast, is obtained by setting the coefficients $a_1, \ldots, a_M$ to 0. The only coefficient to be estimated is the intercept a_0, which is given by $a_0 = \overline{Y}$. This forecast is the baseline upon which all other prospective schemes must improve.
- The *bias-removed individual forecast* U_i, produced by the ith model, is obtained from Eq. (1) by using $a_i = 1$ and setting all other coefficients to 0, except for the intercept $a_0 = -B_i$, where $B_i = \overline{X}_i - \overline{Y}$ is the bias of the ith model.
- The *regression-improved individual forecast* R_i is obtained by linearly regressing the ith forecast against the observations Y. This is equivalent to setting all coefficients in Eq. (1) to 0, except for a_i and a_0. These two coefficients are estimated as $a_i = \mathrm{Cov}(Y, X_i)/\mathrm{Var}(X_i)$ and $a_0 = \overline{Y} - a_i\overline{X}_i$. The regression coefficient a_i rescales the forecast to correct systematic errors in simulating the atmospheric response to the lower boundary conditions and to minimize the effect of climate noise in the model forecast on error variance. The intercept a_0 removes the bias from the rescaled forecast.
- The *bias-removed multimodel ensemble mean forecast* U_{EM} is obtained by using $a_i = 1/M$, $i = 1, \ldots, M$. The intercept, which is set to $a_0 = \overline{Y} - 1/M \sum_{i=1}^{M} \overline{X}_i$, removes the multimodel mean bias. Skill improvements result from the bias removal and from the reduction of the climate noise by ensemble averaging.
- The *regression-improved multimodel ensemble mean forecast* R_{EM} is obtained by linearly regressing the multimodel ensemble mean against the observations. This is equivalent to constraining the coefficients $a_i = a$, $i = 1, \ldots, M$, to be equal. The two unknown coefficients a and a_0 are estimated as $a = \mathrm{Cov}(Y, U_{\mathrm{EM}})/[M\ \mathrm{Var}(U_{\mathrm{EM}})]$ and $a_0 = \overline{Y} - a \sum_{i=1}^{M} \overline{X}_i$. Skill improvement, relative to that of U_{EM}, is achieved by rescaling the ensemble mean forecast to reduce the effect of the climate noise in the ensemble mean and to correct systematic error in the boundary forced response that is common to all models.
- The *regression-improved multimodel forecast* R_{all} is obtained by fitting Eq. (1) to the verifying observations with no constraints on the regression coefficients. Skill improvement in this forecast results from bias removal, climate noise reduction due to ensemble averaging, and the rescaling of the individual forecasts.
- The *regression-improved multimodel subset forecast* R_{subset} is obtained by linearly regressing a subset of models against the observations. To select the best subset of models we use the Akaike information criterion (AIC; Akaike 1974) given by $\mathrm{AIC} = N \log(\overline{Y - R_{\mathrm{subset},K}})^2 + 2(K + 1)$, where N is the sample size, K is the number of models in a subset, and $R_{\mathrm{subset},K}$ is the regression forecast produced by combining the K models. The general idea is to penalize the mean square error by the number of the coefficients that are estimated. Given M models, the total number of all possible subset combinations is 2^M, including a zero-size subset, that is, the climatological forecast. We use the models for which the AIC is the smallest in the training period for constructing a linearly improved forecast in the independent verification period.

These seven forecast variants are summarized in Table 1 for easy reference.

Another potential forecast variant is the ensemble average of regression-improved individual forecasts $(1/M) \sum_{i=1}^{M} R_i$. However, it is not clear that such a forecast is optimal. For example, this forecast variant is asymptotically suboptimal for an ensemble of ''perfect-model'' forecasts

$$Y = \beta + \epsilon, \quad \text{and} \quad X_i = \beta + \epsilon_i, \qquad i = 1, \ldots, M.$$

In this case, the regression-improved individual forecast is given by $R_i = \rho X_i$ where ρ is the correlation between X_i and Y, $\rho^2 = \sigma_\beta^2/(\sigma_\beta^2 + \sigma_\epsilon^2)$. Thus the ensemble average of the regression-improved individual forecasts asymptotically approximates $\rho\beta$ with increasing ensemble size M. If the signal-to-noise ratio is small, as is normally the case for monthly and seasonal individual forecasts in midlatitudes, this forecast variant would substantially underestimate the amplitude of the optimal forecast β. As might be anticipated, the ensemble average of regression-improved individual forecasts did not perform as well as the regression-improved ensemble mean R_{EM}

TABLE 1. The constraints and the number of adjustable coefficients in each forecast variant. The simplest configurations have 12 adjustable parameters, that is, a different value of a_0 for each month of the year. Here M is the ensemble size. Details are given in the text.

Symbol	Description	Constraints on a_0, $i = 1, \ldots, M$	No. Parameters adjusted
C	Climatological	$a_i =$, $i = 1, \ldots, M$	12
U_i	Bias-removed individual	$a_i = 1$, $a_j = 0$, $j \neq i$	12
R_i	Regression-improved individual	$a_j = 0$, $j \neq i$	13
U_{EM}	Bias-removed ensemble mean	$a_i = 1/M$, $i = 1, \ldots, M$	12
R_{EM}	Regression-improved ensemble mean	$a_i = a$, $i = 1, \ldots, M$	13
R_{all}	Regression-improved multi-model	None	$12 + M$
R_{subset}	Regression-improved multimodel subset	$a_i = 0$, $i \notin$ subset of size K (objectively determined)	$12 + K$, $0 \leq K \leq M$

for an ensemble of AMIP simulations in our preliminary tests. We therefore dropped it from consideration.

The performance of the various forecast variants was evaluated with the error variance skill score:

$$S = 1 - \frac{\text{Var}(F - Y)}{\text{Var}(Y)}.$$

This score is a natural measure of performance of an unbiased forecast F. It is proportional to error variance $\text{Var}(F - Y)$ (and mean square error) rescaled in such a way that $S = 1$ for the perfect forecast and $S = 0$ is for the climatological forecast C. Thus, positive (negative) values of S indicate that the forecast is more (less) skillful than the climatological forecast.

3. Results

We used the 10 10-yr AMIP simulations listed in Table 2. Eight of the models are the same as in Krishnamurti et al. (2000). Documentation of the models is found in Phillips (1994). For validation purposes we use 500-hPa geopotential heights Z_{500} from the European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis (Gibson et al. 1997). All data are interpolated onto a 96×48 Gaussian grid. The various versions of Eq. (1) are fitted at each point in the grid.

The skill scores are estimated with a cross-validation procedure in which one year of data is repeatedly withheld from the dataset and the regression coefficients are estimated from the retained nine years. The best subset of models in the forecast R_{subset} is also selected independently of the withheld year. We use monthly data for the whole year and assume that the regression coefficients, except for the intercept a_0, are independent of the annual cycle. Allowing the intercept a_0 to depend on the calendar month in the linear regression technique is equivalent to taking into account the climatological annual cycle that may be present in the data. If the coefficient a_0 were constant, the forecast skill would also include contributions associated with the annual cycle, which should not be considered as very useful. In principle, the dependence on the annual cycle could be built into the linear regression technique for all regression coefficients. However, this approach is impractical for the relatively short AMIP integrations. Indeed, given nine years of verifying data in the training period and an ensemble of eight models or more, the total number of fitted coefficients in the multimodel regression technique would be equal to, or greater than, the number of data points. Thus, it is imperative to reduce the number of coefficients, for example, by assuming that the regression coefficients for each model are independent of the calendar month.

In the following we show skill scores averaged over two regions, the Tropics (30°S–30°N) and the Pacific–North America sector (PNA; 20°–80°N, 180°–45°W). Boundary forcing accounts for more than one-half of the total variability on seasonal timescales in the Tropics (Zwiers 1996; Rowell and Zwiers 1999; Zwiers et al.

TABLE 2. Selected AMIP models.

Acronym	AMIP group
BMRC	Bureau of Meterology Research Centre, Melbourne, Australia
CCC	Canadian Centra for Climate Modelling and Analysis, Victoria, British Columbia, Canada
CSIRO	Commonwealth Scientific and Industrial Research Organization, Mordialloc, Australia
ECMWF	European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom
GFDL	Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey
LMD	Laboratoire de Météorologie Dynamique, Paris, France
MPI	Max-Planck-Institut für Meteorologie, Hamburg, Germany
NMC	National Centers for Environmental Prediction, Suitland, Maryland
UGAMP	The U.K. Universitites' Global Atmospheric Modelling Programme, Reading, United Kingdom
UKMO	Met Office, Bracknell, Berkshire, United Kingdom

2000). The extratropical atmosphere is much less predictable and has a much lower signal-to-noise ratio.

Figure 1 shows cross-validated skill scores S for Z_{500} for the individual AMIP models. In the Tropics, the performance of the individual unbiased forecasts U_i varies substantially from one model to another. Skill score values range from about -0.4 to 0.1, with an average value of about 0. Skill scores for the regression-improved forecasts R_i, which are substantially better, are all positive, with an average value of about 0.2. Skill is lower in the PNA sector (Fig. 1, lower panel). The individual regression-improved forecasts, although substantially better than the raw bias-corrected forecasts, still have small negative skill scores, indicating that they are not able to outperform the climatological forecast.

Figure 2 shows cross-validated skill scores for the biased-removed ensemble mean (U_{EM}) forecast, the regression-improved ensemble mean (R_{EM}) forecast, the regression-improved multimodel (R_{all}) forecast, and the regression-improved multimodel subset (R_{subset}) forecast as a function of the multimodel ensemble size. An ensemble of size M is constructed by using the first M models in Table 2. The results shown in Fig. 2 are not very sensitive to the order in which models are included in the ensemble.

In the Tropics, the regression-improved ensemble mean R_{EM} has the highest skill scores for ensembles of fewer than five models. The bias-corrected multimodel ensemble mean U_{EM} becomes a better deterministic forecast in larger ensembles. As the ensemble becomes larger, it is evident that the uncertainty introduced by estimating an additional adjustable coefficient overcomes any gain in skill due to forecast rescaling. The skill of R_{EM} and U_{EM} increases steeply with increasing ensemble size up to about six but then levels off at a value slightly over 0.3 for larger ensembles. The skill of the regression-improved multimodel forecast R_{all} initially increases with increasing ensemble size but then decreases for the larger ensemble sizes when the number of regression coefficients becomes too large. Overall, skill scores for R_{all} are lower than those for R_{EM}.

In the PNA sector, the regression-improved ensemble mean R_{EM} is the best. It barely outperforms the climatological forecast, however. Note that we use monthly data for the whole year. We expect seasonal means to be more skillful in some seasons. Deviations from the monotonic skill increase as a function of the ensemble size are likely due to sampling variability and differences in the individual model performance. Skill scores of the bias-corrected multimodel ensemble mean U_{EM} increase nearly monotonically with increasing ensemble size as more and more climate noise is filtered out from the ensemble mean. The performance of the regression-improved multimodel forecast R_{all} decreases monotonically with increasing ensemble size.

The attempt to select the ''best'' forecast subset objectively with the AIC was not very successful. The average number of models used for constructing R_{subset} is indicated by the numbers on the blank bars in Fig. 2. There is some skill improvement for R_{subset} in the PNA sector over that for R_{all}, which comes about mainly because fewer, rather than ''better,'' models are used in the linear regression (we found no strong preference for any particular subset of models). The data record apparently is too short to distinguish reliably among the performances of the AMIP models in Table 2. In the Tropics we found some preference for the ECMWF, Geophysical Fluid Dynamics Laboratory (GFDL), Laboratoire de Météorologie Dynamique (LMD), and Max-Planck-Institut für Meteorologie (MPI) models,

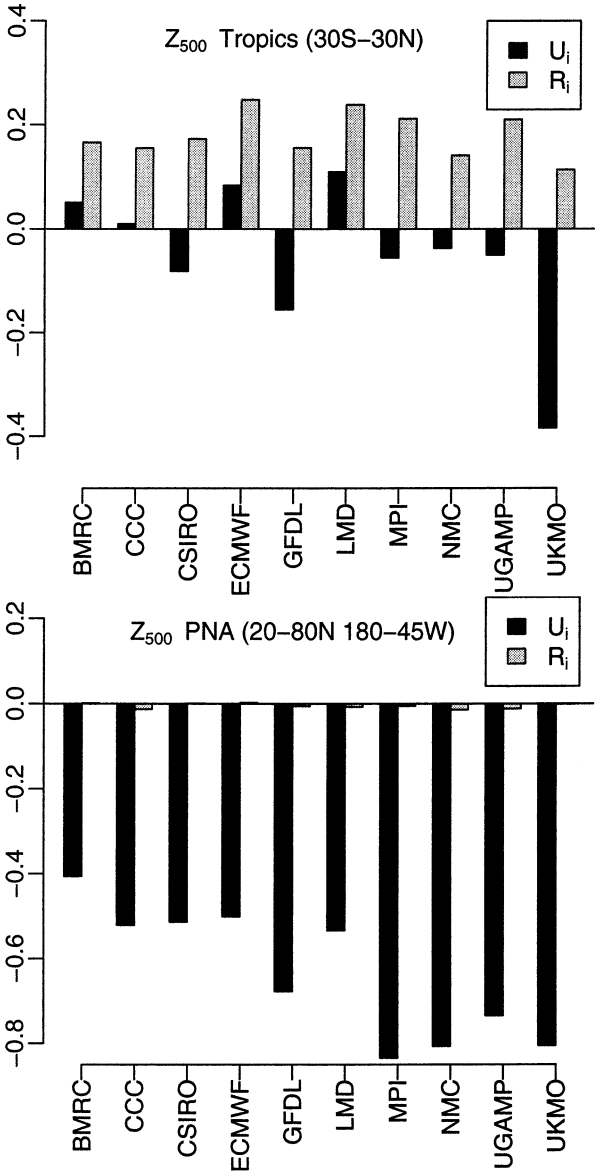

FIG. 1. Cross-validated error variance skill score S for Z_{500} in the (top) Tropics and (bottom) PNA sector for individual AMIP models. Black bars are skill scores of bias-removed individual forecasts U_i. Gray bars are skill scores of regression-improved individual forecasts R_i.

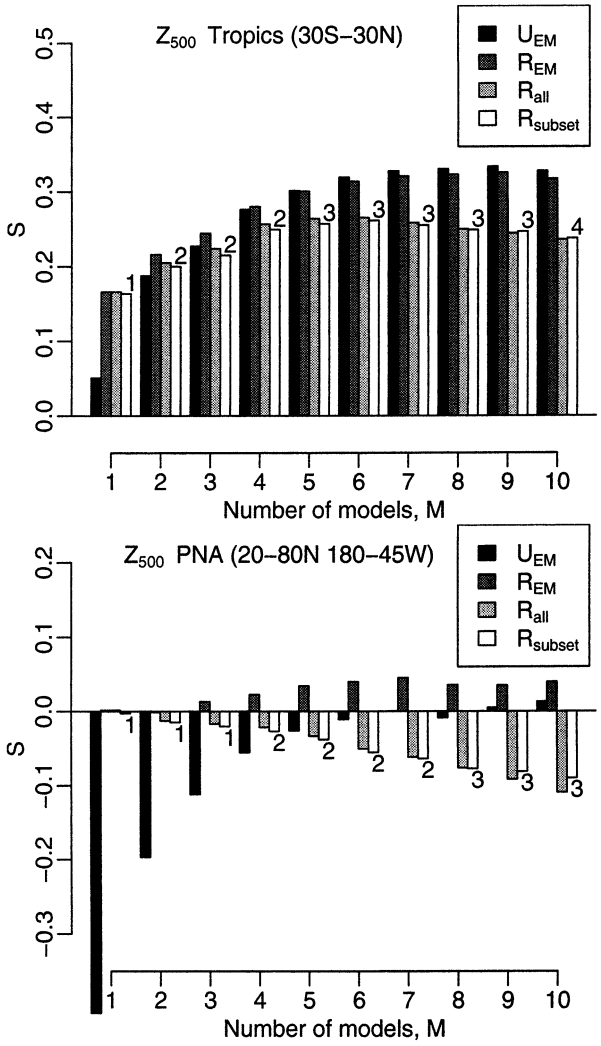

FIG. 2. Cross-validated error variance skill score S for several Z_{500} forecast variants in the (top) Tropics and (bottom) PNA sector as a function of the ensemble size. Skill scores are displayed for the bias-removed ensemble mean U_{EM} (black bars), the regression-improved ensemble mean R_{EM} (dark gray bars), the regression-improved multimodel forecast R_{all} (light gray bars), and the regression-improved multimodel subset forecast R_{subset} for which the best subset of models is determined objectively with AIC (blank bars). The average number of models used for constructing R_{subset} is indicated by the numbers on the blank bars.

which were about 2 times as likely to be chosen than the other models. However, the performance of R_{subset} is about the same as that of R_{all} and is below that of R_{EM}.

The performance of R_{subset} is noticeably worse than that of R_{all} for the same ensemble size as the averaged number of models used for constructing R_{subset}. The additional flexibility of selecting the best model subset apparently aggravates the overfitting problem. The procedure finds the combination of models that ''adapts'' best to the available data sample in the training period, at the expense of the accuracy of the fitted regression coefficients. It would appear that the AIC penalty term, $2(K + 1)$, does not guard adequately against overfitting.

4. Summary

We evaluated the performance of several versions of deterministic unbiased monthly forecasts of Z_{500}, as measured by the error variance, based on an ensemble of 10 10-yr AMIP integrations.

The summary of the results is as follows.

1) In the Tropics, where predictability is relatively high, the regression-improved ensemble mean R_{EM} performs best for small ensembles and the bias-removed ensemble mean U_{EM} is better for ensembles with more than six models. For large ensembles, estimating just one regression coefficient to ''improve'' U_{EM} degrades skill, apparently because the coefficient estimate is subject to sampling errors.
2) In the extratropics, where atmospheric predictability is low, the regression-improved ensemble mean R_{EM} yields the deterministic forecast with the smallest error variance. However, this forecast barely outperforms the climatological forecast. There is no substantial skill improvement for ensembles of more than six models.
3) The performance of the regression-improved multimodel (superensemble) forecast R_{all} is generally not as good as that of the regression-improved ensemble mean R_{EM} and becomes increasingly poorer for ensembles with many models.

The main reason for the poor performance of the multimodel linear regression technique in this study is overfitting that results in overly optimistic estimates of skill when the number of parameters estimated from the data is large relative to the sample size. In these circumstances, the fitted model performs poorly on independent data because it has adapted itself to the unpredictable variability within the available data in the training period.

REFERENCES

Akaike, H., 1974: A new look at the statistical model identification. *IEEE Trans. Auto. Control,* **19,** 716–723.

Bengtsson, L., U. Schlese, E. Roeckner, M. Latif, T. P. Barnett, and N. Graham, 1993: A two-tiered approach to long-range climate forecasting. *Science,* **261,** 1026–1029.

Danard, M. B., M. M. Holl, and J. R. Clark, 1968: Fields by correlation assembly—A numerical analysis technique. *Mon. Wea. Rev.,* **96,** 141–149.

Davis, R. E., 1976: Predictability of sea surface temperature and sea level pressure anomalies over the North Pacific Ocean. *J. Phys. Oceanogr.,* **6,** 249–266.

Derome, J., G. Brunet, A. Plante, N. Gagnon, G. J. Boer, F. W. Zwiers, S. Lambert, and H. Ritchie, 2001: Seasonal predictions based on two dynamical models. *Atmos.–Ocean,* in press.

Doblas-Reyes, F. J., M. Déqué, and J.-P. Piedelievre, 2000: Multimodel spread and probabilistic forecasts in PROVOST. *Quart. J. Roy. Meteor. Soc.,* **126,** 2069–2087.

Fraedrich, K., and N. R. Smith, 1989: Combining predictive schemes in long-range forecasting. *J. Climate,* **2,** 291–294.
Gates, W. L., 1992: AMIP: The Atmospheric Model Intercomparison Project. *Bull. Amer. Meteor. Soc.,* **73,** 1962–1970.
Gibson, J. K., P. Kalberg, S. Uppala, A. Hernandes, A. Nomura, and E. Serrano, 1997: ECMWF Reanalysis Report Series 1—ERA Description. ECMWF, Reading, United Kingdom, 72 pp.
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.
——, 1997: Multi-pattern fingerprint method for detection and attribution of climate change. *Climate Dyn.,* **13,** 601–612.
Kharin, V. V., F. W. Zwiers, and N. Gagnon, 2001: Skill of seasonal hindcasts as a function of the ensemble size. *Climate Dyn.,* **17,** 835–843.
Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Willifor, S. Gadgil, and S. Surendran, 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. *Science,* **285,** 1548–1550.
——, ——, Z. Zhang, T. E. LaRow, D. R. Bachiochi, C. E. Willifor, S. Gadgil, and S. Surendran, 2000: Multimodel ensemble forecasts for weather and seasonal climate. *J. Climate,* **13,** 4196–4216.
Lorenz, E. N., 1982: Atmospheric predictability with a large numerical model. *Tellus,* **34,** 505–513.
Michaelson, J., 1987: Cross-validation in statistical climate forecast models. *J. Climate Appl. Meteor.,* **26,** 1589–1600.
Phillips, T. J., 1994: A summary documentation of the AMIP models. Tech. Report PCMDI Rep. 18, PCMDI, Lawrence Livermore National Laboratory, Livermore, CA, 343 pp.
Rowell, D. P., and F. W. Zwiers, 1999: The global distribution of the sources of decadal variability and mechanisms over the tropical Pacific and southern North America. *Climate Dyn.,* **15,** 751–772.
Thompson, P. D., 1977: How to improve accuracy by combining independent forecasts. *Mon. Wea. Rev.,* **105,** 228–229.
Zwiers, F. W., 1996: Interannual variability and predictability in an ensemble of AMIP climate simulations conducted with the CCC GCM2. *Climate Dyn.,* **12,** 825–848.
——, 1999: The detection of climate change. *Anthropogenic Climate Change,* H. von Storch and G. Flöser, Eds., Springer-Verlag, 161–206.
——, X. L. Wang, and J. Sheng, 2000: The effects of specifying bottom boundary conditions in an ensemble of GCM simulations. *J. Geophys. Res.,* **105,** 7295–7315.