JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 113, D04108, doi:10.1029/2007JD008804, 2008

Measuring the potential predictability of ensemble climate predictions

Youmin Tang,[1] Hai Lin,[2] and Andrew M. Moore[3]

Received 13 April 2007; revised 17 September 2007; accepted 4 December 2007; published 29 February 2008.

[1] In this study, ensemble predictions of the El Niño Southern Oscillation (ENSO) and the Arctic Oscillation (AO) were conducted using two coupled models and two atmospheric circulation models, respectively, as well as various ensemble schemes. Several measures of potential predictability including ensemble mean square (EM^2), ensemble spread and the ratio of signal-to-noise were explored in terms of their ability of estimating a priori the predictive skill of the ENSO and AO ensemble predictions. The emphasis was put on examining the relationship between the measures of predictability that do not use observations and the model prediction skill of correlation and mean square error (MSE) that make use of observations. The relationship identified here offers a practical means of estimating the potential predictability and the confidence level of an individual prediction. It was found that the EM^2 is a better indicator of the actual skill of ensemble ENSO and AO prediction than the ratio of signal-to-noise. When correlation-based metrics are used, the prediction skill is likely to be a linear function of EM^2, i.e., the larger the EM^2 the higher skill the prediction; whereas when *MSE*-based metrics are used, a ''triangular relationship'' is suggested between them, namely, that when EM^2 is large, the prediction is likely to be reliable whereas when EM^2 is small the prediction skill is highly variable. In contrast with ensemble weather prediction (NWP), the ensemble spread is not a good predictor in quantifying climate prediction skill in the models used in this study because the forced response may be much larger than the noise in the climate timescales compared to the NWP. A statistical framework was proposed to explain why EM^2 is a good indicator of actual prediction skill in the ensemble climate predictions.

Citation: Tang, Y., H. Lin, and A. M. Moore (2008), Measuring the potential predictability of ensemble climate predictions, *J. Geophys. Res.*, *113*, D04108, doi:10.1029/2007JD008804.

1. Introduction

[2] A crucial aspect of climate predictability studies is to estimate forecast uncertainty originating from uncertainties in initial condition, physical process parameterization, and stochastic forcing by transients in the climate system (e.g., atmosphere and ocean). To address how uncertainties in an initial state of the climate system affect the prediction of a later state is often referred to as climate predictability of the first kind, whereas the predictability of the second kind is essentially a boundary value problem. Predictability of the first kind has attracted a lot of attention because of the critical importance of initial conditions on prediction skill [e.g., *Epstein*, 1969; *Molteni et al.*, 1996; *Palmer*, 1999; *Moore and Kleeman*, 1998; *Kleeman*, 2002; *DelSole*, 2004, 2005; *Tang et al.*, 2005, 2007]. This is particularly interesting from a practical point of view since certain types of climate states are known to be more predictable than others. Predictability of the first kind may offer a practical mean of estimating the confidence that we can place in future predictions using the same climate model.

[3] Predicting the first kind of predictability is equivalent to solving the Liouville equation for the probability density function (pdf) of the climate state [*Epstein*, 1969; *Palmer*, 1999]. However, it is impractical to solve such an equation because of the huge dimensionality of the climate system (e.g., 10^6 variables for a typical climate model) and because the initial pdf is generally unknown. A practical solution is to approximate the pdf using a finite size of ensembles using some specific techniques [e.g., *Toth and Kalnay*, 1993; *Molteni and Palmer*, 1993; *Kleeman and Majda*, 2005]. The ensemble is generated by repeating the prediction many times, each time perturbing the initial conditions of the forecast model.

[4] Two important issues in ensemble prediction are ensemble representation and ensemble verification. For the former, a widely used measure is the ensemble mean. In general, the mean of an ensemble prediction will, on average, have a smaller error than the mean error of any of the individual forecasts [*Leith*, 1974; *Murphy*, 1988].

[1] Environmental Science and Engineering, University of Northern British Columbia, Prince George, British Columbia, Canada.
[2] Recherche Prévision Numérique, Meteorological Service of Canada, Dorval, Quebec, Canada.
[3] Ocean Sciences Department, University of California, Santa Cruz, California, USA.

0148-0227/08/2007JD008804$09.00

When each individual ensemble member has the same error variance, the ensemble mean is the best linear unbiased estimate of the true state. The ensemble mean is also able to greatly alleviate the impact of random noise on a prediction.

[5] An important problem in predictability study is to seek a predictor of forecast skill, by which the degree of confidence that can be placed in an individual forecast can be assessed. A technique widely used in NWP is to utilize the second moment of an ensemble prediction, i.e., the ensemble spread. If the spread of ensemble members is relatively small, the atmospheric/oceanic state we are predicting is probably relatively insensitive to errors and uncertainties in the initial conditions, so that the prediction skill is probably high. If, however, the ensemble members diverge rapidly, the state that is being predicted may be susceptible to error growth in the initial conditions, leading to a poor prediction skill. Thus a priori likely skill (or usefulness) of an individual prediction might be estimated by the ensemble spread. A good relationship between the ensemble spread and the prediction skill has been found in many NWP models [e.g., *Buizza and Palmer*, 1998; *Whitaker and Loughe*, 1998; *Scherrer et al.*, 2004]. However little connection was found between the ensemble spread and the prediction skill in other climate prediction models. Instead, the climate predictability is mostly related to the variations in the amplitude of ensemble mean anomaly or the signal present in initial conditions [e.g., *Kumar and Hoerling*, 1995, 2000; *Kumar et al.*, 2000; *Tippett et al.*, 2004; *Peng and Kumar*, 2005; *Tang et al.*, 2005, 2007]. An interesting question naturally raises, namely, whether the spread-skill connection is only an attribute of ensemble NWP, and if so, what is the predictor of forecast skill for ensemble climate predictions?

[6] The emphasis of this paper will be on the above question. Toward this goal, we will explore several measures of potential predictability including ensemble mean square, ensemble spread and the ratio of signal-to-noise in terms of their ability in estimating actual model prediction skill. In a prefect model scenario, the measure of potential predictability that does not use observations is a good indicator of the actual skill of the model that makes use of observations. In this paper, we will study ensemble predictions from four different climate models for two important modes of climate variability: the ENSO and the AO. The models include two hybrid coupled models (HCMs), a simple atmospheric general circulation model (SGCM) and a full atmospheric GCM (i.e., the second generation general circulation model of the Canadian Center for Climate Modeling and Analysis, refereed to as GCM2). These models exhibit significant differences in both their dynamics and sophistication, allowing us to explore important properties of ensemble climate prediction (ECP) in more general terms, and to confirm the robustness of results across model formulations.

[7] Section 2 briefly describes the models and ensemble schemes used. Section 3 introduces the definition of the metrics measuring prediction skill and predictability. Section 4 presents the ensemble prediction skill using different measures for the four models. The central issue of the predictor of forecast skill is explored in section 5, followed by summary and discussion in section 6.

2. Prediction Models and Ensemble Schemes

2.1. ENSO HCM1 and HCM2

[8] The HCM1 is composed of an Ocean General Circulation Model (OGCM) coupled to a statistical atmosphere, whereas the HCM2 is the same ocean model coupled to a dynamical atmospheric model of intermediate complexity. The ocean model used is based on the OPA version 8.1 [*Madec et al.*, 1998], a primitive equation OGCM. The model uses an Arakawa *C* grid, and is configured for the tropical Pacific ocean between 30°N–30°S and 120°E–75°W. The horizontal resolution in the zonal direction is 1°, while that in the meridional direction is 0.5° within 5° of the equator, smoothly increasing to 2.0° at 30°N and 20°S. There are 25 vertical levels with 17 concentrated in the top 250m of the ocean. The time step of integration is 1.5 hours and all boundaries are closed, with no slip conditions. A turbulent closure hypothesis is used to parameterize subgrid-scale physical processes, where small-scale horizontal and vertical transports are evaluated in terms of diffusion coefficients and derivatives of the large-scale flow as described by *Blanke and Delecluse* [1993]. The detailed formulation and configuration of the ocean model and its performance in simulating the tropical Pacific are given by *Vialard et al.* [2002].

[9] The statistical atmospheric model is a linear model, which predicts the contemporaneous surface wind stress anomalies from sea surface temperature anomalies (SSTA). The seasonal variations of the responses of wind stress to SST are also included so that for each month there is essentially a different atmospheric model. The model is trained using the NCEP atmospheric reanalysis wind products and the Reynolds-Smith SST observations [*Smith et al.*, 1996] from 1951 to 1980. Therefore the ensemble experiments performed for the period 1981–1998 in this study are completely independent of the construction of the atmospheric model. This strategy eliminates any artificial skill when evaluating the hindcast skills.

[10] The dynamical atmospheric model consists of a Gill-type steady state model which has been used for routine ENSO predictions and for the study of climate predictability, developed by *Kleeman* [1989] (referred as the Kleeman model hereafter). The model computes global anomalies relative to the observed seasonal cycle of surface wind and mean atmospheric wind at 850 mbar. When the Kleeman model is coupled to the OGCM, the OGCM provides SST anomalies to the atmospheric model. The atmosphere is heated by a Newtonian cooling/relaxation to the SST anomaly, and by a latent heating due to deep penetrative convections through a simple moist static energy dependent convection scheme.

[11] In both coupled models, the OGCM is forced by the sum of the associated wind anomalies computed by the atmospheric model and the observed monthly mean climatological winds. The ENSO prediction skill and predictability in the two HCMs have been documented by *Tang et al.* [2003, 2005].

[12] The stochastic optimals (*SOs*) [*Farrell and Ioannou*, 1993; *Kleeman and Moore*, 1997] are used to construct ensemble predictions in the two HCMs. The *SOs* represent

uncertainties associated with stochastic events in the coupled ocean-atmosphere system that can be amplified by the dynamical model during the forecast interval T, which in turn leads to forecast error growth. The *SOs* are defined by the eigenvectors of the operator S [*Farrell and Ioannou*, 1993; *Kleeman and Moore*, 1997]

$$S = \int_0^T \mathbf{A}^*(t,0)\mathbf{U}\mathbf{A}(t,0)\,dt. \quad (1)$$

Here T is the forecast interval of interest and is assumed to be 12 months in this study, $\mathbf{A}(t, 0)$ is the forward tangent linear propagator of the linearized dynamical model that advances the state vector of the system from time 0 to time t, $\mathbf{A}^*(t, 0)$ is the adjoint of $\mathbf{A}(t, 0)$, and the matrix $\mathbf{U}$ defines the norm of interest. In this study, we use a seminorm defined as the square of the Niño3 SSTA index. (The time series of SST anomalies averaged over the Niño3 (150–90°W, 5°N–5°S) region, which is often used to evaluate model skill as is in this study.)

[13] A detailed description the *SOs* of the two HCMs and the construction of ensemble predictions using them are given by *Moore et al.* [2006] and *Tang et al.* [2005]. The ensemble size for the two HCMs is 31 including a control run. The initial conditions are taken from a 3D-var assimilation system developed by *Tang et al.* [2003], which assimilated NCEP reanalysis subsurface temperature. With the assimilation system, both HCMs have a predictive skill of Niño3 SSTA index that is comparable with some best ENSO prediction models in the world at a leading time of the first 12 months [*Tang et al.*, 2004].

2.2. SGCM

[14] The SGCM is a primitive equation dry atmospheric model, initially designed by *Hoskins and Simmons* [1975] to study the life cycle of baroclinic waves, and further developed by *Hall* [2000]. It has a global domain with a horizontal resolution of T21 and 5 levels in the vertical. An important feature of this model is that it uses a time-averaged forcing calculated empirically from observed daily data. By computing the dynamical terms of the model, together with a linear damping, with daily global analyses and averaging in time, the residual term for each time tendency equation is obtained as the forcing. The collective effect of these forcing terms represents all processes that are not resolved by the model's dynamics such as diabatic heating (including latent heat release related to the transient eddies) and the deviation of dissipative processes from linear damping. This atmospheric model has been used to perform seasonal predictions, and was found to be similar in prediction skill to a more complex GCM [*Derome et al.*, 2005].

[15] Global ensemble forecasts are made for the 51 boreal winters (December–January–February (DJF) from 1948/1949 to 1998/1999 with an ensemble size of 70. The initial conditions for the seasonal forecasts are the 0000 UTC 1 December analyses from the National Centers for Environmental Prediction (NCEP) [*Kalnay et al.*, 1996]. Each ensemble run is constructed by adding to the initial condition (i.e., all fields) a small-amplitude perturbation, which is the anomaly (with respect to the 51-year winter climatology) of a random winter day in the 51-year NCEP data set (excluding the winter being predicted) multiplied by 0.1. The forcing used for a given winter is the November mean forcing anomaly of the NCEP data of that year added to the NCEP DJF mean climatological forcing averaged over 50 remaining winters. As such, there is no observed information used for the prediction period.

[16] The skill of the ensemble mean prediction has been evaluated in detail by *Derome et al.* [2005]. It was found that the SGCM has a statistically significant skill in forecasting the AO variability, actually even better than a more complex GCM (Canadian GCM2). In the present study, the AO is defined as the first empirical orthogonal function (EOF) mode of the wintertime (DJF) mean sea level pressure anomalies (MSLPA) north of 20°N from the NCEP reanalysis. The observed and each individual forecast DJF MSLPA field over the 51 winters are projected onto the AO pattern to obtain the corresponding observed and SGCM-predicted principal component time series, i.e., AO indices, which are used in the following discussions.

2.3. GCM2

[17] The GCM2 is a primitive equation model with T32 horizonal resolution corresponding to a 3.75° Gaussian grid, and 10 vertical levels. It contains a comprehensive package of physical parameterizations of subgrid-scale processes. The roles of clouds, water vapor, carbon dioxide, oxygen, and ozone are included in the calculation of solar and terrestrial radiation. The model has been widely used for climate simulations and seasonal climate prediction [e.g., *Boer et al.*, 2000; *Flato et al.*, 2000; *Derome et al.*, 2001].

[18] The seasonal forecasts using GCM2 are part of the Canadian Historical Forecasting Project (HFP), which is designed to test the extent to which the potential predictability of mean seasonal conditions could be achieved [*Derome et al.*, 2001]. The global SSTA of the month prior to the forecast period is maintained throughout the 3-month forecast added to a monthly varying climatological SST field. Ice conditions are specified from the climatology at all times, and the sea ice extent is specified to be that observed during the previous months, and then relaxed to climatology over a period of 15 days.

[19] Global ensemble predictions were performed for 26 boreal winters (DJF) from 1969/1970 to 1994/1995, with an ensemble size of 24. Each ensemble member was initialized from the NCEP reanalysis fields (*Kalnay et al.*, 1996) at 6h intervals prior to the forecast season. The skill of the ensemble prediction by GCM2 has also been analyzed in previous studies [e.g., *Derome et al.*, 2001].

[20] The AO index defined in GCM2 is similar to that in SGCM but using 500 mbar geopotential height anomalies instead of the sea level pressure anomalies. Accordingly, the observed AO index against which the GCM2 AO index is verified is also obtained from the 500 mbar height. Since the extratropical large-scale low-frequency variability has an equivalent barotropic vertical structure, the AO index is insensitive to the choice of vertical level.

[21] A summary on ensemble size, prediction period, and prediction target is listed in Table 1.

3. Measures of Prediction Skill and Predictability

3.1. Measures of Prediction Skill

[22] Two measures will be used to evaluate prediction skill. Denoting by T the anomaly value of the variable of

Table 1. Information of Models and Ensemble Predictions

Item	HCM1	HCM2	SGCM	GCM2
Prediction period	1981–1998	1981–1998	1948–1998	1969–1994
Prediction target	El Niño3 SSTA	El Niño3 SSTA	AO index	AO index
Ensemble size	31	31	70	24
Ensemble method	stochastic optimal	stochastic optimal	random	random

interest (Niño3 SSTA for HCM1 and HCM2, and AO index for AGCM2 and SGCM); superscripts p for prediction, and o for observation, N the number of initial time in the sample, t the lead time of the prediction, subscript i the initial time of prediction, μ^p and μ^o the mean of the prediction and observations over N, we can define these measures as follows.

3.1.1. Anomaly Correlation Skill *r*

$$r(t) = \frac{\sum_i T_i^p(t) T_i^o(t)}{\sqrt{\sum_i T_i^p(t)^2} \sqrt{\sum_i T_i^o(t)^2}} \tag{2}$$

[23] The contribution of each prediction to $r(t)$, denoted as C, can be measured by

$$C_i(t) = \frac{\frac{1}{N} T_i^p(t) T_i^o(t)}{r(t)} \times 100\% \tag{3}$$

where T^p is the ensemble mean anomaly of Niño3 SSTA or AO index, and T^o is corresponding observed counterpart. In linear regression, if observation T^o is statistically predicted by T^p, the correlation coefficient r is proportional to the total observed variance [*von Storch and Zwiers*, 1999]. Thus the C_i can be also interpreted as the observed variance explained by an individual prediction starting from the initial time i (i = 1, ..., N). We compared the C_i and the explained variance, and found both displaying very similar features with a correlation coefficient over 0.95 for all models except SGCM which has a 0.70 value of correlation coefficient.

3.1.2. Mean Square of Error (*MSE*) and Skill Score (*F* and *FF*)

[24] The aforementioned correlation-based measure quantifies the skill of the phase and trend in the prediction, and has been the most widely used measure of skill in NWP and climate prediction. Another important measure of skill is *MSE*, which quantifies skill of the prediction of amplitude. However, the *MSE* is a measure of absolute error, and does not necessarily give an indication of the importance of an error. The absolute error is often dominated by structures with large natural variance and can have little to do with predictability. It has been argued that absolute error variance is not a measure of predictability [*DelSole*, 2004]. Rather, error variance relative to its saturation value is a more appropriate measure of predictability and has been widely used. The saturation value is often taken from the *MSE* of climatological mean forecasts (MSE^{clim}) in atmosphere (weather) predictability study, i.e., the climatological variance over the period of prediction [e.g., *Lorenz*, 1969; *Boer et al.*, 2000; *Tribbia and Baumhefner*, 1988; *Collins and Allen*, 2002]. Thus the ratio of prediction *MSE* over MSE^{clim} is used in this study as the other measure to quantify the skill score of an individual prediction starting from the initial time i, defined as below:

$$F_i = \frac{MSE_i}{MSE_i^{clim}} \tag{4}$$

$$MSE_i = \frac{1}{M} \sum_{t=1}^{M} \left(T_i^p(t) - T_i^o(t) \right)^2 \tag{5}$$

M is leading time of prediction, which is equal to 12 (months) for the ENSO predictions. For AO predictions, T^p is the ensemble mean of the averaged AO index of DJF, therefore the size of M is equal to 1 (season). When the prediction T^p in (5) is replaced by the climatological mean, (5) will give MSE^{clim}. Thus F measures the skill of an individual prediction compared with the climatological prediction. A value of 0 of F indicates a perfect forecast whereas a value of 1 indicates a model prediction equivalent to a climatological forecast. When F_i is larger than 1, a model prediction has no skill since it is worse than a climatological forecast. F is also highly related to the mean square skill score (*MSSS*), a widely used measure in the field of prediction verification proposed by *Murphy and Epstein* [1989], with the relationship of $MSSS = 1 - F$.

[25] Another way to obtain the saturation *MSE* is to use persistence prediction instead of climatological mean prediction in (4). We found that all results and conclusions presented in following sections keep unchanged if the saturation *MSE* is taken from the persistence prediction.

[26] Equation (4) is used for evaluating individual predictions. A similar measure for evaluating the overall skill of all predictions as a function of lead time t is as follows,

$$FF(t) = \frac{\frac{1}{N} \sum_{i=1}^{N} \left(T_i^p(t) - T_i^o(t) \right)^2}{\frac{1}{N} \sum_{i=1}^{N} \left(T^{clim} - T_i^o(t) \right)^2} \tag{6}$$

3.2. Measures of Potential Predictability

[27] Several measures will be explored for quantifying potential predictability of the climate models, including the ensemble mean square (EM^2), ensemble spread (*ES*) and the ratio of signal-to-noise (*ER*). The *ES* and *ER* have been widely used for measuring predictability in NWP and climate predictions as discussed in section 1. The EM^2 will be explored because they give an indicative of the signal size, which is highly associated with the amount of the information provided by predictions *Kleeman*, 2002]. The possible underlying mechanisms for the relationship between EM^2 and prediction skills will be further discussed in section 6.

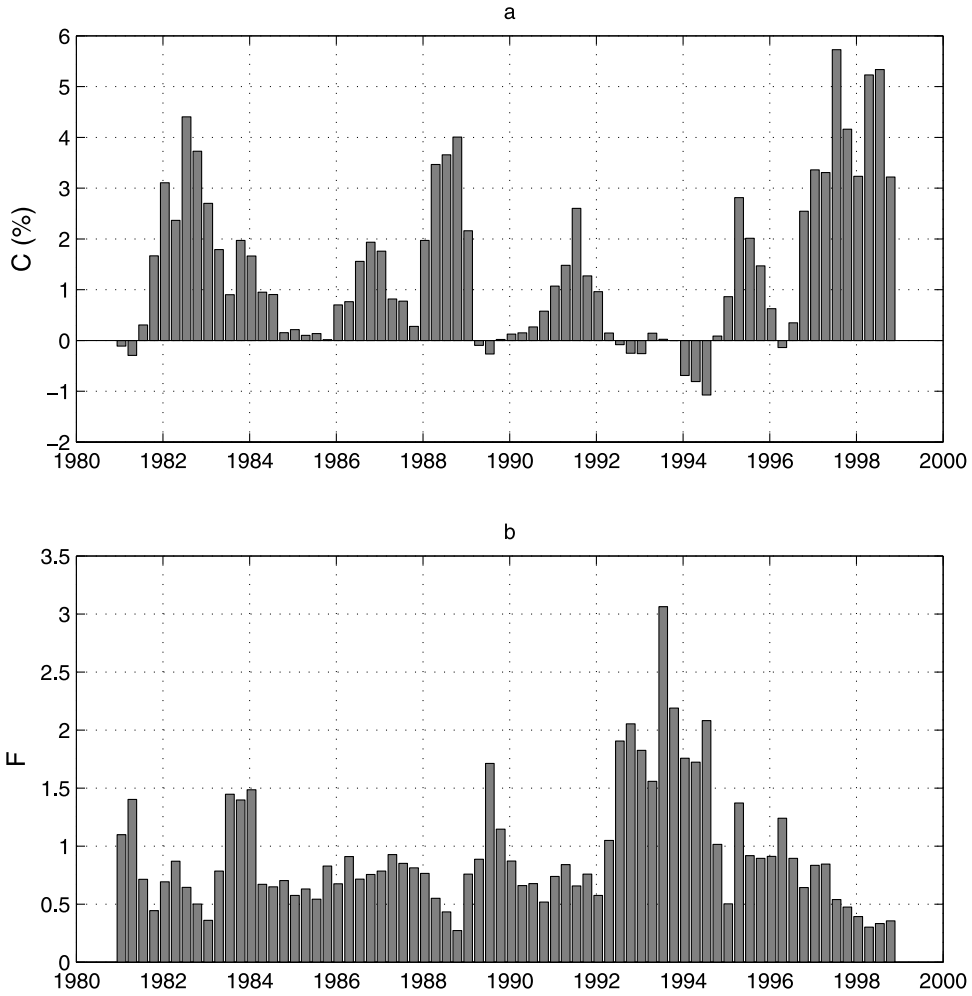

Figure 1. (a) Correlation contribution C averaged over 12 months of lead time, as a function of prediction from 1981 to 1998 for HCM1. (b) Same as Figure 1a but for F.

[28] These measures are defined as follows:

$$EM_i(t) = E\left(T_{ij}^p(t)\right) = \frac{1}{L}\sum_{j=1}^{L} T_{ij}^p(t), \tag{7}$$

where L is the ensemble size of an ensemble prediction.

$$ES_i(t) = STD\left(T_{ij}^p(t)\right) = \sqrt{\frac{1}{L-1}\sum_{j=1}^{L}\left(T_{ij}^p(t) - EM_i(t)\right)^2}. \tag{8}$$

$$ER_i(t) = \left|\frac{EM_i(t)}{ES_i(t)}\right| \tag{9}$$

4. Prediction Skill Scores

[29] The skill scores, defined by (3) and (4), were calculated for the ensemble prediction of each model. Figures 1 and 2 show the skill for HCM1 and HCM2, evaluated by Niño3 SSTA index. As can be seen in Figures 1a and 2a, there is a large variation of correlation contribution C with initial conditions. While some initial conditions lead to good predictions that account for significant contributions to r, many initial conditions correspond to very small values of C.

[30] Shown in Figures 1b and 2b are the variations in F skill. A large F usually corresponds with a negative or very small C, and a large C generally corresponds with a small F. On the other hand, however, a small F is not always associated with a large C. In later discussions, we will see that such relationships between F and C lead to different prospectives when a predictor of forecast skill is evaluated by the two different measures of skill.

[31] A feature demonstrated in Figures 1 and 2 is that the large C is mainly associated with predictions of large amplitude ENSO events. For many predictions, C is relatively small. This implies that ENSO prediction skills displayed in Figures 1 and 2 may be mainly due to a few predictions. This feature was also found in other ENSO models [*Tang et al.*, 2004].

[32] Figures 3 and 4 show the AO ensemble prediction skill for SGCM and GCM2, which demonstrate some features similar to Figures 1 and 2. These include the following:

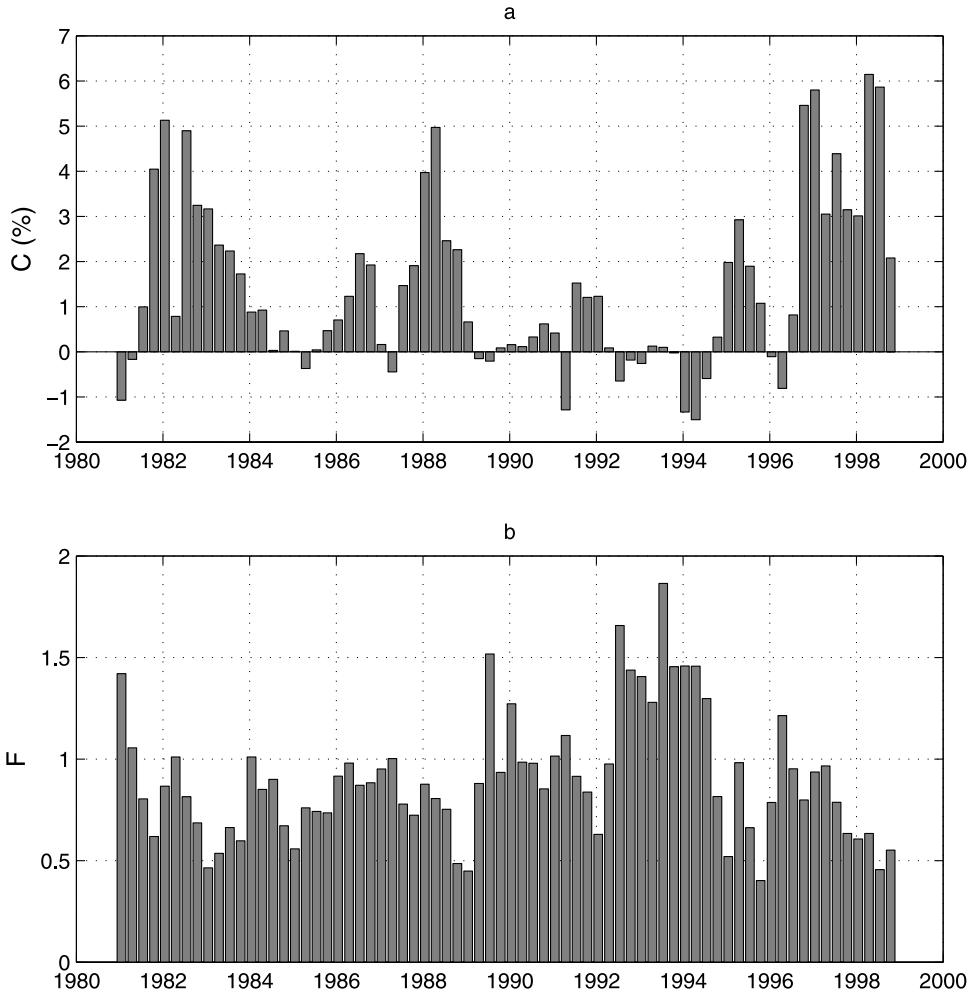

Figure 2. Same as Figure 1 but for HCM2.

[33] 1. There is a large variation of C among predictions. While some winters have good predictions that account for significant contributions to correlation skill and observed variances, others have a very small C.

[34] 2. F variations are less sensitive to predictions compared with large variations of C.

[35] 3. The large C appears only in a few predictions. For most predictions, C and explained variances are relatively small, suggesting that model skill of AO predictions is probably contributed by several successful AO predictions.

[36] 4. The relationship between C and F is very similar to that in HCM1 and HCM2.

5. Measures of Potential Predictability

[37] In this section, we will investigate respectively the ability of EM^2, ES and ER to measure actual forecast skill. One central issue that is addressed is which of these metrics is a more appropriate measure of the actual ensemble prediction skill of climate predictions in the models studied.

5.1. Ensemble Mean Square EM^2

[38] Figures 5 and 6 show the average variation of the measures of predictability for the prediction from 1981 to 1998 for HCM1 and HCM2. Ensemble mean square (also referred to as ensemble mean magnitude sometimes) (Figures 5a and 6a) has a structure similar to the ensemble ratio (Figures 5c and 6c), i.e., large values only appearing in a few predictions, and small values occupying many prediction cases. The similarity between them indicates that the ensemble mean square plays a much more dominant role than the ensemble spread in the ensemble ratio.

[39] Comparing Figures 5a and 6a with Figures 1a and 2a reveals a good relationship between C and ensemble mean square. As the ensemble mean square is large, the corresponding predictions have large Cs, whereas a small ensemble mean square usually corresponds with a small C. This is especially true for several typical ENSO events such as 1983/1984, 1988/1989 and 1997/1998. For example, significantly large values of EM^2 appear in these events, and correspondingly C is large for these predictions. The accumulated contributions C to correlation coefficient $r(t)$ from these predictions actually exceeds 40% for lead times of 1–6 months.

[40] Figure 7 compares the correlation skills of Niño 3 SSTA between two groups of predictions for both models. The first has the prediction samples with EM^2 greater than the median value of all predictions (dashed line), and the second has samples with EM^2 less than the median value (dotted line). (However, some cautions should be taken in

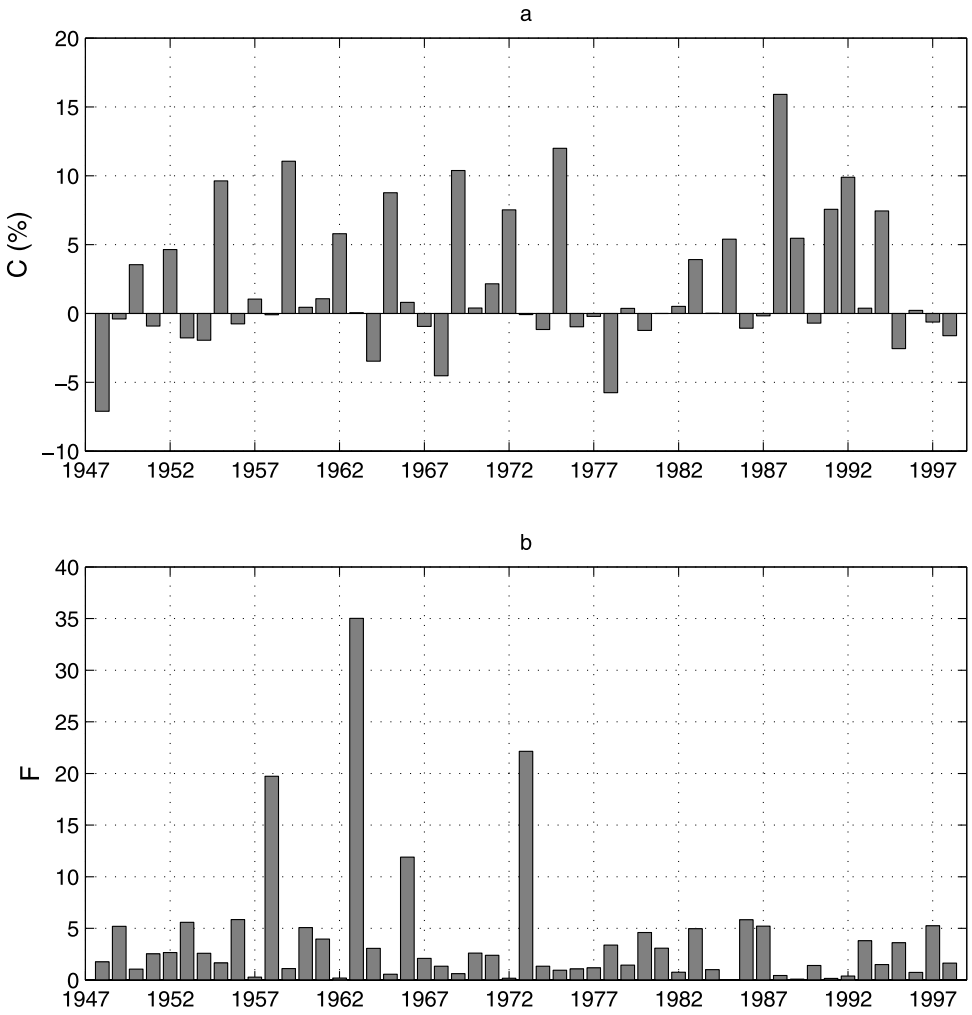

Figure 3. Same as Figure 1 but for SGCM for AO prediction.

interpreting the results since there might be somewhat misleading in the correlations that are calculated from subsets of data, in particular when the subset is made up of high-amplitude cases.) The median value is chosen at each lead time to obtain a sufficient sample size in both groups at the corresponding lead time, (i.e., nearly equal members in each group). It is apparent that the prediction skill with large EM^2 is significantly larger than that with the small EM^2. It is possible that the change in sample size is responsible for the variation in skill with EM^2. To evaluate this, we used a bootstrap method to measure the extent of the uncertainty in the computed correlation due to the finite sample size. This is shown in Figure 7 as vertical bars. (A 1000-member ensemble correlation was computed. Each correlation was obtained using randomly chosen sample pairs of predicted and observed Nino3 indices with the same sample size as that used in the two groups. The standard deviation of ensemble correlation was used to represent the extent of the uncertainty.) As can be seen, the difference in the correlation skill shown in Figure 7 significantly exceeds the correlation error due to the uncertainty of the finite sample size.

[41] Displayed in Figures 8 and 9 are the measures of predictability for the AO prediction, for SGCM and GCM2 respectively, as a function of initial time. As can be seen, it is also apparent that a large EM^2 mainly resides in a few predictions in both models such as winters of 1955/1956, 1959/1960, 1969/1970, 1975/1976, 1983/1984 and 1994/1995 in SGCM and 1969/1970, 1985/1986, 1987/1988, and 1994/1995 in GCM2. For many other predictions, EM^2 is small in both models. A comparison of Figure 8a with Figure 3a and Figure 9a with Figure 4a reveals that a large C generally corresponds to a large EM^2 for both models, except the 1988/1989 case. 1988/1989 has the strongest AO activity during the winters from 1948/1949–1998/1999, leading to a very large C but only a moderate EM^2 from both models. Such a good relation between C and EM^2 is especially obvious for large EM^2. We calculated the accumulated C over predictions with the three largest EM^2s (i.e., M > 5 and 0.26 for SGCM and GCM2 respectively), and found 44% and 46% of the correlation skill r being from these three predictions. Table 2 shows correlation skills between the predicted and observed AO indices, obtained using different samples classified by EM^2. As can be seen, the predictions with a larger EM^2 lead to better skills than those with a smaller EM^2. A bootstrap experiment indicates that the increase in the correlation skill in Table 2 results from the contribution of more skillful predictions with

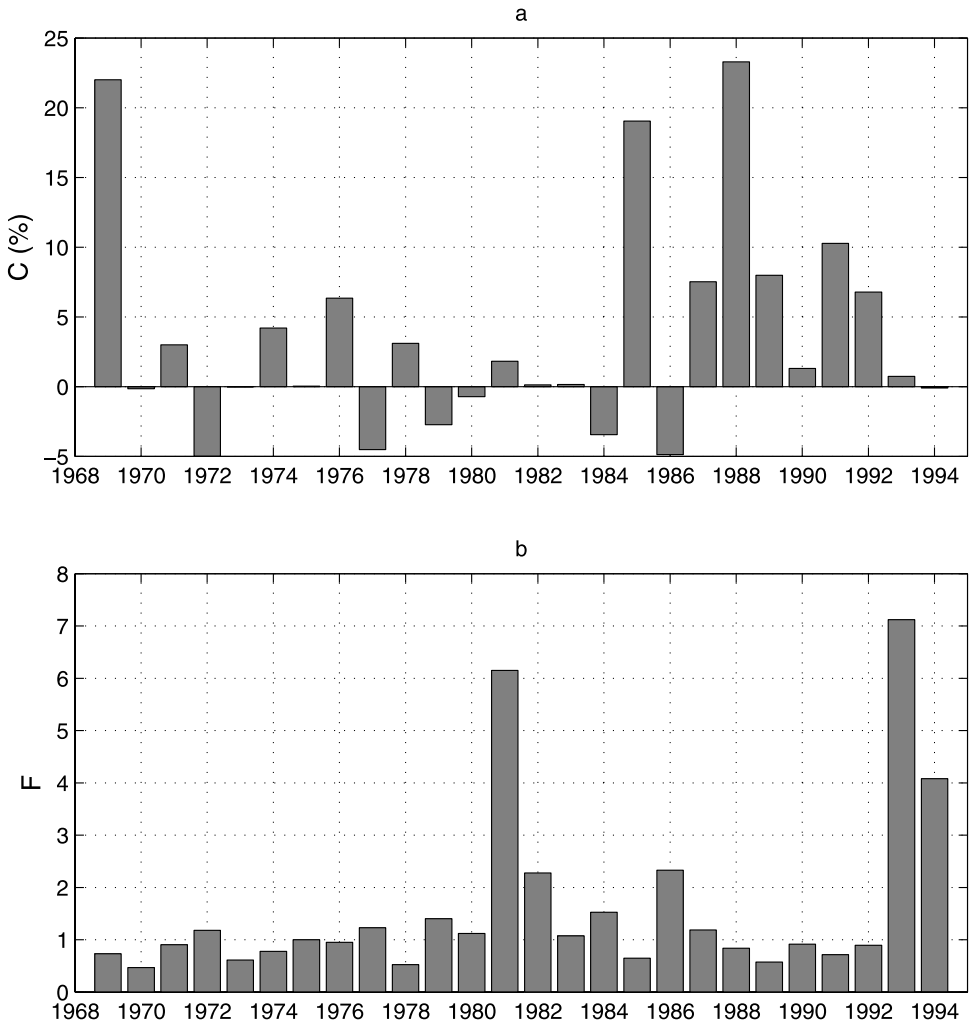

Figure 4. Same as Figure 3 but for GCM2.

larger EM^2, rather than from the uncertainty of the finite sample size (not shown).

[42] The relationship between EM^2 and C can be summarized by Figure 10. As can be seen, a good linear relationship does exist between them. The correlation coefficients are 0.84, 0.89, 0.62 and 0.48 respectively, all of which are statistically significant at a confidence level of 99%. It should be noted that if the outlier point in Figure 10d is removed, the correlation coefficient raises to 0.59.

[43] C is the product of EM multiplied by the observed time series, thus one may intuitively think a large value of EM always corresponding to a large C, resulting in a good relationship between them. In other words, the relationship between C and EM^2 shown in Figure 10 may only be a result of a trivial statistical property, not necessarily indicative of predictive skill. To explore this, we let the prediction T_{pi} is expressed as

$$T_{pi} = T_{oi} + \varepsilon_i \tag{10}$$

the sum of observation and error. Then note that

$$C_i(t) \sim T_{pi}T_{oi} = T_{pi}^2 - T_{pi}\varepsilon_i \tag{11}$$

[44] As seen from (11), when ε_i is small, namely, that the prediction is close to the observation, a good linear relationship exists between them as expected. However when ε is not small, the relationship between C and T_{pi}^2 is complicated, which can be further explored through a Monte Carlo experiment. Suppose the forecast T_{pi} and the observed T_{oi} are two normal-distributed random variables, with ε_i defined by (10). C is calculated from (11). The correlation between C and T_{pi}^2 are simulated by 10000 times. With the same sample size as in Figure 10 (i.e., 72, 72, 51 and 26 respectively), we found that only 0.25–1% of the 10000 runs produce the correlation coefficient greater than 0.48, the minimum correlation coefficient between C and T_{pi}^2 in the four models. Thus large correlation coefficients between T_{pi}^2 and C shown in Figure 10 reflect the fact that the predictions are skillful in these models, and both C and T_{pi}^2 measure the skill.

[45] Two other Monte Carlo experiments were also performed. The first is to repeat the above process but uses T_{oi} to replace a randomly generated time series. The second one, in contrast, uses T_{pi} instead of a random time series. The correlation coefficient between T_{pi}^2 and C was simulated by 10000 times in each experiment. We obtained the same results as above; that is, there is no significant

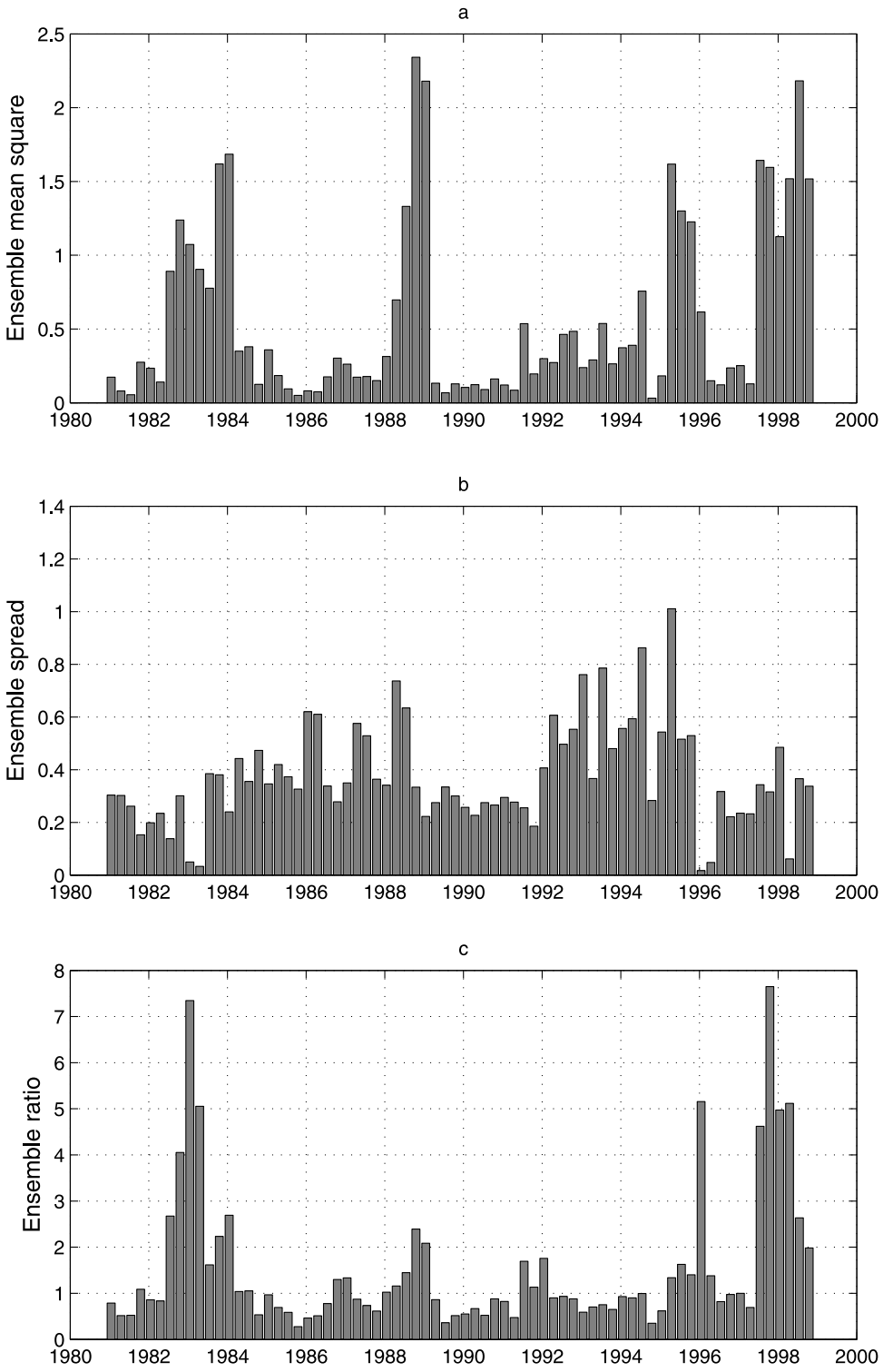

Figure 5. (a) Ensemble mean square EM^2 averaged over 12 months of lead time, as a function of predictions and lead times from 1981 to 1998 for HCM1. (b) Same as Figure 5a but for ensemble spread. (c) Same as Figure 5a but for ensemble ratio.

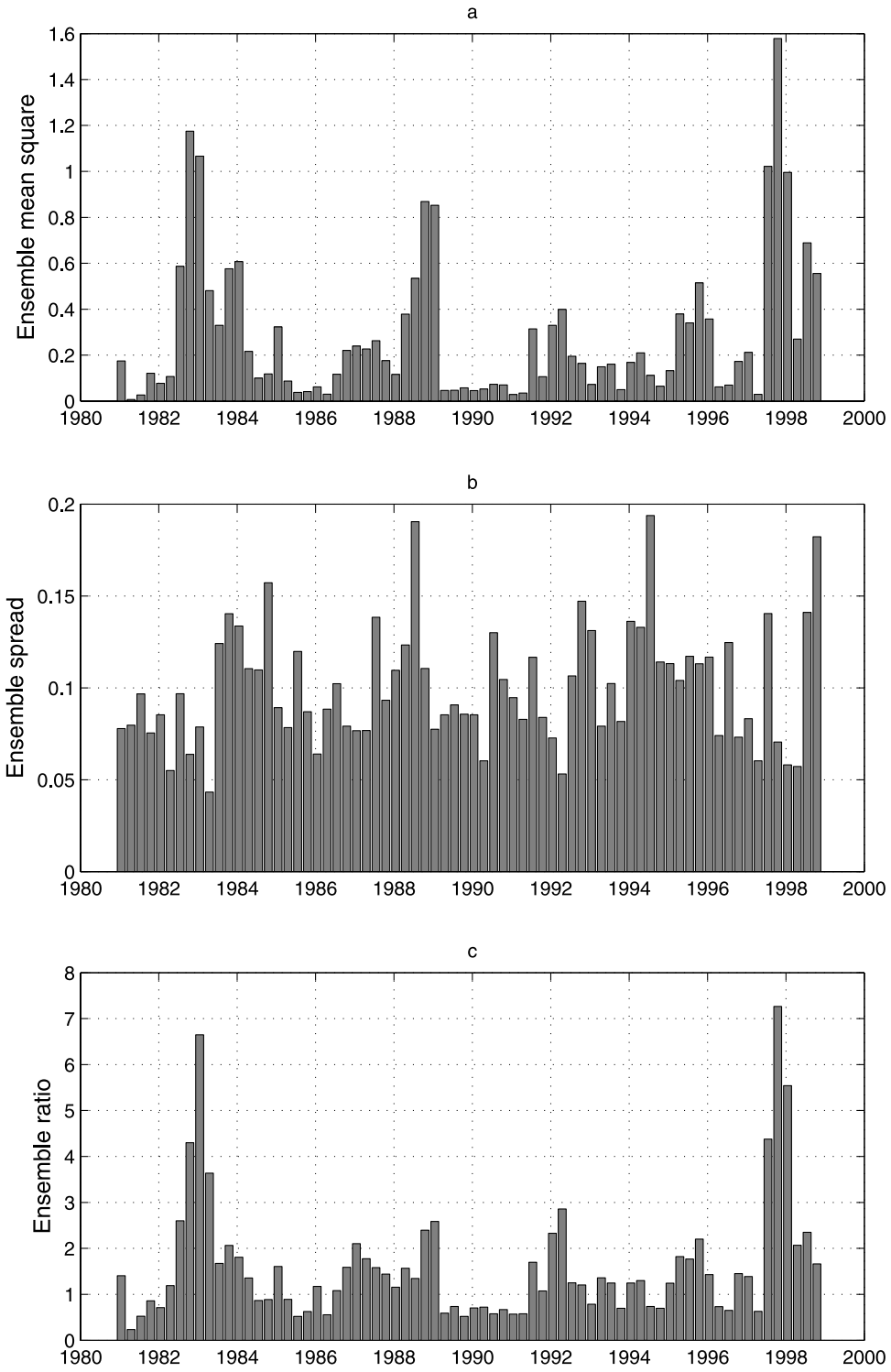

Figure 6. (a) Same as Figure 5 but for HCM2.

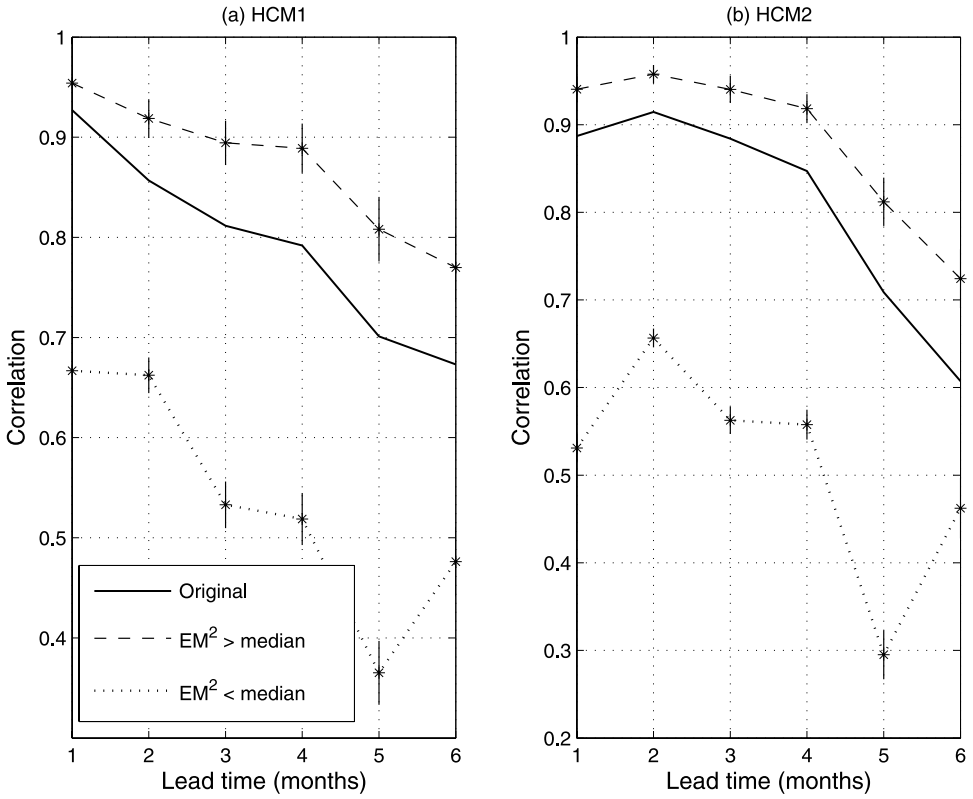

Figure 7. Correlation between the observed and predicted Nino3 indices as a function of lead time for (a) HCM1 and (b) HCM2, calculated respectively using two groups of samples classified by ensemble mean square. The median value of EM^2 at each lead time was used to classify the prediction samples in order to obtain an equal sample size in each group. The dashed line is for the group which has all samples with EM^2 greater than the median value of all predictions whereas the dotted line is for the group with EM^2 less than the median value. The skill from all predictions is shown as the solid line. The vertical lines are the uncertainty extent due to the finite sample size estimated by a 1000-member bootstrap experiment.

correlation existing between T_{pi}^2 and C under the confidence level of 99% and the test threshold value of 0.48.

[46] We have explored the relationship between the ensemble mean square EM^2 and the prediction skill in the four climate models. It is found that when the skill is quantified by the correlation-based measures (C), EM^2 is in general a good linear indicator of the ENSO and AO prediction skill. This is especially true for the prediction of short lead time. Since the correlation-based measures only assess the ability to predict phase and trends, we will next examine the relationship between EM^2 and F.

[47] A comparison of F and ensemble mean square in the four models (Figure 1b versus Figure 5a, Figure 2b versus Figure 6a, Figure 3b versus Figure 8a, and Figure 4b versus Figure 9a) reveals that a large EM^2 often leads to a good prediction skill (i.e., small F); whereas when the EM^2 is small, the F seems much more variable. This is very similar to a so-called "triangular relationship" that has been found to characterize the relationship between ensemble spread and skill in ensemble NWP and climate prediction [e.g., *Buizza and Palmer*, 1998; *Xue et al.*, 1997; *Moore and Kleeman*, 1998]. Specifically when the ensemble spread is small the skill is invariably good, whereas when it is large, the skill can be much more variable. Thus we also use the "triangular relationship" to describe the relationship between the EM^2 and F. Such a "triangular relationship" is more visible in the scatterplot of ensemble mean square with F, as shown in Figure 11.

[48] Figure 12 is the same as Figure 7 but uses FF instead of the correlation to measure prediction skill score. As can been seen, the predictions with large EM^2 (dashed line) are much more skillful than those with small EM^2 (dotted line). Generally the prediction with large EM^2 is about 90–40% better than climatological mean forecast, whereas the prediction with small EM^2 is only 30–10% better than climatological mean forecast. In Figure 12, the FF of large EM^2 is very close to original skill (solid line), indicating the significant impact of the predictions with large EM^2 on prediction skill. We also performed these analyses for the AO prediction of SGCM and GCM2, and got similar results (not shown).

[49] In summary, the ensemble mean square EM^2 is an effective measure for estimating actual prediction skill of the ENSO and AO prediction. When the skill is quantified by correlation-based measures, EM^2 has a simple linear relation with prediction skill. A large EM^2 often leads to a high correlation skill and vise versa. When the skill is quantified by MSE-based measures, a "triangular relationship is suggested between EM^2 and model skill. When EM^2

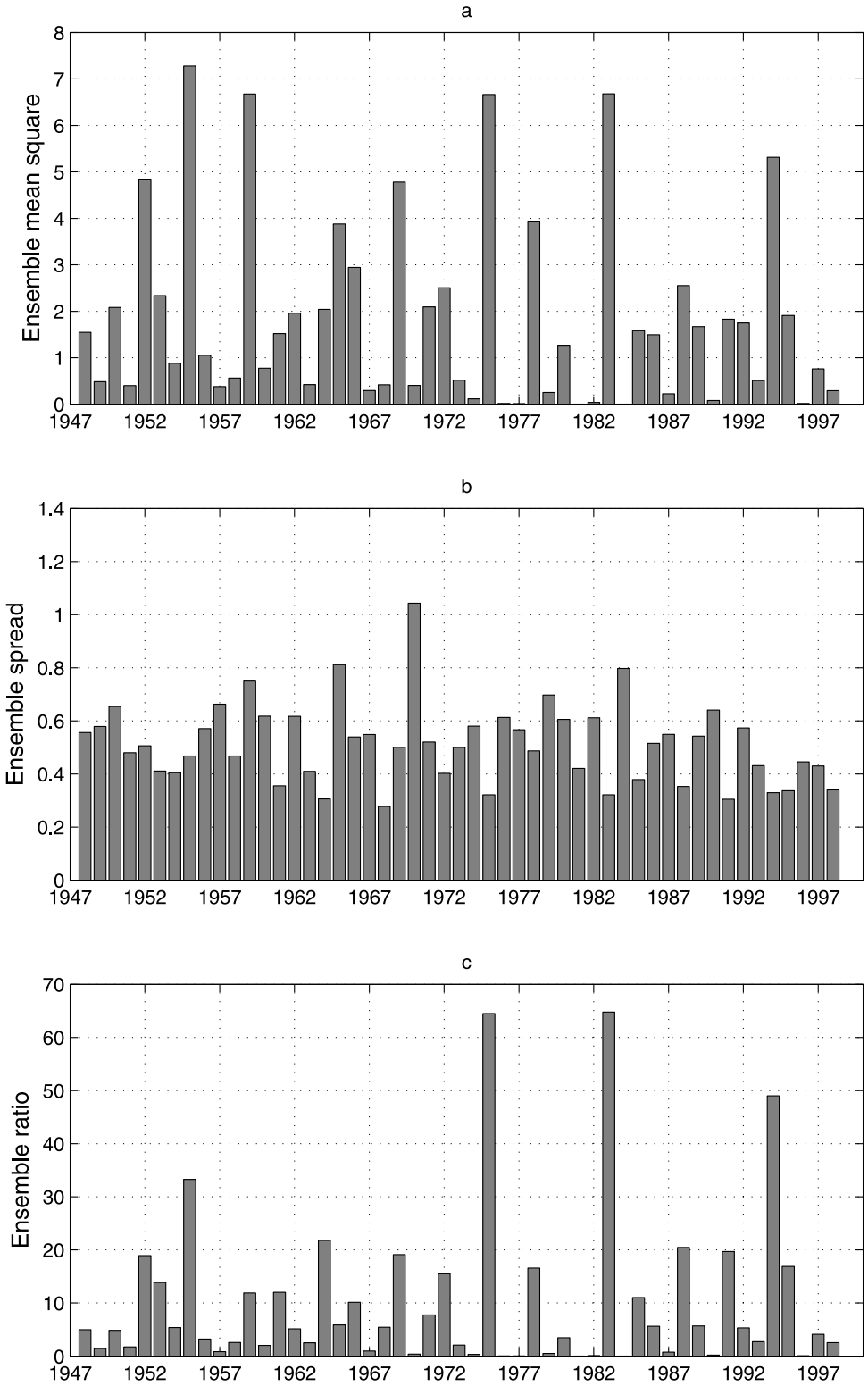

Figure 8. Same as Figure 5 but for SGCM for AO prediction.

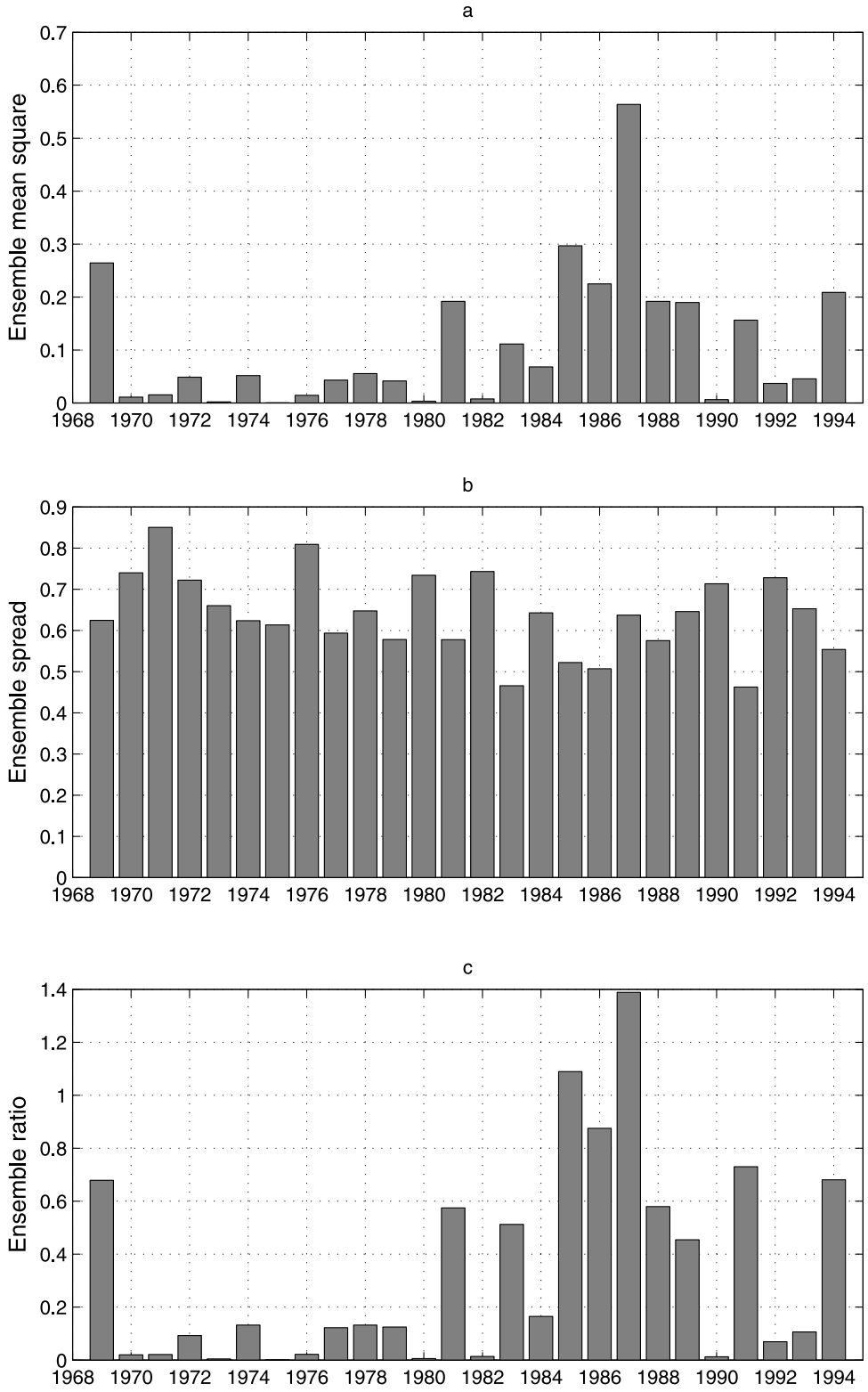

Figure 9. Same as Figure 8 but for GCM2.

Table 2. Correlation Skills Between Predicted and Observed AO Indices as a Function of EM^2[a]

M_{sgcm}^2	M_{gcm2}^2	SGCM ($EM^2 > M_{sgcm}^2$)	GCM2 ($EM^2 > M_{gcm2}^2$)
0.1	0.025	0.45 (44)	0.51 (18)
1.0	0.05	0.57 (27)	0.64 (13)
2.0	0.075	0.66 (16)	0.67 (10)
2.5	0.1	0.79 (12)	0.69 (9)

[a]The number shown in parentheses is the number of samples used. M^2 denotes the threshold values of AO index amplitude.

is large, the prediction is typically good whereas when EM^2 is small, the prediction skill is more variable.

5.2. Ensemble Spread

[50] The ensemble spread is primarily a measure of prediction uncertainty in NWP. One might expect that a large ensemble spread corresponds to a relatively low prediction skill, while a small ensemble spread is associated with a relatively high forecast skill.

[51] Figures 5b and 6b show the variation of ensemble spread for HCM1 and HCM2. The results indicate that both models have a relatively small variation in the ensemble spread compared with that in the ensemble mean square, especially for HCM2. Comparing the ensemble spread with any measures of prediction skill studied here, one might conclude that the ensemble spread is not an effective indicator of prediction skill. Figures 13 and 14 show the variation of two measures of skill (C and F) with ensemble spread for the four models. In Figure 13, there does not exist a statistically significant relationship between the ensemble spread and C in all models. As shown in the top right corner, the correlation between C and the ensemble spread is very small for all cases. In Figure 14, there seems some relationship between the ensemble spread and F for HCM1 but for other three models there is no relationship between them. In fact, there seems to exist an opposite "triangular relationship" between F and the ensemble spread for SGCM and GCM2, namely, that a large ensemble spread corresponds with a small F; whereas when the ensemble spread is small, the F skill is much more variable. Obviously this is not consistent with the notion of ensemble spread and F, suggesting that the ensemble spread is not a good predictor in quantifying climate prediction skill.

[52] The above finding provides a striking counterexample to the widespread perception that ensemble spread is the main determinant of potential forecast skill for NWP models. This is most probably due to two reasons: (1) small ensemble spread and (2) weak variation in ensemble spread.

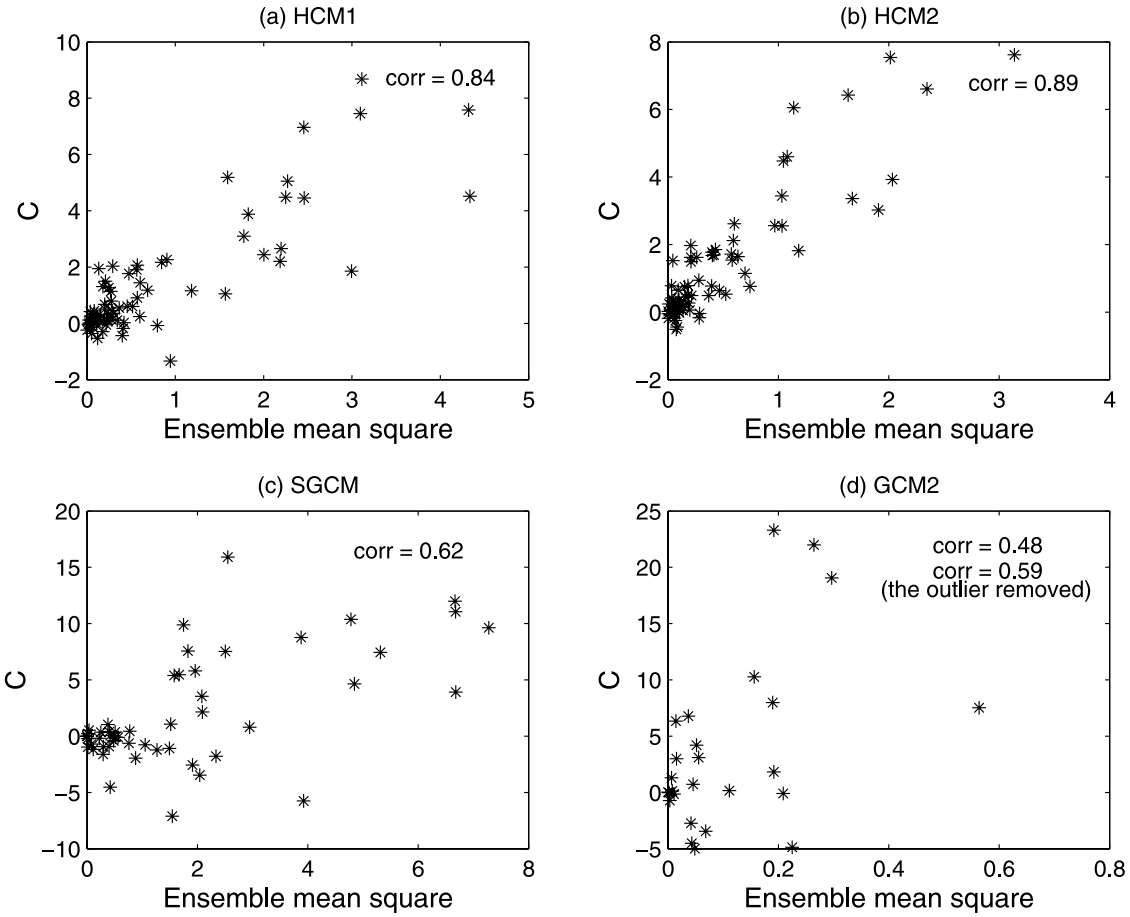

Figure 10. Scatterplots of ensemble mean square (EM^2) with C for (a) HCM1, (b) HCM2, (c) SGCM and (d) GCM2. For both HCM1 and HCM2, the average EM^2 and average C over the 12 months of lead time are plotted.

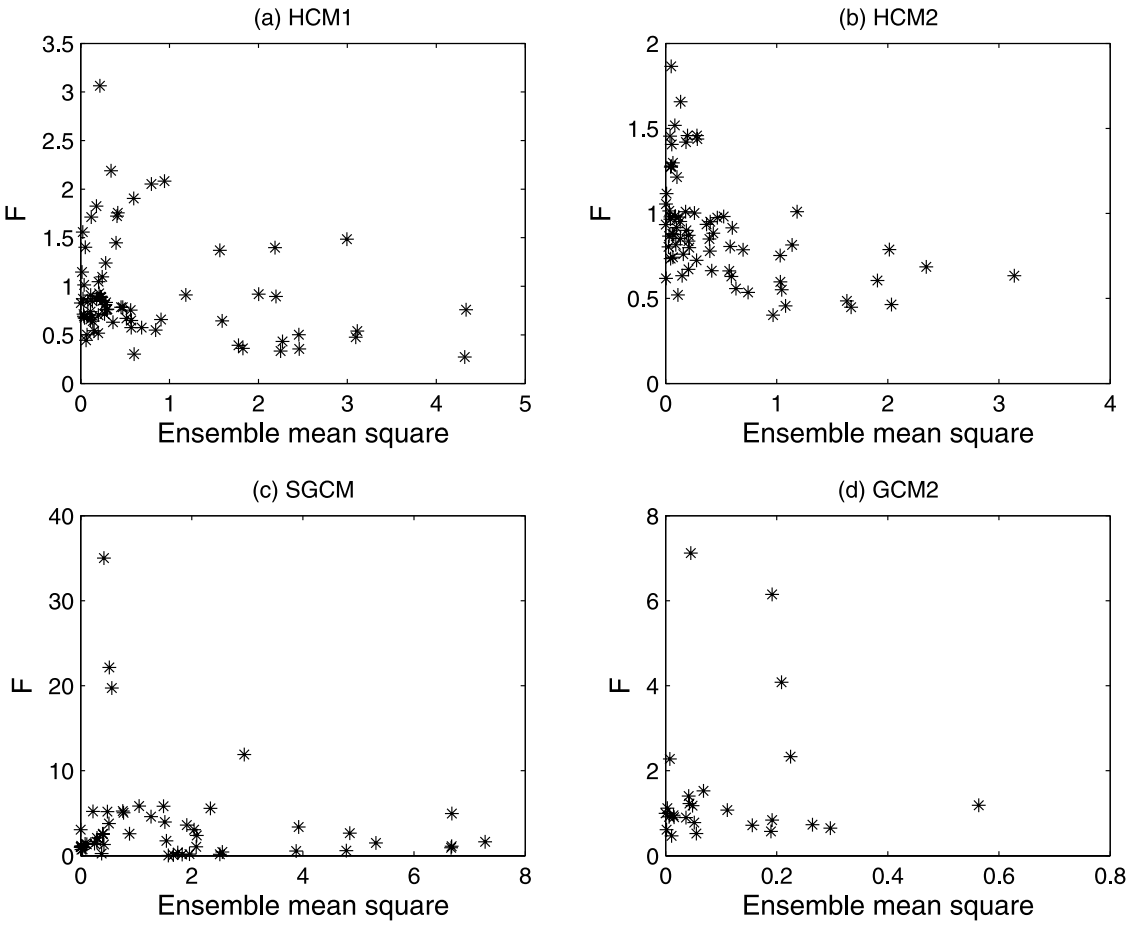

Figure 11. Scatterplots of ensemble mean square (EM^2) with F for (a) HCM1, (b) HCM2, (c) SGCM and (d) GCM2. For both HCM1 and HCM2, the average EM^2 over the 12 months of lead time is plotted.

As shown in Figures 5b, 6b, 8b, and 9b, the ensemble spreads found in these models are either relatively much small or vary little from prediction to prediction (also see section 5.4). Under the assumption of a perfect ensemble, *Houtekamer* [1993] derived an analytical relationship between the ensemble spread and the model skill that is a function only of the variability of ensemble spread (β). When β is very small, the spread is approximately constant, and the prediction error is a random draw from a fixed distribution, so the correlation approaches zero [*Whitaker and Loughe*, 1998].

[53] One interesting question is whether the feature of the ensemble spread found in this study is common for all ECPs? A complete answer would require us to explore other climate models. However, a close inspection of the models and ensemble methods used in this study might be able to shed some light on this question. As argued before, the variation of spread usually depends on two terms: (1) initial perturbation that generates the ensemble and (2) the growth of initial perturbation with time associated with model dynamics and thermodynamics. In NWP, both terms play an important role and are highly dependent on initial conditions. However in ECP, the initial conditions have a relatively small contribution to the growth of initial perturbations, and the growth of initial perturbations is probably dominated by the model behavior, leading to little variation of the ensemble spread with predictions. This might be true since for ECPs, the impact of initial conditions on perturbations (random forcing) could be dissipated by chaotic components of model system in a relative long-term climate prediction runs (also see next section). In addition, climate prediction is usually represented using a long time average (e.g., 1 month mean for HCMs and 3-month mean for SCGM and GCM2), which can largely remove the uncertainty related to initial conditions, and all the members converge to a state that is more determined by the model climate under a specific boundary condition. Therefore the suggestion that a small ensemble spread is likely a common feature for all ECPs is well supported by our results. In this study, we used four models with different complexity, ranging from a simple AGCM, hybrid coupled models, to a full AGCM. The ensemble methods also cover a broad degree of variation, from a well-designed random scheme to stochastical optimal perturbation method. However, the ensemble spread in all these ECPs shows a very similar feature, i.e., little variation with initial conditions. Thus the nature of ensemble spread in a NWP may be different from that in a ECP. The former may be greatly affected by initial conditions whereas the latter is dominated by model dynamics.

5.3. Ensemble Ratio

[54] The ratio of signal over noise has also been a widely used measure in quantifying predictability. We examined

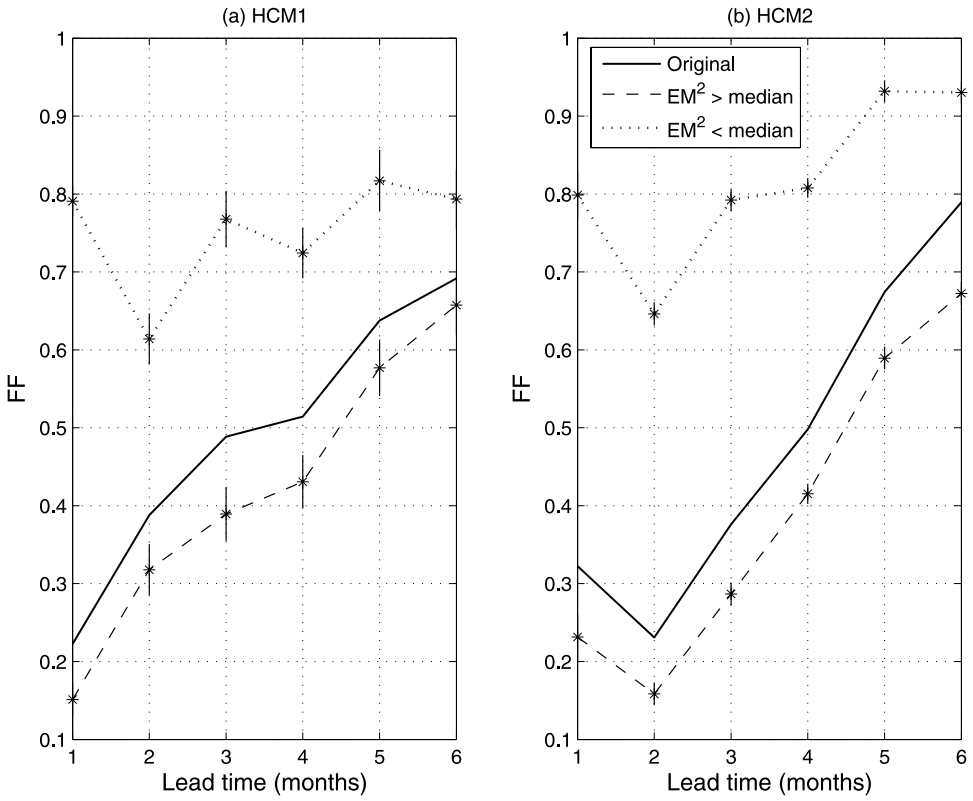

Figure 12. Same as Figure 7 but for the MSE-based measure FF.

the relationship between the ensemble ratio and the model skill following the same framework as above. Results show that the ensemble ratio is a useful measure of the forecast skill, but not as good as the ensemble mean square. This can be concluded by comparing Figures 15 and 16 with Figures 10 and 11. A probable reason is that the noise component (ensemble spread) degrades the capability of the ensemble ratio in quantifying the forecast skill.

5.4. Further Statistical Interpretation

[55] In this subsection, we will show that the above results can be characterized fairly well using a Gaussian framework with constant variances. First, we will examine if ensemble spread is little varied from prediction to prediction in the statistical sense. To answer this, A statistical F test of variance was performed for each model. For an individual model, the median value of all ensemble variances, denoted by σ, was chosen as a reference value. The F test, examining if there is significant difference between one individual variance s and the σ, was carried out for all ensemble variances and for all models. With the significance level of 0.1, the results are shown in Table 3. The percentage shown is the ratio of the number of ensemble variances that have no significant difference from the reference variance σ to the number of all ensemble variances. For HCM1 and HCM2, the averaged ensemble variance over 12-month leading time was used for the F test. As can be seen, most of ensemble variances have no statistical significant difference from the σ for each model, especially for HCM1, HCM2 and GCM2, indicating little ensemble spread variability in the four models.

[56] The ENSO and AO ensemble predictions could be thought of as an approximate Gaussian process, which has been validated in HCM1, HCM2 and SGCM [*Tang et al.*, 2005, 2007]. (A Kolmogorov-Smirnov normality test was also performed for ensemble predictions of GCM2. The result shows that all ensemble predictions pass the test at the significance level of 0.1.) For normally distributed variables, the contribution C is proportional to $\mu(\mu+\varepsilon)$ where μ is the ensemble mean EM and the quantity $\mu+\varepsilon$ is the observation. The observation is the ensemble mean plus a noise term with mean zero, $\langle\varepsilon\rangle = 0$. (This is valid under a perfect model approximation, where the observation can be treated as a member of the integration.) The variance $\langle\varepsilon^2\rangle$ of the noise term determines the correlation between observation and ensemble mean. Similar to the relationship between C and relative entropy [*Tang et al.*, 2007], the square of the correlation between EM^2 and correlation contribution C is

$$
\begin{aligned}
\rho c^2 &= \frac{\langle(\mu^2-\langle\mu^2\rangle)(\mu(\mu+\varepsilon)-\langle\mu^2\rangle)\rangle^2}{\langle(\mu^2-\langle\mu^2\rangle)^2\rangle\langle(\mu(\mu+\varepsilon)-\langle\mu^2\rangle)^2\rangle} \\
&= \frac{\langle(\langle\mu^2\rangle-\mu^2)^2\rangle^2}{\langle(\mu^2-\langle\mu^2\rangle)^2\rangle\langle\langle\mu^2\rangle^2+\varepsilon^2\mu^2-2\langle\mu^2\rangle\mu^2+\mu^4\rangle} \\
&= \frac{\langle(\langle\mu^2\rangle-\mu^2)^2\rangle}{\langle\langle\mu^2\rangle^2+\varepsilon^2\mu^2-2\langle\mu^2\rangle\mu^2+\mu^4\rangle} \\
&= \frac{\langle\mu^4\rangle-\langle\mu^2\rangle^2}{\langle\mu^4\rangle-\langle\mu^2\rangle^2+\langle\varepsilon^2\rangle\langle\mu^2\rangle} \\
&= \frac{1}{1+\dfrac{\langle\varepsilon^2\rangle\langle\mu^2\rangle}{\langle\mu^4\rangle-\langle\mu^2\rangle^2}} = \frac{1}{1+\dfrac{1}{2}\dfrac{\langle\varepsilon^2\rangle}{\langle\mu^2\rangle}}
\end{aligned}
$$

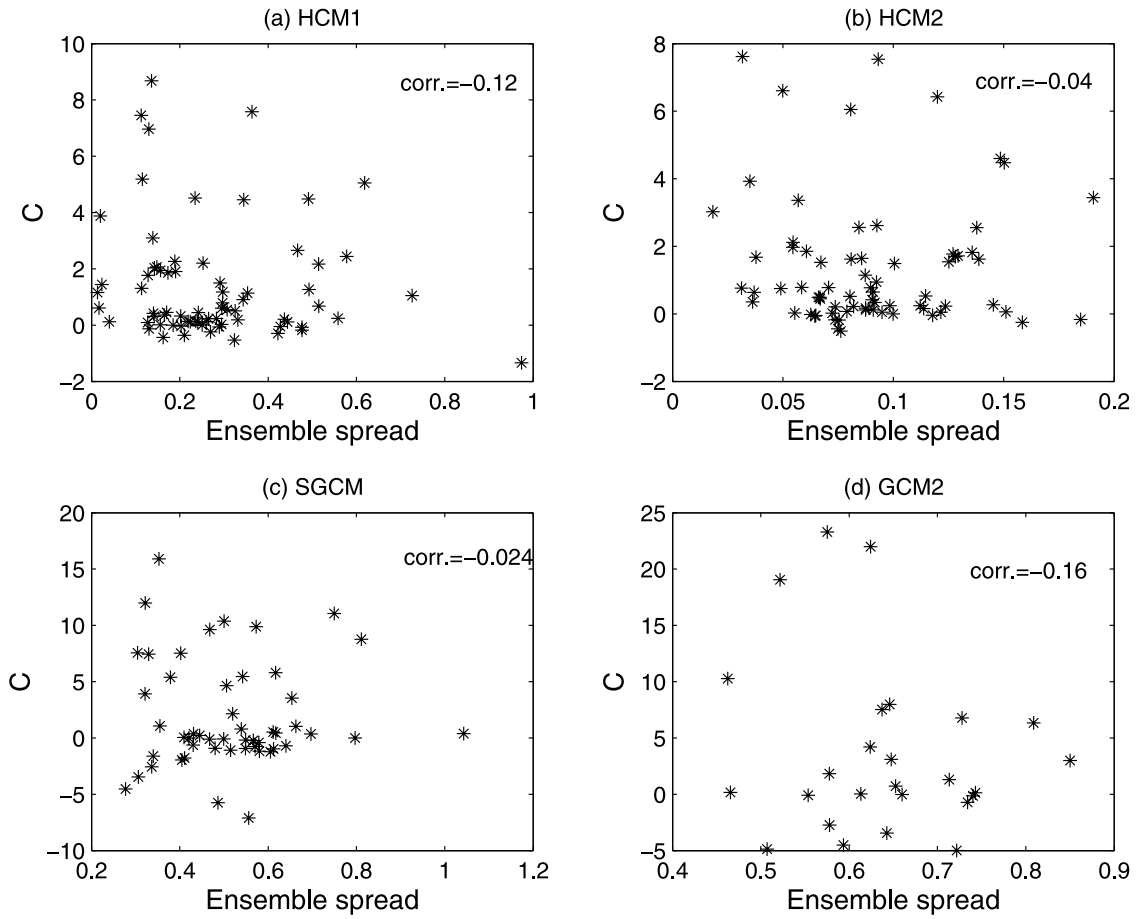

Figure 13. Scatterplots of ensemble spread with *C* for (a) HCM1, (b) HCM2, (c) SGCM and (d) GCM2. For both HCM1 and HCM2, the average *ES* over the 12 months of lead time is plotted.

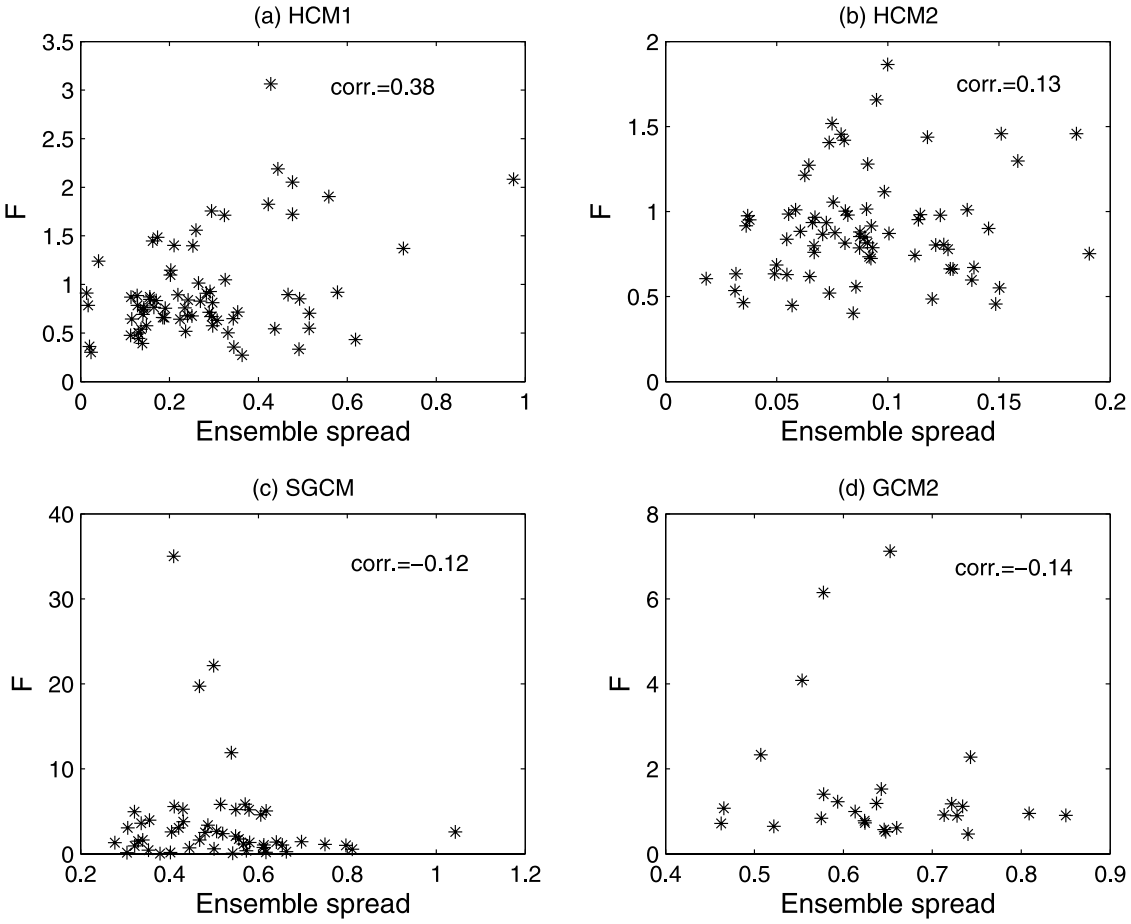

Figure 14. Same as Figure 13 but F instead of C.

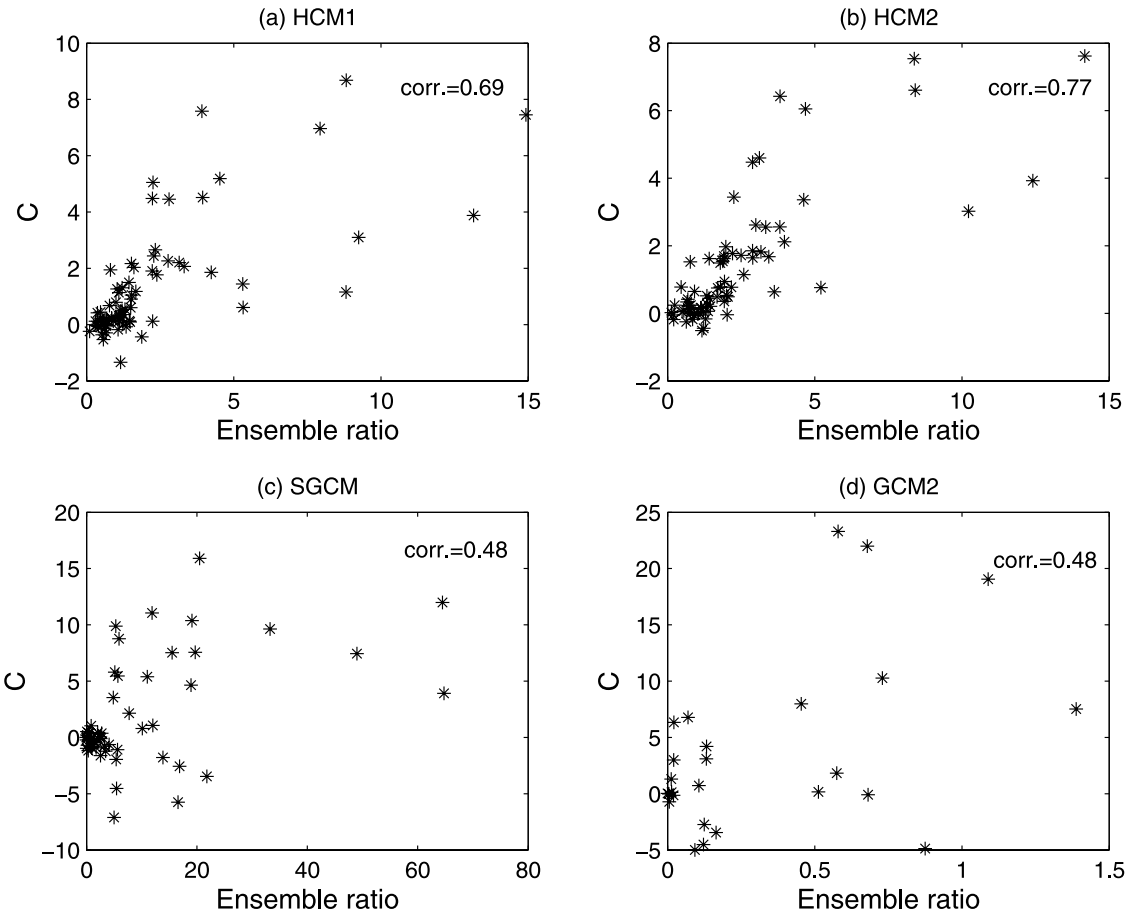

Figure 15. Scatterplots of ensemble ratio to correlation contribution (C) for (a) HCM1, (b) HCM2, (c) SGCM and (d) GCM2. For both HCM1 and HCM2, the average EM^2 over the 12 months of lead time is plotted.

where $\langle \ldots \rangle$ denotes the expectation, and we use the fact that $\langle \mu^4 \rangle = 3\ \langle \mu^2 \rangle^2$ for normally distributed variables. A similar calculation shows that the correlation r is related to the signal-to-noise ratio $\langle \mu^2 \rangle / \langle \varepsilon^2 \rangle$ by *Kleeman and Moore* [1997],

$$r^2 = \frac{1}{1 + \frac{\langle \varepsilon^2 \rangle}{\langle \mu^2 \rangle}}.$$

[57] Since

$$\frac{\langle \varepsilon^2 \rangle}{\langle \mu^2 \rangle} = \frac{1}{r^2} - 1,$$

$$\rho c^2 = \frac{1}{1 + \frac{1}{2}\left(\frac{1}{r^2} - 1\right)} = \frac{2r^2}{2r^2 + 1 - r^2} = \frac{2r^2}{r^2 + 1}. \qquad (12)$$

[58] Equation (12) produces a theoretical relationship between ρc (the correlation between C and EM^2) and r (the correlation skill of prediction). When the r is small the ρc is also small, and verse visa. Therefore (12) builds a theoretical base for the analyses performed in section 5.1, namely, that the relationship between C and EM^2 is able to measure model prediction skills.

[59] It is interesting to compare ρc from (12) with its actual counterpart calculated in section 5.1, by which we can see how well the predictability of these forecast models can be described by a Gaussian framework with constant variances. The actual ρc obtained in section 5.1 was 0.84 for HCM1, 0.89 for HCM2, 0.62 for SGCM, 0.59 for GCM2 (with the outlier point removed), as shown in Figure 10. Correspondingly, the ρc from (12) is respectively 0.90, 0.94, 0.54 and 0.60 for the four models, which are in good agreement with their actual counterparts. Such a good consistence between theoretical ρc and their actual values suggests that the predictability of these forecast models can be characterized fairly well using a Gaussian framework with constant variances.

5.5. Ensemble Size Sensitivity

[60] All the results reported above involve varied ensembles. It is interest from a practical perspective to see how robust the results are when smaller ensembles are used. Our finding that only the ensemble mean is responsible for variations of model skill suggests that a large ensemble size is not required to generate predictions and measure the potential predictability. Usually one can estimate reasonably well the mean of the distribution with few samples if the distribution has small variance as found in the four models. To examine the sensitivity of the predictability measure of EM^2 to the ensemble size, we calculated EM^2 using several different ensemble sizes for each model. The results showed

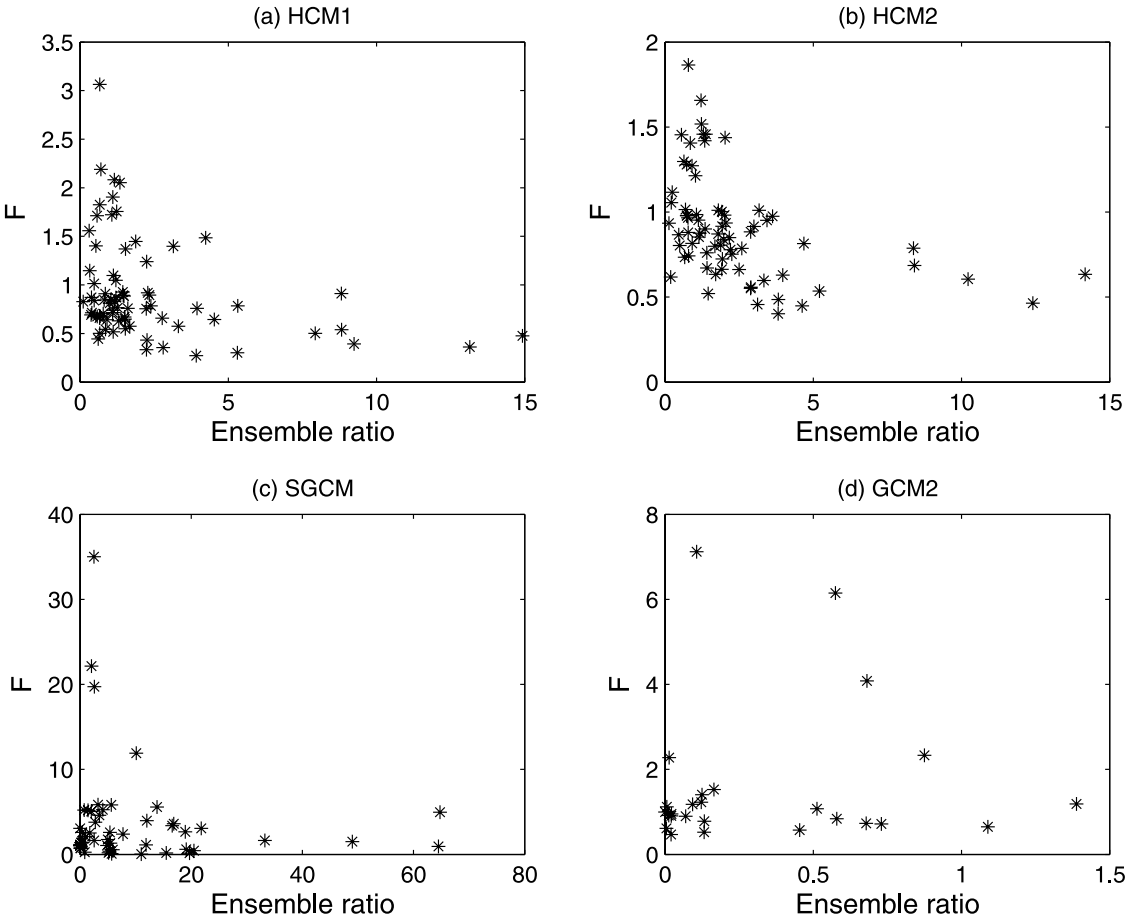

Figure 16. Same as Figure 15 but F instead of C.

that the EM^2 is little varied with the ensemble size when it is over 10–15 for all models. Shown in Figure 17 are the EM^2s estimated respectively by randomly chosen 15 ensemble members and by all members for the four models, where the vertical line superimposed onto the bar is the extent of uncertainty in the computed EM^2 due to the finite sample size that was estimated by the bootstrap method with 1000 members. As can be seen, the EM^2 estimated by the two ensemble sizes are very close, and their differences do not exceed the error scope due to the uncertainty of the finite sample size. We also repeated the other calculations performed in this study using different ensemble sizes, and found that the skills of predicted ENSO/AO index against observed ENSO/AO index vary little with the change of ensemble size when the ensemble members are over 10–15 (not shown). The relationship between the measures of potential predictability and prediction skill scores identified in this study was also little changed when the ensemble size was changed from all members to 15. For example, Figure 18 shows the scatterplot of correlation contribution C against the EM^2 using 15 ensemble members. As can be seen, Figure 18 is very similar to Figure 10, suggesting that the ensemble member of 10–15 is an appropriate size to generate predictions and measure the potential predictability in the four models. This is also consistent with an earlier finding by *Kumar and Hoerling* [1995] that a 10–15 member ensemble may be enough to infer the ensemble mean signal when the noise is quasi-invariant.

6. Summary and Discussion

[61] An important task of predictability studies is to measure the forecast uncertainty by ensemble prediction, i.e., to use ensemble predictions to determine a priori the likely skill of an individual prediction. In this study, we have explored the ENSO and AO predictability using multiple climate models with different complexity including a simple and a full global atmospheric general circulation model, and two hybrid coupled models. It was found that the ensemble mean square EM^2 can measure reasonably well the actual prediction skill of the ENSO and AO predictions. The relationship between EM^2 and skill depends on the measure of prediction skill. When the correlation-based measures are used, there is an identifiable

Table 3. F Statistical Test of Variances

Models	F Test Results
HCM1	85%
HCM2	94%
SGCM	59%
GCM2	92%

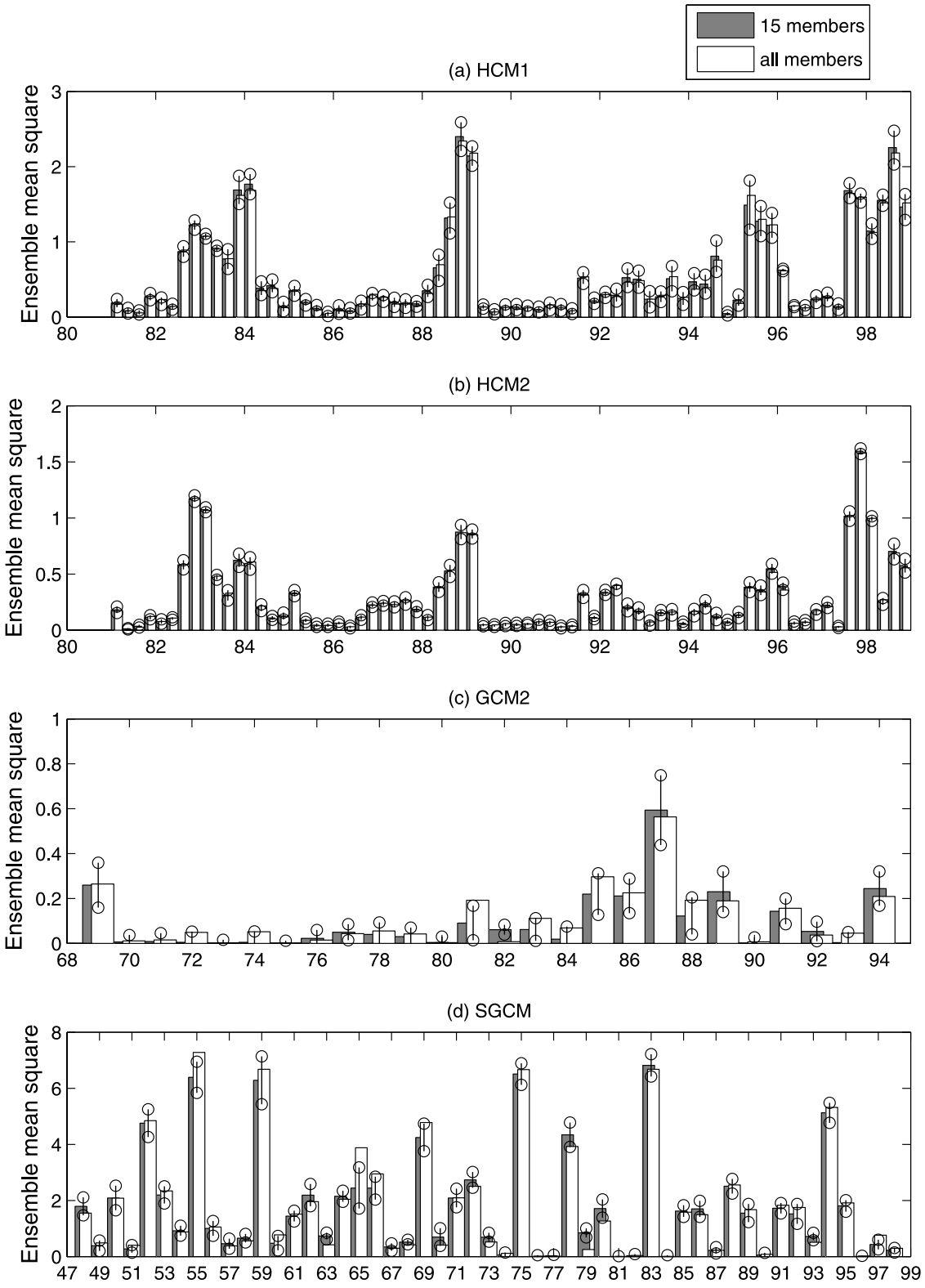

Figure 17. EM^2 estimated by randomly chosen 15 ensemble members (shaded bar) and by all ensemble members (open bar). The short line connected by circles is the uncertainty estimated by a bootstrap method with 1000 members.

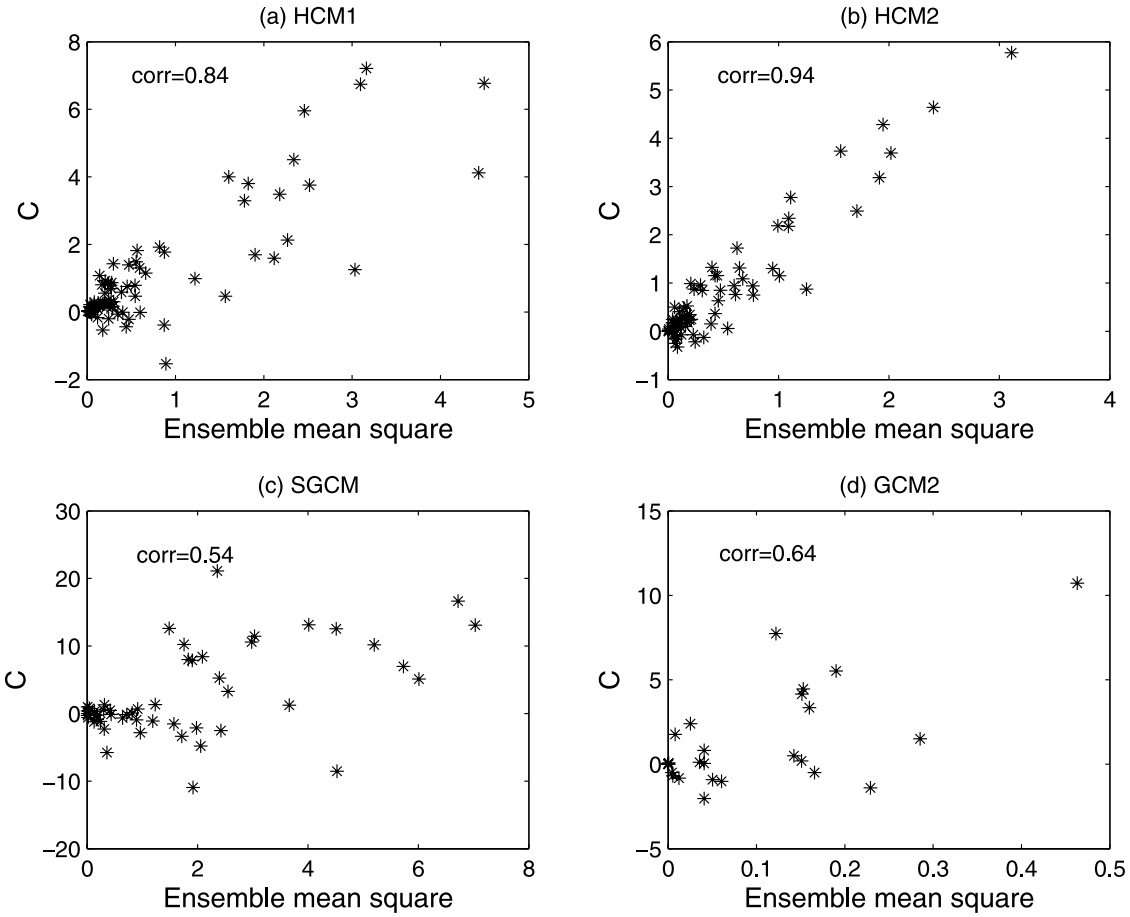

Figure 18. Same as Figure 10 but using a randomly chosen 15 members instead of all members to estimate EM^2 and C.

linear relationship between prediction skill and EM^2. When *MSE*-based measures are used a "triangular relationship" between the EM^2 and prediction skill is suggested, i.e., when EM^2 is large, the corresponding prediction is found to be reliable whereas when EM^2 is small, the prediction skill is found to be much more variable. It is also found that the ensemble spread is not an effective indicator to prediction skill in all four climate models. The predictability of these models can be characterized fairly well using a Gaussian framework with constant variances.

[62] Considerable effort has been expended in recent years to determine the predictor of forecast skill in ensemble NWP. The most widely used measures of potential predictability so far have been ensemble spread and the ensemble ratio of signal over noise. Thus the finding that the ensemble mean square is a very effective predictor of forecast skill in ECP has considerable practical significance for improving our capability of predicting climate variability and utilizing climate prediction, in particular for the ENSO interannual prediction and the AO seasonal prediction.

[63] It is of interest to further explore the underlying physical interpretation for the relationship between EM^2 and prediction skills. As found by *Kleeman* [2002] and *Tang et al.* [2005], the extra information that is provided by a prediction, defined by relative entropy, is dominated by the ensemble mean square when the ensemble spread is little varied. As the ensemble mean square is larger, more information will be produced compared with the climatological prediction, leading to a prediction that is likely reliable. The other interpretation is associated with signals present in the initial conditions. It has been found that the ensemble mean square is highly related to the eigenmode signals (i.e., the singular vector) of initial fields in many climate models [*Kleeman and Moore*, 1997; *Tang et al.*, 2004]. It has been well argued that the periods during which the slowly decaying eigenmodes are present with large amplitude are periods which should be intrinsically more predictable because such modes are able to "resist" dissipation by the more chaotic components of the system.

[64] A practical significance of this work is to use climate models to issue probabilistic forecasts operationally. It was found that only the ensemble mean is responsible for variations of model skill, and that the ensemble member of 10–15 is an appropriate size to generate predictions and measure the potential predictability in the four models. The ECPs studied in the four models can be thought of an approximate Gaussian process as argued in section 5.4. For normally distributed variables, their probability density function is only determined by the mean and variance. This suggests that as few as 10–15 members may be appropriate to approach probabilistic forecasts operationally using these models. A detailed analysis and study on probability climate prediction using these models is underway.

[65] **Acknowledgments.** This work is supported by the Canadian Foundation for Climate and Atmospheric Sciences (CFCAS) through grant GR-523 and Canada Research Chair Program. Y.T. thanks M. Tippett for his valuable contribution to equation (12).

References

Blanke, B., and P. Delecluse (1993), Variability of the tropical Atlantic ocean simulated by a general circulation model with two different mixed-layer physics, *J. Phys. Oceanogr.*, *23*, 1363–1388.

Boer, G. J., G. Flato, M. C. Reader, and D. Ramsden (2000), A transient climate change simulation with greenhouse gas and aerosol forcing: Experimental design and comparison with the instrumental record for the 20th century, *Clim. Dyn.*, *16*, 405–425.

Buizza, R., and T. N. Palmer (1998), Impact of ensemble size on ensemble prediction, *Mon. Weather Rev.*, *126*, 2503–2518.

Collins, M., and M. R. Allen (2002), Assessing the relative roles of initial and boundary conditions in interannual to decadal climate predictability, *J. Clim.*, *15*, 3104–3109.

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

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

Derome, J., G. Brunet, A. Plante, N. Gagnon, G. J. Boer, F. W. Zwiers, S. Lambert, J. Sheng, and H. Ritchie (2001), Seasonal predictions based on two dynamical models, *Atmos. Ocean*, *39*, 485–501.

Derome, J., H. Lin, and G. Brunet (2005), Seasonal forecasting with a simple general circulation model: Predictive skill in the AO and PNA, *J. Clim.*, *18*, 597–609.

Epstein, E. S. (1969), Stochastic dynamic prediction, *Tellus*, *21*, 739–759.

Farrell, B. F., and P. J. Ioannou (1993), Stochastic dynamics of baroclinic waves, *J. Atmos. Sci.*, *50*, 4044–4057.

Flato, G. M., G. J. Boer, W. Lee, N. McFarlane, D. Ramsden, and A. Weaver (2000), The CCCma global coupled model and its climate, *Clim. Dyn.*, *16*, 451–467.

Hall, N. M. J. (2000), A simple GCM based on dry dynamics and constant forcing, *J. Atmos. Sci.*, *57*, 1557–1572.

Hoskins, B. J., and A. J. Simmons (1975), A multi-layer spectral model and the semi-implicit method, *Q. J. R. Meteorol. Soc.*, *101*, 637–655.

Houtekamer, P. L. (1993), Global and local skill forecasts, *Mon. Weather Rev.*, *121*, 1834–1846.

Kalnay, E., et al. (1996), NCEP/NCAR 40-year reanalysis project, *Bull. Am. Meteorol. Soc.*, *77*, 437–471.

Kleeman, R. (1989), A modeling study of the effect of the Andes on the summertime circulation of tropical South America, *J. Atmos. Sci.*, *46*, 3344–3362.

Kleeman, R. (2002), Measuring dynamical prediction utility using relative entropy, *J. Atmos. Sci.*, *59*, 2057–2072.

Kleeman, R., and A. M. Moore (1997), A theory for the limitation of ENSO predictability due to stochastic atmospheric transients, *J. Atmos. Sci.*, *54*, 753–767.

Kleeman, R., and A. J. Majda (2005), Predictability in a model of geostrophic turbulence, *J. Atmos. Sci.*, *62*, 2864–2879.

Kumar, A., and M. P. Hoerling (1995), Prospects and limitations of seasonal atmospheric GCM predictions, *Bull. Am. Meteorol. Soc.*, *76*, 335–345.

Kumar, A., and M. P. Hoerling (2000), Analysis of a conceptual model of seasonal climate variability and implications for seasonal predictions, *Bull. Am. Meteorol. Soc.*, *81*, 255–264.

Kumar, A., A. B. Barnston, P. Peng, M. P. Hoerling, and L. Goddard (2000), Changes in the spread of the variability of the seasonal mean atmospheric states associated with ENSO, *J. Clim.*, *13*, 3139–3151.

Leith, C. E. (1974), Theoretical skill of Monte Carlo forecasts, *Mon. Weather Rev.*, *102*, 409–418.

Lorenz, E. N. (1969), The predictability of a flow which possesses many scales of motion, *Tellus*, *21*, 289–307.

Madec, G., P. Delecluse, M. Imbard, and C. Levy (1998), OPA 8.1 Ocean General circulation model reference manual, 91 pp., Institut Pierre Simon Laplace, Paris.

Molteni, R., and T. N. Palmer (1993), Predictability and finite-time instability of the northern winter circulation, *Q. J. R. Meteorol. Soc.*, *119*, 269–298.

Molteni, R., R. Buizza, and T. N. Palmer (1996), The ECMWF ensemble prediction system: Methodology and validation, *Q. J. R. Meteorol. Soc.*, *122*, 73–119.

Moore, A. M., and R. Kleeman (1998), Skill assessment for ENSO using ensemble prediction, *Q. J. R. Meteorol. Soc.*, *124*, 557–584.

Moore, A., J. Zavala-Garay, Y. Tang, R. Kleeman, J. Vialard, A. Weaver, K. Sahami, L. T. Anderson, and M. Fisher (2006), Optimal forcing patterns for coupled models of ENSO, *J. Clim.*, *19*, 4683–4699.

Murphy, J. M. (1988), The impact of ensemble forecasts on predictability, *Q. J. R. Meteorol. Soc.*, *114*, 463–493.

Murphy, A. H., and E. S. Epstein (1989), Skill scores and correlation coefficients in model verification, *Mon. Weather Rev.*, *117*, 572–581.

Palmer, T. N. (1999), Predicting uncertainty in forecast of weather and climate, *ECMWF Tech. Memo. 294*, Eur. Cent. for Med.-Range Weather Forecasts, Reading, U. K.

Peng, P., and A. Kumar (2005), A large ensemble analysis of the influence of tropical SSTs on seasonal atmospheric variability, *J. Clim.*, *15*, 1068–1085.

Scherrer, S., C. Appenzeller, P. Eckert, and D. Cattani (2004), Analysis of the spread-skill relations using the ECMWF ensemble prediction system over Europe, *Weather Forecasting*, *19*, 552–565.

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

Tang, Y., R. Kleeman, A. M. Moore, A. Weaver, and J. Vialard (2003), The use of ocean reanalysis products to initialize ENSO predictions, *Geophys. Res. Lett.*, *30*(13), 1694, doi:10.1029/2003GL017664.

Tang, Y., R. Kleeman, and A. M. Moore (2004), A simple method for estimating variations in the predictability of ENSO, *Geophys. Res. Lett.*, *31*, L17205, doi:10.1029/2004GL020673.

Tang, Y., R. Kleeman, and A. M. Moore (2005), On the reliability of ENSO dynamical predictions, *J. Atmos. Sci.*, *62*, 1770–1791.

Tang, Y., H. Lin, J. Derome, and M. K. Tippett (2007), A predictability measure applied to seasonal predictions of the Arctic Oscillation, *J. Clim.*, *20*, 4733–4750.

Tippett, M. K., R. Kleeman, and Y. Tang (2004), Measuring the potential utility of seasonal climate predictions, *Geophys. Res. Lett.*, *31*, L22201, doi:10.1029/2004GL021575.

Toth, Z., and E. Kalnay (1993), Operational ensemble prediction at the National Meteorological Center. Practical aspects, *Bull. Am. Meteorol. Soc.*, *74*, 2317–2330.

Tribbia, J. J., and D. P. Baumhefner (1988), Estimates of the predictability of low-frequency variability with a spectral general circulation model, *J. Atmos. Sci.*, *45*, 2306–2317.

Vialard, J., P. Delecluse, and C. Menkes (2002), A modeling study of salinity variability and its effects in the tropical Pacific Ocean during the 1993–1999 period, *J. Geophys. Res.*, *107*(C12), 8005, doi:10.1029/2000JC000758.

von Storch, H., and F. Zwiers (1999), *Statistical Analysis in Climate Research*, 484 pp., Cambridge Univ. Press, Cambridge, U. K.

Whitaker, J. S., and A. F. Loughe (1998), The relationship between ensemble spread and ensemble mean skill, *Mon. Weather Rev.*, *126*, 3292–3302.

Xue, Y., M. A. Cane, S. E. Zebiak, and T. N. Palmer (1997), Predictability of a coupled model of ENSO using singular vector analysis part II: Optimal growth and forecast skill, *Mon. Weather Rev.*, *125*, 2057–2073.

H. Lin, Recherche Prévision Numérique, Meteorological Service of Canada, 2121 Voie de Service nord, Route Trans-Canadienne, Dorval, QC, Canada H9P 1J3.

A. M. Moore, Ocean Sciences Department, University of California, Santa Cruz, CA 95064, USA.

Y. Tang, Environmental Science and Engineering, University of Northern British Columbia, Prince George, BC, Canada V2N 4Z9. (ytang@unbc.ca)