Coupled Ocean–Atmosphere Forecasts in the Presence of Climate Drift

TIMOTHY N. STOCKDALE

European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom

(Manuscript received 18 June 1995, in final form 29 April 1996)

ABSTRACT

Two different coupled atmosphere–ocean GCMs are used to forecast SST anomalies with lead times of up to one year. The initialization procedure does not balance the ocean and atmosphere components, nor is the coupled model flux corrected to maintain the correct mean state. Rather, the coupled model is allowed to evolve freely during the forecast. The inevitable climate drift is estimated across an ensemble of forecasts and subtracted to give the true forecast. Although the climate drift is often bigger than the interannual signal, the method works. This is true for a drift toward both warmer and colder SSTs, as exemplified by the two models.

The best way of establishing the mean bias correction from a small sample of prior forecasts is discussed. In some circumstances the sample median may be a more robust estimator than the sample mean. For the limited set of forecasts here, use of the median bias in the cross-correlated forecasts reduces forecast error, when compared to use of the mean bias.

1. Introduction

Many modeling studies have addressed the problem of dynamical forecasts of El Niño. Often the models are highly simplified and designed to model only the behavior of anomalies in the ocean–atmosphere system; the mean state of the system is simply treated as a given. Examples of this approach are Cane et al. (1986), Kleeman (1993), and Balmaseda et al. (1994). The advantages of these models, simplicity and low cost, can exist together with a moderately high level of forecast skill, especially in the longer range. The disadvantages are an inevitable loss of skill due to the simplified physics and the lack of detail, manifested in the difficulty in beating persistence in the short range (less than 3 months, say). The assumption of a fixed mean state, and only a limited number of relevant physical processes acting around it, may also be a problem in the presence of longer-term shifts in the climate system.

Considerable effort has been expended, therefore, on the development of more comprehensive ocean–atmosphere models, which contain a much wider range of parameterizations, supposedly including all relevant processes. The models are generally set to simulate the full behavior of the system—that is, both the mean seasonal cycle and departures from it. This is a demanding task, and it is unfortunately the case that errors in simulating the mean state are often at least as large as the anomalies we are trying to predict. For a recent study of errors in the climate of coupled GCMs, see Mechoso et al. (1995). Further, these climate errors can often develop on a timescale of several months and therefore have the potential to seriously contaminate seasonal forecasts.

Various ways have been devised to circumvent this problem. The most common has been to abandon the attempt to model all parts of the problem and instead to impose the mean climate in some way. This typically involves substituting the deficient model wind stress climatology with an observed stress climatology, and making some corrections to the net surface heat flux, to ensure that the model temperature balance remains realistic. The result is basically an anomaly model, with an imposed and realistic mean climatology. Ji et al. (1994) use this approach, although to reduce rather than eliminate their coupled model drift. Flux correction of some sort is also common in climate change simulations (Sausen et al. 1988; Manabe et al. 1991; Murphy 1995). A different method was used by Latif et al. (1993). Here, the climatology of the coupled model was not altered or constrained. Instead the analysis of the initial state for the forecast was made using the mean state of the coupled system and the observed anomalies only. In this case the forecasts show no climate drift, nor is there any application of flux correction terms to the coupled model. But the mean state about which the calculation proceeds is wrong, and the observed anomalies may not be compatible with the mean state to which they are applied.

Corresponding author address: Dr. Timothy N. Stockdale, European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, Berkshire RG2 9AX, United Kingdom.
E-mail: net@ecmwf.int

In this paper a different approach is explored. The idea is very simple: it is to apply no corrections or fixes to either the analysis or the forecast of the coupled system, but instead to estimate the climate drift of the coupled system from a control set of forecasts and subtract it as an a posteriori correction. This is not particularly new: a similar concept underlies the use of control experiments in climate change experiments, and Ji et al. (1994) use a mean bias correction in their forecasts in addition to flux corrections. The point at issue is whether estimating and removing forecast bias is, on its own, sufficient to deal with model drift in an ENSO forecasting environment. An associated question is how bias removal may best be done, and what issues are involved.

To the extent that the processes responsible for El Niño are linear, the three possibilities for coping with model drift are equivalent: flux correction, anomaly-only initialization, and bias correction. But there are significant nonlinearities in the tropical system, and these three methods are therefore likely to give different results. A strict comparison of the methods would be a substantial undertaking, and it is not clear whether the results from a single model system would be generally applicable or even statistically significant in their own right, given the sampling possible in our short historical record. The more modest aim of this paper is to establish the credentials of the drift subtraction method, by applying it to two different coupled GCM forecasting systems. The two systems chosen have substantial and opposite biases, thus allowing some assessment of the range of validity of the method. It is anticipated that any operational use of this technique would be in a situation where model errors were smaller than those considered here.

Section 2 describes the coupled models used, while section 3 relates how the models were initialized and the forecasts made. The results from the forecasts are described next, and then in section 5 further consideration is given to methods of estimating the forecast bias. Section 6 concludes.

2. The coupled model

The first system with which forecasts were made uses the first version of a coupled GCM developed at the European Centre for Medium-Range Weather Forecasts (ECMWF) in recent years. The ocean component is HOPE, the Hamburg Ocean Primitive Equation model, as used by Latif et al. (1994). It is a global z-coordinate model with 20 vertical levels, 0.5° meridional resolution in the equatorial waveguide stretching to 2.8° at higher latitudes, 2.8° zonal resolution, continuously variable bottom topography, and various contributions to the parameterization of mixing processes. The atmosphere model used is a modified T21 version of the ECMWF forecast model known as cycle 44, which was in operational use in the latter part of 1993. Modifications were made to the horizontal diffusion scheme and the value of a parameter in the diagnostic cloud scheme. The tuning and performance of the coupled system is described in Stockdale et al. (1994).

The second set of experiments used basically the same atmosphere model, although the parameter for maximum subgrid-scale convective cloud cover was increased from 16% to 28%. Adjusting this particular parameter is an easy way of controlling the overall temperature balance of the Tropics in the coupled ocean–atmosphere system. The ocean used was HOPEv2, a model that has characteristics substantially changed from those of the original HOPE model. The biggest change is that advection of temperature and salinity now uses centered differencing rather than upstream, and this results in a substantially less diffusive system. (It was the corresponding reduction in the mixing of heat down through the thermocline that required the retuning of the atmosphere cloud cover.) Parallel with this change, the parameterizations of viscosity were retuned, primarily with an eye on the Equatorial Undercurrent. The surface mixing scheme was altered to give a smoother decay with depth of wind-driven turbulence in the case of deep mixed layers. Numerous other small changes and corrections were also incorporated. Although HOPEv2 is essentially a global model, in the second set of experiments the ocean domain was restricted to the equatorial Pacific, from 120°E to 90°W and between 24°N and 24°S. This was for compatibility with a set of data assimilation experiments, which were run in a less computationally powerful environment, the results of which will be reported elsewhere. The effect of the domain restriction has only a limited effect on the ocean model itself, but the imposition of fixed rather than variable SST outside of the ocean model domain can impact significantly on the dynamics of the coupled system.

The two coupled models have quite different characteristics when they are run for an extended integration. Figure 1 shows a plot of 12-month running mean Nino-3 SST from both models, compared to the observed value of Nino-3 over recent years. It can be seen that the first model has a cold bias in its equatorial surface temperature. Using years 6–15 to define a climatology, there is an average cold bias of 1.2°C in the tropical Pacific as a whole. The second model, however, is slightly too warm, with an annual mean bias in the same period of plus 0.15°C, averaged over the tropical Pacific. The warming in Nino-3 is almost 0.6°C. As regards the east–west temperature gradient across the Pacific, defined as the temperature difference between 160°E and 120°W, the first model is about 10% too weak, while the second model appears to be about 5% too strong, although the ''observed'' estimate is probably weakened by smoothing contained in the SST climatology. In terms of overall spatial structure of SST, the first model shows an excessive cold tongue along the equator, whereas the second model is fairly realistic. A fuller description of the climatology of the first model can be found in Mechoso et al. (1995).

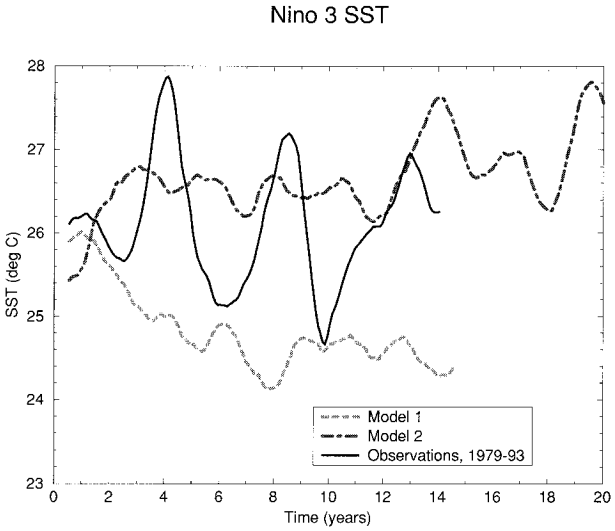

FIG. 1. Twelve-month running means of Nino-3 SST (5°N–5°S, 90°W–150°W) from extended integrations of the two forecast models. Shown for comparison is the observed SST for the period 1979–93.

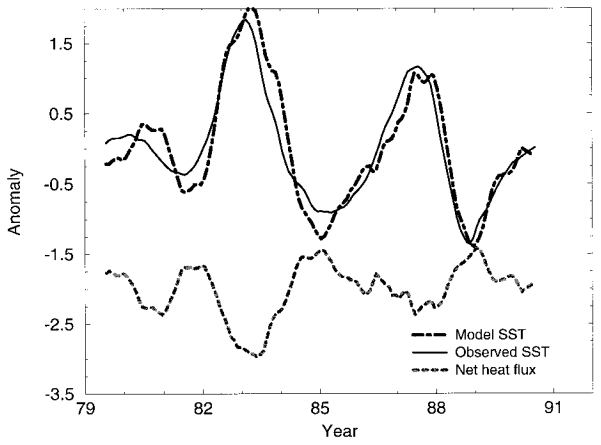

FIG. 2. Twelve-month runing mean of SST anomalies from the ocean model forced by ECHAM3 fluxes, compared to observed SST, for the period 1979–90. Also shown, offset and scaled, is the net heat flux anomaly into the ocean. The heat flux damps the SST anomalies rather than causes them.

The two models differ not only in their mean state, but also in their variability. The first model produces only very weak amplitude ENSO events, while the second model is able to produce relatively large amplitude disturbances. For the second model, in fact, small adjustments to the cloud parameters can give either a system that is highly active from the start, or one that never produces large anomalies. The version chosen for the forecasts, and therefore shown here, is just below the point at which large variability is always present. The apparent switch on of larger variability after 12 years is perhaps due to a slow drift in the ocean model taking the coupled system past a bifurcation point. This is very interesting, but beyond the scope of this paper. The point to be made is that the second model is close to producing large Nino-3 SST anomalies, while experiments with the first never gave any indication of large anomalies being possible.

Given the different model performances in these extended integrations, it can be anticipated that their behavior in forecast mode will also differ substantially. This will provide a good test of the proposed method of drift subtraction.

3. Forecast methodology

Coupled GCM forecasts of ENSO require an initial state. This was provided for the ocean simply by driving the model with appropriate surface forcing up to the start time of the forecast; no assimilation of subsurface ocean data was used. The ocean forcing consists of daily averaged fluxes of momentum, heat, and solar radiation from an extended integration of an atmosphere GCM. The atmosphere model used for this was the T42 ECHAM3 model, integrated from January 1979 up to June 1993. The idea behind using model simulated forcing is that the large-scale variability of the tropical atmosphere is strongly controlled by the SST, and thus the wind stress anomalies produced by the model should match observations, at least for suitably large time and space scales. Advantages of model-derived forcing fields are consistency and completeness. Disadvantages are the risk of bias and other systematic errors, and a reduction in skill in specifying the monthly and seasonal timescale of wind variations. Note that the patchy nature of the observations prior to the TAO array mean that this skill reduction may not be much worse than that found in observationally based forcing products.

In addition to the ECHAM3 fluxes of momentum and heat, the ocean model is given a heat flux that relaxes the model temperature toward the instantaneously observed SST. This feedback term ensures the stability of model SST. If both ocean and atmosphere models are behaving perfectly, the feedback term will equal zero, at least when indeterminacy and observational error are ignored. The first forecast experiments used initial states created with a feedback of 40 W m^{-2} $°C^{-1}$, whereas the second used states created with a value of 400 W m^{-2} $°C^{-1}$. The main result of this is that the second set of forecasts is more accurate in the short range, simply because the initial SSTs match observations more closely.

The results of using ECHAM3 forcing to initialize the ocean models can be anticipated in part by looking at Fig. 2. This shows the simulated interannual SST anomaly in Nino-3 (5°N–5°S, 150°–90°W) compared to the observations, for a model run of HOPEv2 using a negative feedback of 40 W m^{-2} $°C^{-1}$. Although this feedback is a fairly strong constraint on SST in passive areas of the ocean, in active areas such as Nino-3 it has only a limited effect. Figure 2 also shows the net heat flux acting on the ocean model, which is the sum of the ECHAM3 flux and the feedback term. It is clear that

the heat flux is not driving the interannual SST variability, which is rather determined by local and remote wind stress changes. In fact, the heat flux is an almost perfect damping of the model SST, the anomaly correlation for the curves in Fig. 2 being −0.95. A further point is that the feedback term contributes less to the net heat flux variability than do the original ECHAM3 fluxes. It may be concluded both that the ECHAM3 wind stresses contain a good representation of the interannual forcing, and that the heat fluxes contain a sensible interannual signal.

Figure 2 demonstrates that the ECHAM3 wind forcing is able to reproduce the observed large-scale variations in SST really rather well. The central Pacific shows an even better agreement between simulated and observed interannual SST. But it should be remembered that the forcing has little verisimilitude in its monthly fluctuations; the anomaly correlation coefficient for monthly SST is reduced to 0.88 for Nino-3, compared to 0.94 for 12-month running means. It should also be noted that data assimilation experiments suggest that the off-equatorial anomalies produced by ECHAM3 may not be as good as the equatorial fields we see here. Nonetheless, it is reasonable to expect a moderate level of forecast skill to be derivable from these initial conditions.

The forecasts are made by coupling the ocean and atmosphere models together and running the system forward from the ocean initial state. Atmospheric initial conditions (including the condition of the land surface) are taken from the ECMWF operational analysis of 1 January or July 1990, as appropriate; that is, no information specific to the start date is provided via the atmosphere. The coupled system is integrated for 1 year in most cases. Forecasts are made from both 1 January and 1 July for a set of years; this gives some idea of any seasonal variations in the performance of the forecast system while keeping the total number of integrations required small.

The key part of any real-time forecast system that uses this direct ''couple and run'' approach is to estimate the bias of the forecast. The estimated bias can then be combined with the model-produced fields to give an optimal forecast of the actual SST. The true bias of the forecast system may well be a complex and nonlinear function of the actual or forecast state of the ocean–atmosphere system. In practice, however, the bias must be estimated from forecasts based on a relatively small sample of initial conditions, and it is likely to be difficult to constrain anything other than the linear part. Indeed, even estimating the linear bias reliably can be difficult. Further discussion of some of the issues involved in bias estimation is given in section 5. In these experiments, the bias is simply estimated as the mean error in absolute SST from a standard sample of 10 forecasts.

Forecasts with start dates in the years 1981–90 make up the standard sample for each bias calculation. The bias is calculated separately for forecasts started in January or July, and as a function of forecast lead time. Ideally, verification of forecasts would be restricted to cases with a start date after the end of the period used to define the mean bias. But a reasonable number of cases are needed to define the bias, while at the same time a good size sample is needed to assess forecast skill; and since we have only a limited number of forecasts available, we are forced to use the same forecasts both for defining the bias and estimating the forecast skill. This has the risk of inducing artificial skill into the estimates of forecast quality. Cross validation, where the mean bias is estimated from the other nine cases for each forecast being assessed, might seem to remove most of the problem. In fact, it has no effect on the anomaly correlation scores, while leading to a 10% increase in the standard deviation and rms error of the forecasts. It can be shown that the larger rms errors are an overestimate of future forecast errors, while the standard verification procedure (ignoring the fact that the forecast whose error is being measured also contributed to the estimate of the mean correction) underestimates rms errors. It is straightforward to scale the error estimates to be unbiased, and this is done in the next section. The remaining problem is that the mean bias in the forecasts is artificially zero, with or without cross validation. In a real time forecasting situation, a bias of order (rms error)/$n^{1/2}$ would be expected, where n is the number of past forecasts used to estimate the model drift.

4. Forecast results

Figure 3 shows the evolution of SST in the Nino-4 index region from the two sets of forecast experiments, with July start dates. The observed climatology is also shown. The first set of forecasts are typically 1°C too cold within the first month, warm slightly after 3 or 4 months, and then develop a strong cold bias at 6 months and beyond. The overall pattern of behavior is similar in all cases and differs substantially from climatology. The error develops in a nontrivial way. This is presumably because it is contributed to from a number of sources: heat flux imbalances in the coupled system, subsurface anomalies due to imbalances between the initial ocean state and the wind stresses produced by the coupling, and possible coupled interactions between various errors and the seasonal cycle. It is unlikely the error could be anticipated with any accuracy purely on the basis of uncoupled experiments. The error is a function of the climatology of the initial states as well as the model, so long-term integrations of the coupled system will also fail to give quantitative estimates of the forecast drift. The forecast bias can be estimated only from a set of forecasts.

The second set of forecasts also develops a substantial bias but of the opposite sign. At 12-month lead times the SST is typically more than 1°C too warm, compared to about 3°C too cold in the first set. In both cases, the

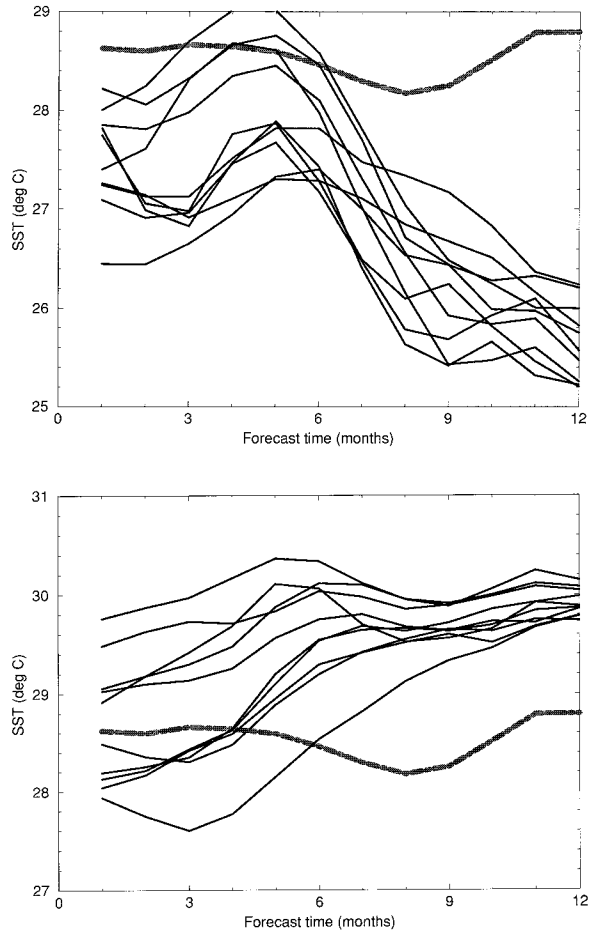

FIG. 3. Development of Nino-4 SST in the uncorrected forecasts, which start in July. The observed mean value is also shown (thick line). (a) Forecasts with model 1. (b) Forecasts with model 2.

bias is greater than or equal to the spread between the different years, at least for time ranges beyond the first few months. These forecast experiments are both ''large bias'' cases and therefore supply a nontrivial test for the concept of mean drift subtraction as an a posteriori correction.

The quality of SST forecasts can be assessed in a variety of ways. Here we restrict our attention to the Nino-3 and Nino-4 indices, and consider anomaly correlations, anomaly amplitudes, and root-mean-square errors. Anomaly correlations are in some sense the easiest of skill scores to do well at. They test whether the forecast system is correctly reproducing the pattern of year to year variability, but they are blind to the mean value of the field and to the amplitude of the variations. The anomaly amplitudes can easily be examined separately (as can the mean forecast bias, although it is by definition zero in the present case). The rms error is perhaps the most comprehensive measure, in that a good score requires the pattern and amplitude of variability and the mean value all to be well reproduced. Unless the correlation is very high, however, the rms errors can generally be reduced by systematically underpredicting the amplitudes of anomalies, that is, by a bias toward climatology. This may or may not be the most useful of forecasts; it certainly does not represent a likely future course for the system. In the context of developing and improving forecast systems, such an amplitude bias is symptomatic of model problems rather than model successes. Overall, then, the rms error must be used carefully if it is not sometimes to be misleading.

The actual anomalies are presented in Fig. 4 for the forecasts from the two different models. The most obvious problem is the small amplitude of anomalies in the first model, more of which will be said shortly. The second model is generally more successful, but it does not do very well in predicting the 1982/83 warm event.

The Nino-3 and Nino-4 anomaly correlation scores for the experiments are shown in Fig. 5, together with

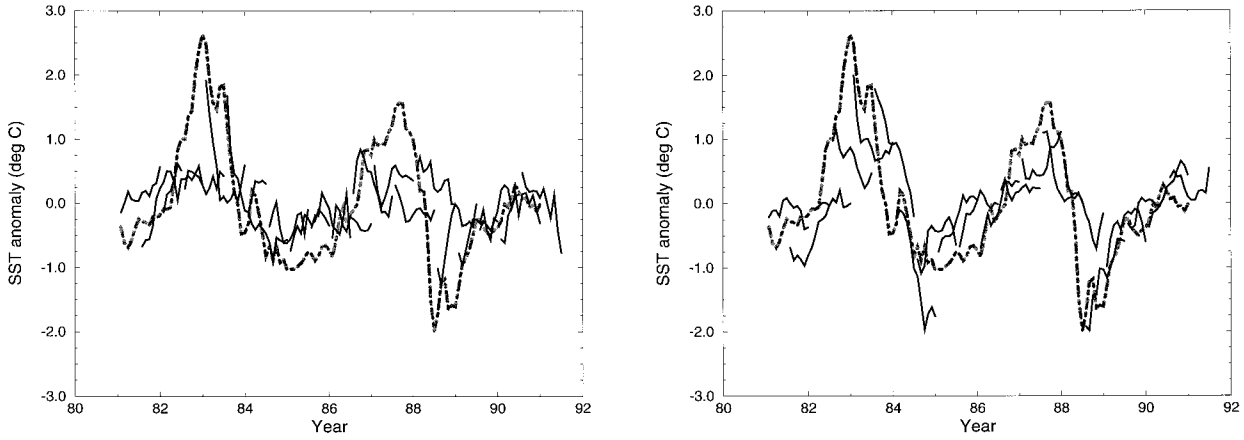

FIG. 4. SST anomalies from individual forecasts (solid lines), plotted together with observed values (thicker dashed line). (a) Forecasts with model 1. (b) Forecasts with model 2.

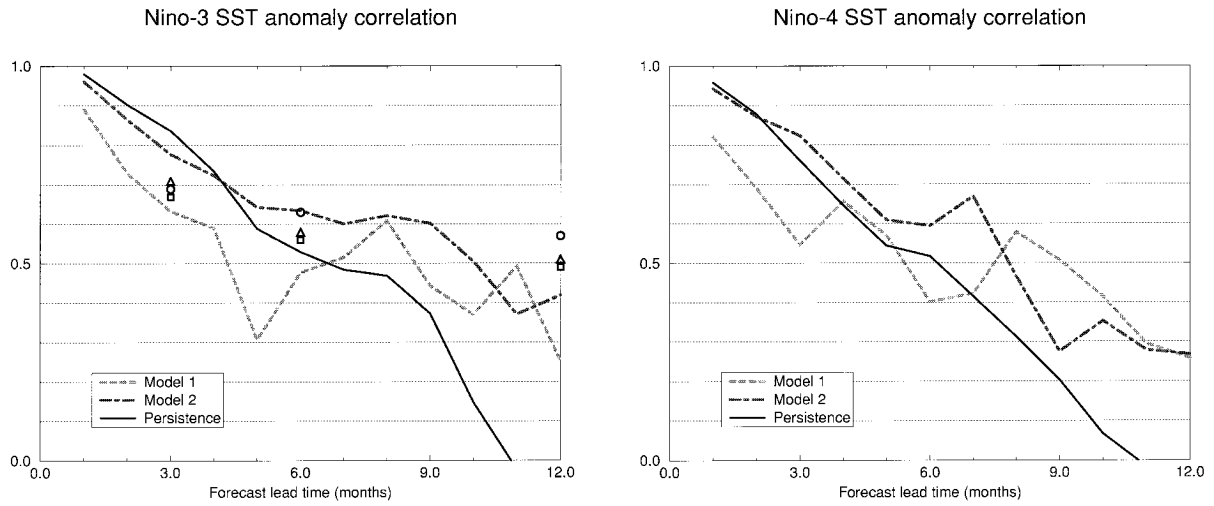

FIG. 5. Anomaly correlation scores for the forecasts. (a) Nino-3. Also shown as symbols are some scores from three other forecasting systems (circle—Cane–Zebiak, triangle—Balmaseda, square—Kleeman). (b) Nino-4.

the score given by a forecast that simply extends the initial SST anomalies (persistence). The first set of experiments has a correlation score, which is only superior to persistence at lead times of greater than 6 months, at which point it is already below 0.6. This system would be of little practical use in El Niño forecasting. In terms of absolute significance, however, the model certainly has some skill: a 95% significance level test for the correlation being greater than zero requires a value greater than 0.38, which the system possesses out to 10 or 11 months. (The 90% level for this test, one-tailed and assuming that forecasts made 6 months apart are independent, is 0.31.) The definition of correlation used here is the usual product-moment correlation coefficient. This measures both the ordering and the spacing of the values, and a disadvantage is that it can be strongly influenced by outlying points. A more robust measure is the rank correlation coefficient, which measures only the relative ordering of points and not their values. As such, it is insensitive to, for example, systematic problems in dealing with large anomalies, or other nonlinear biases. When applied to the two sets of forecasts considered in this paper, the rank correlation is usually (but not always) higher than the product-moment correlation, suggesting that the models are indeed having problems with some of the nonlinearities. The small samples make this difficult to investigate properly. To give some idea of how this may be affecting the scores, note that the significance of the 12-month Nino-4 forecasts is 0.94 for a rank correlation test and only 0.88 for the product-moment test. Average differences in the actual correlation score are a few percentage points.

The second set of forecasts has generally higher correlations than the first. This is due in part to the more accurate initialization of SST, which helps the short-range forecasts especially, but is also likely to be helped by the improvements in the ocean model. A factor that is probably reducing the model skill is the restriction of the ocean model to the Pacific basin only, and the consequent absence of any SST anomalies outside this domain. In particular, all attempts with the Pacific basin model to forecast the 1982/83 El Niño from the end of 1981 fail, while the global model is successful. The significance of the correlation scores of the second forecast set is stronger than the first, and the fact that persistence is now bettered at 5 months (3 months for Nino-4) means that the forecast ''skill'' is more likely to be useful. The relative skill of these two sets of forecasts is not important for this paper, but it should be noted that although the correlation differences would mostly fail standard significance tests, this is largely because such tests assume that the two sets are independent samples of their respective parent populations. In fact, both forecast sets are made for the same period, and so differences between them are more significant than might otherwise appear. Of course, there is still significant sampling error, and the relative performance of models in one decade may not be relevant to a different decade.

A final comment on the correlation scores is to compare them with other published results. Comparable systems are those that do not use any subsurface data and have been tested over a representative set of ENSO conditions; examples are forecasts from the Cane–Zebiak model, and those by Kleeman (1993) and Balmaseda et al. (1994). Results from these models are marked on Fig. 5. There is the inevitable problem that we are comparing different time periods in each case, and the fact that the forecasts in this paper are restricted to the 1980s may mean that their success is overstated. On the other hand, the models in this paper have not been tuned in any way to try and optimize the forecast scores, and do not benefit from any ensemble averaging. Given these caveats, it seems that the results are broadly

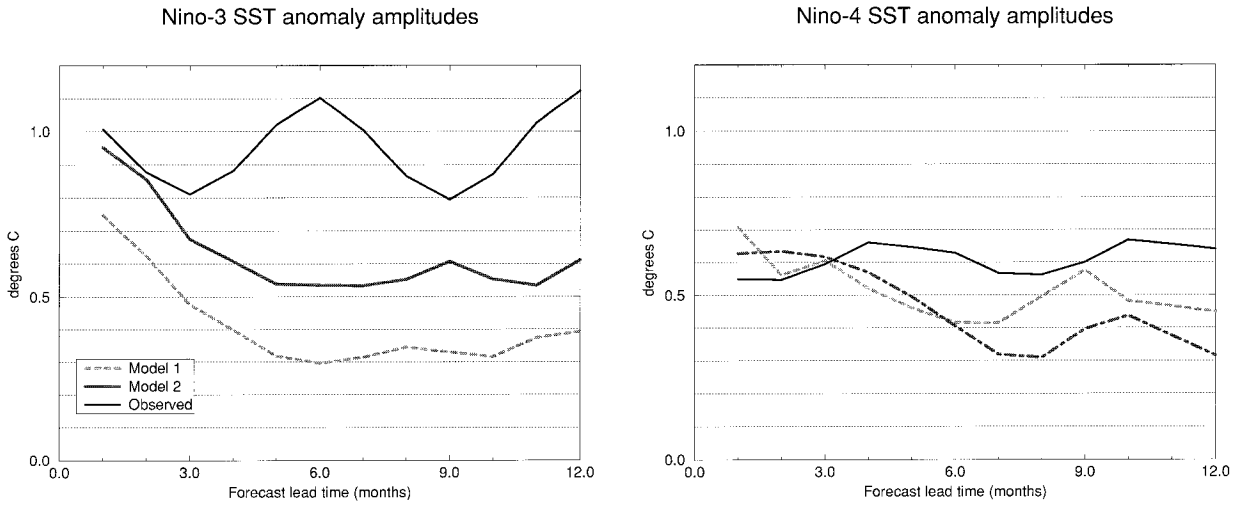

FIG. 6. Anomaly amplitudes for the forecasts and the observations. (a) Nino-3, (b) Nino-4.

comparable, except possibly in the longer range. Difficulties at 12-month lead times are perhaps most likely to be model problems and may be associated with the weak amplitude of the forecast anomalies. Certainly the rank correlation scores, which are unaffected by nonlinear amplitude problems, are appreciably higher at 12 months.

Figure 6 shows the standard deviation of the SST anomalies, both for the forecasts and observations. In Nino-3, both models seriously underestimate the amplitude of anomalies, but the problem is most serious with the first model. Investigations have shown the most likely cause to be inadequacies in the original ocean model. The stronger variability in the second system, which uses a revised version of the ocean model, is consistent with this. Nino-4 variability is also underestimated by both systems for forecasts of more than three months, although it is less of a problem than for Nino-3.

The rms errors from the forecasts are shown in Fig. 7, together with the errors obtained from assuming either persistence or climatology. It can again be seen that the first set of forecasts is not very useful, in that the errors are not much smaller than the best of persistence/climatology. The second forecast set is generally better, and both Nino-3 and Nino-4 show a clear gap between the forecast errors and persistence beyond two months, and between forecast and climatology up to 8 or 9 months. Since the second forecast set has a larger amplitude of variability, its asymptotic value of rms error will be larger than that of the first set; this makes the improved performance in the first 10 months even more laudable.

The scores discussed so far are for the combined set

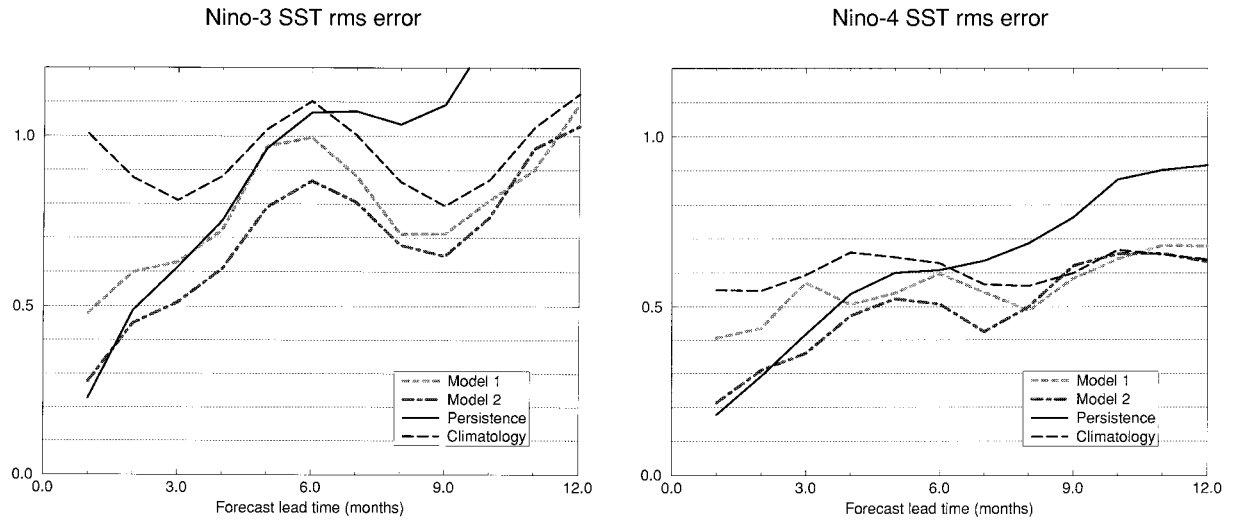

FIG. 7. The rms errors for the forecasts. Also shown are the rms errors that would be obtained from persistence or climatology. (a) Nino-3, (b) Nino-4.

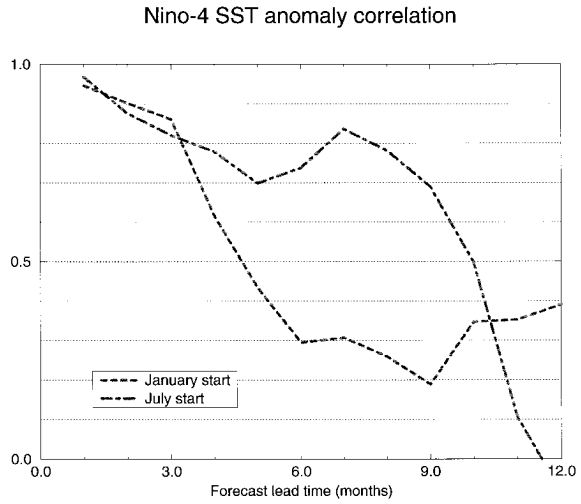

FIG. 8. Anomaly correlation scores for Nino-4 SST for the second set of forecasts, showing the different behavior of January and July starts.

of forecasts, those initialized in both January and July. An assessment of the seasonal behavior of the forecast system is hampered by the small sample sizes, but one result is of some interest. In terms of overall scores, a strong seasonal dependence is not found in these experiments. As in many other studies, the period of boreal spring/early summer produces more erratic forecasts, and there are some sharp temporary dips in the correlation scores. But the skill generally recovers, and for the most part there is little to choose between those forecasts starting in January and those in July. Differences are well within the noise level of the small samples. There is one major exception to this, however, and that is the west-central Pacific temperatures in the second set of forecasts. These show a very strong seasonal dependence, best exhibited by a plot of Nino-4 SST (see Fig. 8).

Forecasts starting in January show a rather rapid drop off in skill after 3 months or so and remain at low values. Forecasts starting in July, however, maintain a high correlation for about 9 months before plunging precipitously toward zero correlation. This predictability barrier is not seen in Nino-3 SST, nor is it seen anywhere in the first set of forecasts. It is seen, however, in a different set of experiments using the same Pacific-basin-only coupled model. There is strong evidence, therefore, that this predictability barrier is caused by something specific to the model. Together with the drop in correlation, the amplitude of the forecast anomalies becomes very small at this time in the second model, and this is probably the cause of the drop in correlation. The anomaly amplitude does not become small either in the observations or the first model. Quite why the anomaly amplitudes should show a marked seasonal cycle in the second model is not clear. It might be due to the limited domain of the ocean model and the imposition of fixed SST around it. Alternatively, it might be the drift of the system toward a warmer state interacting with the seasonal cycle in some nonlinear way. If this latter is the case, then it is an example where the bias correction method is failing.

5. Further considerations concerning bias removal

A key issue in applying this method is having a good procedure to estimate the bias in the forecasts. This is, in fact, a fairly general problem, since even if measures were taken to reduce the climate drift of the forecast system, it is likely that some residual biases in the forecast will remain. What makes the problem nontrivial is the fact that we are only able to generate a small sample, and that both nonlinearities and nonnormal distributions are likely to be involved.

The results so far have used the simplest method of estimating the bias, namely taking the mean over all known cases. If the forecast errors are independent of the forecast state (i.e., the system is linear in its error statistics), and if the error distribution is normal, then the sample mean is the best estimate of the bias. The expected error in the sample mean is $\sigma(n-1)^{-\frac{1}{2}}$, where σ is the sample (not the true) standard deviation of error and n is the sample size. The expected error in a forecast for next year is therefore increased from $[n(n-1)^{-1}]^{\frac{1}{2}}\sigma$ to $[(n+1)(n-1)^{-1}]^{\frac{1}{2}}\sigma$. The extra contribution from uncertainty in the bias is therefore fairly small for $n > 5$ or so. An expected rms error of 0.5°C would be increased by 0.05°C for $n = 5$, and only 0.02°C for $n = 15$. This does not mean that the bias is known to this accuracy. The uncertainty in the bias (and thus in any real-time forecast) might be close to 0.25°C for the $n = 5$ case. The point is that, given the likely errors in a forecast, shifting it by the uncertainty in the mean will have little impact on the *expected* forecast accuracy.

However, it may be that the forecast errors are not normally distributed, but have rather fatter tails. That is, forecasts might often be right but sometimes be very wrong, for example, if they miss completely the development of an event. If the error distribution is in fact nonnormal, then the above estimate of sampling error for the mean bias may be overoptimistic. As an extreme example, the expected error in estimating the mean of a Lorentz distribution from the sample mean remains constant, however large the sample size becomes. In such circumstances, the sample median can often be a more useful and robust estimator of the mean, since it is less influenced by extreme values.

The sample median may be a better estimator of the true mean in the case where the population (true) mean and median are the same. If the error distribution is nonsymmetric, then the median will generally converge to a different value from the mean. This median may in fact be more appropriate for estimating the expected drift in a forecast. For example, suppose that a forecast system can produce strong cold events but only mod-

erate warm events; given typical errors in ocean models this is not an unlikely scenario. In this case, the forecast *errors* are also likely to be skewed, with more large negative errors than large positive ones. The mean error, therefore, will be negative, even if a typical forecast (i.e., one that successfully predicts near-zero anomalies) is unbiased. Assuming the sample size is large enough to estimate both mean and median accurately, then subtraction of the mean error will in most circumstances produce a small artificial warming of the forecasts. This could be seriously misleading if not recognized. The median error, however, is unaffected by the skewness of the forecast anomaly amplitudes and will be approximately zero in the present case. It is probably more appropriate for correcting forecast biases than the mean, because it does not produce false warming events. Note, however, that using the mean would result in lower rms errors: the false warmings are more than compensated by the reduced error in the occasional coolings. Since the mean minimizes rms error, rms error statistics will always favor the mean, regardless of whether or not this is most appropriate. There are other error statistics that can be used, however. A simple one is the mean absolute error, and it is easy to show that this is the measure of error that is minimized by the median. In statistical terms it is a question of choice as to which error is to be minimized, and therefore which definition of the average (mean or median) it is preferred to estimate. Even in the case of estimating the mean, however, it may end up being better to use the sample median, depending on the interaction between nonnormality, skewness, and sample size.

As a preliminary test, the median technique of bias estimation was applied to the four sets of forecasts in this paper (January and July starts for the two different models). Cross validation was used, so that for each forecast the drift was estimated from the nine other forecasts in the same set. The estimation was made using either the mean or the median. Typical differences were in the region 0.1°–0.2°C. The corrected forecasts were then verified against the observations, using both rms and mean absolute error scores. To produce a summary score, the errors were averaged over all lead times (differences in errors are roughly equally distributed across different lead times, so this does not imply a particular bias to large lead times). Table 1 summarizes the mean error scores from the two techniques applied to the four forecast sets.

From our earlier discussion, we might expect that for a nonnormal but symmetric distribution the median would give better results than the mean for small sample sizes, for both mean absolute error and rms error. We also expect that for a skew distribution, use of the median might increase rms error while reducing mean absolute error. This is why both error scores are presented: looking at them both should give a balanced picture as to whether the use of the median is making a real improvement in the sampling. Perhaps surprisingly, the results here seem quite consistent with our expectations for a nonnormal, slightly skewed error distribution. The error scores are improved in 7 out of 8 cases, with an apparently clear-cut improvement in the mean absolute error score. Statistical significance is hard to assess, but the odds of getting a positive difference in all four cases for the mean absolute error by chance are 6.25% and, if we assume the error differences to be independent, the odds of 7 out of 8 being positive are only 3.5%. These results would certainly not discourage us from using the median to estimate the bias from small samples.

TABLE 1. Error scores from cross-validated forecasts using two different methods of removing bias (sample mean and sample median). Sets 1 and 2 are the forecasts with the first model, starting in January and July, respectively. Sets 3 and 4 are the forecasts with the second model. Numbers are errors in degrees Celsius, multiplied by 1000. Positive differences imply that the median method is better than the mean method.

	rms error score			Mean absolute error score		
	Mean	Median	Difference	Mean	Median	Difference
Set 1	858	836	22	668	662	6
Set 2	891	890	1	708	682	26
Set 3	763	752	11	607	594	13
Set 4	773	784	−11	644	642	2

Why is a detailed consideration of the optimal drift estimation worthwhile? In an operational setting, it might be thought straightforward to produce a large enough ensemble of past cases (say 20 or more) for such detail to be irrelevant. Although this is undoubtedly true in the long term, in the short term it is not so easy. This is because the bias in the forecast system is a function of the ocean mean initial state, and the ocean mean initial state is likely to be a function of the observing system used to create it, at least until our model simulations have much smaller biases. Because of this, the drift in forecasts for the 1980s is likely to be different to the drift in the 1990s, due to the improvements in the observing system. In time, observing systems should stabilize and model errors reduce, so that a larger set of consistent forecasts can be evaluated. At the moment, only the last few years of forecasts can be used safely to evaluate the expected drift in a forecast for next year.

6. Conclusions

Two different coupled GCM systems have been used to forecast SST anomalies using a method of drift subtraction. Both models, but especially the first, have considerable failings in their ability to simulate and predict the amplitude and structure of El Niño events. The initialization procedure for the ocean may be adequate but is a long way from capturing a good approximation to the actual initial state. The model systems show substantial drift in their mean state during 12-month fore-

casts. Despite all these obstacles, both systems are able to produce skillful forecasts of equatorial SST. Whether or how much the forecasts might have been better if we had used a technique other than bias subtraction is unknown, but what is clear is that drift subtraction works. That meaningful forecasts can be extracted in the presence of such large drifts as are shown in Fig. 3a is almost surprising and demonstrates the substantial linearity of a major part of ENSO physics.

Both forecast systems worked, establishing the validity of drift subtraction in both warm and cold bias cases. The relatively large drift of the two model systems means that the result is encouraging, in that any actual implementation would be in a system with smaller drifts than those shown here and that the validity of the method should therefore be even stronger than in these experiments.

It should not be thought, however, that errors in the mean state are entirely without consequence. Uncoupled experiments with both ocean and atmosphere have shown that increasing the mean state errors produces a drop in the skill of the models in simulating anomalies. Further, the seasonal drop in skill discussed at the end of section 4 may be an example of model drift impacting on anomaly development. The interaction between mean state and anomalies is not always large, but it does exist. Because of this, it may well be advisable to use some constraints or adjustments in the coupled model to minimize or reduce the climate drift. What these experiments demonstrate, however, is that it is unnecessary to force the models too hard in an attempt to make the climate drift vanishingly small.

A further aspect that is clear from these experiments is that the variability characteristics of the coupled model are very important when it comes to making forecasts. The first model is just not able to produce large SST anomalies, and so its forecasts are strongly biased toward climatology. This not only means that the actual forecasts must be ''interpreted'' or rescaled before use, but that the coupled response of the system and hence its longer-term evolution is wrong. Using such a coupled model to predict atmospheric behavior, which is ultimately more important than the SST anomalies themselves, will also disappoint. Although a climate drift may be more obvious than a defective variability, it is probably a good model ENSO variability that is more important for ENSO forecasting. Model drifts can easily be accounted for; lack of proper El Niños cannot.

Finally, the question of constructing the bias correction needs to be considered. When and if a large set of compatible forecasts are available, the mean error over previous forecasts will probably be an adequate measure, at least if the desire is to minimize rms errors. In the near term, only smaller samples are likely to be coherent, due to the continued evolution of the observing system. The results from this set of experiments, preliminary though they must be, suggest that even with a 9-yr reference dataset the use of a sample median may be more appropriate for estimating the true mean bias. Given the arguments in favor of the true median bias being more appropriate than the true mean as a reference point for El Niño forecasting, the case for using the median is significant. Other possibilities might be a winsorized mean (removing outliers, and taking the mean of the central points), which is something of a hybrid method. Too much sophistication is unlikely to be justified, however.

More pragmatically, most forecast systems will operate in some sort of ensemble mode. The uncertainty in the model bias can then be accounted for by taking several different estimates of it, spanning the range of plausible values. In fact this could be a useful method of augmenting other ensemble generation techniques. It is also possible that clever use of ensembles over past cases can refine estimates of likely model errors in a particular forecast situation.

Bias correction techniques are flexible and powerful in their application to ENSO forecasting. While it is strongly desirable to minimize all sources of error in coupled GCM forecasts, the existence of climate drift should not cause undue despair.

REFERENCES

Balmaseda, M. A., D. L. T. Anderson, and M. K. Davey, 1994: ENSO prediction using a dynamical ocean model coupled to statistical atmospheres. *Tellus,* **46A,** 497–511.

Cane, M. A., S. E. Zebiak, and S. C. Dolan, 1986: Experimental forecasts of El Niño. *Nature,* **321,** 827–832.

Ji, M., A. Kumar, and A. Leetmaa, 1994: An experimental coupled forecast system at the National Meteorological Center. Some early results. *Tellus,* **46A,** 398–418.

Kleeman, R., 1993: On the dependence of hindcast skill on ocean thermodynamics in a coupled ocean–atmosphere model. *J. Climate,* **6,** 2012–2033.

Latif, M., A. Sterl, E. Maier–Reimer, and M. M. Junge, 1993: Structure and predictability of the El Niño/Southern Oscillation phenomenon in a coupled ocean–atmosphere general circulation model. *J. Climate,* **6,** 700–708.

——, T. Stockdale, J. Wolff, G. Burgers, E. Maier–Reimer, M. Junge, K. Arpe, and L. Bengtsson, 1994: Climatology and variability in the ECHO coupled GCM. *Tellus,* **46A,** 351–366.

Manabe, S., R. J. Stouffer, M. J. Spelman, and K. Bryan, 1991: Transient responses of a coupled ocean–atmosphere model to gradual changes of atmospheric CO_2. Part I: Annual mean response. *J. Climate,* **4,** 785–818.

Mechoso, C. R., 1995: The seasonal cycle over the tropical Pacific in general circulation models. *Mon. Wea. Rev.,* **123,** 2825–2838.

Murphy, J. M., 1995: Transient response of the Hadley Centre coupled ocean–atmosphere model to increasing carbon dioxide. Part I: Control climate and flux adjustment. *J. Climate,* **8,** 36–56.

Sausen, R., K. Barthel, and K. Hasselmann, 1988: Coupled ocean–atmosphere models with flux correction. *Climate Dyn.,* **2,** 145–163.

Stockdale, T., M. Latif, G. Burgers, and J.-O. Wolff, 1994: Some sensitivities of a coupled ocean–atmosphere GCM. *Tellus,* **46A,** 367–380.