Retrospective ENSO Forecasts: Sensitivity to Atmospheric Model and Ocean Resolution

EDWIN K. SCHNEIDER,* DAVID G. DEWITT,+ ANTHONY ROSATI,# BEN P. KIRTMAN,* LINK JI,@ AND JOSEPH J. TRIBBIA&

*George Mason University, Fairfax, Virginia, and Center for Ocean–Land–Atmosphere Studies, Calverton, Maryland
+International Research Institute for Climate Prediction, Palisades, New York
#NOAA, Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey
@Department of Oceanography, Texas A&M University, College Station, Texas
&National Center for Atmospheric Research, Boulder, Colorado

(Manuscript received 19 September 2002, in final form 16 June 2003)

ABSTRACT

Results are described from a series of 40 retrospective forecasts of tropical Pacific SST, starting 1 January and 1 July 1980–99, performed with several coupled ocean–atmosphere general circulation models sharing the same ocean model—the Modular Ocean Model version 3 (MOM3) OGCM—and the same initial conditions. The atmospheric components of the coupled models were the Center for Ocean–Land–Atmosphere Studies (COLA), ECHAM, and Community Climate Model version 3 (CCM3) models at T42 horizontal resolution, and no empirical corrections were applied to the coupling. Additionally, the retrospective forecasts using the COLA and ECHAM atmospheric models were carried out with two resolutions of the OGCM. The high-resolution version of the OGCM had 1° horizontal resolution (1/3° meridional resolution near the equator) and 40 levels in the vertical, while the lower-resolution version had 1.5° horizontal resolution (1/2° meridional resolution near the equator) and 25 levels. The initial states were taken from an ocean data assimilation performed by the Geophysical Fluid Dynamics Laboratory (GFDL) using the high-resolution OGCM. Initial conditions for the lower-resolution retrospective forecasts were obtained by interpolation from the GFDL ocean data assimilation.

The systematic errors of the mean evolution in the coupled models depend strongly on the atmospheric model, with the COLA versions having a warm bias in tropical Pacific SST, the CCM3 version a cold bias, and the ECHAM versions a smaller cold bias. Each of the models exhibits similar levels of skill, although some statistically significant differences are identified. The models have better retrospective forecast performance from the 1 July initial conditions, suggesting a spring prediction barrier. A consensus retrospective forecast produced by taking the ensemble average of the retrospective forecasts from all of the models is generally superior to any of the individual retrospective forecasts. One reason that averaging across models appears to be successful is that the averaging reduces the effects of systematic errors in the structure of the ENSO variability of the different models. The effect of reducing noise by averaging ensembles of forecasts made with the same model is compared to the effects from multimodel ensembling for a subset of the cases; however, the sample size is not large enough to clearly distinguish between the multimodel consensus and the single-model ensembles.

There are obvious problems with the retrospective forecasts that can be connected to the various systematic errors of the coupled models in simulation mode, and which are ultimately due to model error (errors in the physical parameterizations and numerical truncation). These errors lead to initial shock and a ''spring variability barrier'' that degrade the retrospective forecasts.

1. Introduction

In June 1999, groups from several institutions, including the Center for Ocean–Land–Atmosphere Studies (COLA), the Geophysical Fluid Dynamics Laboratory (GFDL), the International Research Institute for Climate Prediction (IRI), and the National Center for Atmospheric Research (NCAR), met and agreed on a framework for a baseline comparison of the sensitivity of coupled GCM (CGCM) retrospective forecasts (hereafter referred to as ''forecasts'' for brevity) to the choice of atmospheric GCM (AGCM). A prime motivation for undertaking these coordinated computations was to determine the impact of multimodel ensembles on the forecast skill of El Niño–Southern Oscillation (ENSO) variability in the tropical Pacific. The participants agreed to use a specific configuration of the GFDL Modular Ocean Model version 3 (MOM3) OGCM as the ocean component of their coupled model. Among other features, the OGCM would have a near-global domain, 1° horizontal resolution (1/3° meridional resolution near the equator), and 40 levels in the vertical. Ocean initial conditions were to be provided from an ocean data assimilation (ODA) carried out at GFDL using the agreed-

Corresponding author address: Dr. Edwin K. Schneider, Center for Ocean–Land–Atmosphere Studies, 4041 Powder Mill Road, Suite 302, Calverton, MD 20705-3106.
E-mail: schneide@cola.iges.org

on ocean model. Each group would use its own chosen AGCM. It was also agreed that the AGCM would have approximately T42 horizontal resolution and 18 levels and that ad hoc initialization and coupling schemes designed to reduce initial shock and climate drift (e.g., anomaly initialization, flux correction, anomaly coupling) would not be used.

There were several reasons for choosing the experimental design for the comparison of coupled models using the same ocean model and ocean initial conditions but different AGCMs. One was that it was felt that sensitivity to choice of AGCM was likely to be large, based both on understanding of the sensitivity of ENSO behavior in intermediate coupled models to the strength of the wind stress coupling (summarized in Neelin et al. 1998) and on experience with the sensitivity of simulations of ENSO variability in coupled GCMs to AGCM parameterizations (e.g., Schneider 2002). Also, the use of a common ocean model together with initial conditions generated using that model would allow a controlled evaluation of the forecast sensitivity to AGCM, which might then be related to the systematic errors of the different coupled models and eventually traced to the different physical parameterizations in the AGCMs. Finally, the choice of a common ocean model and data assimilation system for producing the ocean initial condition data would be important for practical reasons, eliminating the duplication of effort that would have been required to configure and test different ocean models and to perform separate ocean data assimilations with different ocean models and data assimilation systems.

In the course of carrying out the forecasts, it was realized that the coupled model was placing very high demands on the computing resources available to several of the groups, primarily because of the high resolution of the OGCM. As an aid to making design decisions for future experiments, an evaluation of the effect on the forecasts due to reducing the OGCM resolution to one that was significantly less resource intensive was undertaken. A version of the OGCM with medium resolution and no other changes was constructed, the initial conditions from the GFDL ODA were interpolated to the new resolution, and the series of forecasts using two of the AGCMs was repeated.

Results comparing the skill of the forecasts from the individual models and the consensus forecasts obtained by simple averaging of all of the forecasts (across AGCM and OGCM resolution) are presented below. The expectation is that averaging across ensembles of cases will raise correlations for two reasons. First, noise present in the individual forecasts will be reduced. This will occur when the ensemble is performed with a single model or multiple models. Second, when multiple models are used, there is the possibility that the effects of the systematic errors of the different models will be reduced by the averaging.

There are several major points to be noted: 1) forecast skill did not depend strongly on the choice of an AGCM, despite markedly different AGCM-dependent systematic errors; 2) there was no apparent degradation of the forecast skill due to reducing the OGCM resolution; and 3) the multimodel approach can contribute to improvement of skill. We also demonstrate that there is potential for substantial improvement in coupled-model forecasts through reduction of the systematic errors of the coupled models. These systematic errors can be evaluated and addressed in simulation mode.

2. The coupled models

The coupled models consist of various AGCMs at similar spatial resolution: COLA version 2 (V2), ECHAM4.5 (GCM based on European Centre for Medium-Range Weather Forecasts forecast models, modified and extended in Hamburg; version 4.5), or NCAR Community Climate Model version 3.6 (CCM3) coupled to MOM3. The AGCMs are summarized in Table 1. The ocean model (section 2a), coupling (section 2b), initial conditions (section 2c), cases (section 2d), and comparison of AGCM surface fluxes when forced by observed SST (section 2e) are described briefly below. The coupled models are referred to by their designations in Table 1: COLA high, COLA medium, CCM medium, ECHAM high, and ECHAM medium.

a. Ocean model

The ocean model is version 3 of the GFDL MOM (Pacanowski and Griffies 1998), a finite-difference treatment of the primitive equations of motion using the Boussinesq and hydrostatic approximations in spherical coordinates. The domain is that of the World Ocean between 74°S and 65°N. The coastline and bottom topography are realistic except that ocean depths less than 100 m are set to 100 m, and the maximum depth is set to 6000 m. The artificial high-latitude meridional boundaries are impermeable and insulating. The vertical mixing scheme is the nonlocal K-profile parameterization of Large et al. (1994). The horizontal mixing of tracers and momentum is Laplacian. The momentum mixing uses the space–time-dependent scheme of Smagorinsky (1963) and the tracer mixing uses Redi (1982) diffusion along with Gent and McWilliams (1990) quasi-adiabatic stirring.

The two versions of the ocean model used in this comparison differ only in resolution and are referred to in the following as the ''high''- and ''medium''-resolution versions. The medium-resolution version was also used in the coupled simulations described by Kirtman et al. (2002). The zonal resolution of the high-resolution version is 1.0°, while that of the medium-resolution version is 1.5°. In the high-resolution version, the meridional grid spacing is 1/3° between 10°S and 10°N, gradually increasing to 1.0° at 30°N and 30°S and fixed at 1.0° in the extratropics. In the medium-resolution ver-

TABLE 1. Coupled-model configurations.

Model	Cola		ECHAM		CCM medium
	high	medium	high	medium	
OGCM	Resolution (lat × lon, levels)	1° × 1°, 40 levels	1.5° × 1.5°, 25 levels	1° × 1°, 40 levels	1.5° × 1.5°, 25 levels	1.5° × 1.5°, 25 levels
AGCM properties	Version	COLA V2 (Schneider 2002)		ECHAM4.5 (Roeckner et al. 1996)		CCM3.6 (Kiehl et al. 1996)
	Structure	spectral T42 18 levels		spectral T42 19 levels		spectral T42 18 levels
	Vertical coordinate	sigma		hybrid sigma–pressure		hybrid sigma–pressure
	Radiative transfer (solar; terrestrial)	Lacis and Hansen (1974), Davies (1982); Harshvardahn et al. (1987)		Fouquart and Bonnel (1980); Morcrette et al. (1986)		Briegleb (1992); Ramanathan and Downey (1986)
	Deep convection	Moorthi and Suarez (1992); DeWitt (1996)		Tiedtke (1989); Nordeng (1994)		Zhang and McFarland (1995)
	Shallow convection	Tiedtke (1984)		Tiedtke (1989); Nordeng (1994)		Hack (1994)
	Turbulent diffusion	Mellor and Yamada (1982), level 2.0		Louis (1979); Brinkop and Roeckner (1995)		Holtslag and Boville (1993)
	Cloud water	diagnostic		prognostic		diagnostic
	Cloud radiative properties	Kiehl et al. (1994, 1996); DeWitt and Schneider (1996)		Rockel et al. (1991); Roeckner (1995)		Kiehl et al. (1996, 1998)

sion, the meridional grid spacing is 0.5° between 10°S and 10°N, gradually increasing to 1.5° at 30°N and 30°S and fixed at 1.5° in the extratropics. The high-resolution version has 40 levels in the vertical, with 27 levels in the upper 450 m, while the medium-resolution version has 25 levels in the vertical, with 17 levels in the upper 450 m. The diffusivities provided by the Smagorinsky momentum mixing scheme are functions of the horizontal resolution.

b. Coupling and computational aspects

The atmospheric model provides heat, momentum, fresh water, and surface solar flux to the ocean model. The ocean model provides SST to the atmosphere. Information is exchanged between the atmosphere and ocean once daily. No empirical corrections were applied to the fluxes provided by the atmosphere to the ocean. The COLA and CCM coupled models use the coupling software provided with MOM3 for interpolation between the atmosphere and ocean model grids. The ECHAM coupled model uses the Ocean Atmosphere Sea Ice Soil (OASIS) coupling software (Terray et al. 1999) that is produced by the European Centre for Research and Advanced Training in Scientific Computation (CERFACS).

In the COLA coupled model, the SST seen by the atmosphere is the SST of the ocean model plus a spatially dependent, time-independent term that corrects for the spectral truncation error that produces nonzero elevation of the lower boundary of the atmosphere over the ocean and associated unrealistic features in the precipitation distribution (K. Campana and M. Kanamitsu 1987, personal communication). The correction has the value $(-6.5 \times z_{\text{atm}})$°K, where z_{atm} is the height of the atmospheric lower boundary in kilometers. Corrections are not applied explicitly to the SST in the ECHAM and CCM coupled models but are included implicitly in the representation of the topography.

Other CGCM forecast models that do not use empirical corrections at the atmosphere–ocean interface have been developed at the European Centre for Medium-Range Weather Forecasts (ECMWF; Stockdale 1997; Stockdale et al. 1998) and the Bureau of Meteorology Research Centre (BMRC; Wang et al. 2002).

c. Initial conditions

Initial conditions for the high-resolution ocean were taken from a 1980–99 ocean data assimilation produced at GFDL using variational optimal interpolation (Derber and Rosati 1989). The ocean model used for the data assimilation was identical (except for changes introduced for coupling) to the high-resolution version of the ocean model used in the forecasts. The OGCM was spun up for 20 yr using observed climatological boundary conditions to initialize the assimilation. The initial conditions for the medium-resolution ocean were pro-

duced by linear spatial interpolation of the high-resolution initial conditions. All forecasts used the unperturbed initial states directly from the analysis.

Atmospheric initial conditions were taken from long simulations made with the respective AGCMs forced by the observed time-dependent SST. The upper-ocean heat content is thought to contain the memory of the coupled system, at least in the equatorial ocean (e.g., Schneider et al. 1995), which provides a justification for using analyses of the ocean, together with a ''balanced'' atmosphere, as the initial states for the forecasts.

d. Cases and verification

Forty cases of forecasts were made with each of the five coupled models with 1 January and 1 July 1980–99 initial conditions, one 12-month forecast for each model and initial condition. Since the skill of the predictions depends strongly on the time of year of the initial state, results from the 1 January and 1 July initial conditions are presented separately, as are summary results for both initial times taken together. Results from the COLA model forecasts are also described in Schneider et al. (2001).

In addition to the cases described in the previous paragraph, two additional ensembles of 6-month forecasts were generated, one using COLA medium and the other using ECHAM medium. Each ensemble consisted of four perturbed initial condition cases for each of the 1 January initial states. The perturbations were in the atmospheric initial states only, obtained successively by integrating the AGCM for a day, and then resetting the internal clock of the AGCM to the initial time. These additional cases allow comparison of the effects of averaging across five-member multimodel and five-member single-model ensembles.

Long simulations were made with each of the medium-resolution coupled models. Multidecadal samples from these coupled simulations (50 yr for COLA medium, 40 yr for CCM medium, and 20 yr for ECHAM medium) were used to determine features of the simulated model climates.

The Smith et al. (1996) SST and the GFDL ODA were used for verification of the forecasts. Anomalies were calculated relative to the 1980–99 climatology. Skill was measured by anomaly correlation (correlation of the time series of observed and forecast anomalies) and rms error.

e. Atmospheric model surface fluxes

Comparison of the surface heat and momentum fluxes of the different AGCMs when forced by observed SST helps to understand the different systematic errors in the coupled models. Schneider (2002, his Fig. 11) compared the annual mean surface stresses and net heat fluxes from simulations with COLA V2 and CCM3 forced by the climatological annual cycle of SST. COLA V2 produced an easterly wind stress on the ocean that was about 0.1 dyn cm^{-2} weaker (more westerly) in the central and western equatorial Pacific than that produced by CCM3. The net heat flux into the ocean produced by COLA V2 was larger than that produced by CCM3 in the eastern equatorial Pacific, with differences of $+40$ W m^{-2} near 120°W, and smaller than CCM3's in the western equatorial Pacific, with differences of -20 W m^{-2} near 160°E. COLA V2 also produced weaker equatorial Pacific easterlies than the ECHAM4.5 (not shown) by about 0.3 dyn cm^{-2} near 160°W and larger net heat flux throughout the equatorial Pacific, with differences reaching $+50$ W m^{-2} near 160°W and $+20$ W m^{-2} near 160°E. The major differences between CCM3 and ECHAM4.5 were stronger easterlies in the central equatorial Pacific for ECHAM4.5 and larger net heat flux into the western equatorial Pacific for CCM3. These flux differences led to several degrees difference between the equatorial Pacific annual mean SST in the long coupled simulations (COLA medium SST was warm relative to CCM medium and ECHAM medium), differences in the mean east–west slope of the equatorial Pacific thermocline (COLA medium slope was weaker than CCM medium and ECHAM medium), and differences in the annual cycle of equatorial Pacific SST. Effects of the flux differences on the simulated ENSO variability in the long coupled simulations were also large (see section 4).

3. Results

In this section, the systematic errors (section 3a) and forecast skill of the SST anomalies (SSTAs) (sections 3b, c) and heat content anomalies (section 3d) from the different models are described and compared. Additionally, the skill of a multimodel averaged forecast, generated as the arithmetic mean of the results from the five models and referred to as ''Consensus,'' is included in the comparisons. Evaluation of more sophisticated statistical approaches for combining the results from the different models, such as those described by Krishnamurti et al. (2000) and Kharin and Zwiers (2002), is not considered here. The multimodel Consensus is compared to the ensemble means from five-member single-model ensembles for two of the models, and these single-model ensembles are also used to compare the influence of the different AGCMs in the two models (section 3e).

a. Systematic error of SST

Each coupled model has large systematic SST errors. These systematic errors are initially zero and evolve with time. The systematic error behavior also depends on the date of the initial condition. The average of each ensemble of forecasts as a function of model and start day (the average forecast evolution), which is equiva-

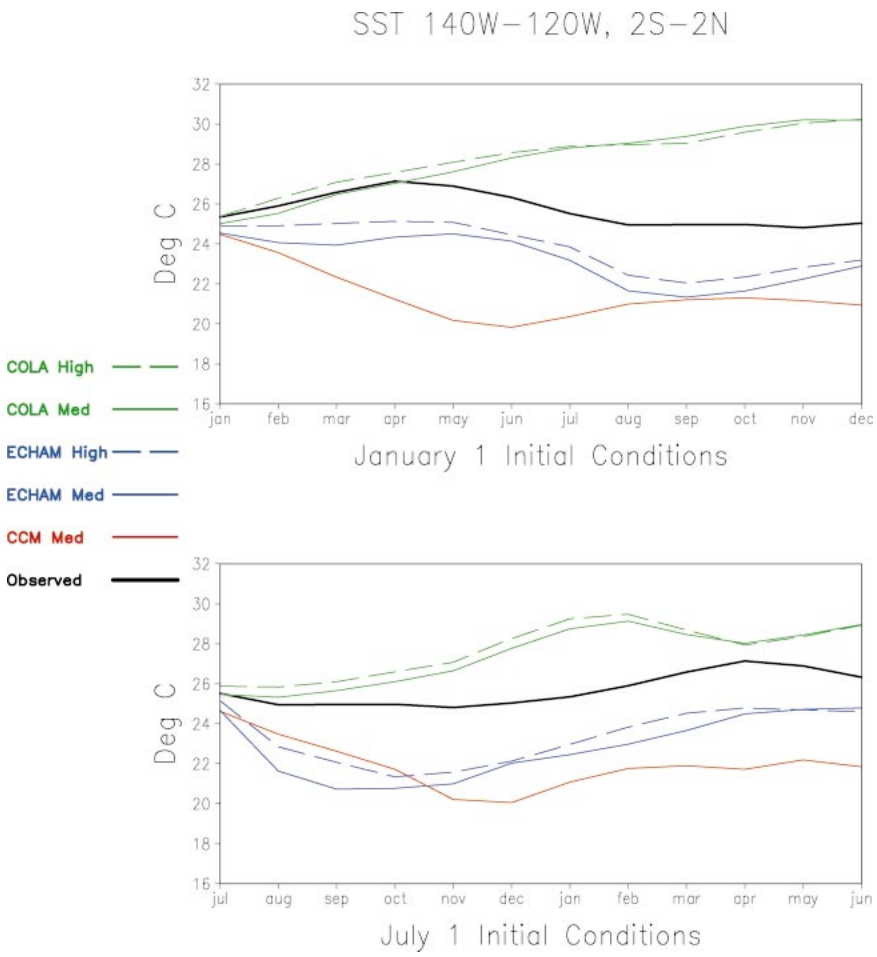

FIG. 1. Evolution of the SST in the region 2°S–2°N, 140°–120°W for each of the models and for observations. (top) Average over the 1 Jan initial condition retrospective forecasts and (bottom) average over the 1 Jul initial condition retrospective forecasts.

lently the sum of the systematic error and the annual cycle, is removed to obtain the predicted SSTA.

The average forecast evolutions of SST in a region in the eastern Pacific near 130°W at the equator (2°S–2°N, 140°–120°W) are shown in Fig. 1. This region was chosen because the largest systematic errors occur there and because the index is representative of the equatorial SST systematic errors throughout the central and eastern Pacific for all of the models. The model evolution is compared to the observed mean evolution over the same period (i.e., the annual cycle of SST for 1980–99). The difference of forecast minus observed evolution is interpreted as the systematic error of the model mean. Figure 1 shows that there is a strong dependence of this systematic error on AGCMs, but little dependence on ocean model resolution.

The COLA models have a warm bias in the eastern Pacific. For 1 January initial conditions, the COLA models warm throughout the 12-month forecast period, but the systematic error is small for the first few months since the observed annual cycle is also an increasing SST in the eastern Pacific. Later in the year, the COLA models continue to warm, while the observed SST cools, and the error becomes large. For 1 July initial conditions, both COLA model SSTs peak in the winter, and then the systematic error begins to decrease. This behavior can be simply described as a drift toward the model's somewhat weak in amplitude climatological annually varying SST (shown in Kirtman et al. 2002). The drift has about a 6-month timescale. Contrasting with the behavior in the Pacific, the average forecasts in the Indian and Atlantic Oceans (not shown) are phase locked to the calendar year more or less as observed, independent of start date. This is consistent with the smaller Indian and Atlantic Ocean errors in the coupled model climatology found in simulation mode (the long simulation described in section 2e) and consequently smaller drift in the forecasts.

In contrast to the COLA models, the ECHAM and CCM coupled models cool initially relative to observations and remain cold throughout the forecast period. After the initial few months, the ECHAM SSTs roughly

parallel the observed SST for both the 1 January and 1 July initial conditions, while the average CCM forecast is not phase locked to the annual cycle, cooling for the first 6 months and then warming slightly for both sets of initial conditions.

The spatial structure of the SST systematic error (not shown) is an El Niño–like wedge of warm water in the central and eastern Pacific for the COLA models. The ECHAM and CCM models have cold errors confined to a narrow band along the equator, extending from the eastern Pacific into the western Pacific. The broader structure of the COLA model errors could be indicative of errors in the AGCM heat flux, while the equatorially confined structure of the ECHAM and CCM errors could indicate equatorial easterly wind stresses that are too strong (Schneider 2002).

The different forecast models have a wide range of systematic SST errors and systematic error evolution, with the COLA models having a warm bias, and the ECHAM and CCM models having cold biases. However, as will be seen below, after the systematic evolutions are removed, it becomes more difficult to distinguish the forecasts of the anomalies by the different coupled models. Drifts occur not only in the mean state and annual cycle but also in the magnitude of the variability (see section 4).

b. Niño-3 SSTA retrospective forecasts

Anomaly correlations for the forecast SSTA in the Niño-3 region are shown in Fig. 2, and rms errors are shown in Fig. 3. These scores and others presented below are calculated using monthly mean data, so that what is referred to as the 1-month lead time forecast is the prediction of the first-month average. The anomaly correlations and rms errors for persistence forecasts are also shown. Persistence forecasts (called ''Persistence'' hereafter) are made from the monthly mean temperature averaged over the month prior to the date of initiation of the dynamic model forecast (i.e., December mean temperatures for 1 January initial conditions and June mean temperature for 1 July initial conditions). Since the ENSO variability is quasiperiodic, negative correlations for Persistence at lead times comparable to half the ENSO period (1–2 yr) contain useful information. This can be taken into account below by using the absolute value of the Persistence correlation as the measure of Persistence forecast skill against which to test the dynamical forecasts. However, the issue of negative correlation is not important for the evaluation of these forecasts.

The anomaly correlations show a strong seasonal dependence, with evidence of what has been termed the ''spring prediction barrier'' (Webster 1995; Chen et al. 1995). The anomaly correlations of Niño-3 SSTA forecasts with 1 January initial conditions (Fig. 2b) decay rapidly, losing ''useful'' skill (anomaly correlation decreasing below 0.6) in 3–6 months, depending on atmospheric model, whereas the forecasts from 1 July initial conditions (Fig. 2c) are skillful for all individual models for more than 9 months.

The forecast performance appears to depend more on the atmospheric model than the ocean resolution, especially when stratified by initial conditions. However, the sample size for the individual initial conditions (20 cases) is small, so that results are statistically distinguishable only at relatively weak significance levels. Detailed results from significance tests are described in section 1a of appendix A.

As shown in Fig. 2, the Consensus anomaly correlation is always the highest or very close to the highest. There are some limited time periods when the Consensus anomaly correlation is exceeded slightly by one or another of the models or by Persistence.

The Niño-3 SSTA rms errors (Fig. 3) show little to distinguish one model from another. After the first 3 months, the model forecasts have smaller errors than the Persistence or Climatology (zero anomaly) statistical forecasts (choosing whichever has smaller error) for either all initial conditions taken together or for July initial conditions. However, for January initial conditions the individual model forecasts have comparable rms errors to the statistical forecasts at all lead times out to 12 months. The rms error of Consensus is consistently smaller than that of any of the individual models or statistical forecasts. Significance tests are described in section 1b of appendix A.

c. Regional skill for SST anomalies

Point correlations (anomaly correlation with the analysis evaluated at each point in space, averaged over time) for SSTAs are described in this subsection (significance tests in appendix A, sections 3a,b) and for heat content in section 3d (significance tests in appendix A, sections 3d,e). Also, the individual forecasts by each model for four specific cases at 6-month lead are presented in appendix B.

The SSTA correlations for all initial conditions, computed as a function of lead time, are shown in Figs. 4 and 5 for 3- and 6-month leads, respectively. Correlations with magnitude >0.3 are significantly different from zero at the 10% level. Correlations <-0.4 and >0.4 are shown by colored shading, with negative values allowed in order to show regions where the forecasts convey information concerning quasiperiodic processes or mismatch of spatial structure. However, large negative correlations are rarely found.

The 3-month-lead correlations for the model forecasts are similar for all model versions (Figs. 4c–g) in both the regional distribution and magnitude. Correlations are above 0.6 for much of the near-equatorial central and eastern Pacific.

The model forecasts for 6-months lead time (Fig. 5) have correlations exceeding 0.6 in a region restricted to

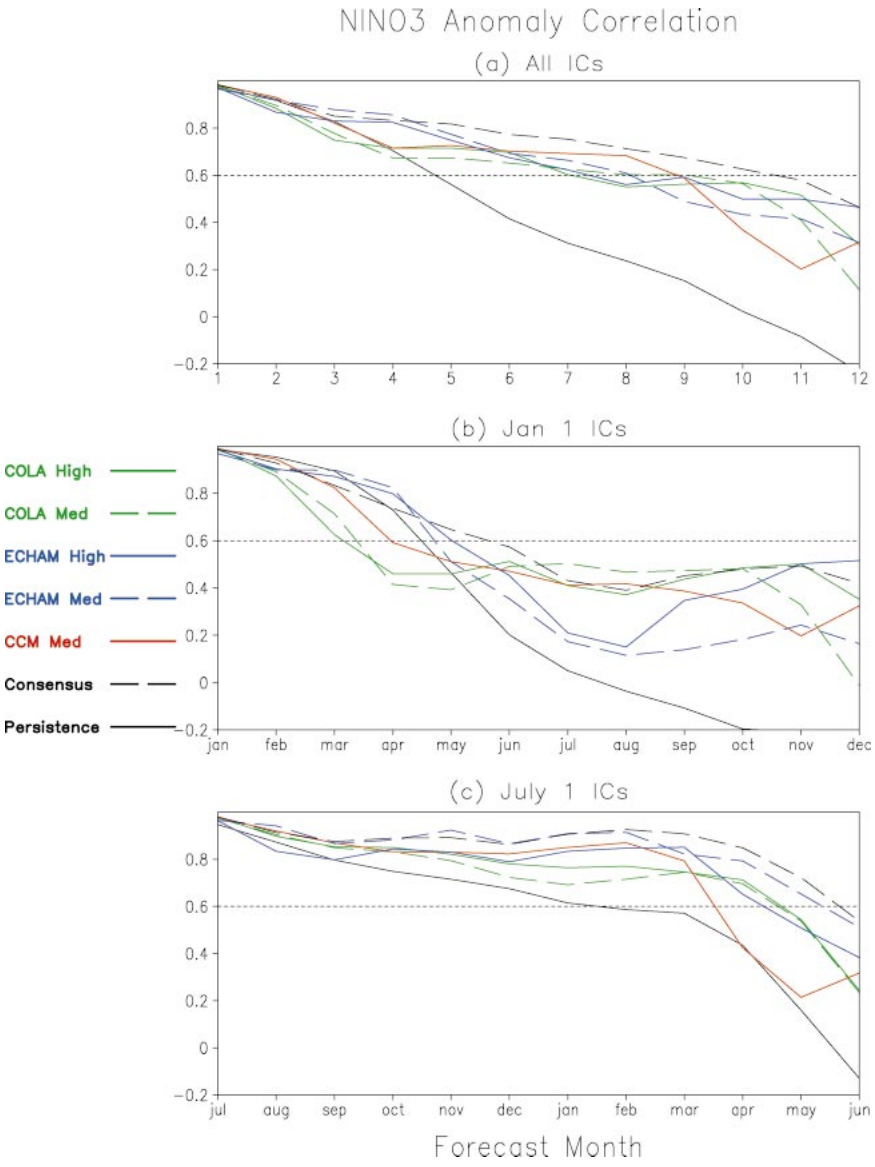

FIG. 2. Niño-3 SSTA correlations between retrospective forecasts and analysis for the models, Consensus, and Persistence as a function of lead time: (a) All initial conditions combined, (b) 1 Jan initial conditions, and (c) 1 Jul initial conditions.

the eastern equatorial Pacific and slightly weighted toward the Southern Hemisphere.

The 3- and 6-months-lead January and July initial state Consensus SSTA correlations are shown in Fig. 6. All of the models and Persistence have similar patterns (not shown) for the corresponding initial conditions and lead times.

While some statistically significant differences are found for SSTA correlations from models with different AGCMs, results from models with the same AGCM (COLA high versus COLA medium and ECHAM high versus ECHAM medium) are statistically indistinguishable.

The significance of a spring prediction barrier for SSTAs is examined by testing the hypothesis that the July initial condition Consensus correlation is higher than the January initial condition Consensus correlation for equivalent lead times (appendix A, section 2c). The statistical test provides support for the notion of a spring prediction barrier for SST in this set of forecasts.

d. Evaluation of heat content predictions

The evolution of the predicted upper-ocean heat content (defined here as the vertically averaged temperature, a measure of the thermocline depth) is primarily the response of the ocean to the evolution of the wind stress. Heat content anomalies were available from the verti-

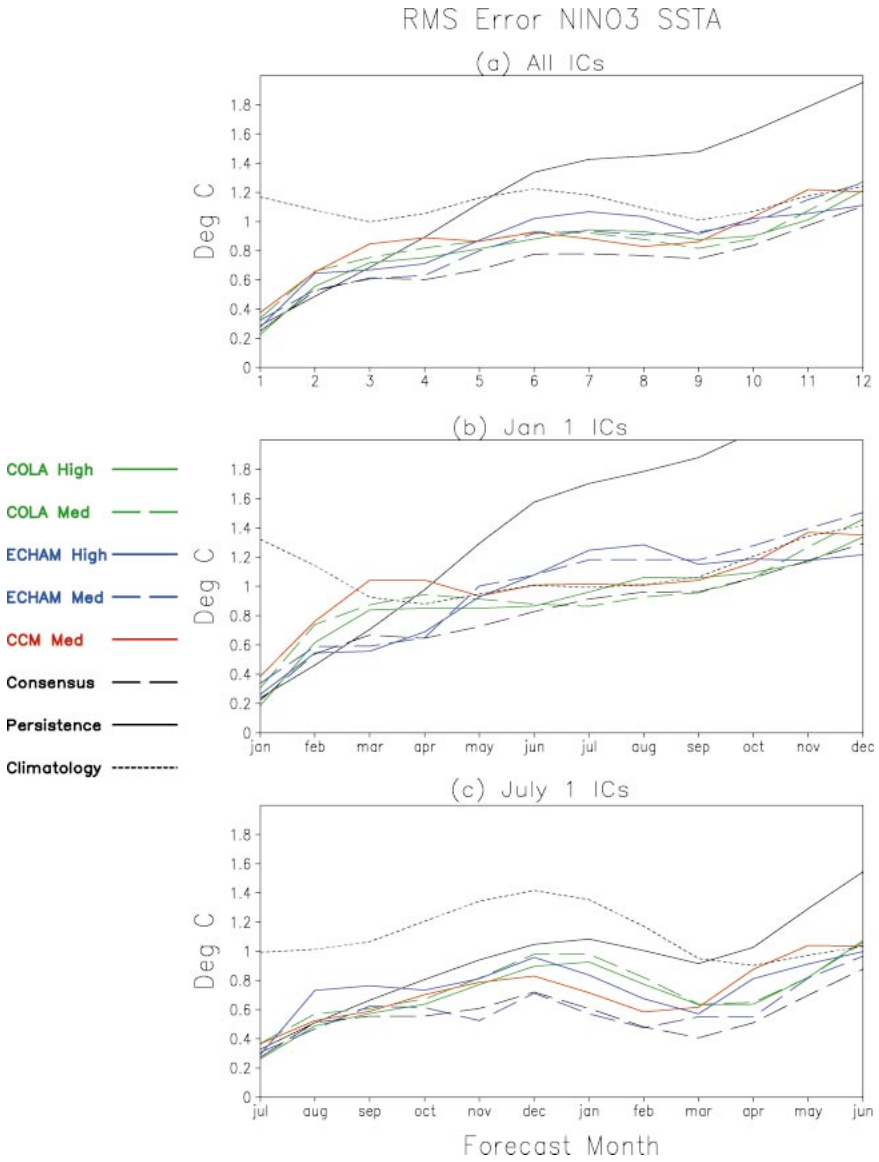

FIG. 3. Rms errors of retrospective forecasts of Niño-3 SSTAs for the models, Consensus, Persistence, and Climatology (zero-anomaly retrospective forecast), as a function of lead time: (a) All initial conditions combined, (b) 1 Jan initial conditions, and (c) 1 Jul initial conditions.

cally averaged temperature over the top 250 m for COLA medium, CCM medium, and the GFDL ODA; over the top 300 m for COLA high; and over the top 200 m for the ECHAM models. Heat content anomalies were produced for each by removal of the annual cycle.

Figures 7 and 8 show the correlations of the heat content anomaly forecasts for all initial conditions at 3- and 6-months lead, respectively. The individual models are able to make predictions of the heat content anomalies equal to or better than Persistence in similar regions. In the Tropics at 3 and 6 months, these regions include the tropical Pacific south of the equator, the eastern equatorial Pacific, the western tropical Pacific to the north of the equator, the tropical Indian Ocean to the south of the equator, and the equatorial Atlantic. All of the models as well as Persistence have difficulty forecasting the equatorial Pacific heat content near the date line; however, the COLA coupled models appear to have the most difficulty forecasting this quantity at 6-months lead time. The regions of high correlation of heat content and SST correspond in the far eastern Pacific, but not in the equatorial central Pacific, where at 3 months for all models and 6 months for the COLA models, SST but not heat content correlations are high. The coupling of heat content and SST anomalies in the far eastern Pacific is similar to that usually assumed in simple intermediate coupled models, while the decoupling of heat content and SST anomalies in the central equatorial Pa-

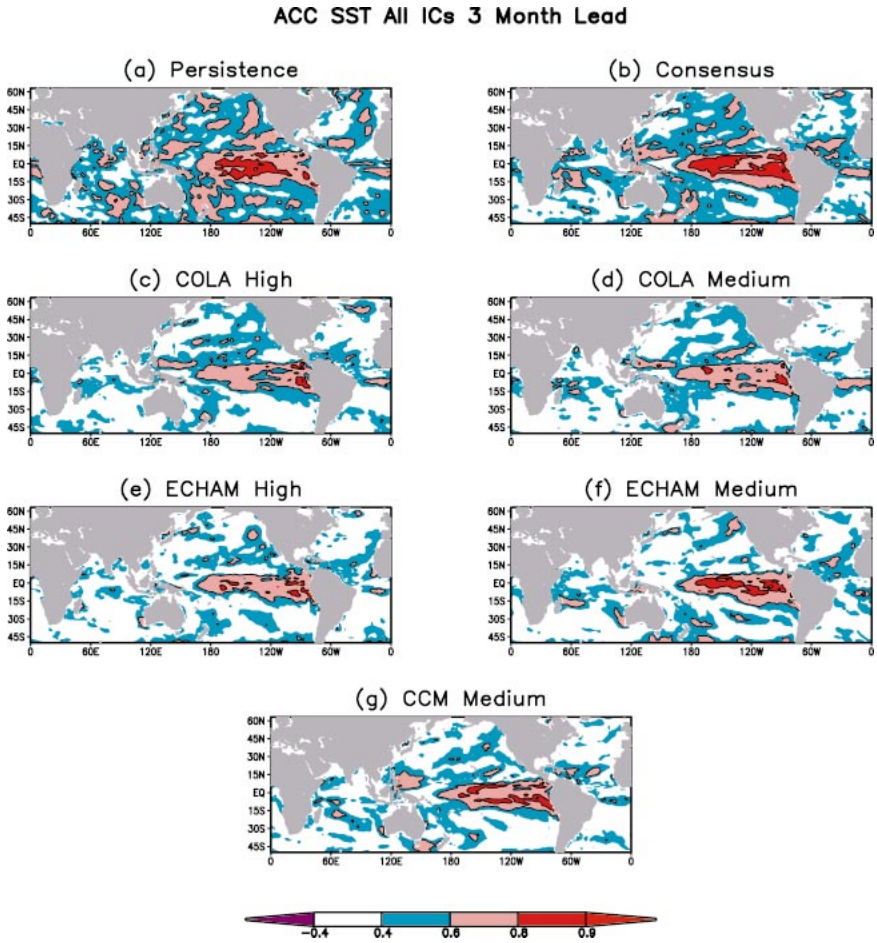

FIG. 4. Point correlation of SST retrospective forecasts with analysis for all initial conditions at 3-months lead time for (a) Persistence, (b) Consensus, (c) COLA high, (d) COLA medium, (e) ECHAM high, (f) ECHAM medium, and (g) CCM medium.

cific is consistent with an important role for zonal advection in the SSTA budget there (Huang and Schneider 1995; Picaut et al. 2001). As in the case of SST, Consensus provides the best and most coherent heat content forecast statistics. The 6-month lead time Consensus forecast (Fig. 7b) has correlations >0.6 over most of the Pacific within 15° latitude of the equator, with the region of high correlation centered on the equator to the east of the date line and in bands flanking the equator in the western Pacific. This structure also occurs in the separate model forecasts, although it is less coherent.

The 3- and 6-months-lead January and July initial state Consensus heat content anomaly correlations are shown in Fig. 9. The models and Persistence have similar patterns (not shown) for the corresponding initial conditions and lead times.

While some statistically significant differences in heat content correlations are found from models with different AGCMs, results from models with the same AGCM (COLA high versus COLA medium and ECHAM high versus ECHAM medium) are statistically indistinguishable.

The significance of a spring prediction barrier for heat content is examined by testing the hypothesis that the July initial condition Consensus correlation is higher than the January initial condition correlation for equivalent lead times (appendix A, section 2f). The statistical test provides support for the notion of a spring prediction barrier for heat content as well as SST in this set of forecasts.

e. Single-model ensembles

Data for a comparison of the effect of single-model versus multimodel ensembling on forecast skill was produced by making five-member ensembles out to 6-months lead time from 1 January initial conditions for both the ECHAM medium and COLA medium models. These ensembles consisted of the original case plus four additional forecasts for the respective models. Initial

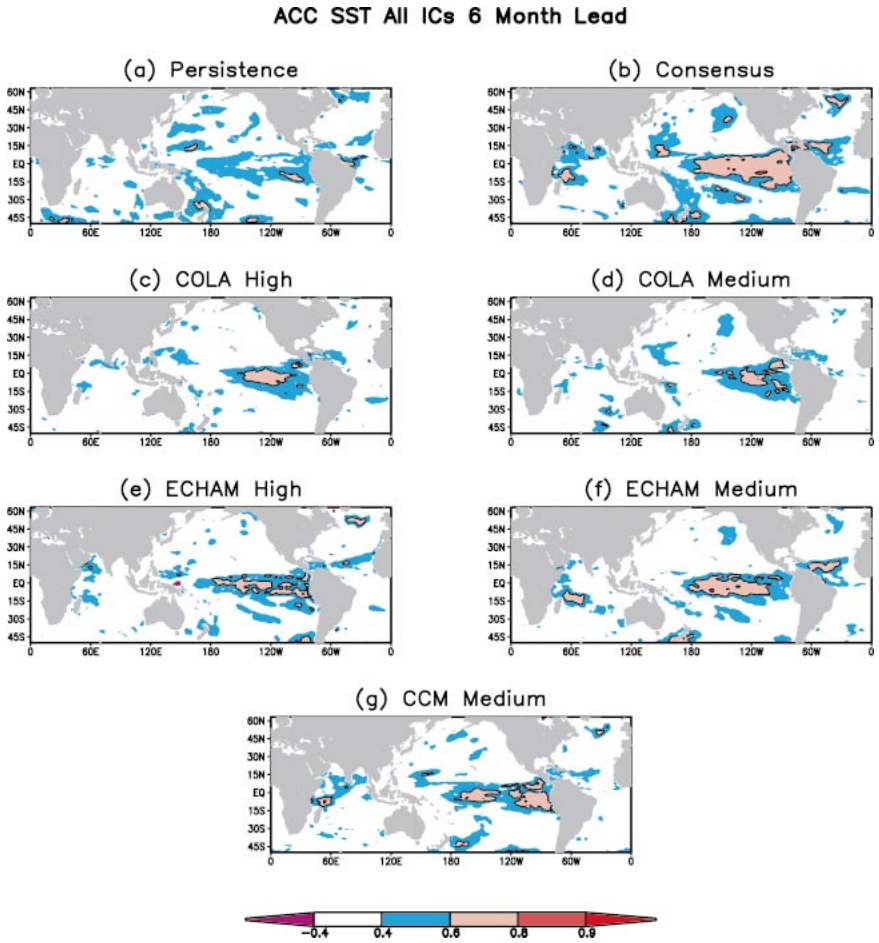

FIG. 5. As in Fig. 4, but for 6-months lead time.

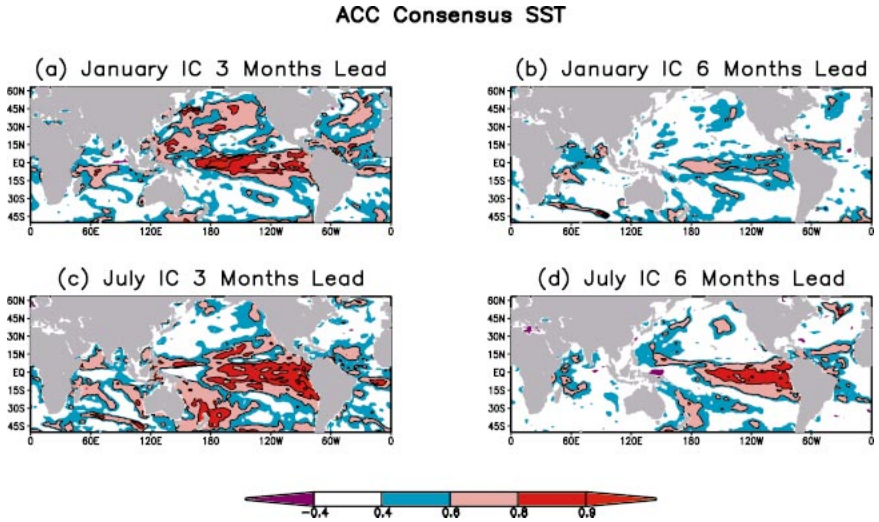

FIG. 6. Point correlation of Consensus SSTA retrospective forecasts with analysis for (a) Jan initial conditions at 3-months lead, (b) Jan initial conditions at 6-months lead, (c) Jul initial conditions at 3-months lead, and (d) Jul initial conditions at 6-months lead.

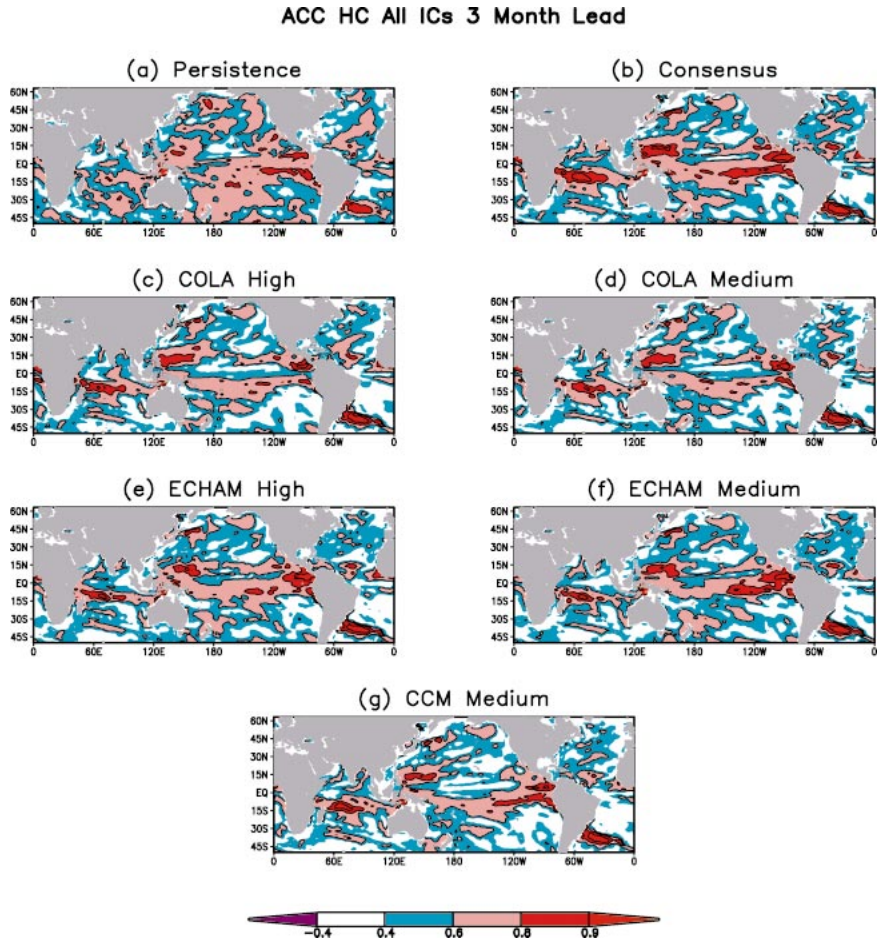

FIG. 7. As in Fig. 4, but for heat content anomalies.

state perturbations for the AGCMs were generated as described in section 2e. The Niño-3 correlations for the multimodel consensus and single-model ensemble means are shown in Fig. 10. Results are described in detail in appendix C. Only one significant difference in skill between the five-model multimodel ensemble and the five-member single-model ensembles was found: the 3-month-lead western Pacific heat content was better predicted by Consensus than the ECHAM ensemble. The comparison between multimodel and single-model ensembles demonstrates that both approaches lead to improvement. Although there is no clear statistical distinction between the two types of ensembles, perhaps because of the small sample size, there is a suggestion that the benefits can occur in different regions for each approach, as would be expected if the effect of multimodel ensembling was to reduce the systematic forecast errors.

4. Relationships between systematic error and forecast error

An evaluation of the effects of systematic model errors on the forecasts is made from a comparison of statistical properties of the forecasts with those of long simulations made with several of the coupled models.

The evolution of the standard deviation of anomalies from the mean annual cycle in the equatorial Pacific for analysis, Consensus, and the separate models is shown in Fig. 11 for SST and in Fig. 12 for heat content. The results are stratified into January and July initial conditions. This measure of model variability is computed by averaging the square of the anomalies over the 20 January and July initial condition cases as a function of lead time (i.e., month of year) and then taking the square root. Anomalies are calculated by removing the mean annual cycle as before. If the model and initial conditions were perfect, then this measure of the model variability would have the same annual cycle as the analysis. Differences between the forecasts and analysis can be due to either errors in the initial conditions or errors in the model formulation (model error), or both.

For comparison, the annual cycle of the standard deviation of SST and heat content anomalies calculated from multidecadal samples taken from long integrations with the COLA medium, CCM medium, and ECHAM medium models are shown in Fig. 13. If the models

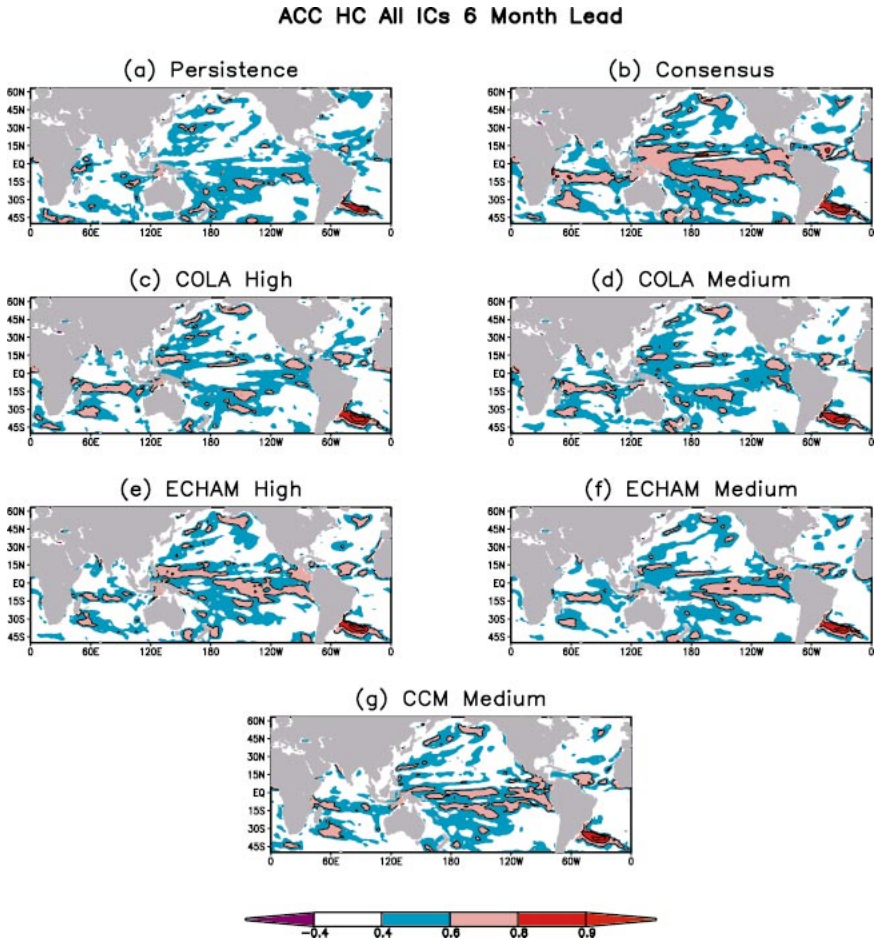

FIG. 8. As in Fig. 4, but for heat content anomalies at 6-months lead time.

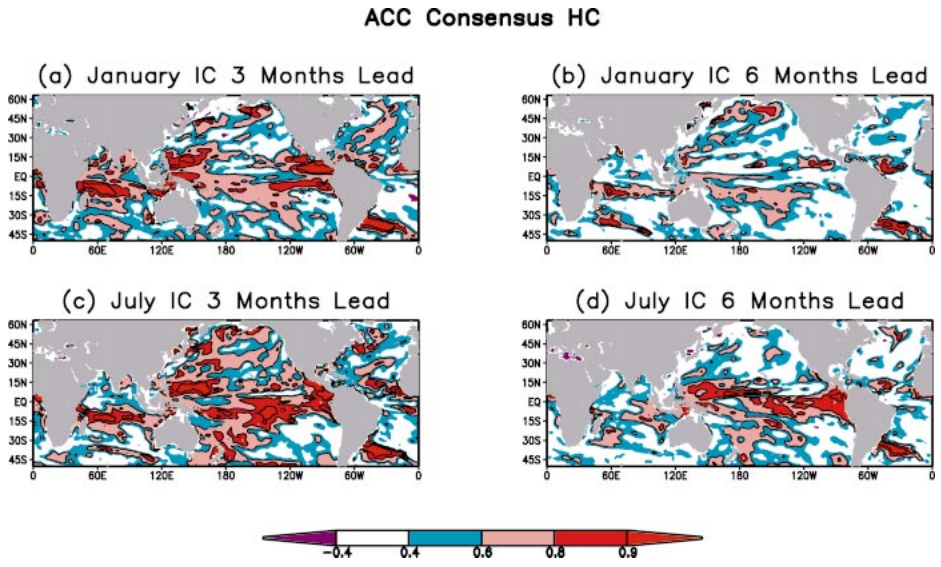

FIG. 9. As in Fig. 6, but for Consensus heat content anomaly retrospective forecasts.

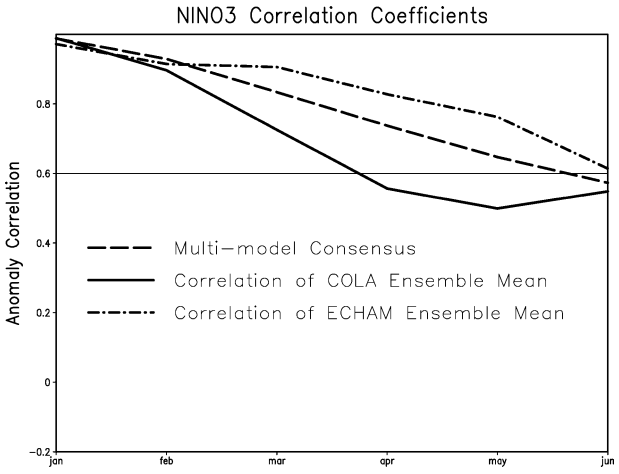

FIG. 10. As in Fig. 2, but for the five-case single-model and multimodel ensemble mean retrospective forecasts.

were perfect, then this measure of the model variability would be close to that of the analysis shown in Figs. 11a and 12a. Large differences between the analysis and the simulations are presumed to be due only to model error. When the model is started from observed initial conditions, which are far from the model's simulated climate, the expectation is that there will be an ''initial shock'' followed by some period of adjustment, after which the properties of the variability will approach those found in the long simulations. Exactly how fast and in what manner this adjustment proceeds will determine in large part how model error evolves and degrades forecasts made from analyzed initial conditions. Comparison of model variability in forecast mode with variability in simulation mode then gives information on how model error degrades potential forecast skill but does not address the role of initial condition error.

None of the model simulations produces particularly realistic equatorial Pacific variability in simulation mode. Comparison of the SST variability of Figs. 11a and 13 (left column) shows that the models' SST variability is weak relative to the analysis, especially for CCM medium. Also, COLA medium simulates a boreal summer maximum near 120°W instead of a spring maximum near 90°W, as observed. The location and phasing of the ECHAM variability compares best with the analysis, but the amplitude is weak. Comparison of Figs. 12a (left) and 13 (right) also shows that the heat content variability in simulation mode is also unrealistically weak, especially for CCM medium. Initial SST and heat content anomalies with the observed magnitude are then not compatible with the coupled models' lower level of equilibrium variability.

All of the models experience an initial shock in SST to varying degrees for January initial states (Figs. 11c–g, left). Maximum SSTA variability occurs near the beginning of the forecasts (January–February) and decays strongly thereafter. Examination of Fig. 13 (left-hand column) confirms that this initial variability is far larger than found in the models' equilibrium level of internal variability. The initial shock is particularly prominent for COLA medium and CCM medium. The effects of the initial shock extend through at least several-months lead time. The COLA models' forecast SSTA variability (Figs. 11c,d) begins to resemble that from the simulation (Fig. 13a) after about 6-months lead time, with a secondary maximum in September–October–November (SON), which, as is the case in simulation mode, is weaker, and several months and several tens of degrees longitude displaced from the analyzed April maximum. Effects of the initial shock in the ECHAM models (Figs. 11e,f) also seem to last several months, and the evolution of variability begins to resemble that in the long simulation (Fig. 13b) in the latter part of the forecast. In the case of CCM medium (Fig. 11g), however, the model does not relax to near its simulated climatology (Fig. 13c) out to 12-months lead, and an unrealistic secondary maximum of SST variability develops in the eastern Pacific at about 6-months lead. Consensus (Fig. 11b) also contains a strong initial shock and does not resemble the observed variability.

The forecast heat content variability for January initial conditions (Figs. 12c–g, left) also shows evidence of initial shock when compared to the variability in simulation mode (Fig. 13, right). The variability is initially larger than in simulation mode. The COLA and ECHAM variability appears to adjust towards the simulation pattern after about 10-months lead. CCM medium maintains a higher level of variability than in the simulation out to 12-months lead. Consensus variability merely decays toward zero.

The July initial condition forecasts appear to suffer less initial shock than the January initial condition cases. The SST variability (Figs. 11c–g, right) near the July initial conditions is not as large as in the January initial condition cases, although all models still exceed the observed variability at 1-month lead. The COLA models (Figs. 11c,d) approach the weaker simulation mode pattern at about 3-months lead, while the ECHAM (Figs. 11e,f) and CCM (Fig. 11g) models maintain observed levels of variability, substantially larger than that in simulation mode, out to about 8-months lead. CCM medium develops spurious variability in the west-central Pacific at about 5-months lead, similar to the behavior in the January initial condition case. Consensus variability also resembles the observed before decaying strongly at about 8-months lead. The July initial conditions heat content variability (Figs. 12c–g, right) behaves much more realistically than that for the January initial conditions out to the following spring, with the models showing less evidence of initial shock than in the January cases, despite having larger variability than in simulation mode and each producing an eastward propagating maximum in the eastern Pacific in boreal winter of amplitude close to 2°C. Consensus variability is maintained out to about 8-months lead.

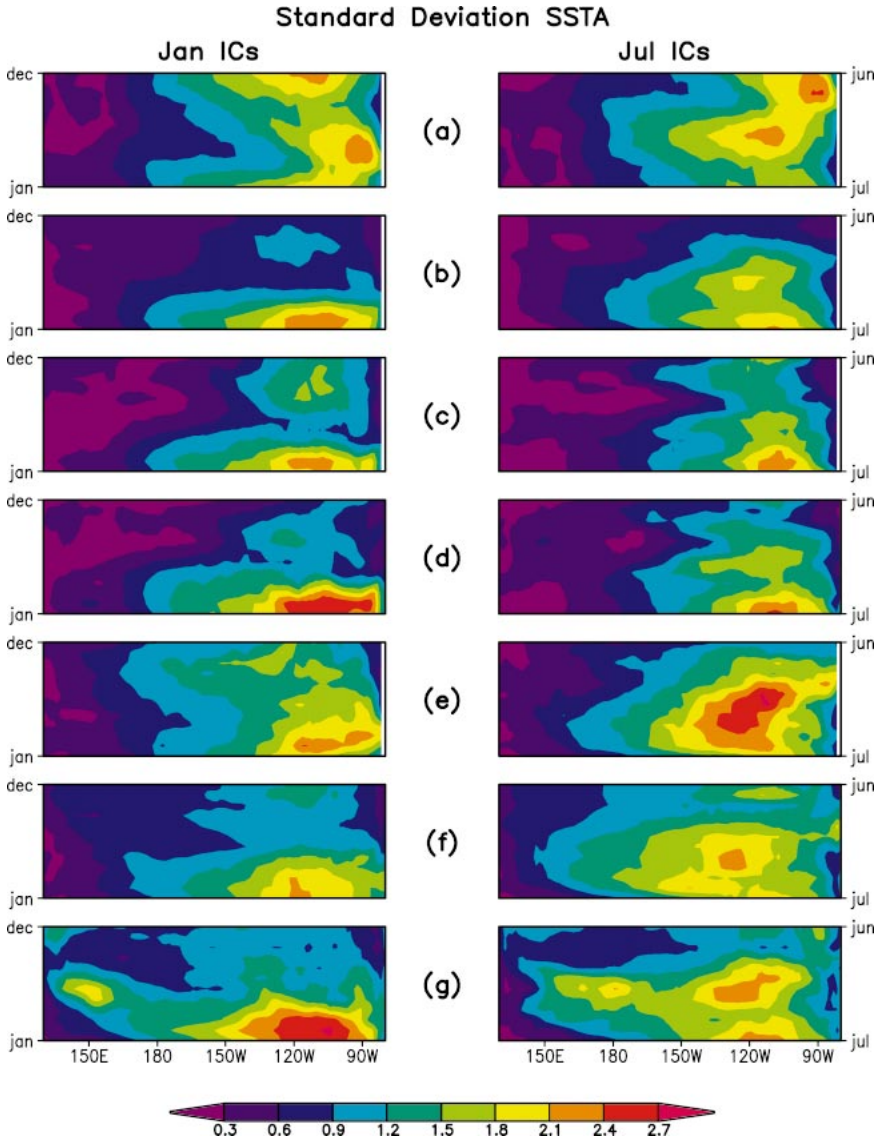

FIG. 11. Longitude–time sections of the evolution of the std dev of SST anomalies (°C) in the equatorial Pacific as a function of month of year for analysis and retrospective forecasts: (a) analysis 1980–99, (b) Consensus, (c) COLA high, (d) COLA medium, (e) ECHAM high, (f) ECHAM medium, (g) CCM medium. Model results in left-hand column are averaged over cases with Jan initial conditions, and in right-hand column the averages are over the cases with Jul initial conditions. Time increases upward in (a)–(g).

The July initial condition forecasts suffer less from initial shock than the January initial condition forecasts. The model climate seems to be more sensitive to initial imbalances and amplifies them more strongly in January than July. However, spring appears to be a barrier to variability as well as forecast skill, so that the variability is damped to the simulation amplitude only during that part of the year. This property of the forecast variability does not appear not to depend strongly on the sign of the drift error of the mean state SST.

Another problem that could have implications for making forecasts with the coupled models is that the 3-month lead time forecasts (Fig. 4) are inferior to persistence over much the World Ocean (although the number of cases is not large enough for many of these differences to be statistically significant). The persistence correlations of the observations can be considered to be an intrinsic statistic of the climate system (e.g., 3-month lag autocorrelation relative to December and June). The inferior performance of the models could a priori be due to either problems with the initial state/initialization or to errors in the coupled-model physics that would cause the models to have different statistical properties than nature. To examine the latter possibility, the persistence

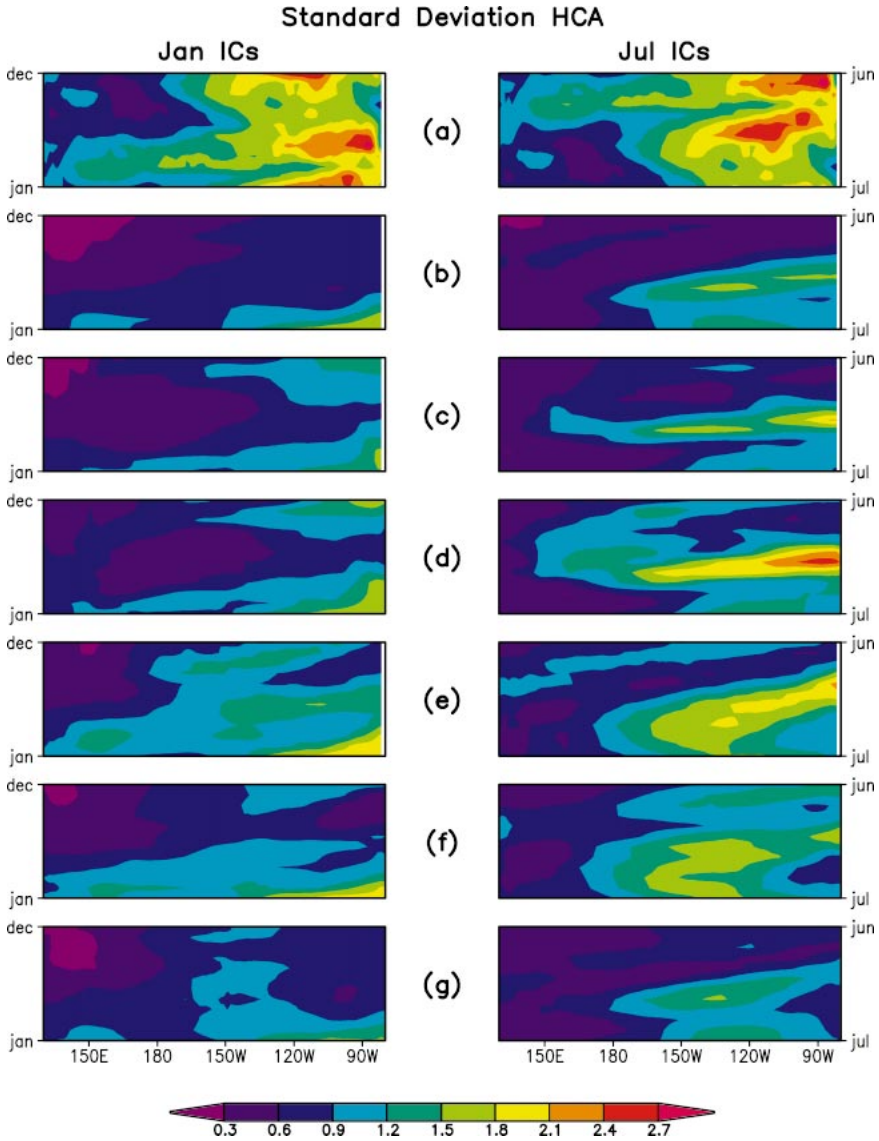

FIG. 12. As in Fig. 11, but for std dev of heat content anomalies (°C).

statistics were calculated from the multidecade simulations with the COLA medium, ECHAM medium, and CCM medium models. The correlations of persistence forecasts of SSTAs in the long integration were found for December and June initial states for 3-month lead time, using data from the model evolutions (i.e., the same start and end months as in the Persistence forecasts shown in Fig. 4a). This procedure is equivalent to finding the 3-month lag autocorrelations of the model SSTA. If the model is correctly simulating the statistical properties of the natural system, then these lag autocorrelations should be the same as those found from evaluation of the observed SST evolution. The results are shown in Fig. 14. Comparison of Figs. 14 and 4a shows that the coupled models are less persistent than nature throughout the Tropics. This is a (statistically significant) systematic error in the higher-order statistics of the model that is probably responsible, at least in part, for the poor performance of the model predictions with respect to persistence. We do not know at this time whether this lack of simulated persistence is due to problems with the AGCM (e.g., too much wind stress or heat flux variability), the OGCM (e.g., mixed-layer depth too shallow), or the coupled system.

5. Conclusions

Sets of forty 12-month retrospective forecasts of SST, 20 starting on 1 January 1980–99 and 20 starting on 1 July 1980–99, have been performed using several coupled atmosphere–ocean GCMs differing in the AGCM or the resolution of the OGCM. The systematic errors

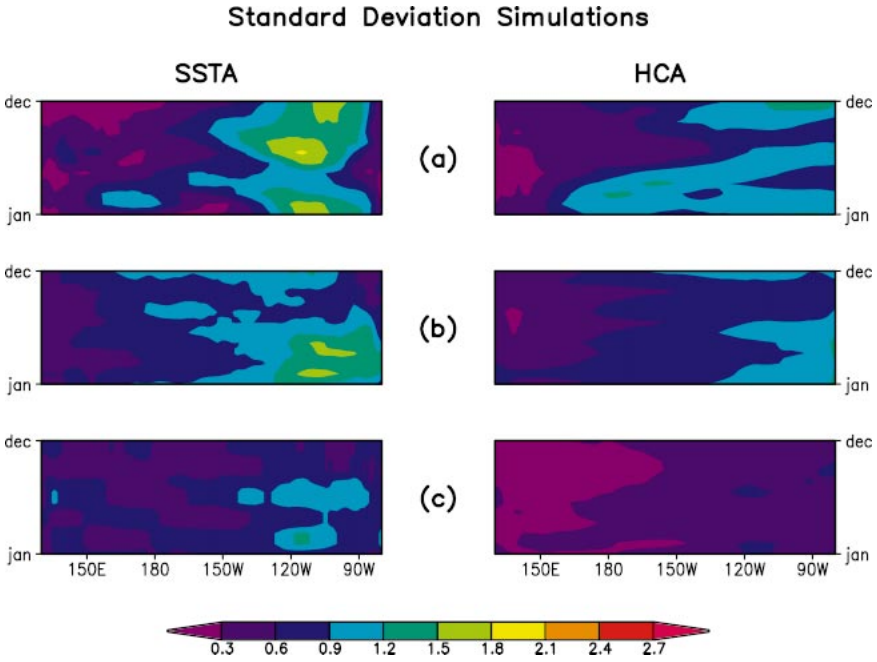

FIG. 13. Longitude–time sections of the annual cycle of the std dev of (left) SST anomalies (°C) and (right) heat content anomalies (°C) in the equatorial Pacific from long simulations with the coupled models: (a) COLA medium, (b) ECHAM medium, (c) CCM medium. Time increases upward in (a)–(c).

of the coupled predictions are large, seasonally dependent, and clearly a function of the AGCM—the COLA models have a strong warm bias, CCM a strong cold bias, and ECHAM a moderate cold bias. Although the sample size is too small to distinguish between the results from different models with a high level of significance, the skill in predicting the anomalies appears to depend mostly on the AGCM and not the resolution of the OGCM. Retrospective forecast results from ensembles made with the same AGCM but different OGCM resolution are statistically indistinguishable from each other at the 10% level. Results from ensembles made

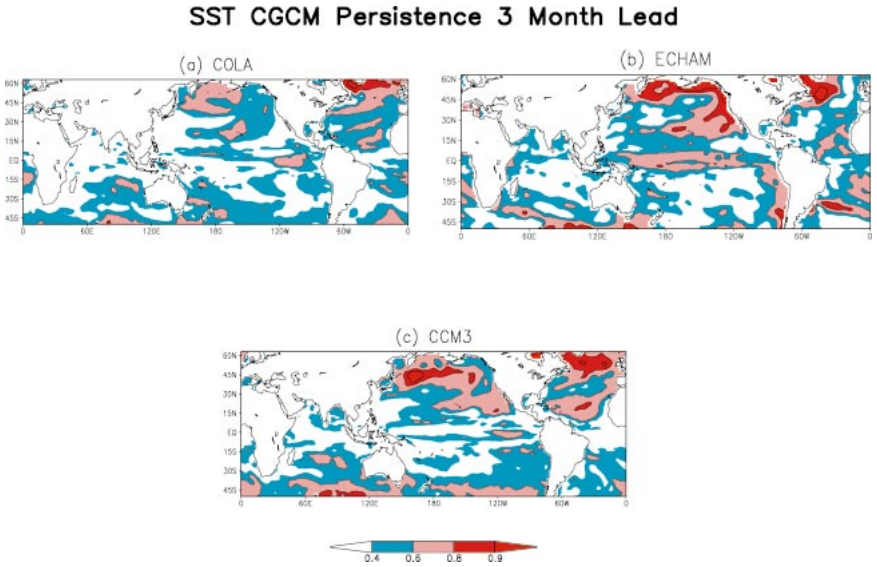

FIG. 14. Three-month lag autocorrelation for long runs of the coupled GCMs: (a) COLA medium, (b) ECHAM medium, (c) CCM medium, for Dec (–Mar) and Jun (–Sep) base (–end) months. These are the statistics of perfect-model Persistence retrospective forecasts corresponding to Fig. 4a.

with different AGCMs, regardless of the OGCM resolution, can be distinguished from each other at the 10% level in some situations, although the percentage area covered by regionally coherent significant differences is small.

No single model is consistently better than all of the others—different models perform best, based both on anomaly correlation and rms error, depending on lead time, initial time, and location. A summary of significant differences found between models for SST is that the ECHAM models tended to be better than the COLA models in the eastern and central equatorial Pacific, whereas the COLA models and CCM medium tended to be better than the ECHAM models in the western Pacific. For equatorial Pacific heat content, the ECHAM models tended to be better than the COLA models (eastern and central Pacific) and CCM (eastern and western Pacific), and there were situations in which CCM was better than ECHAM or COLA in the central Pacific. If the number of cases was extended to order 100 or greater and the correlation levels found in the extended ensembles remained similar to those found here, statistical significance would be greatly improved and more precise and useful conclusions could be made.

The Consensus forecast, defined as the arithmetic average of the five separate model forecasts, is consistently (but not always) superior to any single model. In many situations, the improvement over individual models from the multimodel Consensus is statistically significant. The improvements from multimodel averaging are expected to be due to the reduction of systematic errors in the variability caused by model deficiencies, as well as reduction in noise, while improvements from averaging over an ensemble made with a single model are due to improving the signal-to-noise ratio. A limited comparison between Consensus and five-member single-model ensembles made with two of the models for January initial conditions yielded some statistically significant relationships. However, the sample size was not large enough to clearly distinguish between the multimodel Consensus and the single-model ensembles. While simple averaging is an effective way of combining forecasts, more sophisticated statistical techniques could help even more. One obvious approach to pursue for producing improved forecasts is then to optimally and intelligently combine ensembles of forecasts made with different models (e.g., Krishnamurti et al. 2000; Kharin and Zwiers 2002).

The forecast skill is also seasonally dependent, with the ''spring prediction barrier'' playing an important role. A spring prediction barrier in both SST and heat content was found to be supported by statistically significant differences between point correlations from January and July initial conditions at equivalent lead times.

In any retrospective forecast study, there are several potential sources of artificial skill because of explicit and implicit use of dependent data in the forecast evaluation. The sources of artificial skill that we are aware of in the results described here include 1) prior tuning of the atmospheric models and ocean models using observed data from the forecast period, 2) computing anomalies relative to the average of all forecasts independently for each model and initial condition date, and 3) verification against the ocean data assimilation in regions of sparse observational data and, consequently, comparison of model against model. We feel that the artificial skill introduced by these procedures is probably minor. Real-time prediction and dissemination of the predictions and verification of a sufficient sample of the real-time forecasts is the only sure way to eliminate problems of artificial skill or worse. This approach, however, will necessarily require a reasonable sample of future ENSO events to produce definitive results.

There are several problems that stand out in the coupled models. First, the models have severe systematic errors in the mean SST that develop on the timescale of months. These systematic errors can lead to errors in the forecasts in several ways. One possible effect is that mean state errors in tropical SST can cause erroneous coupling between the atmosphere and ocean. For example, SST errors in key regions will lead to convection occurring in the wrong locations, producing erroneous surface wind stress and heat fluxes that will further amplify the SST errors.

Additional problems appear in higher-order statistics. The coupled models were shown to produce erroneous SST and heat content variability (primarily too little) in the tropical Pacific in simulation mode, as well as SST lag autocorrelations that were erroneously small. The coupled system will therefore necessarily be out of balance when observed initial conditions are used for the ocean, since this initial state is not compatible with the climate of the coupled model. The coupled model will then experience an initial shock that could be transmitted rapidly to remote locations by oceanic wave propagation and lead to degradation of the forecasts both locally and remotely through coupled interactions. Evidence was presented from the forecasts that there is a more severe initial shock for January than July intial conditions. From this analysis, it is apparent that the spring predictability barrier is associated with a spring variability barrier. The adjustment of the models to their own (weak) variability climatology is not uniform, but rather occurs mostly in the boreal spring. If this association is important for prediction, then reducing model error has a large potential for reducing the spring prediction barrier effect and improving skill.

Future research will have to address the problems enumerated above in order to eliminate potential causes of forecast error. There are some ad hoc fixes that have already been developed to address the problems of systematic error of the mean SST and initial shock. The procedure of anomaly coupling has been used for some time (Ji et al. 1994; Kirtman et al. 1997) to minimize mean SST errors, and initial shock has been reduced by the technique of anomaly initialization (Latif et al. 1993;

Schneider et al. 1999). However, the problems of the role of erroneous equilibrium variability and of errors in the lag autocorrelation of simulated SST and their relationship to the model predictions have not previously been emphasized. Consequently, ways of reducing these errors have not yet been developed.

Acknowledgments. There are several scientists who made major contributions in the design and implementation of this project. These include Steve Zebiak (IRI); J. Shukla, Bohua Huang, Yun Fan, and Zhengxin Zhu (COLA); Ants Leetmaa and Matt Harrison (GFDL); Peter Gent and Jeff Lee (NCAR); and Dave Behringer and Ming Ji (NCEP). Support for this research for Schneider and Kirtman was provided by the National Science Foundation (ATM 98-14295 and ATM 01-22859), the National Oceanic and Atmospheric Administration (NA 96-GP0056), and the National Aeronautics and Space Administration (NAG 5-8202). DeWitt's participation was funded by a grant/cooperative agreement from the National Oceanic and Atmospheric Administration (NA 07-GP0213). The views expressed herein are those of the authors and do not necessarily reflect the view of NOAA or any of its subagencies. Computer resources were provided by the CSL at NCAR.

APPENDIX A

Statistical Significance Tests

The results were tested for significance using standard *t* and *F* tests, with the number of degrees of freedom taken to be the number of cases in the sample. Since the number of cases is small (either 20 or 40), a significance level of 10% was used to test the hypotheses that statistics are different or that one statistic is a priori larger than another.

1. Significance of Niño-3 SSTA correlations and rms errors

a. Niño-3 SSTA

Statistical significance tests were applied to the statistical hypotheses that correlations in Fig. 2 are different from zero, that correlations are different from each other, that Consensus is higher than other correlations, and that Persistence is lower than other correlations. At 3-months lead for January initial conditions, the only differences that are significant are between the ECHAM medium and both COLA models, ECHAM high and COLA high, and Persistence and COLA high. Consensus is better than COLA high. At 6-months lead, none of the models are distinguishable from the others or Consensus, although Consensus is significantly better than Persistence. Consensus remains statistically higher than zero out to 12-months lead time. For the July initial condition cases, none of the correlations are statistically distinguishable from each other at 3-months lead. At 6- and 9-months lead, the models and Consensus are not statistically different; however, the Consensus correlation is higher than Persistence. At 12-months lead, only the Consensus and ECHAM medium correlations are significantly different from zero. When the ensemble size is increased to 40 by combining the January and July initial conditions, ECHAM medium and COLA high are different at 3-months lead. At 6- and 9-months lead, the models and Consensus are statistically indistinguishable, and the models and Consensus are better than Persistence.

b. Niño-3 rmse

Differences of the rms errors of the Niño-3 SSTA predictions shown in Fig. 3 were tested for significance. For all initial conditions Consensus error is statistically smaller than that for the other curves in the following situations: CCM medium and COLA medium 1–5-months lead and ECHAM high 5–9-months lead. For January initial conditions, Consensus produces no significant improvement. For July initial conditions, Consensus is better than Persistence/Climatology 4–11-months lead, ECHAM high 2–10-months lead, COLA high and COLA medium 7–9-months lead, CCM medium 9–11-months lead.

2. Significance of regional skill

a. SSTA: All initial conditions

All initial states are taken together for a sample size of 40 in significance testing. Results from stratifying the cases into 1 January and 1 July initial states (sample sizes of 20) are also considered. The statistical significance level of the hypotheses that the correlations differ between models and that Consensus is better than other correlations have been evaluated. Comparisons are described only when there are subjectively coherent regions significant at the 10% level.

Significant differences in the 3-month-lead correlations (Fig. 4) are as follows: Consensus, COLA high, COLA medium, and CCM medium correlations are higher than the ECHAM high or medium correlations in a region in the far western Pacific (0°–12°N, 120°–150°E). Consensus is higher than COLA high in the eastern equatorial Pacific (5°S–5°N, 170°–110°W).

At 6-months lead (Fig. 5), Consensus has larger correlations than Persistence in the eastern equatorial Pacific, consistent with the higher Niño-3 anomaly correlations of the models shown in Fig. 2. Consensus is also better than Persistence in the western Indian Ocean (40°–60°E near 10°N and 10°S). However, Persistence is competitive with Consensus, even for 6-months lead time, in the central equatorial Pacific and the tropical Atlantic. A small region of significant negative corre-

lation is found in the far western Pacific in the ECHAM high forecasts. This results from an SSTA structure that has equatorial anomalies of the same sign in the western Pacific as those in the central and eastern Pacific, rather that of the opposite sign, as in the well-known observed structure (e.g., Kirtman and Shukla 2002). Consensus as well as both ECHAM models and CCM medium are different from (higher than) COLA high and COLA medium in the equatorial Pacific near the dateline (0°–7°S, 170°E–160°W). Consensus is superior to ECHAM medium in the western Pacific (0°–10°N, 140°–160°E).

b. SSTA: January and July initial conditions

For January and July initial conditions separately (Fig. 6), regions where Consensus gives significant improvement and where models are significantly different from each other are also regions of high correlation in the relevant Consensus panel of Fig. 6, since differences between correlations turn out to be significant only when at least one of the correlations is high. At 3-months lead for January initial conditions, Consensus (Fig. 6a) is better than ECHAM high or medium in a region in the far western Pacific (0°–12°N, 120°–150°E), and CCM medium is better than the ECHAM models in these regions. At 6-months lead for January initial conditions, all cases are statistically indistinguishable. At 3-months lead for July initial conditions, the COLA models are better than the ECHAM and CCM models in the western equatorial Pacific (140°–170°E, near 5°N). At 6-months lead for July initial conditions, Consensus (Fig. 6d) is better than the COLA models in the central equatorial Pacific (0°–7°S, 170°E–160°W), the ECHAM and CCM models are better than the COLA models in this region, and Consensus is better than Persistence in the eastern tropical Pacific.

c. Spring prediction barrier for SSTA

Referring to Fig. 6, at 3-months lead, the July initial condition correlation (Fig. 6c) is not significantly higher than the January initial condition correlation (Fig. 6a) in the tropical Pacific. At 6-months lead, the July initial conditions correlation (Fig. 6d) is significantly higher than the January initial conditions correlation (Fig. 6b) throughout much of the eastern equatorial Pacific (date line–South American coast, 10°S–10°N).

d. Heat content anomalies: All initial conditions

At 3-months lead (Fig. 7), Consensus is better than Persistence in the bands of high correlation in Fig. 7b (10°S in the western Indian Ocean, 10°N in the western Pacific, and the equatorial far eastern Pacific). The ECHAM models and Consensus are better than the COLA models in the equatorial eastern Pacific (date line to South America near the equator). At 6-months lead (Fig. 8), Consensus is better than Persistence over most of the tropical Pacific. All of the other models are better than the COLA models in the central equatorial Pacific (5°S–5°N, 170°E–160°W), and Consensus is better than the COLA models in a similar region that extends farther east–west.

e. Heat content anomalies: January and July initial conditions

For January and July initial conditions separately (Fig. 9), regions where Consensus gives significant improvement and where models are significantly different from each other are also areas of high correlation in the corresponding Consensus panel of Fig. 9. At 3-months lead for January initial conditions, Consensus (Fig. 9a) is better than the COLA models in the eastern equatorial Pacific (140°W to the South American coast in a narrow band near the equator), CCM medium in the eastern and western Pacific (140°W to the South American coast and 130°–170°E in narrow bands near 7°N), ECHAM high in the western Pacific (120°–140°E in a band near 10°N), and Persistence in the Indian Ocean (40°–80°E in bands near 7°N and 7°S). The ECHAM models are better than the other models in the respective regions where correlations for the COLA and CCM models are lower than Consensus. For January initial conditions at 6-months lead, Consensus (Fig. 9b) is better than Persistence in the central equatorial Pacific (160°E–150°W at the equator) and in the Indian Ocean (5°S–5°N, 70°–100°E and African coast–80°E near 10°S). ECHAM medium is better than the COLA models in the central equatorial Pacific (160°E–140°W in a narrow band along the equator). For July initial conditions at 3-months lead, Consensus (Fig. 9c) is better than Persistence in the high-correlation regions of Fig. 9c in the western Pacific near 10°N, in the eastern Pacific along the central American coast, and in the Indian Ocean near 10°S. At 6-months lead for July initial conditions, Consensus (Fig. 9d) is better than the COLA models in the central equatorial Pacific (5°S–5°N, 180°–120°W), CCM in a narrow band 130°E–160°W near 10°N, ECHAM medium in the western equatorial Pacific (5°S–5°N, 140°–160°E), and Persistence over the regions in the tropical Pacific in Fig. 9d where Consensus correlation > 0.8. The ECHAM and CCM models are better than the COLA models in the central equatorial Pacific region where Consensus is better than the COLA models. The ECHAM models are better than CCM medium in the region where Consensus is better than CCM medium, and CCM medium is better than ECHAM medium in the western Pacific region where Consensus is better than ECHAM medium.

f. Spring prediction barrier for heat content anomalies

Referring to Fig. 9, at 3-months lead, the July initial condition correlation (Fig. 9c) is significantly higher

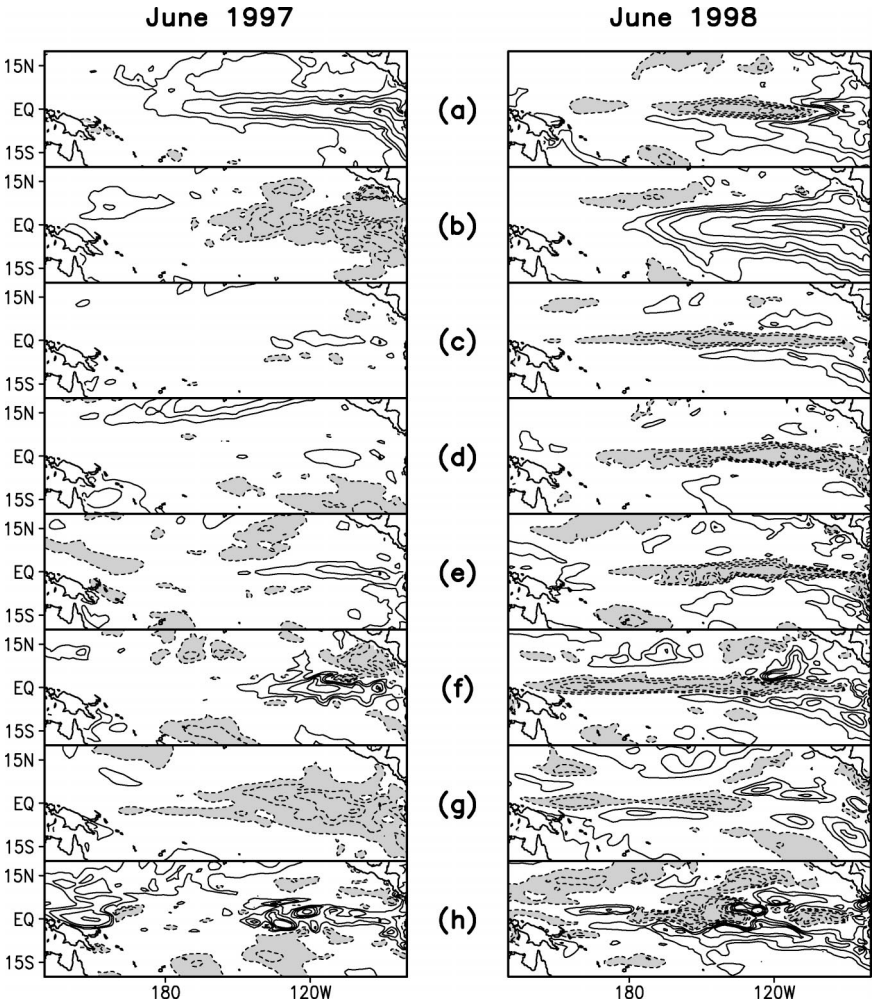

FIG. B1. (left) Jun 1997 and (right) Jun 1998 SSTAs for (a) analysis of observations and for 6-months-lead retrospective predictions (Jan 1997 and Jan 1998 initial conditions), (b) Persistence, (c) Consensus, (d) COLA high, (e) COLA medium, (f) ECHAM high, (g) ECHAM medium, and (h) CCM medium. Contour levels are at integer values (°C), with additional contours at ±0.5°C and zero contour suppressed.

than the January initial condition correlation (Fig. 9a) in several regions in the central Pacific between 15°S and 15°N (regions of correlation >0.8 in Fig. 9c located 150°–180°E and 140°–160°W). At 6-months lead, the July initial conditions correlation (Fig. 9d) is significantly higher than the January initial conditions correlation (Fig. 9b) in bands that extend across the whole tropical Pacific (following the bands of correlations >0.8 in Fig. 9d from 120°E–150°W near 7°N and the band from 180°W to South America along the equator).

APPENDIX B

Case Studies

The 6-month lead time forecast SSTA for 1997 and 1998 are shown in Figs. B1 (1 January initial conditions, June mean verification) and B2 (1 July initial conditions, December mean verification). These case studies illustrate the spread in the forecasts from the individual models and show how they combine in the Consensus forecasts.

A strong warm event was observed in June 1997 (Fig. B1a, left) that developed from a negative anomaly 6 months before (Fig. B1b, left). Persistence (Fig. B1b, left) was not a successful predictor. The multimodel Consensus (Fig. B1c, left) was for a weak/near-neutral warm event. Consensus was made up of individual forecasts (Figs. B1d–h, left) ranging from moderately strong warming (two models), weak warming (two models), to moderate cooling (one model). The event observed in June 1998 (Fig. B1a, right), with cold SSTAs in the central equatorial Pacific and moderate warm anomalies in the far eastern equatorial Pacific, developed from a strong warm anomaly 6 months before (Fig. B1b, right).

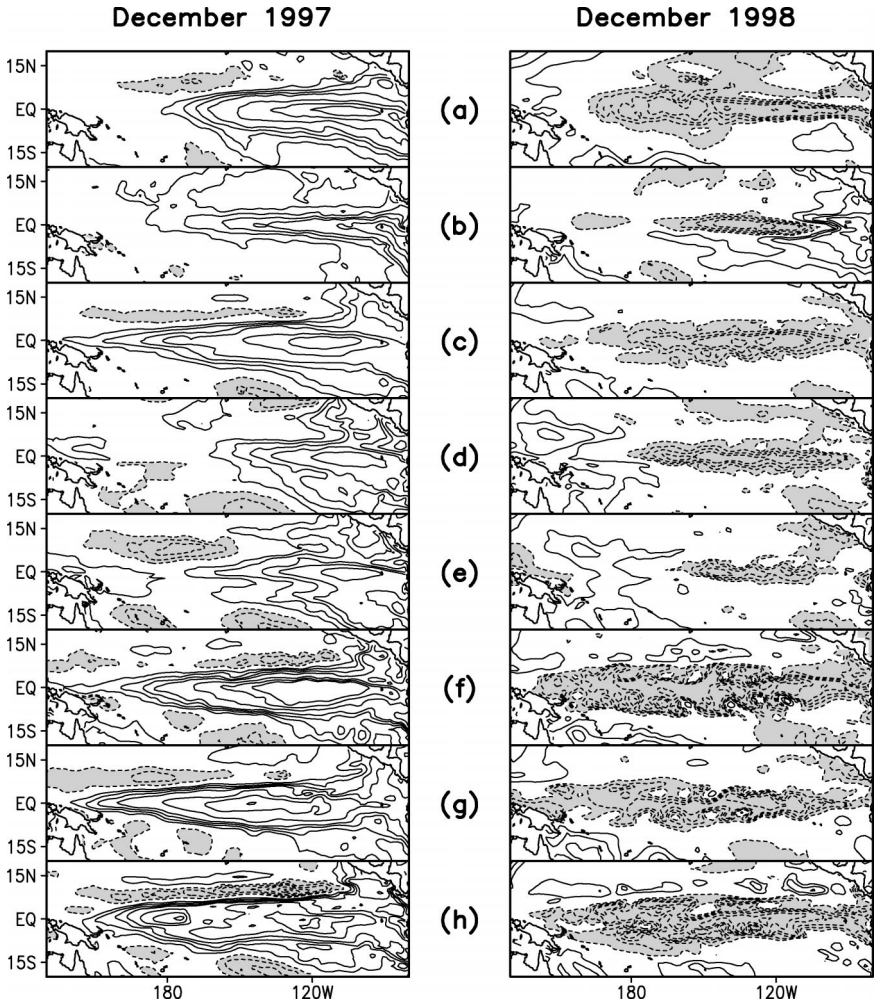

FIG. B2. As in Fig. B1, but for (left) Dec 1997 and (right) Dec 1998 SSTAs.

Persistence (Fig. B1b, right) was a good predictor for the far eastern Pacific, but not the central Pacific. Consensus and all of the individual models (Figs. B1c–h, right) produced central equatorial Pacific SSTAs of the observed sign and more or less the observed magnitude. The models produced SSTAs of mixed sign in the far eastern equatorial Pacific. Several of the models (Figs. B1f–h, right) produced anomalies that were too strong in the far western equatorial Pacific. Consensus produced an intermediate structure that tends to average out the structural errors in the individual models.

The strong warm event was near peak in December 1997 (Fig. B2a, left). The pattern persisted from 6 months earlier (Fig. B2b, left) and grew slightly in amplitude. The forecasts (Figs. B2c–h, left) were all reasonably successful in maintaining and increasing the warm SSTA in the eastern Pacific. However, the SSTA was restricted too far to the east in the COLA models (Figs. B2d,e, left) and extended too far to the west in the ECHAM and CCM models. Similar comments, but with the sign reversed, apply to the cold event of December 1998 (Figs. B2a–h, right). In this case, the forecasts were all reasonably successful in maintaining and increasing the cold SSTA in the eastern Pacific. However, the SSTA was restricted too far to the east in the COLA models and extended too far to the west in the ECHAM and CCM models. The restriction of the COLA anomalies to the far eastern Pacific is consistent with extended coupled simulations (Kirtman et al. 2002).

APPENDIX C

Single-Model Ensembles

At 3- and 6-months lead, the single-model ensemble mean correlations shown in Fig. 10 are higher than those found for the single realizations shown in Fig. 2. The improvement for the ECHAM medium model is more than that for the COLA medium model, which seems to be related to a larger spread in the correlations for the individual ECHAM medium forecasts. However, the statistical relationship between the individual models

and Consensus is not changed from that described in section 1a of appendix A for the single-model realizations. At 3-months lead, differences in the correlations between the ECHAM medium and COLA medium single-model ensemble means are significant at the 10% level, while neither single-model ensemble mean can be distinguished from the multimodel Consensus. At 6-months lead, none of the correlations shown in Fig. 10 are statistically different.

Differences in point correlations between the ensemble means were also examined and tested for significance at the 10% level. At 3-months lead for SST, Consensus is better than ECHAM medium in the same region in the western Pacific as was found in section 2b of appendix A (0°–12°N, 120°–150°E) for the single-member set of forecasts. At 3-months lead for heat content, ECHAM medium is different from (higher than) COLA medium in the same regions, as was the case for the single-member set of forecasts. No coherent regions of significant difference are found for SST correlation at 6-months lead. At 6-months lead for heat content, ECHAM medium is different from (higher than) COLA medium in the near-equatorial central South Pacific (0°–10°S, 180°–140°W).

REFERENCES

Briegleb, B. P., 1992: Delta-Eddington approximation for solar radiation in the NCAR Community Climate Model. *J. Geophys. Res.,* **97,** 7603–7612.

Brinkop, S., and E. Roeckner, 1995: Sensitivity of a general circulation model to parameterization of cloud–turbulence interactions in the atmospheric boundary layer. *Tellus,* **47A,** 197–220.

Chen, D., S. E. Zebiak, A. J. Busalacchi, and M. A. Cane, 1995: An improved procedure for El Nino forecasting. *Science,* **269,** 1699–1702.

Davies, R., 1982: Documentation of the solar radiation parameterization in the GLAS climate model. NASA Tech. Memo. 83961, 57 pp. [Available from NASA Center for AeroSpace Information (CASI), 7121 Standard Dr., Hanover, MD 21076-1320.]

Derber, J., and A. Rosati, 1989: A global oceanic data assimilation system. *J. Phys. Oceanogr.,* **19,** 1333–1347.

DeWitt, D. G., 1996: The effect of the cumulus convection scheme on the climate of the COLA general circulation model. COLA Rep. 27, 58 pp. [Available from COLA, 4041 Powder Mill Rd., Suite 302, Calverton, MD 20705.]

——, and E. K. Schneider, 1996: The earth radiation budget as simulated by the COLA GCM. COLA Rep. 35, 39 pp. [Available from COLA, 4041 Powder Mill Rd., Suite 302, Calverton, MD 20705.]

Fouquart, Y., and B. Bonnel, 1980: Computations of solar heating of the earth's atmosphere: A new parameterization. *Beitr. Phys. Atmos.,* **53,** 35–62.

Gent, P. R., and J. C. McWilliams, 1990: Isopycnal mixing in ocean circulation models. *J. Phys. Oceanogr.,* **20,** 150–155.

Hack, J. J., 1994: Parameterization of moist convection in the National Center for Atmospheric Research Community Climate Model (CCM2). *J. Geophys. Res.,* **99,** 5551–5568.

Harshvardhan, R. Davies, D. A. Randall, and T. G. Corsetti, 1987: A fast radiation parameterization for atmospheric circulation models. *J. Geophys. Res.,* **92** (D1), 1009–1016.

Holtslag, A. A. M., and B. A. Boville, 1993: Local versus nonlocal boundary-layer diffusion in a global climate model. *J. Climate,* **6,** 1825–1842.

Huang, B., and E. K. Schneider, 1995: The response of an ocean general circulation model to surface wind stress produced by an atmospheric general circulation model. *Mon. Wea. Rev.,* **123,** 3059–3085.

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

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

Kiehl, J. T., J. J. Hack, and B. P. Briegleb, 1994: The simulated earth radiation budget of the National Center for Atmospheric Research community climate model CCM2 and comparisons with the Earth Radiation Budget Experiment (ERBE). *J. Geophys. Res.,* **99,** 20 815–20 827.

——, ——, G. B. Bonan, B. A. Boville, B. P. Briegleb, M. D. L. Williamson, and P. J. Rasch, 1996: Description of the NCAR Community Climate Model (CCM3). NCAR Tech. Note NCAR/TN-420+STR, 152 pp.

——, ——, ——, ——, D. L. Williamson, and P. J. Rasch, 1998: The National Center for Atmospheric Research Community Climate Model: CCM3. *J. Climate,* **11,** 1131–1149.

Kirtman, B. P., and J. Shukla, 2002: Interactive coupled ensemble: A new coupling strategy for CGCMs. *Geophys. Res. Lett.,* **29,** 1367, 10.1029/2002GL014834.

——, ——, B. Huang, Z. Zhu, and E. K. Schneider, 1997: Multiseasonal predictions with a coupled tropical ocean–global atmosphere system. *Mon. Wea. Rev.,* **125,** 789–808.

——, Y. Fan, and E. K. Schneider, 2002: The COLA global coupled and anomaly coupled ocean–atmosphere GCM. *J. Climate,* **15,** 2301–2320.

Krishnamurti, T. N., C. M. Kishtawal, Z. Zhang, T. E. LaRow, D. R. Bachiochi, C. E. Willifor, S. Gadgil, and S. Surendran, 2000: Multimodel ensemble forecasts for weather and seasonal climate. *J. Climate,* **13,** 4196–4216.

Lacis, A. A., and J. E. Hansen, 1974: A parameterization for the absorption of solar radiation in the earth's atmosphere. *J. Atmos. Sci.,* **31,** 118–133.

Large, W. G., J. C. McWilliams, and S. C. Doney, 1994: Oceanic vertical mixing: A review and a model with a nonlocal boundary layer parameterization. *Rev. Geophys.,* **32,** 363–403.

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.

Louis, J. F., 1979: A parametric model of vertical eddy fluxes in the atmosphere. *Bound.-Layer Meteor.,* **17,** 187–202.

Mellor, G. L., and T. Yamada, 1982: Development of a turbulence closure model for geophysical fluid processes. *Rev. Geophys. Space Phys.,* **20,** 851–875.

Moorthi, S., and M. J. Suarez, 1992: Relaxed Arakawa–Schubert: A parameterization of moist convection for general circulation models. *Mon. Wea. Rev.,* **120,** 978–1002.

Morcrette, J.-J., L. Smith, and Y. Fouquart, 1986: Pressure and temperature dependence of the absorption in longwave radiation parameterizations. *Beitr. Phys. Atmos.,* **59,** 455–469.

Neelin, J. D., D. S. Battisti, A. C. Hirst, F.-F. Jin, Y. Wakata, T. Yamagata, and S. E. Zebiak, 1998: ENSO theory. *J. Geophys. Res.,* **103,** 14 261–14 290.

Nordeng, T. E., 1994: Extended versions of the convective parameterization scheme at ECMWF and their impact on the mean and transient activity of the model in the Tropics. ECMWF Research Department Tech. Memo. 206, European Centre for Medium-Range Weather Forecasts, 41 pp. [Available from ECMWF, Shinfield Park, Reading, Berkshire, RG29AX, United Kingdom.]

Pacanowski, R. C., and S. M. Griffies, 1998: MOM 3.0 manual. NOAA/Geophysical Fluid Dynamics Laboratory, Princeton, NJ, 668 pp.

Picaut, J., M. Ioualalen, T. Delcroix, F. Masia, R. Murtugudde, and J. Vialard, 2001: The oceanic zone of convergence on the eastern edge of the Pacific warm pool: A synthesis of results and im-

plications for El Niño–Southern Oscillation and biogeochemical phenomena. *J. Geophys. Res.,* **106,** 2363–2386.
Ramanathan, V., and P. Downey, 1986: A nonisothermal emissivity and absorptivity formulation for water vapor. *J. Geophys. Res.,* **91,** 8649–8666.
Redi, M. H., 1982: Oceanic isopycnal mixing by coordinate rotation. *J. Phys. Oceanogr.,* **12,** 1155–1158.
Rockel, B., E. Raschke, and B. Weyres, 1991: A parameterization of broad band radiative transfer properties of water, ice and mixed clouds. *Beitr. Phys. Atmos.,* **64,** 1–12.
Roeckner, E., 1995: Parameterization of cloud radiative properties in the ECHAM4 model. *Proc. WCRP Workshop on Cloud Microphysics Parameterizations in Global Atmospheric Circulation Models,* Kananaskis, AB, Canada, WCRP, 105–116.
——, and Coauthors, 1996: The atmospheric general circulation model ECHAM4: Model description and simulation of present day climate. Rep. 218, Max-Planck-Institut fur Meteorologie, 90 pp. [Available from MPI fur Meteorologie, Bundesstr. 55, 20146 Hamburg, Germany.]
Schneider, E. K., 2002: Understanding differences between the equatorial Pacific as simulated by two coupled GCMs. *J. Climate,* **15,** 449–469.
——, B. Huang, and J. Shukla, 1995: Ocean wave dynamics and El Niño. *J. Climate,* **8,** 2415–2439.
——, ——, Z. Zhu, D. G. DeWitt, J. L. Kinter III, B. Kirtman, and J. Shukla, 1999: Ocean data assimilation, initialization, and predictions of ENSO with a coupled GCM. *Mon. Wea. Rev.,* **127,** 1187–1207.
——, B. P. Kirtman, Y. Fan, and Z. Zhu, 2001: Retrospective ENSO forecasts: The effect of ocean resolution. COLA Tech. Rep. 109, 27 pp. [Available from COLA, 4041 Powder Mill Rd., Suite 302, Calverton, MD 20705.]
Smagorinsky, J., 1963: General circulation experiments with the primitive equations: I. The basic experiment. *Mon. Wea. Rev.,* **91,** 99–164.
Smith, T. M., R. W. Reynolds, R. E. Livezey, and D. C. Stokes, 1996: Reconstruction of historical sea surface temperature using empirical orthogonal functions. *J. Climate,* **9,** 1403–1420.
Stockdale, T. N., 1997: Coupled ocean–atmosphere forecasts in the presence of climate drift. *J. Climate,* **10,** 809–818.
——, D. L. T. Anderson, J. O. S. Alves, and M. A. Balmaseda, 1998: Global seasonal rainfall forecasts using a coupled ocean–atmosphere model. *Nature,* **392,** 370–373.
Terray, L., A. Piacentini, and S. Valcke, 1999: OASIS 2.3, Ocean Atmosphere Sea Ice Soil: User's guide. CERFACS Tech. Rep. TR/CMGC/99/37, Toulouse, France, 82 pp. [Available online at http://www.cerfacs.fr/globc/publication.html.]
Tiedtke, M., 1984: The effect of penetrative cumulus convection on the large-scale flow in a general circulation model. *Beitr. Phys. Atmos.,* **57,** 216–239.
——, 1989: A comprehensive mass flux scheme for cumulus parameterization in large-scale models. *Mon. Wea. Rev.,* **117,** 1779–1800.
Wang, G., R. Kleeman, N. Smith, and F. Tseitkin, 2002: The BMRC coupled general circulation model ENSO forecast system. *Mon. Wea. Rev.,* **130,** 975–991.
Webster, P. J., 1995: The annual cycle of and the predictability of the tropical coupled ocean–atmosphere system. *Meteor. Atmos. Phys.,* **56,** 33–55.
Zhang, G. J., and N. A. McFarlane, 1995: Sensitivity of climate simulations to the parameterization of cumulus convection in the Canadian Climate Centre general circulation model. *Atmos.–Ocean,* **33,** 407–446.