An Assessment of the CFS Real-Time Seasonal Forecasts

WANQIU WANG, MINGYUE CHEN, AND ARUN KUMAR

Climate Prediction Center, National Centers for Environmental Prediction, Camp Spring, Maryland

(Manuscript received 13 August 2009, in final form 12 January 2010)

ABSTRACT

This study assesses the real-time seasonal forecasts for 2005–08 with the current National Centers for Environmental Prediction (NCEP) Climate Forecast System (CFS). The forecasts are compared with retrospective forecasts (or hindcasts) for 1981–2004 to examine the consistency of the forecast system, and with the Atmospheric Model Intercomparison Project (AMIP) simulations forced with observed sea surface temperatures (SSTs) to contrast the realized skill against the potential predictability due to the specification of the observed sea surface temperatures. The analysis focuses on the forecasts of SSTs, 2-m surface air temperature (T2M), and precipitation.

The CFS forecasts maintained a good level of prediction skill for SSTs in the tropical Pacific, the western Indian Ocean, and the northern Atlantic. The SST forecast skill is within the range of hindcast skill levels calculated with 4-yr windows, which can vary greatly associated with the interannual El Niño-Southern Oscillation (ENSO) variability. Overall, the SST forecast skill over the globe is comparable to the average of the hindcast skill. For the tropical eastern Pacific, however, the forecast skill at lead times longer than 2 months is less than the average hindcast skill due to the relatively weaker ENSO variability during the forecast period (2005–08). The forecasts and hindcasts show a similar level of precipitation skill over most of the globe. For T2M, the spatial distribution of skill differs substantially between the forecasts and hindcasts. In particular, the T2M skill of the forecasts for the Northern Hemisphere during its warm seasons is lower than that of the hindcasts.

Comparison with the AMIP simulations shows similar levels of precipitation skill over the tropical Pacific. Over the tropical Indian Ocean, the CFS forecasts show a substantially higher level of skill than the AMIP simulations for a large part of the period. This conforms with the results from previous studies that while interannual variability in the tropical Pacific atmosphere is slaved to the underlying SST anomalies, specification of SSTs (as for the AMIP simulations) in the Indian Ocean may lead to incorrect simulation of the atmospheric variability. Over the tropical Atlantic, the precipitation skill of both the CFS forecasts and AMIP simulations is low, suggesting that SSTs have less control over the atmospheric anomalies and the predictability is low.

The analysis reveals several deficiencies in the current CFS that need to be corrected for improved seasonal forecasting. For example, the CFS tends to consistently forecast larger ENSO amplitude and delayed transition between the ENSO phases. Forecasts of T2M also have a strong cold bias in Northern Hemisphere mid-to high latitudes during warm seasons. This error is due to initial soil moisture anomalies, which appear to be too wet compared with two other observational analyses. The strong impacts of soil moisture on the seasonal forecasts, and large discrepancies among the soil moisture analyses, call for more accurate specification of soil moisture. Furthermore, average forecast SST and T2M anomalies for 2005–08 show a cold bias over the entire globe, indicating that the model is unable to maintain the observed long-term warming trend.
the adoption of single-tier atmosphere–ocean prediction systems, and the removal of surface flux adjustment methodologies (Wang et al. 2002; Anderson et al. 2003; Graham et al. 2005; Gueremy et al. 2005; Saha et al. 2006; Anderson et al. 2007). As a consequence of these modeling advances, dynamical forecast systems have become an integral tool in the operational prediction of tropical El Niño–Southern Oscillation (ENSO) variability and seasonal climate prediction (Anderson et al. 2007; O’Lenic et al. 2008).

The improvements in coupled forecast systems have led to a more satisfactory representation of various observed phenomena, such as the tropical ENSO and principle modes of atmospheric variability including the North Atlantic Oscillation (NAO) and the Pacific–North American (PNA) pattern (Anderson et al. 2003; Saha et al. 2006). Saha et al. (2006) showed that the forecast skill of the National Centers for Environmental Prediction (NCEP) Climate Forecast System (CFS) for tropical sea surface temperatures (SSTs) is competitive with other statistical methods used at the Climate Prediction Center (CPC) and is significantly better than the previous version of the NCEP coupled model. The recent version of the European Centre for Medium-Range Weather Forecasts (ECMWF) forecast models was also found to be better than statistical models at forecasting the onset of ENSO events in boreal spring–summer (Van Oldenborgh et al. 2003). For the extratropical surface air temperature and precipitation, the forecast skill of dynamical prediction systems is comparable with (and complements) the skill of statistical prediction methods (Van Oldenborgh et al. 2003; Saha et al. 2006).

The skill assessments of seasonal forecast systems are generally based on retrospective histories of forecasts (also referred to as hindcasts). Hindcasts are necessary not only for determining systematic errors of the forecast system, which can often be as large as or even larger than the climate signal (Stockdale 1997), but also for assessing the performance of the real-time forecast systems in providing estimates of skill information to the user community. Evaluation of the forecast performance based on hindcasts is also necessary for objective consolidation with other seasonal forecast tools (Van den Dool and Rukhovets 1994; Peng et al. 2002; Barnston et al. 2003; Doblas-Reyes et al. 2005).

A different aspect of the assessment of the models performance is the analysis of the real-time operational forecasts. The NCEP CFS was implemented in 2004 for operational forecasting. Prior to its operational implementation, an extensive set of hindcasts was made for 1981–2004, and the real-time forecasts were started in October 2004. In this paper, we assess the seasonal forecast performance of the real-time CFS forecasts for January 2005–December 2008. The practical and scientific rational and the scope of the present analysis include the following:

(a) *A diagnosis of the consistency of real-time forecast skill against the skill estimated based on the hindcasts.* While diagnoses of hindcasts are helpful for an overall assessment of the forecast systems, because of various reasons, such estimates may not necessarily be consistent with the skill of real-time operational forecasts. Possible reasons include (i) the forecast skill calculated with the hindcasts utilizes information that may be different from that used for the real-time forecasts (e.g., the use of hindcast runs from the future for the computation of climatology); (ii) the configuration of real-time forecasts can be different from that of hindcasts [e.g., forecast ensemble size, which has strong impacts on forecast skill (Kumar and Hoerling 2000; Kumar et al. 2001), is generally larger in real-time forecasts than for the hindcasts (Saha et al. 2006; Anderson et al. 2007)]; (iii) since interannual and decadal variabilities have strong impacts on the seasonal forecast skill, the hindcast skill estimated based on a relatively long period may not be representative of the skill of real-time forecasts for a shorter period (Nakaegawa et al. 2004; Grimm et al. 2006; Tang et al. 2008); and (iv) climate predictability is largely regime dependent (particularly for ENSO) and the period of real-time forecasts may have specific climate features that may also lead to differences in the real-time prediction skill compared to the skill based on hindcasts.

(b) *An assessment of changes in systematic biases that need to be taken into account for making real-time forecasts.* Although the real-time CFS biases and initialization system is the same as for the hindcasts, systematic biases may be introduced by various other, often unavoidable, inconsistencies between hindcasts and the real-time operational forecasts. Factors that may contribute to these differences include (i) delayed availability of observational data that are included in the hindcasts, but could not be included in the real-time forecasts due to time constraints; (ii) changes in the data platforms, such as the inclusion in the NCEP Global Ocean Data Assimilation System (GODAS) of Argo data starting in 2001, which has been suggested to be critical for improved climate monitoring and seasonal prediction by providing an accurate subsurface analysis (Graham et al. 2006; Balmaseda and Anderson 2009); and (iii) use of fixed or variable external forcing such as the greenhouse gas concentration.
(c) A comparison of prediction skill with estimates of potential predictability. Interannual variations in SSTs, particularly in the tropical Pacific related to ENSO, are the dominant source of predictability on a seasonal time scale. The potential predictability of seasonal climate anomalies due to SST is traditionally estimated based on Atmospheric Model Intercomparison Project (AMIP) simulations forced with the observed evolution of SSTs (Rowell 1998; Kumar and Hoerling 1998; Folland et al. 2001; Mathieu et al. 2004; Schubert et al. 2008). For the real-time seasonal forecasts based on the coupled models, the interannual variability of predicted SSTs is still expected to play a dominant role. However, as the SST in a coupled forecast system is predicted, errors in SST prediction may influence the realized prediction skill. It is therefore of interest to compare the realized skill with the estimates of potential predictability based on the AMIP simulations. Further, as the AMIP simulations are far removed from the observed atmospheric initial conditions, and do not include coupled ocean–atmosphere evolution consistent with the air–sea interaction, the comparison of coupled forecasts and AMIP simulations also provides an assessment of (i) the improvement in skill due to the inclusion of atmospheric and land surface initial conditions in the CFS, particularly for soil moisture, and (ii) the possible impacts of coupled versus uncoupled air–sea interactions. Implicit in this discussion is the fact that although the AMIP simulations provide an estimate of the potential predictability due to SSTs, there are other factors in the initialized coupled seasonal predictions that could lead to somewhat higher skill.

(d) A summary of verification skill for a set of key variables of interest to a large user community. In addition to their use in the CPC’s operational climate prediction, the CFS forecasts are widely accessed in real time for various applications such as the ENSO outlook at the International Research Institute for Climate and Society (information online at http://iri.columbia.edu/climate/ENSO/currentinfo/SST_table.html), the hydrologic forecast based on CFS precipitation at the Princeton University (http://hydrology.princeton.edu/∼luo/research/FORECAST/multimodel.php), and the University of Washington (http://www.hydro.washington.edu/forecast/westwide). A regular assessment of the real-time CFS verification skill provides further information to the user community.

This paper is organized as follows. Section 2 describes the forecast and AMIP simulation data. Section 3 presents an analysis of the CFS forecast skill and consistency in skill between the hindcasts and forecasts. Section 4 compares the CFS forecasts with the AMIP simulations. Section 5 analyzes systematic errors in the real-time forecasts. Section 6 provides a summary of the analysis.

2. The forecast and verification data

Our analysis is based on forecasts for target seasons in 2005–08. Data used include the real-time forecasts from the NCEP operational CFS, hindcasts for 1981–2004, AMIP simulations for 2005–08 with the atmospheric component of the CFS, and observations. Anomalies for forecasts–hindcasts, AMIP simulations, and observations are defined as the departure from the respective seasonal climatologies taken as a 1981–2004 average. The seasonal climatology for the CFS forecasts and hindcasts is calculated from the 1981–2004 hindcasts and as a function of lead time.

The atmospheric component of the CFS is the 2003 version of the NCEP atmospheric Global Forecast System (GFS) model at T62 horizontal resolution with 64 vertical layers. The oceanic component of the CFS is version 3 of the Geophysical Fluid Dynamics Laboratory’s Modular Ocean Model (MOM3; Pacanowski and Griffies 1998) with a zonal resolution of $1^\circ$ and a meridional resolution of $1/8^\circ$ between $10^\circ$S and $10^\circ$N, gradually increasing through the tropics until becoming fixed at $1^\circ$ poleward of $30^\circ$S and $30^\circ$N. The atmosphere and ocean are coupled once per day. Sea ice is prescribed as climatology. Greenhouse gas concentrations are fixed at the 1988 level. Initial conditions for both hindcasts and forecasts with the CFS are taken from the NCEP–Department of Energy (DOE) Reanalysis-2 (R2; Kanamitsu et al. 2002) for the atmosphere and land, and from the NCEP Global Ocean Data Assimilation System (GODAS) for the ocean. A more detailed description of the model and analyses of its performance based on the hindcasts can be found in Saha et al. (2006).

For the real-time prediction, the CFS produced one forecast run each day from September 2004 to January 2005, two forecast runs from February 2005 to December 2007, and four forecast runs since January 2008. Each forecast run covers the partial month after the initial date and the nine subsequent full target months. This study uses forecast ensembles of 20 forecasts from the last 20 days of the initial months from September 2004 to January 2005, 40 forecasts from the last 20 days of the initial months from January 2005 to December 2007, and 40 forecasts from the last 10 days of the initial months starting January 2008. Forecast lead time is defined as the time difference (in months) between the initial month and the beginning of the target season. For example, the 0-month-lead forecast for January–March (JFM) 2006 and the 2-month-lead forecast for March–May (MAM) 2006 are both taken from an aggregate of the December 2005 initial conditions.
Hindcasts produced with the CFS for 1981–2004 are used for analyzing the consistency between the forecasts and hindcasts. The hindcasts from each initial month consist of an ensemble of 15 runs initialized across the last 20 days or so (Saha et al. 2006).

The AMIP simulations with the atmospheric component of the CFS consist of 18 simulations forced with the observed SSTs (Reynolds et al. 2002). Each AMIP run started with different atmospheric initial conditions without reinitialization during the integration. The ensemble mean of the AMIP simulations represents the atmospheric response to the observed SST forcing, and provides an estimate of the predictability of the seasonal means due to near-perfect knowledge of SSTs. Predictability estimation, however, does not include the possible effects of initializing atmospheric and land conditions and, further, could be adversely influenced by the errors in the air–sea interaction.

The CFS forecasts, hindcasts, and AMIP simulations are compared with observations to diagnose the model’s performance. Observational data used in this study includes the SST analysis of Reynolds et al. (2002), the surface 2-m air temperature (T2M) from the Climate Anomaly Monitoring Systems (CAMS) at the Climate Prediction Center (CPC) (Ropelewski et al. 1985), precipitation from Janowiak and Xie (1999), the 200-mb height (Z200) from R2 (Kanamitsu et al. 2002), and soil moisture from R2, from the NCEP North American Regional Reanalysis (Mesinger et al. 2006), and from a CPC analysis with a one-layer leaky-bucket model (Fan and van den Dool 2004).

3. Analysis of the CFS forecast skill

In this section, we assess the performance of the CFS forecasts and their consistency with the hindcasts. We first look at the forecasts of SSTs and then diagnose the skill levels for other fields. The forecast skill is calculated as an anomaly correlation based on the ensemble mean of individual forecast runs. The level of significance of the anomaly correlation is estimated based on the Monte Carlo approach whereby correlations after randomizing the forecasts are first computed. This procedure is repeated 10,000 times, and significance is estimated based on the fraction of times the actual correlation exceeds the correlations achieved with the randomized set. The significance level of the correlation differences is calculated in a similar way.

a. Forecasts of SSTs

1) TROPICAL SST INDICES

Figure 1 shows the 0-month-, 3-month-, and 6-month-lead forecasts together with the observed patterns of evolution for three tropical SST indices: the Niño-3.4 ($5°S–5°N, 190°–240°E$) index, the Indian Ocean dipole mode index (DMI), and the average of the SST anomalies over the major hurricane development region (MDR; $10°–20°N, 280°–340°E$). The Niño-3.4 index represents tropical Pacific El Niño–Southern Oscillation (ENSO) variability (Barnston et al. 1997). The DMI index is defined as the SST difference between the western tropical Indian Ocean ($10°S–10°N, 50°–70°E$) and the eastern tropical Indian Ocean ($10°S–0°, 90°–110°E$), and has been shown to impact the climate over Australia and Asia (Saji and Yamagata 2003). Warm-season MDR SST anomalies influence the interannual variability of tropical hurricane activity (Goldenberg et al. 2001; Saunders and Lea 2008).

While the CFS captures the overall interannual Niño-3.4 SST variations between the warm and cold phases, there are substantial errors in the forecasting of the ENSO phase and amplitude (Fig. 1a). For the 0-month-lead forecast, the CFS tended to persist and amplify the observed initial anomalies. At longer leads, the CFS consistently predicted a delayed transition between the warm and cold phases. For example, 3- and 6-month lead-time CFS forecasts for the ENSO transition were delayed by 3–6 months. The phase errors in the forecasts of the tropical SST Niño indices remain a major impediment for skillful seasonal forecasts of atmospheric variability (e.g., hurricane seasonal outlooks) with longer lead times.

For the Indian Ocean DMI, the CFS captured the observed positive phase during the boreal summer to fall period in 2006 and 2007 (Fig. 1b), which was also well predicted with another coupled model (Luo et al. 2008). The CFS forecasts at lead times of 3 and 6 months for 2005 and 2008 were not successful. The CFS produced a positive (negative) DMI for the 2005 (2008) boreal summer when the observed DMI was negative (positive). The failures of the DMI forecasts in 2005 and 2008 were due to errors in the forecast SSTs in the eastern Indian Ocean (not shown). These results suggest large year-to-year variations in the performance of the CFS in forecasting the Indian Ocean DMI.

In the Atlantic, the observation shows a positive MDR index during most of 2005–08 (Fig. 1c). The CFS reproduced the warmth during this period but with a weaker amplitude, especially for longer lead times. A recent study by Cai et al. (2009) suggests that colder SSTs in the CFS forecast are possibly due to the use of greenhouse gas concentrations that were fixed at the 1988 level, as the corresponding incorrect radiative forcing is not sufficient to maintain the observed warming SST trends.
We will now discuss the temporal correlation of the observed and the CFS-predicted SSTs. As expected, the temporal correlation of the seasonal mean SST decreases with increasing lead time (Fig. 2). At 0-month lead time, high correlation values are found in the tropical Pacific and northern tropical Atlantic. At lead times of 3 and 6 months, relatively higher correlation skill is located only in the tropical central Pacific, the northern tropical Atlantic, and the western Indian Ocean. The high skill in the northern tropical Atlantic is largely due to the correct forecast of the sign of the anomalies but with weaker amplitude as indicated by the MDR index in Fig. 1c.

The corresponding temporal correlation for the hindcasts is shown in Fig. 3. While the forecasts show higher skill values in some local areas (Fig. 2), the hindcasts show larger spatial coverage of the significant skills. For example, the forecasts have higher skill at all lead times over the northwestern tropical Atlantic (around 20°N, 60°W), parts of the North Pacific, and to the northeast of Madagascar, but they show almost no skill at 3- and 6-month lead times in the North Atlantic around 30°N and at 6-month lead time in the tropical central to eastern Indian Ocean, regions where hindcasts have a significant level of skill. In particular, the forecast
skill in the tropical eastern Pacific is lower than the hindcast skill.

Average correlation skills over the globe and the Niño-3.4 region are compared in Fig. 4. For comparison, hindcast skill for a moving 4-yr window is also included in Fig. 4. We can see that the correlations vary greatly among individual 4-yr segments and the forecast skill for both the globe and the Niño-3.4 region is within the range of the skill of the hindcasts. For the global average, the 4-yr (2005–08) forecast skill (Fig. 4a, red curve) is close to the 20-yr (1981–2004) average hindcast skill for all lead times (Fig. 4a, blue curve). For the Niño-3.4

---

**Fig. 2.** Temporal correlation of seasonal SST between the observations and the forecast at (a) 0-, (b) 3-, and (c) 6-month leads for JFM 2005–OND 2008. The correlation is shaded at an interval of 0.1 with areas with values <0.2 in white. Areas where forecast skill is higher (lower) than the maximum (minimum) of the correlations calculated with 4-yr windows of the hindcasts are shown with dark (blue) hatching. Values that do not pass the significance level of 99% are omitted.
average, the forecast skill (Fig. 4b, red curve) is comparable to the 1981–2004 average hindcast skill for lead times of 0–2 months (Fig. 4b, blue curve). Beyond 2 months, the forecast skill of the Niño-3.4 SSTs is substantially smaller than the 1981–2004 average hindcast skill (Fig. 4b).

There are multiple factors that contribute to the skill differences between the forecasts and hindcasts. First, the forecasts use larger ensemble size (20–40 members) than the hindcasts (15 members). Second, while the forecast runs for 2005–07 were from the last 20 days of the initial months, the forecasts starting January 2008 were from the last 10 days and, thus, have shorter effective lead times compared to the hindcasts, which were initialized from the last 20 days or so of each month (Saha et al. 2006). Third, oceanic initial conditions for the forecasts may also have improved due to the assimilation of additional observational platforms such as

![Fig. 3. As in Fig. 2, but for hindcasts for 1981–2004.](image-url)
the Argo data, which started to be assimilated into GODAS in 2001. All these factors would have led to an improved level of forecast skill compared to the hindcasts. However, the result that the skill over the ENSO region is substantially lower than the hindcast skill (Fig. 4) suggests an influence from some other factor, for example, regime dependence in the expected level of prediction skill.

To pursue this hypothesis further, the correlation is calculated for 4-yr sliding windows for the entire period of hindcasts and forecasts (1981–2008) and is shown in Fig. 5. It is seen that there exists substantial interannual variability in both the global average of skill (Fig. 5a) and Niño-3.4 skill (Fig. 5b). The global mean of the skill generally follows the variation of the Niño-3.4 correlation skill, with relatively lower skill in 1990–93 and after 2000, suggesting that the long-term variation in the global SST forecast skill is dominated by the skill variation in the ENSO prediction. A comparison with the evolution of the observed Niño-3.4 SST amplitude (Fig. 5c) suggests that the high (low) skill of ENSO prediction is associated with the strong (weak) ENSO regimes, and the lower forecast skill for the eastern Pacific during the real-time forecast period (2005–08) at longer lead times is likely due to the relatively weak ENSO variability during this period (Van den Dool and Toth 1991; Kumar 2009); that is, low signal-to-noise regimes lead to lower expected values of skill.

The above results suggest that the forecast skill variation is largely dominated by the ENSO variability. However, the impacts of other factors, for example, differences in ensemble size, the natural decadal variability, and changes in the input data in the oceanic initialization, may have also contributed to the differences in the skill. Calculations with different ensemble sizes for forecasts show that the use of a 40-member ensemble in the forecasts enhances the skill by approximately 0.02 over a large part of the globe compared to the skill calculated based on a 15-member ensemble (similar to that for the hindcasts) (not shown).

Implications about the influence of decadal variability may be inferred from the differences between the hindcast and forecast skill. Figure 2 shows that the maximum forecast skill in the eastern tropical Pacific at lead times of 3 and 6 months is located off the equator, compared to the hindcast skill, which shows maximum skill near the equator (Fig. 3). Further, areas of forecast skill higher than the maximum hindcast skill calculated with 4-yr windows (dark hatching in Fig. 2) are located in the northern Pacific, and appear to be collocated with

---

![Figure 4](image_url)

**Fig. 4.** Averaged SST temporal correlation for (a) the entire globe and (b) the Niño-3.4 region. Red curves are for 2005–08 forecasts, blue curves for 1981–2004 hindcasts, and gray curves are for 4-yr windows of the hindcasts.
the areas of large amplitude in the spatial pattern of the Pacific decadal oscillation (PDO). It will be interesting to further analyze if the CFS has better prediction skill for the PDO during the period of real-time forecasts compared to the period of the hindcasts for 1981–2004.

Changes in the input data in the oceanic initialization such as the assimilation of the Argo data in the oceanic initial conditions may also have influenced forecast skill. However, the impacts of such changes cannot be assessed with the comparison between forecasts and

Fig. 5. (a) Variations in the global average of the SST temporal correlation between the CFS forecast and observations. (b) Variations in the temporal correlation of Niño-3.4 SSTs between the CFS forecast and observations. (c) Variations in the standard deviation of the observed Niño-3.4 SSTs. Both the correlation and standard deviation are calculated based on the seasonal mean from 4-yr sliding windows. The date on the x axis is the beginning time of the 4-yr windows.
hindcasts. Offline forecast experiments using oceanic initial conditions without assimilating the Argo data would be needed to diagnose this impact (Balmaseda and Anderson 2009).

b. Forecasts of atmospheric fields

The fact that the CFS 0-month-lead forecasts have a good level of skill for SST prediction in the tropical eastern Pacific (Figs. 2a and 3a), a region with well-documented impacts on the global climate (Ropelewski and Halpert 1987; Halpert and Ropelewski 1992), provides a sound basis for using the CFS for predicting seasonal atmospheric climate variability. It is also expected that the forecasting of atmospheric variables with 0-month lead time, because of higher skill in SST predictions and because of the initialization of atmospheric and land conditions, would be more skillful than forecasts for longer leads. For this reason, in the rest of this study we focus on the CFS forecast with 0-month lead time only.

Temporal correlations of land surface T2M, precipitation, and Z200 for the 0-month-lead forecasts are shown in Fig. 6. In general, the T2M correlation skill is quite low. Regions of appreciable skill are confined only to a few areas including central and eastern Australia, southeastern Africa, central South America, northern Mexico and southwestern United States, western and eastern Canada, and Eurasia around 40°N (Fig. 6a). The negative seasonal
forecast skill over Russia and central North America is surprising and is not seen in the hindcasts (Fig. 7a). Possible causes for this negative prediction skill are investigated later and are related to the abnormally wet initial soil moisture conditions resulting in an unrealistic forecast of cold anomalies during warm seasons.

The precipitation skill is highest over the tropical Pacific and relatively high over the tropical Indian Ocean, northern tropical Atlantic, eastern Australia, southern Africa, and southwestern Asia (Fig. 6b). Precipitation correlations exceeding 0.4 are seen over central North America where the T2M skill is low. For Z200, the largest correlation is confined to the tropics in response to the interannual variability of tropical SST anomalies (Fig. 6c). Low Z200 correlation is seen over central North America and western-central Russia. The spatial structure of skill for rainfall and Z200 conforms with the expected influence of SST variability related to ENSO documented in earlier studies, namely, high skill for rainfall in the tropical eastern Pacific and high skill for the interannual variability Z200 in tropical latitudes (Peng et al. 2002).

Temporal correlations of T2M, precipitation, and Z200 for 1981–2004 hindcasts are shown in Fig. 7. There are notable differences in the T2m skill between the forecasts and hindcasts over various local areas with higher forecast skill over eastern Australia and central South America, and lower forecast skill in northern South America (Figs. 6a and 7a). In particular, the skill of the T2m forecasts over Russia and central North America is lower than that for the hindcasts (Figs. 6a and 7a). The precipitation skill distribution of the hindcasts is similar to that of the forecasts (Figs. 6b and 7b), with differences over some local areas. For example, the forecast skill over the

![Fig. 7. As in Fig. 6, but for 1981–2004 hindcasts.](image-url)
United States is mostly confined to the northwest while the hindcasts show significant skill over large parts of the western, central, and southern areas. The hindcast Z200 skill for 1981–2004 is comparable to the forecast skill in the tropics (Figs. 6c and 7c). The most distinct difference in Z200 skill between the forecasts and hindcasts is that no real-time forecast skill is found over the contiguous United States where there exists small yet significant skill in the hindcasts.

These results indicate that although the overall distributions are similar between the forecasts and hindcasts, there are local differences. In particular, and as will be further discussed later, the low T2M skill over Russia and central North America suggests a possible deficiency in the real-time forecasts. An implication of this analysis is that the estimates of the hindcast skill, at times, may not be representative of the expected skill of the real-time forecast.

4. Comparing real-time prediction skill and potential predictability due to SSTs

In this section we present an analysis comparing the prediction skill based on the 0-month-lead CFS real-time forecasts and the skill based on the ensemble mean of the AMIP simulation runs forced by observed SSTs. The model for the AMIP simulations is the same as the atmospheric component of the CFS. This comparison helps analyze differences in the realized skill and the estimate of the potential predictability of seasonal atmospheric anomalies due to SSTs. This analysis also provides an assessment about the realism of the estimates of seasonal predictability based on the AMIP simulations. This is an important consideration since AMIP simulations are a useful tool for understanding different facets of atmospheric variability, and only a comparison with coupled simulations can provide an assessment of their fidelity. As the SST forecast skill at 0 months is generally high over the globe (Fig. 2a), especially for the tropics, a comparison with the AMIP simulations also allows a diagnosis of the role of air–sea coupling and the impacts of atmospheric and land initial conditions.

There are three fundamental differences between the CFS forecasts and the AMIP simulations. First, the coupled ocean–atmospheric evolution consistent with the air–sea heat flux is included in the CFS but not in the AMIP simulations that are forced by the observed evolution of SSTs. Second, the CFS is initialized with the observed information for the atmosphere and the land surface, while the integration of the AMIP simulations is far removed from the initial conditions. Third, the oceanic surface condition (i.e., SST) during the CFS forecast is predicted, and even though the skill of the SST predictions for 0-month lead is high, nonetheless the SST is less accurate than the observed SSTs specified in the AMIP simulations. The inclusion of coupled evolution and the observed information in the atmospheric and land surface initial conditions in the CFS is expected to result in better performance than that of the AMIP simulations while the errors in the forecast SSTs could lead to worse performance. Therefore, the regions where the CFS prediction skill exceeds the skill based on the AMIP may indicate that the influence of the coupling and the initial conditions is important.

a. Spatial distributions of temporal correlation skill

Spatial distributions of temporal correlation skill for the AMIP simulations are presented in Fig. 8. The AMIP simulations have a higher level of Z200 skill across the tropical Atlantic, Africa, and Indian Ocean, and a lower level of skill in Northern Hemisphere high latitudes than the forecasts (Figs. 6c and 8c). The precipitation skill distribution in the tropical Pacific is similar between forecasts and the AMIP simulations (Figs. 6b and 8b). However, the AMIP precipitation skill over the Indian Ocean, and over most of the global land, is lower than the forecast skill. For T2M, the skill distribution of the AMIP simulations is very similar to that of the forecasts, except for Russia and central North America, where the forecast skill is negative while the AMIP simulation skill is either near zero (over Russia) or positive (in central North America) (Figs. 6a and 8a).

b. Temporal evolution of pattern correlation skill

The time evolution of the pattern correlation skill is examined in this subsection. We first analyze the skill for tropical precipitation. We then look at the skill of T2M and precipitation over the extratropics.

For reference, the spatial correlation of the predicted SSTs for the tropics (20°S–20°N) is shown in Fig. 9 (black curves). The average SST correlation is 0.78 for the Pacific and 0.54 for the Indian Ocean. For the Atlantic, the correlation is about 0.8 before FMA 2007 and thereafter drops to lower values. This drop may be related to the reduced amplitude of the observed SST anomalies (e.g., observed MDR SSTs shown in Fig. 1, bottom), leading to a low signal regime for which the predictive skill is also expected to be low (Van den Dool and Toth 1991; Kumar 2009).

Consistent with the spatial structure of the temporal correlation (Fig. 6b), the forecast precipitation spatial correlation skill (Fig. 9, red curves) is highest for the Pacific compared to those for the Indian and Atlantic Oceans. There is a clear seasonality in the skill over the tropical Pacific with smaller values during the boreal summer period (Fig. 9a). This is due to the northward shift of convective activities during boreal summer and also due to the
fact that the ENSO-related SST anomalies peak in boreal winter (see Fig. 1). Such seasonality is less clear in the Indian and Atlantic Oceans (Figs. 9b and 9c). The overall level of prediction skill is generally poor over the Atlantic even though prior to FMA 2007 the level of prediction skill for SSTs themselves is fairly high.

The temporal evolution of the spatial correlation for the AMIP tropical precipitation over oceanic areas is also shown in Fig. 9 (blue curves). For the tropical Pacific, correlation values for the AMIP simulations are slightly smaller than those for the forecasts (Fig. 9a). This suggests that precipitation anomalies in the tropical Pacific are largely a result of interannual SST variability, and the real-time forecast skill is close to the potential predictability achieved due to specification of the observed (perfect) SSTs with small additional skill from other sources (e.g., the atmosphere initialization and air–sea coupling). Forcing of the atmosphere by the ocean in the tropical Pacific is also confirmed by the success of hybrid coupled models (where atmospheric anomalies are parameterized in terms of anomalous SSTs) in predicting ENSO-related SST variability (Syu and Neelin 2000; Tang et al. 2008).

Over the Indian Ocean, the AMIP precipitation correlation is clearly smaller than that for the CFS forecasts for a large part of 2005–08, especially for the boreal spring and summer in 2006 and 2007 (Fig. 9b). This is the case even though the CFS prediction skill for SSTs is lower than that for the Pacific Ocean; that is, the accuracy of CFS-predicted SSTs is much worse than the near-perfect SSTs specified in the AMIP simulations. Although possible influences of the observed initial conditions in the CFS cannot be discounted, the differences between the CFS forecasts and the AMIP simulations for

![Figure 8](image-url)
the tropical Indian Ocean are consistent with previous studies that have shown that the atmospheric variability in the Indian Ocean is related to the atmosphere–ocean coupled response to remote forcings from the Pacific through the atmospheric bridge (Wang et al. 2005; Wu and Kirtman 2005; Krishna Kumar et al. 2005). The superiority of real-time prediction skill compared to the predictability estimated based on the AMIP simulations suggests the possibility that over the Indian Ocean the AMIP setup for estimating the potential predictability may not be a suitable approach. However, additional experiments with the atmospheric component of the CFS initialized from observations and forced with observed SSTs are needed to determine the roles of initial atmosphere–land initial conditions.

Over the Atlantic Ocean, the correlation for both the CFS forecasts and AMIP simulations is low (Fig. 9c), suggesting that the atmospheric variability over this region may not be forced by the local SSTs (as measured by the AMIP simulations), and is less predictable compared to the Pacific and Indian Ocean. Thus, in contrast to the precipitation variability in the Indian Ocean, inclusion of coupled ocean–atmosphere evolution also does not lead to improved predictions. It remains to be seen if further improvements in the coupled SST predictions, and ocean data assimilation, would result in improved skill above that estimated from the AMIP simulations.

Figure 10 compares the spatial correlation for Northern Hemisphere land surface T2M and precipitation. For precipitation, although the skill is low, the CFS forecasts are consistently better than the AMIP simulations for almost the entire period (Fig. 10b). For T2M (Fig. 10a), the forecast skill is substantially lower than the AMIP simulation skill for a large portion of the Northern Hemisphere warm seasons (JAS and ASO 2005, AMJ and MJJ 2006, JAS and ASO 2007, and JAS to SON 2008). The forecast skill is also lower than the AMIP skill for OND 2008. Since the predictable component of the

---

**Fig. 9.** Spatial anomaly correlation for the tropics (20°S–20°N) between the observations and the model output for the CFS 0-month-lead forecast SST (black curves), CFS 0-month-lead forecast precipitation (red curves), and precipitation from the AMIP simulations (blue curves) for the (a) Pacific, (b) Indian, and (c) Atlantic Oceans. For SST, the dots indicate that the correlation values are at the 99% significance level. For precipitation, the dots indicate that the values of the correlation difference between the CFS forecasts and AMIP simulations are significant at the 99% level.
extratropical variability is primarily from the tropical Pacific where the CFS and AMIP skill for precipitation is comparable, the improvement in the CFS forecasts of precipitation and T2M (aside from summer and fall) over Northern Hemisphere land is likely due to the observed information in the initial conditions and, therefore, adds to the potential predictability estimated from the specification of the SSTs alone. We note that both the T2M and precipitation correlation values during NDJ 2007/08 to FMA 2008 are relatively high due to the strong forcing associated with the La Niña conditions.

5. Systematic errors in the CFS seasonal forecasts

In this section, we analyze systematic errors in the CFS real-time forecasts. We will focus on two major errors: 1) a warm-season cold T2M bias in the Northern Hemisphere and 2) a mean cold bias during the entire period of 2005–08.

a. The cold summers in the forecasts: Impacts of initial soil moisture

Generally, one would expect the use of the observed land surface initial conditions to enhance the seasonal forecast skill. In particular, inclusion of observed initial soil moisture is considered to lead to better forecasts of T2M (Huang et al. 1996; Liu 2003). It is therefore surprising that the CFS T2M correlation for the summer and fall seasons is even lower than that for the AMIP simulations that do not include information about the observed land surface conditions. One possibility for the inferior CFS forecasts of T2M is that the initial soil moisture is erroneous. To investigate this further, the average JJA T2M anomalies for 2005–08 from CAMS observation, CFS forecasts, and AMIP simulations for Northern Hemisphere are shown in Fig. 11 together with the R2 soil moisture in May. The observations have above normal JJA T2M anomalies over most of the land areas (Fig. 11a), consistent with the recent warming trends. The AMIP simulations reproduced the warmth over large parts of the Northern Hemisphere with weaker amplitude over Eurasia and comparable amplitude for the central and eastern United States (Fig. 11c). The CFS forecasts failed to capture the overall warmth. In particular, the CFS produced large cold anomalies over central North America, eastern Europe, and Russia (Fig. 11b). Such a forecast for cold anomalies during the warm season occurred in all individual years of 2005–08 (not shown). The cold JJA T2M anomalies are consistent with the large initial positive local moisture anomalies in May (Fig. 11d), suggesting that the erroneous T2M anomalies at Northern Hemisphere mid- to high latitudes during the warm seasons of 2005–08 were related to the initial soil moisture anomalies.

Soil moisture in May averaged over North America between 40° and 60°N for 1981–2008 from R2, which has
been used to initialize the CFS for both hindcasts and forecasts, is compared with that from the Climate Prediction Center analysis with a one-layer leaky-bucket model (LB; Fan and van den Dool 2004) and the NCEP North American Regional Reanalysis (RR, Mesinger et al. 2006) to examine the uncertainties among different analyses (Fig. 12). There are large differences in the interannual variations among the three analyses. For example, the change in soil moisture in 2000 from the previous year is positive in R2, negative in RR, and almost zero in LB. In addition, a unique feature in R2 is the overall upward trend since 1988, resulting in consistently above normal anomalies after 1995 with relatively large positive values for 2005–08, especially compared to the LB analysis.

These results suggest the possibility that initial soil moisture anomalies in the CFS forecasts for the warm seasons are too wet, which lead to too cold T2M anomalies in the forecasts. The strong impacts of initial soil moisture (in the CFS) and the large uncertainties in soil moisture analyses call for a more accurate specification of initial soil moisture for improved seasonal forecasts.

b. Systematic mean bias in the forecast

In this section, we diagnose the mean bias in the real-time forecasts for 2005–08. Shown in Fig. 13 are 2005–08 averages of the zonal-mean monthly anomalies of SST, T2M, and Z200 from observations and CFS 2-month-lead forecasts. The anomalies are calculated with respect to
the average of the hindcast period (1981–2004). Positive SST anomalies are seen in the observations at most latitudes except for near the equator, where the anomalies are close to zero, and to the south of 40°S, where the anomalies are negative. The CFS forecasts show a cold bias at most latitudes with an error of $-0.1$ K at almost all latitudes and $-0.4$ K around 59°N. The differences in T2M between the CFS forecasts and the observations are similar to that in SST but with a larger bias amplitude of $-0.3$ K for most latitudes. In particular, the CFS cold bias at high latitudes around 70°N is more than $-1$ K. For Z200, the CFS failed to produce the observed positive anomalies with a negative bias of 5–10 m to the south of 40°N and a negative bias larger than 10 m to the north of 40°N, indicating a mean tropospheric cold temperature bias.

A consistent bias in the CFS prediction may be related to various causes. Cai et al. (2009) suggest that the use of fixed greenhouse gas concentrations is responsible for the weaker warming trend in the CFS. In addition, as shown in the previous subsection, the initial wet soil moisture anomalies also contributed to the cold bias during the Northern Hemisphere warm seasons. The lack of interannual variability in sea ice (together with lowering sea ice trends in recent years) is another possible reason for the cold bias at high latitudes in the CFS operational real-time forecasts.

6. Summary

The NCEP dynamical seasonal climate forecasts for 2005–08 are analyzed to assess the real-time performance of the CFS and to diagnose the factors that may impact its real-time performance. Real-time forecasts are compared with the 1981–2004 retrospective forecasts (or hindcasts) to examine the consistency of the forecast system. Simulations of the AMIP type are also used to compare the realized skill against the potential predictability due to SSTS and to examine the role of air–sea interaction and initial conditions in the CFS (which might lead to prediction skill that is better than that estimated from the AMIP simulations forced by the SSTS). The analysis focuses on the forecasts of SST anomalies that represent a forcing of the atmospheric variability, extratropical surface 2-m temperature (T2M), and precipitation.

For tropical SSTS, the CFS performs well with a correlation skill higher than 0.6 in the Pacific, northern Atlantic, and western Indian Oceans. However, there are substantial errors in the forecasts as revealed by the SST indices. For the Niño-3.4 index, the model tended to amplify the initial SST anomalies and consistently forecasted delayed transitions between warm and cold ENSO phases. The CFS model’s performance in forecasting the Indian Ocean dipole mode index (DMI) varied from year to year with erroneous forecasts for 2005 and 2008 but more reasonable forecasts for 2006 and 2007. For the main Atlantic hurricane development region (MDR), the
model captured the observed warm anomalies over most of the period but with weaker amplitude.

The SST forecast skill is within the range of hindcast skills calculated with 4-yr windows, which can vary greatly because of ENSO variability. The global average of the SST forecast skill is comparable to the average skill of the hindcasts. However, for the tropical eastern Pacific, where the El Niño–Southern Oscillation (ENSO) SST variability is maximum, the forecast SST skill is lower than that of the average hindcast skill for lead times longer than 2 months. This lower forecast skill over the tropical eastern Pacific at longer lead times is consistent with the weak ENSO variability during the last few years.

Diagnoses of the forecasts for other fields (T2M, precipitation, and Z200) focus on the 0-month lead time. The highest level of Z200 skill for both the forecasts and hindcasts is confined to the tropics, indicating the dominance of ENSO-related variability. For precipitation, the forecast skill distribution is similar to that for the hindcasts. For T2M, the skill differences between the forecasts and hindcasts are found in various local areas with higher forecast skill over eastern Australia and central South America, and lower forecast skill in northern South America. In particular, the skill of the T2M forecasts over Russia and central North America is lower than that of the hindcasts. The skill differences between the hindcasts and forecasts suggest that the hindcast skill, at times, may not be representative of the skill of the real-time forecasts.

The CFS 0-month-lead forecasts are further compared with the AMIP simulations to examine the impacts of air–sea coupling and atmospheric and land initial conditions. The precipitation skills of the CFS forecasts and AMIP simulations for the tropical Pacific are similar and both show a distinct seasonality with lower correlation values for the boreal summer period. Over the tropical Indian Ocean, the CFS forecasts have a substantially higher level of skill than the AMIP simulations for a large part of the analysis period. This is consistent with the results from previous studies that the tropical Pacific atmosphere responds to the underlying SST anomalies, while specification of the SSTs over the Indian Ocean could lead to erroneous atmospheric variability. Over the tropical Atlantic, the precipitation skill of both the CFS forecast and the AMIP simulation is low, suggesting that SST is not the dominant forcing for the atmospheric anomalies and that the predictability is low. Further analyses based on observations and other coupled forecast systems will be helpful to our understanding of whether or not the low prediction skill of the atmospheric anomalies over the tropical Atlantic in the CFS is due to the lack of an accurate representation of certain important coupled processes.

In the Northern Hemisphere, the CFS forecast skill for precipitation is consistently better than that of the AMIP simulations. For T2M, the CFS forecasts also show better skill than the AMIP simulations, except during the boreal summer and fall seasons, during which the forecast skill is significantly lower than the AMIP skill. The improvement in the CFS forecasts of precipitation and T2M (excluding the boreal warm seasons) compared to that in the AMIP simulations indicates that the observed information in the atmospheric and initial land conditions adds positively to the potential predictability estimated from the AMIP simulations.

The lower skill in the CFS T2M forecasts for northern summer and fall seasons, compared to the hindcasts and the AMIP simulations, is related to the consistent forecasts of erroneous cold anomalies over eastern Europe, Russia, and central North America. These cold T2M anomalies are found to be related to the initial soil moisture anomalies from the NCEP Reanalysis-2 (R2), which produced the largest positive soil moisture anomalies over central North America during 2005–08. In addition, a comparison with two other soil moisture analyses shows large differences among the observational soil moisture estimates, with R2 soil moisture anomalies during 2005–08 being the wettest among the three analyses. The strong impacts of the soil moisture on the seasonal forecasts, and the large discrepancies among the soil moisture analyses, call for more accurate specification of the soil moisture for improved seasonal forecasts.

There is also a systematic cold bias in the CFS during the real-time forecast period. When averaged for the entire period of 2005–08, the CFS 2-month-lead forecast SSTs are about 0.1 K too cold at most latitudes and about 0.4 K cooler than the observations at 59°N. For T2M, the cold bias is about −0.3 K for most latitudes and is as large as −1 K around 70°N. The forecast for 200-mb height (Z200) is also consistently lower than the observed result, indicating a cooler troposphere. There may be multiple reasons for this cold bias in the CFS forecasts, including the use of fixed greenhouse gas concentrations, lack of sea ice changes in the model, and a too wet initial soil moisture.

While the CFS has been shown to be superior to the previous NCEP coupled model in forecasting the ENSO variability and comparable to statistical tools in forecasting the land surface precipitation and T2M (Saha et al. 2006), the analysis based on the real-time forecasts also points to deficiencies in the current coupled CFS that need to be corrected for improved forecasts. In particular, the too strong ENSO amplitude during the beginning of the forecasts and delayed transition of the ENSO phases in the forecasts could induce an erroneous atmospheric response and lead to unsatisfactory forecasts of atmospheric variability. The strong impacts of the initial soil moisture and its uncertainty in the analyses also necessitate reliable observational estimates of
the soil moisture. Furthermore, the cold bias during the forecast period indicates that the model is unable to capture the observed long-term warming trend and its correction is highly desirable.

Acknowledgments. The AMIP simulations used in this study were performed by Dr. Bhaskar Jha. We also wish to thank Dr. Yun Fan for making available the CPC leaky-bucket model soil moisture, and Dr. Wanru Wu for providing the NCEP North American Regional Reanalysis soil moisture. The authors greatly appreciate the valuable comments of three anonymous reviewers. Their comments have led to a significant improvement of this paper.

REFERENCES


