Clim Dyn (2013) 41:1615–1633
DOI 10.1007/s00382-013-1845-2

Predictions of Nino3.4 SST in CFSv1 and CFSv2: a diagnostic comparison

Anthony G. Barnston · Michael K. Tippett

Received: 18 December 2012 / Accepted: 12 June 2013 / Published online: 25 July 2013

Abstract Diagnostic evaluations of the relative performances of CFSv1 and CFSv2 in prediction of monthly anomalies of the ENSO-related Nino3.4 SST index are conducted using the common hindcast period of 1982–2009 for lead times of up to 9 months. CFSv2 outperforms CFSv1 in temporal correlation skill for predictions at moderate to long lead times that traverse the northern spring ENSO predictability barrier (e.g., a forecast for July made in February). However, for predictions during less challenging times of the year (e.g., a forecast for January made in August), CFSv1 has higher correlations than CFSv2. This seeming retrogression is caused by a cold bias in CFSv2 predictions for Nino3.4 SST during 1982–1998, and a warm bias during 1999–2009. Work by others has related this time-conditional bias to changes in the observing system in late 1998 that affected the ocean reanalysis serving as initial conditions for CFSv2. A posteriori correction of these differing biases, and of a similar (but lesser) situation affecting CFSv1, allows for a more realistic evaluation of the relative performances of the two CFS versions. After the dual bias corrections, CFSv2 has slightly better correlation skill than CFSv1 for most months and lead times, with approximately equal skills for forecasts not traversing the ENSO predictability barrier and better skills for most (particularly long-lead) predictions traversing the barrier. The overall difference in correlation skill is not statistically field significant. However, CFSv2 has statistically significantly improved amplitude bias, and visibly better probabilistic reliability, and lacks target month slippage as compared with CFSv1. Together, all of the above improvements result in a highly significantly reduced overall RMSE—the metric most indicative of final accuracy.

Keywords Coupled ocean–atmosphere models · NOAA CFSv1 and CFSv2 · ENSO prediction · Skill diagnosis · Model hindcasts · Nino3.4 SST index · Target month slippage · Statistical field significance

1 Introduction

The first version of the Climate Forecast System coupled model (CFSv1; Saha et al. 2006) was run operationally by NOAA's Climate Prediction Center (CPC) between 2004 and 2011.[1] In April 2011 the second version, CFSv2 (Saha et al. 2013), was implemented and used operationally. In both model versions, an ensemble of forecasts is run from each start time, each starting from a different initial analysis, and each resulting in a different realization of the predicted seasonal mean, together defining a predicted

This paper is a contribution to the Topical Collection on Climate Forecast System Version 2 (CFSv2). CFSv2 is a coupled global climate model and was implemented by National Centers for Environmental Prediction (NCEP) in seasonal forecasting operations in March 2011. This Topical Collection is coordinated by Jin Huang, Arun Kumar, Jim Kinter and Annarita Mariotti.

A. G. Barnston (✉) · M. K. Tippett
International Research Institute for Climate and Society, The Earth Institute at Columbia University, Lamont Campus, Palisades, NY 10964, USA
e-mail: tonyb@iri.columbia.edu

M. K. Tippett
Department of Meteorology, Center of Excellence for Climate Change Research, King Abdulaziz University, Jiddah, Saudi Arabia

[1] CFSv1 continued to be run, however, through most of 2012 in parallel with CFSv2.

probability distribution. Some basic characteristics of the CFSv1 and CFSv2 model versions are shown in Table 1. The CFSv2 represents an improved version of CFS in many respects. Additional to changes in the model dynamics and an enhancement in forecast resolution and ensemble size, notable differences in CFSv1 and CFSv2 include, first, that the CO_2 concentration in CFSv2 evolves over time with the initial CO_2 concentration prescribed as the global mean observed CO_2 value at the beginning of the forecast, while for CFSv1 the CO_2 value is fixed at the observed 1988 concentration. A second difference between CFSv1 and CFSv2 is in the initial conditions: In CFSv2, initial conditions come from the Climate Forecast System Reanalysis (CFSR; Saha et al. 2010), while in CFSv1 they come from NCEP/DOE Reanalysis-2 (R-2; Kanamitsu et al. 2002). It is documented by Saha et al. (2010) that the atmospheric analysis (and hence the initial conditions) based on the CFSR is more realistic than for the R-2.

Given the improvements in CFSv2 compared with CFSv1, one would expect relatively better predictive skill in CFSv2 in most fields and over many regions of the globe. However, a discontinuity at year 1999 in the CFSR, related to a change in the atmospheric observing system, induced a change in the characteristics of the SST used for the initial conditions for the CFSv2 integrations beginning that year—especially those in the tropical Pacific (Xue et al. 2011; Kumar et al. 2012; Xue et al. 2013). Here we compare the skill of predictions of Nino3.4 SST in the tropical Pacific by CFSv2 to those of CFSv1, and examine which features of the skill differences may be related to CFS model improvement, and/or to the 1999 discontinuity in the initial conditions due to the CFSR. More background about the 1999 discontinuity will be provided in the context of the initial presentation of results below, in Sect. 3.1.

The Nino3.4 region is selected as the focus of this study because it is closely associated with the ENSO state (Barnston et al. 1997), which influences seasonal climate through well known teleconnections (e.g., Ropelewski and Halpert 1987; Mason and Goddard 2001; Hoerling and Kumar 2002; among many others). We focus on prediction of the Nino3.4 index, and examine the significant performance differences between CFSv1 and CFSv2, and between each model version with and without corrections for their discontinuous climatologies.[2] The ultimate interest is in model version comparisons following the corrections. The significance of the discontinuities themselves is assessed, given the 28-year hindcast records. The data and methods are described in Sect. 2, followed by results in Sect. 3 and a discussion and some conclusions in Sect. 4.

2 Data and methods

The retrospective forecasts (i.e., hindcasts) of CFSv1 and CFSv2 were initialized from their respective Reanalysis data, producing ensembles run on a time-staggered schedule within each month (e.g., 4 members at 5-day intervals for CFSv2). The hindcasts of both versions begin in 1982, and are run 9 months into the future (Table 1). Here, for simplicity the lead time is defined by the lead month order of the hindcast, ranging from 1 to 9, despite that lead time is often defined to be one less. The CFSR Reanalysis from which CFSv2 runs are initialized is at T382 ($\sim$38 km) horizontal resolution, while that for CFSv1, the NCEP/DOE Reanalysis (Kanamitsu et al. 2002), is at T62 ($\sim$2°).

The observed SST data against which the CFS hindcasts are verified are the monthly mean of the optimum interpolation version 2 (OIv2; Reynolds et al. 2002), at 1° resolution. Here we use the mean SST over the Nino3.4 region (5°N–5°S, 120°–170°W), and use 1-month averages for both predictions and observations.

For the deterministic verifications, only the ensemble mean of the model predictions is used, and treated as a single best guess forecast. For the probabilistic reliability analysis the distribution of the individual ensemble members are used to define the model forecast probabilities for the tercile-based categories. Those categories are defined with respect to the model's climatological distribution, using individual members, which varies as a function of the start time and lead time. Tercile-based categories are also defined for the observations.

The verification measures include basic performance diagnostics for deterministic forecasts: temporal correlation with observations for a given season and lead time, root mean squared error (RMSE), ratio of interannual standard deviation of model predictions to those observed, and a lesser known measure called target month slippage. The latter is an indication of biases in the timing of the predictions, such as that in which predictions verify better on target months occurring earlier than the intended month (Tippett et al. 2012; Barnston et al. 2012). A final deterministic diagnostic is a comparison of linear trends in the model predictions to that observed. Prediction bias is not examined in the usual manner, because the bias in the Nino3.4 SST predictions of both models changes abruptly around a specific year within the hindcast history, and corrective measures are taken that largely eliminate model bias. Specifically, the differing biases observed over two portions of the hindcast period are removed individually

[2] A discontinuity in CFSv1 is found also noted for Nino3.4 forecasts, but it is smaller in magnitude than that of CFSv2 and has a different cause.

Table 1 Some basic specifications for CFSv1 and CFSv2

	CFSv1	CFSv2
Horizontal resolution	T62 (~2°)	T126 (~1°)
Vertical resolution	64 levels	64 levels
Atmospheric model	GFS from 2003	GFS from 2009
Ocean model	MOM3	MOM4
No. ensemble members/month	15	24
Initial conditions for 0.5 month outlook (example shown is for a seasonal mean forecast for DJF)	R2 and GODAS: five initial conditions each from near the 1st and 11th of Nov., and 21st of Oct.	CFSR: four initial conditions each from the 17th, 12th, 7th, 2nd of Nov., and 27th of Oct.
Climatological base period	1982–2004	1982–2004
Maximum forecast lead time	9 months	9 months
Source of initial condition data (horizontal resolution)	NCEP/DOE reanalysis (T62)	Climate Forecast System Reanalysis, or CFSR (T382)
Sea ice	Climatology	Predicted
Carbon dioxide concentration setting	Fixed at 1988 level	Evolving with time

using the two sub-period climatologies, so that each portion becomes bias-free.

A probabilistic verification analysis—reliability analysis (Murphy 1973; Wilks 2006)—is used to detect probabilistic confidence levels as described by the distribution of the models' ensemble members associated with each prediction, and probabilistic biases. Reliability is a measure of the correspondence between the forecast probabilities and their subsequent observed relative frequencies, spanning the full range of issued forecast probabilities. Model probability forecasts are defined on the basis of the proportion of ensemble members falling into each of the three defined climatologically equiprobable categories. Perfect reliability would be achieved if, for example, the above normal Nino3.4 SST category were assigned a probability of 40 % in 20 instances over all of the issued forecasts, and the later observed seasonal mean anomalies were in the above normal category in exactly 8 (i.e., 40 %) of those instances. Here we analyze reliability for the 6-month lead forecasts, representing a moderate to long lead time. Because our sample size of predictions is small, we combine all target months, and form eleven 10 %-wide forecast probability bins centered on 0, 10, …, 90 and 100 % probability. Then there are (28) (12) = 336 predictions, resulting in an average of about 31 predictions per probability bin. However, as will be discussed in Sect. 3.1, the 336 predictions are not independent cases, because the ENSO state changes slowly so that forecasts of adjacent start times or adjacent lead times are strongly mutually correlated. The reliability analysis will be described further in the context of its application, in Sect. 3.6.

To help provide statistical support for the reality of the two separate bias periods for each CFS model version, t-tests for the mean difference are applied. Additionally, differences in skill between Nino3.4 predictions of CFSv1 and CFSv2, stratified by season and lead time, are assessed statistically under several arrangements of bias correction status. Comparisons when both model versions are corrected are of greatest interest. Given that there are 12 seasons and 9 lead times, collective (or field) significance tests (Livezey and Chen 1983) for the overall skill difference for the entire matrix of 108 skill differences are conducted to determine the likelihood that the individually significant cells in the matrix are significant by chance. Determining field significance requires accounting for the effective number of statistically independent cases in the data set. Here we estimate this number—the statistical degrees of freedom—based on the autocorrelation structure in the forecast and observed SST data.

3 Results

The comparative performance diagnostics for CFSv1 and CFSv2 are given first using deterministic verification measures in Sects. 3.1–3.5, followed by a probabilistic diagnostic measure (reliability analysis) in Sect. 3.6.

3.1 Anomaly correlation

The anomaly correlations between predictions and observations of Nino3.4 SST are shown in the left column of Fig. 1 as a function of target month and lead time for CFSv1 and CFSv2. The most noticeable skill difference is found in forecasts for northern late spring/summer at medium to long lead times, where CFSv1 has relatively low skill (correlations of 0.5 or lower) while CFSv2 shows higher skill (0.6–0.7). These forecasts are for target months

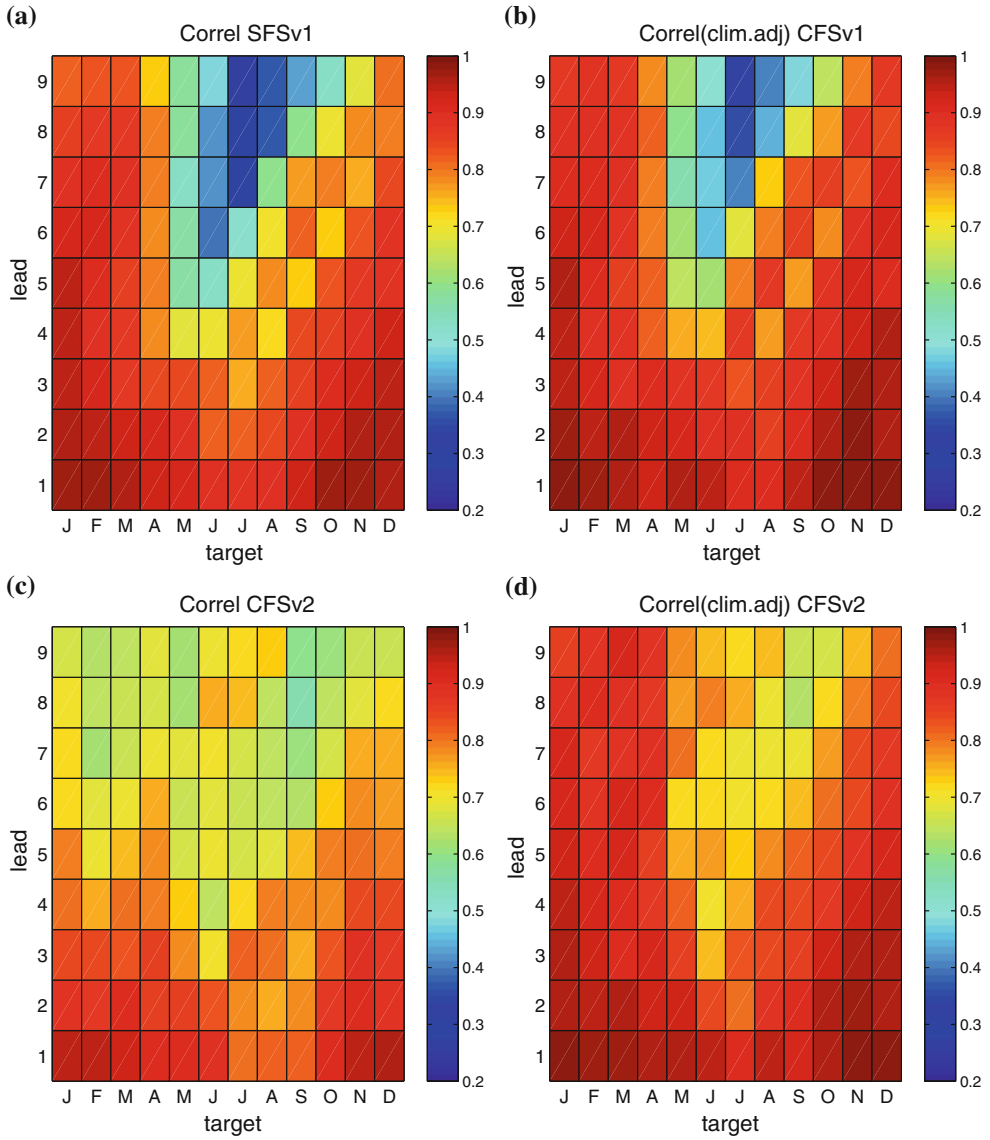

Fig. 1 Temporal correlation between **a** CFSv1 and **c** CFSv2 predictions of Nino3.4 SST and verifying observations over the 1982–2009 period. The target month is indicated on the *horizontal axis*, and lead time on the *vertical axis*. A lead time of 1 month implies a prediction made at the very beginning of the target month using data up to the end of the previous month. *Right column* shows temporal correlation for CFSv1 (**b**) and CFSv2 (**d**) following elimination of discontinuities in the predictions of each model by using two separate climatologies (see text)

beyond the northern spring ENSO predictability barrier that are made much earlier than that barrier—the condition known to present a large challenge (e.g., Jin et al. 2008). Another conditional skill difference, in the opposite direction, is found for predictions for times near the mature stage of an ENSO episode made from start times after the onset of the episode (e.g., a forecast for February made in July). These "easier" predictions are better made by CFSv1 than CFSv2. Why would this be the case for a model that outperforms its predecessor under more difficult prediction circumstances?

Figure 2 shows the error of CFSv1 and CFSv2 predictions as a function of start time for all seasons and leads through the 28 year hindcast period. A discontinuity in the CFSv1 errors appears near 1991, and a larger one is seen in CFSv2 errors near 1999. Such discontinuities would be expected to degrade all verification measures relative to discontinuity-free errors, including temporal correlation. The source of the 1991 change in CFSv1 error has been attributed to a problem in the use of bathythermograph (XBT) measurements prior to 1991 (Berringer and Xue 2004), and is not examined closely

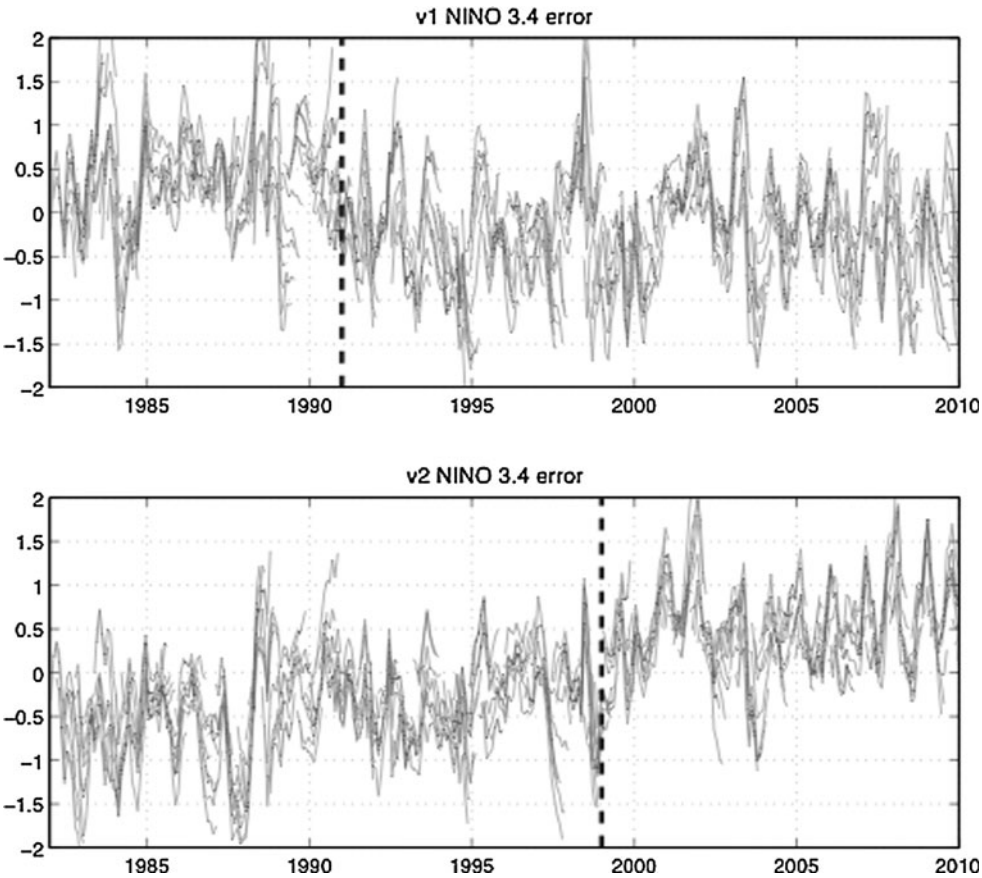

Fig. 2 Error (°C) in Nino3.4 SST predictions of CFSv1 (*top*) and CFSv2 (*bottom*) for start times (indicated on *horizontal axis*) over the course of the 1982–2009 period. Errors for predictions at all lead times are shown. *Vertical lines* are drawn at the beginning of 1991 (for CFSv1) and 1999 (for CFSv2) to highlight the points of error discontinuity

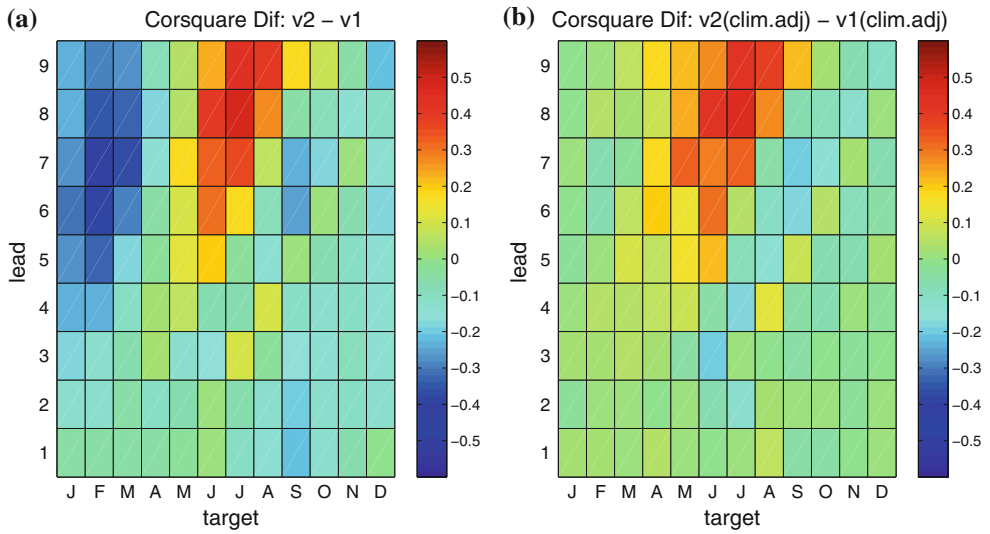

Fig. 3 Difference in squared correlation (of predictions vs. observations) of CFSv2 and CFSv1. **a** without treatment for discontinuities and **b** following treatment using dual climatologies for each model version (*right*). Negative sign is retained *upon squaring*. The target months and lead times are as described above in caption of Fig. 1

here. The CFSv2 error discontinuity, on the other hand, has been related to a discontinuity at year 1999 in the CFSR reanalysis data (Saha et al. 2010) that induced a change in the characteristics of the SST—particularly in the tropical Pacific (Kumar et al. 2012; Xue et al. 2013). This SST change has been attributed to the introduction

Table 2 Significance category for discontinuity in CFSv1 bias (for 1982–1990 vs. 1991–2009) as a function of target month and lead time

Lead (months)	Target month—CFSv1											
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
9										**1**	**1**	5
8									**1**	**1**		
7								5	**1**	5		
6								**1**	5			
5									5			
4												
3												
2												
1												

Entries of "5" denote 2-sided significance at the 5 % level, and "**1**" likewise but at the 1 % level. Changes from the earlier to later period are negative for all cells. Field significance for a downward discontinuity over the set of 108 cells is $p = 0.05$ with a 2-sided test

Table 3 Significance category for discontinuity in CFSv2 bias (for 1982–1998 vs. 1999–2009) as a function of target month and lead time

Lead (months)	Target month—CFSv2											
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
9	**1**	**1**	**1**	**1**	**1**	5	5			5	5	5
8	5	5	**1**	**1**	5	5	5			5	5	**1**
7	5	**1**	**1**	5	5	5			5	5	**1**	5
6	5	5	5	5	5			5	5	**1**	5	
5	5	5	5	5				5	5	5		5
4		5									5	5
3										5	5	
2				–					**1**	**1**	5	
1		–	–	–				5	**1**	5		

Entries of "5" denote 2-sided significance at the 5 % level, and "**1**" likewise but at the 1 % level. Changes from the earlier to later period are positive for all cells except where noted by minus sign. Field significance for an upward discontinuity over the set of 108 cells is $p = 0.002$ with a 2-sided test

of the ATOVS[3] radiance data in the atmospheric assimilation beginning in late 1998 (Zhang et al. 2012), due to forcing from the atmospheric to the oceanic aspects of the Reanalysis (Xue et al. 2011). The positive change in central tropical Pacific SST in 1999 does not coincide with observed SST trends documented in other studies, which have been slightly downward (e.g., Kumar et al. 2010, 2012; Deser et al. 2010; Lyon and DeWitt 2012; Lyon et al. 2013), and is therefore seen as spurious. A change in tropical Pacific SST behavior around 1999, if real, would be important because of the implied ENSO state, with its known remote teleconnections to seasonal climate (Hoerling and Kumar 2002; among many others). A change in the climatology of reanalyzed tropical Pacific SST in 1999 implies a change in the initial conditions used to begin a prediction run of CFSv2. Changes beginning in 1999 in the CFSv2 predictions have indeed been noted in SST and related oceanic and atmospheric fields (e.g., subsurface ocean temperature, low-level zonal winds, and precipitation), and appear most strongly in the general vicinity of the tropical Pacific (Wang et al. 2011; Chelliah et al. 2011; Ebisuzaki and Zhang 2011). Kumar et al. (2012) found that these changes are not replicated when using independent oceanic initialization data, such as the NCEP global ocean data assimilation system (GODAS) (Behringer and Xue 2004) or the National Oceanographic Data Ceter (NODC) (Levitus et al. 2009). It will be shown below that the impact of the 1999 discontinuity on the predictions of Nino3.4 SST is evident from the shortest lead time, propagates to longer lead times, and exhibits seasonal dependence.

[3] ATOVS refers to the Advanced Television and Infrared Observation Satellite (TIROS) Operational Vertical Sounder radiation data system.

To free the evaluation from the effects of discontinuities in both CFS versions, dual climatologies from which to form anomalies are developed (1982–1990 and 1991–2009 for CFSv1; 1982–1998 and 1999–2009 for CFSv2), and the evaluations are repeated. Results following this modification (to be called "correction" or "adjustment") are shown in the right column of Fig. 1. Improvements are noted in the cases of both model versions, but are more substantial in CFSv2. In CFSv2, higher correlations are seen in most seasons and leads, but most notably for predictions for late northern autumn and winter made during summer or later—forecasts considered least challenging, but less skillful than those of CFSv1 before the correction. The correlation differences between CFSv2 and CFSv1 before and after the discontinuity corrections for both models are shown in Fig. 3 in terms of the difference in squared correlation (where negative signs are retained upon squaring). The relative superiority of CFSv2 for long lead predictions through the northern spring predictability barrier is present with or without the correction, but with the correction CFSv2 no longer falls short of CFSv1 for moderate and long lead predictions for northern winter made from earlier within the same ENSO cycle. It is noted, however, that even following the correction CFSv1 performs about as well for these predictions as CFSv2. Following the correction, the equal or better performance of CFSv2 applies to most seasons and leads, although an exception is noted for moderate to long lead predictions for northern summer/fall, made from March or April through the northern spring barrier.

The brevity of the 28-year hindcast record, and particularly of the subperiods that define the dual climatologies, raises questions about the statistical significance of the skill differences between the model versions before and after the climatology correction, and even of the existence of the forecast discontinuities themselves. Statistical assessments are carried out to address these issues. First, t-tests are conducted for the differences between the means of the prediction errors before and after the discontinuities for each start time and lead time (i.e., the lines in Fig. 2). Significance results are shown in Tables 2, 3 for the errors of CFSv1 and CFSv2, respectively. The downward step in CFSv1 bias beginning in 1991 is statistically significant only at moderate to long lead times for months in the second half of the year. The significance of the upward step in CFSv2 bias beginning in 1999 is more pervasive and appears from the very shortest leads for forecasts starting in the latter half of the calendar year, suggesting an associated discontinuity in the initial conditions provided by the CFSR during those months. This seasonal preference for the initial condition discontinuity was noted by Kumar et al. (2012).

Because some month/lead combinations are expected to be statistically significant by chance when a multiplicity of tests are conducted, we assess the field significance of the set of 108 cells collectively, each having its own "local" significance. One approach to evaluating field significance, and the one used here, is to estimate the number of effective degree of freedom, or statistically independent forecast cases, that exists within the full 108-cell matrix. This estimate is based on the interplay of the autocorrelation as a function of temporal lag time between the Nino3.4 observations and predictions. Within the predictions there is autocorrelation both with respect to time for any fixed lead time (forming one row in the skill matrix such as that shown in Fig. 1), and with respect to lead time for fixed target month (a column in the skill matrix). Using the autocorrelation structure spanning the first n lags (within which autocorrelations are beyond those mainly associated with sampling variability), the effective time required to gain one additional degree of freedom is determined. Within one dimension (months for a fixed lead, or lead for a fixed target month) the effective time t is estimated using

$$\tau = 1 + 2\sum_{lag=1}^{n} \left(autocor1_{lag}\right)\left(autocor2_{lag}\right) \qquad (1)$$

where autocor1 is the autocorrelation at a given lag time for a first variable, autocor2 is that of the second variable and the lag time spans up to a chosen stopping value n beyond which autocorrelations become insignificant. For example, for a statistical test involving CFSv2 predictions against observations, those would be the two variables. Equation (1) was used by Davis (1976) and later used in Chen (1982) and Livezey and Chen (1983) in the context of independent sampling times for climate variables whose anomalies have long decay times. The larger the sum of the cross-products of the autocorrelations over the included lag times, the longer the effective time $\boldsymbol{t}$, and the fewer the resulting statistical degrees of freedom (i.e., sample size) to be used in statistical tests.

Applying (1) to the observed and predicted Nino3.4 SST data, we first note that the autocorrelation in both model and observed data at 1 year lag is near zero. This result, independently confirmed elsewhere for many ENSO-related variables, implies that for any single month/lead-time combination, the full 28 years can be used as the effective sample size. For lags smaller than 12 months, autocorrelations for the monthly observations during 1982–2009 are roughly 0.95, 0.87, 0.76, 0.65, 0.53, 0.40 and 0.29 for lags of 1 through 7 months, respectively. These autocorrelations are influenced equally by all times of the year, including times of relatively low or high autocorrelation. Autocorrelations in CFSv1 and CFSv2 for fixed lead times, while not identical to those of the observations, are approximately equivalent when aggregated over all lead times. Application of (1) to these autocorrelations for either

model version versus observations, or a model version versus itself before and after correction, yields 1 temporal degree of freedom per 7.4 months, resulting in 1.60 degrees of freedom per year. Application of the same approach to the lead dimension pertains to autocorrelations between model predictions for fixed targets, where lag now represents differences in lead time. The result is identical for autocorrelations for the observations, but slightly stronger model autocorrelations in the lead dimension for fixed target than in the time dimension for fixed lead. The outcome is 1 degree of freedom per 9.8 months, and consequently the 9 lead times yield 1.82 degrees of freedom per forecast integration. Because the months within a year and the leads for forecasts for a given targeted month represent two separate dimensions, the entire matrix of 108 months/lead-time cells produces (1.60) (1.82) = 2.91 degrees of freedom per year of predictions over the 12 months and 9 lead times. Thus, while a single cell in the matrix provides 28 degrees of freedom over the 28 year period, the matrix of 108 time series, with 28 forecast-observation pairs each, supplies about 81 (2.91 times 28).

With an estimate of 81 effective degrees of freedom for all target months and leads over the 28 year period, a field significance test for the model bias discontinuities is applied to the average t-statistic across over the 108 target-month/lead combinations for CFSv1 (Table 2) and CFSv2 (Table 3). Although the physical underpinnings of both discontinuities have been identified (an XBT issue for CFSv1, and an ATOVS impact for CFSv2), and these causes may provide expectations of the directions of discontinuity in the case of each model version, we use a 2-sided test to be cautious. The result is a significance p value of 0.05 for CFSv1, and 0.002 for CFSv2. Although the discontinuity in the bias of CFSv1 is significant, it has not been a major issue in CFSv1 research. For example, while the performance of CFSv1 is examined in detail in Jin and Kinter (2009), the discontinuity is not discussed. In the case of CFSv2 the discontinuity is more widely recognized (Wang et al. 2011; Xue et al. 2011, 2013; Zhang et al. 2012; Kumar et al. 2012).

The correlation skill differences between CFSv1 and CFSv2, or between either CFS version before and after correction for the changing bias, are similarly tested for statistical significance. The Fisher r-to-Z transformation (Hayes 1973) is used for significance tests of differences between two correlations. Although we do not show significance results for individual target-month/lead combinations, we assess the field significance of the matrix of skill differences as a whole. Field significance of the set of correlation differences are tested for (1) CFSv1 versus CFSv2, (2) CFSv1 before versus after correction, (3) CFSv2 before versus after correction, (4) CFSv1 versus corrected CFSv2, and (5) corrected CFSv1 versus corrected CFSv2. One-tailed significance tests are used for all comparisons, because one expects a priori for CFSv2 to be more skillful than its CFSv1 predecessor and for corrected models to be more skillful than uncorrected ones. Results of these field significance tests, and the percentages of individual cells showing improvements or degradations (and correlation differences that are significant), are provided in Table 4. Among the five comparisons, the only one having field significance is the CFSv2 versus corrected CVSv2 correlation set, meaning that the re-definition of anomalies using the dual climatology in CFSv2 significantly increases the correlation skill. Skill differences between CFSv1 and CFSv2 fail to achieve significance in a collective sense over all months and leads, even when the corrected CFSv2 is compared with the uncorrected CFSv1. It is noted that when neither model version is corrected, CFSv2 has slightly lower overall skill then CFSv1 (top row of Table 4), as noted also in Xue et al. (2013). With corrections, this ranking is reversed.

While the above results may seem discouraging, one must keep in mind that they may reflect the modest sample size more than a lack of real incremental improvements in model quality between CFSv1 and CFSv2. Additionally, the tests weight all cells in the matrix equally, regardless of lead time and season. Such equal weighting ignores the existence of features considered of relatively greater importance, such as performance in predictions traversing the northern spring predictability barrier that suggests mostly higher skill in CFSv2.

3.2 RMSE

A similar skill comparison is conducted for RMSE using standardized anomalies,[4] with results shown in Fig. 4. The results for RMSE differ noticeably in pattern from those of correlation because biases in both mean and in amplitude contribute to RMSE but not to correlation. RMSE scores are greatly reduced with the dual climatology correction for both model versions, indicating the importance of the changing sub-period biases that can greatly exacerbate the squares of the largest errors in the direction of the bias—especially at short to medium leads for early northern autumn when large errors already occur in predicting new ENSO events. Correction substantially improves these prediction errors, particularly for late northern summer target months made early in the calendar year.

The statistical significance of differences between RMSE for the model versions, and for each version before

[4] Here the RMSE is standardized for each season individually to scale it so that climatology forecasts (zero anomaly) would result in the same RMSE-based skill (of zero) for all seasons, and all seasons' RMSE would contribute equally to a seasonally combined RMSE.

Table 4 Local and field significance evaluation for various correlation skill comparisons involving uncorrected and corrected versions of CFSv1 and/or CFSv2

Model versions for comparison, and their overall correlation over 12 seasons and 9 leads ("adj" = dual climatol.)	% Cells with increased (significantly increased) correlation	% Cells with decreased (significantly decreased) correlation	108-Cell field significance: z-statistic	108-Cell field significance: 1-sided p value for improvement
CFSv1–CFSv2	26 (5)	74 (15)	−1.00	0.84
0.833–0.775				
CFSv1–CFSv1 (adj)	99 (0)	1 (0)	0.81	0.21
0.833–0.870				
CFSv2–CFSv2 (adj)	100 (33)	0 (0)	2.06	0.02
0.775–0.880				
CFSv1–CFSv2 (adj)	85 (6)	15 (0)	1.06	0.15
0.833–0.880				
CFSv1 (adj)–CFSv2 (adj)	54 (5)	46 (0)	0.25	0.40
0.870–0.880				

A one-sided test is used for field significance shown in last column. The $p = 0.05$ level (one-sided) is used to count significantly changed individual cells for the percentages shown in parentheses in the first two columns

and after the climate adjustment, is determined using the F test for each of the month/lead combinations. A summary of the outcome of the significance tests is shown in Table 5, along with field significance for overall differences between the two versions being compared. Local and field significance for reductions of RMSE with the dual climatology correction are strong for both model versions, and exceed those for the corresponding correlation skill improvements (Table 4), indicating that RMSE is particularly strongly impacted by the changing model biases. Overall differences in RMSE between CFSv1 and CFSv2 are not statistically significant when the predictions of neither model are corrected by the dual climatology, but are significant when they both are corrected. This significance, and a much larger degree of significance when CFSv1 is not corrected and CFSv2 is corrected, did not appear in the same tests for correlation skill differences.

3.3 Standard deviation ratio

Figure 5 shows the ratio of the interannual standard deviation of the model ensemble mean predictions to that of the corresponding observations for each model version for each target month and lead time, both before and after correcting biases by forming two climatologies in place of a single discontinuous one. Ideally the ratio would be no higher than unity throughout all seasons and leads, and lower to the extent that predictive skill is imperfect (in theory, the ratio should equal the square root of the fraction of the observed variance explained by the forecasts).

The climatology correction results in small changes in the ratios for CFSv1, but a noticeable decrease toward unity is found in the case of CFSv2 for short to intermediate lead times for target months in the second half of the year. More importantly, the ratio of CFSv1 is noted to be too high (>1.5) even following the correction for intermediate lead predictions for northern spring season when the observed standard deviation is near its seasonal minimum. CFSv2 lacks this weakness and, following the bias correction, shows ratios fairly close to unity throughout many seasons and leads. In keeping with the lower skill expected for forecasts traversing the northern spring predictability barrier, ratios of less than unity are noted in CFSv2 for predictions for June–October made at medium and long leads.

Significance and field significance test results for the standard deviation ratios are shown in Table 6 (Here, counts of individual matrix cell increases and decreases are not shown because decreases in ratios initially less than unity may or may not be desirable). In contrast with significance results for correlation and RMSE skills, here the differences between CFSv1 and CFSv2 are field significant regardless of the climatology correction status of the models, while differences related to the dual climatology corrections themselves are not field significant. The conclusion is that the standard deviation ratio is an attribute in which CFSv2 shows better performance than CFSv1—namely, the predictions of CFSv1 have higher amplitude than warranted, while those of CFSv2 are substantially more in keeping with realistic signal to noise ratios. This

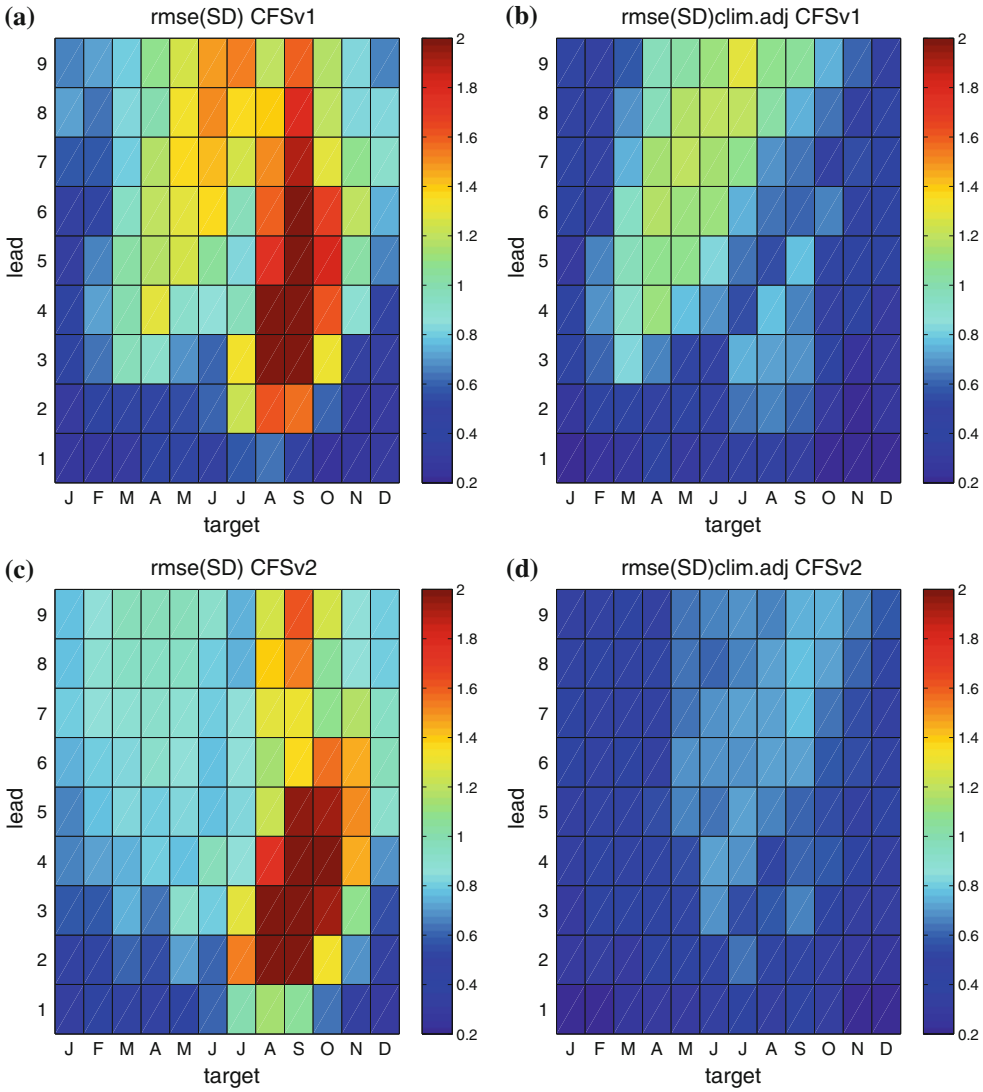

Fig. 4 Root mean squared error of predicted versus observed standardized anomalies of **a** CFSv1 and **c** CFSv2 without treatment for discontinuities and **b**, **d** following treatment using dual climatologies for each model version. In the absence of any skill, RMSE of 1.41 is expected. The target months and lead times are as described above in caption of Fig. 1

characteristic will be corroborated below in a probabilistic reliability analysis.

3.4 Target month slippage

"Target month slippage" occurs when predictions verify with higher skill for target months earlier or later than those intended, such as a 4-month lead prediction intended for July verifying better using observations of May or June instead of July. Slippage typically occurs when predictions are late in reproducing observed changes, such as onsets or endings of ENSO episodes. In an extreme case, a prediction for a new event may not be made until the event is already present in the initial conditions. Slippage cannot be diagnosed by comparing forecasts with the verifying observations only for the intended target time. Although slippage is a systematic error, it is indistinguishable from a random error when forecasts at different leads are evaluated independently. It is most likely to occur when prediction is most difficult, such a prediction made in March for targets of July and beyond.

Slippage is expressed in plots of skill as a function of the lag time between the measured target period and the intended one. Typically, due to sampling considerations, the diagnosis is made for all seasons together. In the absence of slippage, correlation skill maximizes for the intended target (lag = 0), and drops off with increasing positive or negative lags. When slippage exists, skill is greatest for a nonzero lag time such as one or more months earlier than the intended month. To the extent that slippage is systematic, it can be corrected using statistical methods, such as multiple regression, that define optimum shifts of

Table 5 Local and field significance evaluation for various RMSE skill comparisons involving uncorrected and corrected versions of CFSv1 and/or CFSv2

Model versions for comparison, and their overall RMSE over 12 seasons and 9 leads ("adj" = dual climatol.)	% Cells with decreased (significantly decreased) RMSE	% Cells with increased (significantly increased) RMSE	108-Cell field significance: average MSE ratio for F-test	108-Cell field significance status (significant means $p < 0.05$)
CFSv1–CFSv2 1.14–1.09	44 (18)	56 (24)	1.14	Not significant
CFSv1–CFSv1 (adj) 1.14–0.58	86 (68)	14 (0)	3.97	Significant $p < 0.001$
CFSv2–CFSv2 (adj) 1.09–0.49	100 (90)	0 (0)	5.60	Significant $p \ll 0.001$
CFSv1–CFSv2 (adj) 1.14–0.49	89 (72)	11 (1)	5.96	Significant $p \ll 0.001$
CFSv1 (adj)–CFSv2 (adj) 0.58–0.49	64 (31)	36 (3)	1.72	Significant $p < 0.01$

A one-sided test is used for field significance shown in last column. The $p = 0.05$ level (one-sided) is used to count significantly changed individual cells for the percentages shown in the first two columns

the model's forecasts to targets different from those originally intended (Tippett et al. 2012). Here we apply such a multiple regression-based correction to the forecasts of CFSv1 and CFSv2, to increase a skill metric based on the mean squared error (MSE):

$$MSEskill = 1 - \frac{MSE}{SD_{obs}^2} \quad (2)$$

In (2), constant forecasts for the climatological mean results in a score of 0. Figures 6, 7 show slippage and skill results using (2) for CFSv1 and CFSv2, respectively, before and after the regression correction. Slippage is seen in CFSv1 (top left panel of Fig. 6), and it increases with increasing lead times to about 3 months for 9-month lead predictions. The MSE-based skill score (bottom left panel) indicates sub-zero skill for long-lead CFSv1 forecasts for northern summer. After the regression correction (right panels) slippage is decreased and the skill of the long-lead summer forecasts is improved. The same diagnostics for CFSv2 (Fig. 7) indicate little original slippage, and the multiple regression correction does little to improve the already good performance. Reduction of slippage may be a way that the performance of CFSv1 could have been improved in addition to the improvements related to the dual climatology correction (Figs. 1, 4; Tables 4, 5).

3.5 Trend bias and its seasonality

The time-conditional biases in the CFSv1 and CFSv2 predictions (Fig. 2) create trend biases in the sense that a linear trend fit to the predictions exhibits slopes that do not appear in the observations. However, each model also exhibits more gradual trends within each of its sub-periods—particularly CFSv2, and notably for start months around northern late summer and autumn. Figure 8 shows Nino3.4 predictions for the first month from each model version, along with the corresponding observations, for start times of 1 August, 1 September and 1 October for each year of the hindcast period. As expected from earlier analyses, CFSv1 exhibits a positive bias before 1991 and negative bias from 1991 onward, while CFSv2 shows negative bias before 1999 and positive bias from 1999 onward. Additionally, the magnitude of the negative biases in CFSv2 appears to decrease with time up to 1999, and positive biases to increase with time from 1999 forward.

At the earliest lead time, predictions are influenced heavily by the initial conditions (Kumar et al. 2012; Xue et al. 2013). The systematic discrepancies between the short-lead predictions and the observations shown in Fig. 8 are thus mainly indicative of biases in the SST initial conditions in the case of CFSv2, and here these are most prominent for the August, September and October start times. Figure 9 shows biases in the slope of the least-squares linear trend for predictions of CFSv1 and CFSv2 for each target month and lead time. The bias profile for CFSv2 (right panel) resembles the inter-period difference in forecast climatology shown in Kumar et al. (see their Fig. 2c), as would be expected. The positive trend biases in CFSv2 for the shortest lead predictions of August, September and October are noted in the bottom row of cells in Fig. 9b; these northern autumn biases amplify as they propagate to predictions for later target months with

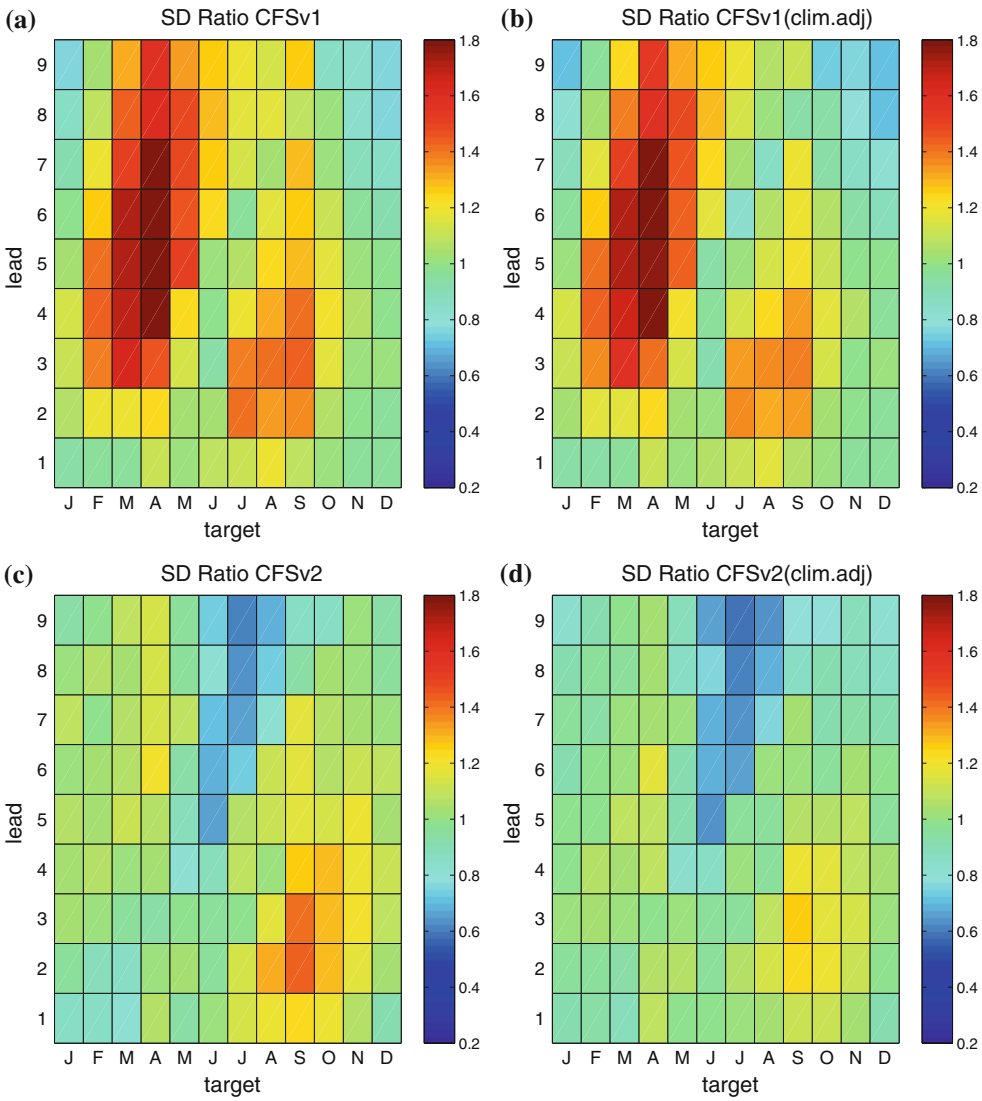

Fig. 5 Ratio of interannual standard deviation of predicted vs. observed anomalies of **a** CFSv1 and **c** CFSv2 without treatment for discontinuities and **b**, **d** following treatment using dual climatologies for each model version. Ideally, the ratio is unity or less. The target months and lead times are as described above in caption of Fig. 1

increasing lead times. This initial condition bias is thus suggested to be partly responsible for the initially noted lower skills of CFSv2 than CFSv1 for predictions made during the less challenging seasons of the year if the data are not corrected by using two separate climatologies.

A reason for a remaining gradual positive trend in CFSv2 predictions relative to observations even after the discontinuity correction using dual climatologies is unknown. A problem involving radiation balance may be a candidate explanation, but additional study is required to explore such a hypothesis.

The trend bias in CFSv1 is negative for most months and leads, partly because of the discontinuity in 1991 but also due to gradual trends within the sub-periods. In contrast to CFSv2, trend biases in CFSv1 do not appear at short leads, indicating a likely lack of major biases in initial conditions. However, CFSv1 has the disadvantage of a non-evolving CO_2 concentration setting, which could result in the slowly declining Nino3.4 SST predictions relative to observed SST.

Significance and field significance test results for the linear trend biases relative to trends in the observations are shown in Table 7. A Fisher Z test is applied to the differences between the observations and the model predictions in their SST-versus-time correlation (which directly determines the slope of the linear trend for standardized data), where the observational data (usually having a near-zero correlation, or trend) is treated as a population so that a 1-sample test is conducted. Results indicate field significance in the downward (upward) temporal trends in CFSv1 (CFSv2), and trends are sufficiently pervasive in CFSv2 that 82 % of the cells in the month/lead-time matrix are individually significant.

3.6 Reliability analysis

We assess the reliability and sharpness of the probabilistic predictions of Nino3.4 SST from the two CFS versions. For any prediction, probabilities for the below-, near- and above-normal categories are defined by counting the proportion of ensemble members whose predictions are in each respective category. The three categories are defined such that each has a one-third probability of occurring during the 28-year hindcast period (i.e., tercile cutoffs are used). For the models, terciles are defined using the individual ensemble members over the study period. The observations are also categorized. The categories may be thought to loosely represent El Nino, neutral and La Nina, although many ENSO classification systems are not tercile-based. Reliability analysis is carried out for each forecast category separately, but plotted together. For simplicity, here we focus only on the 6-month lead predictions. Furthermore, we ignore the near-normal category, which has been demonstrated to have weak performance (Van den Dool and Toth 1991).

As mentioned in Sect. 2, reliability analysis examines the correspondence between the forecast probabilities and their corresponding later observed relative frequencies. Ideally, the two should match. Over- and under-forecasting of the probability for a given category are specific forms of imperfect reliability. Forecast probability biases may depend on the probability level itself, or may be fairly constant over all forecast probabilities. The reliability diagram permits examination of such attributes of the set of probability forecasts. Because the forecast probabilities for

Table 6 Local and field significance evaluation of various standard deviation ratio comparisons involving uncorrected and corrected versions of CFSv1 and/or CFSv2

Model versions for comparison, and their overall SD ratio over 12 seasons and 9 leads ("adj" = dual climatol.)	108-Cell field significance: avg variance ratio for F-test	108-Cell field significance status (significant means $p < 0.05$)
CFSv1–CFSv2 1.21–1.03	1.51	Significant $p < 0.05$
CFSv1–CFSv1 (adj) 1.21–1.18	1.07	Not significant
CFSv2–CFSv2 (adj) 1.03–0.97	1.13	Not significant
CFSv1–CFSv2 (adj) 1.21–0.97	1.66	Significant $p < 0.05$
CFSv1 (adj)–CFSv2 (adj) 1.18–0.97	1.56	Significant $p < 0.05$

A one-sided test is used for field significance shown in last column

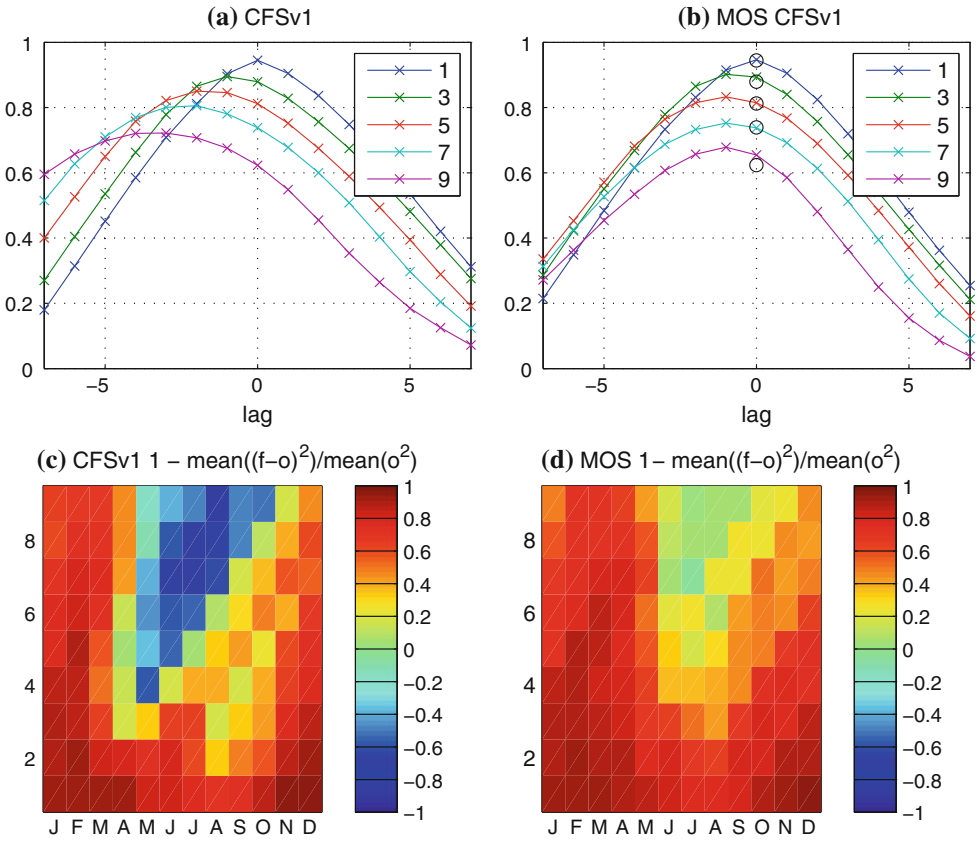

Fig. 6 Target period slippage, and its correction, in CFSv1: (*top*) Correlation between predictions and observations as a function of lag time between verified target month and intended target month, for leads of 1, 3, 5, 7 and 9 months before (*left*) and after (*right*) a MOS correction for slippage based on multiple regression. Predictions free of slippage should have maximum correlation at zero lag. The *hollow circles* in the *right* figure show the correlation at zero lag prior to the correction. (*bottom*) Mean squared error skill score as a function of target month and lead time before (*left*) and after (*right*) the MOS correction

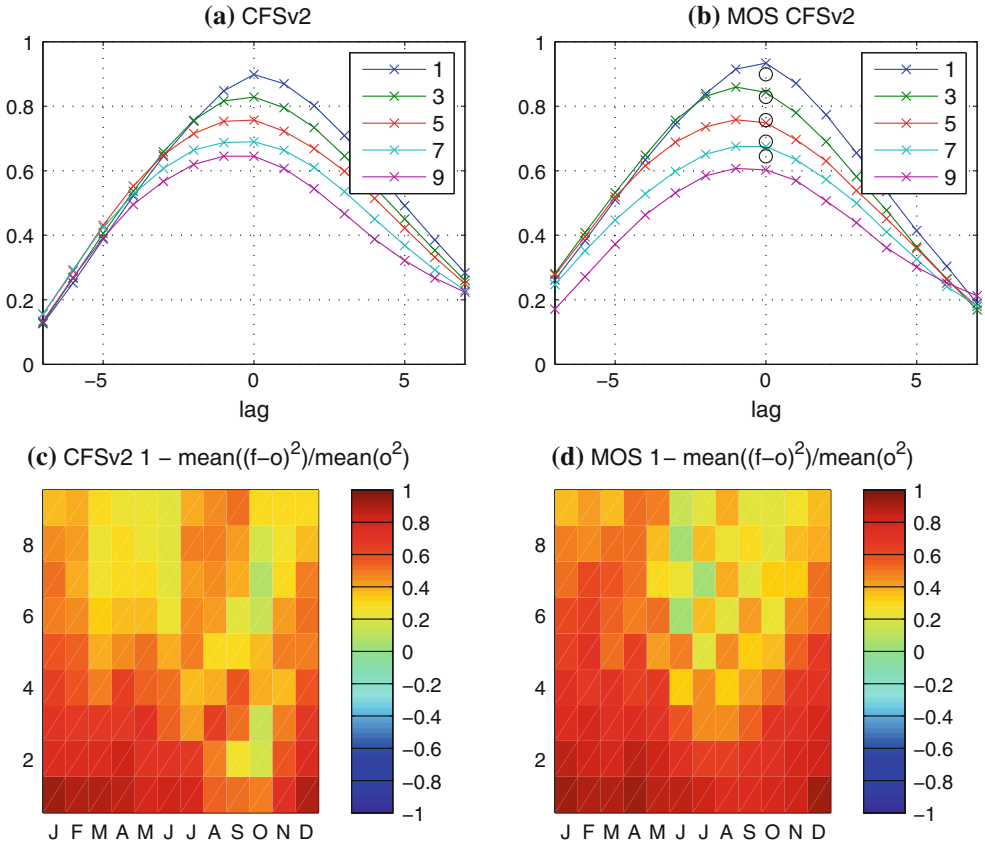

Fig. 7 As in Fig. 6, except for CFSv2 slippage and its correction

each of the categories are binned into an array of probability intervals, reliability analysis requires a large sample of forecasts for each bin to be populated sufficiently for statistical robustness. Here we combine all target months, and form eleven 10 %-wide forecast probability bins centered on 0, 10, …, 90 and 100 % probability, to average about 31 predictions per probability bin, or (28) (12) /11. As indicated in Sect. 2 and discussed in Sect. 3.1, lack of independence among the forecasts results in far fewer than 336 independent forecast cases, so that the results are expected to paint a largely qualitative picture—a "sanity check" for probabilistic reliability.

The reliability diagrams for the above and below normal categories are shown for the two CFS model versions, with uncorrected climatologies, in Fig. 10 as the red and blue curves, respectively. For each category, forecasts are binned for increasing forecast probability intervals (x-axis), and are compared to their corresponding observed relative frequencies of occurrence (y-axis). The diagonal line (y = x) represents perfectly reliable forecasts. The plot insets below the main panel show the percentage of forecasts having probabilities in each of the probability bins.

For CFSv1 (Fig. 10a), positive skill is evidenced by the fact that predictions with increasing probabilities for both below and above normal SST tend to be associated with increasing observed relative frequencies of occurrence. The curves are not smooth because of sampling variability, but the average slope of both curves is seen to be somewhat less than unity. Thus, forecasts are "overconfident", as very low (high) probabilities are not matched by comparably low (high) frequencies of observed occurrence. Overconfidence is particularly noticeable for probabilities between 0.7 and 0.9 for both categories, and for probabilities of 0.0 for above normal predictions. The inset plot at the bottom shows that the lowest bin (0–0.05) is by far the most frequently issued probability, followed by the highest bin (0.95–1.00) and the second lowest bin (0.05–0.15). The U-shaped curve described by the histogram bars indicates high forecast sharpness (i.e., probabilities deviating strongly and frequently from climatology), and the fact that the slope of the lines is <1 indicates that this degree of sharpness is not warranted, given the level of predictive skill achieved at the 6-month lead time.

The reliability result for the uncorrected CFSv2 (Fig. 10b) is somewhat similar to that of CFSv1, except that overconfidence appears milder, as the curves have slope closer (but still less than) unity, with smaller deviations below the ideal reliability (45°) line for bins for 0.50

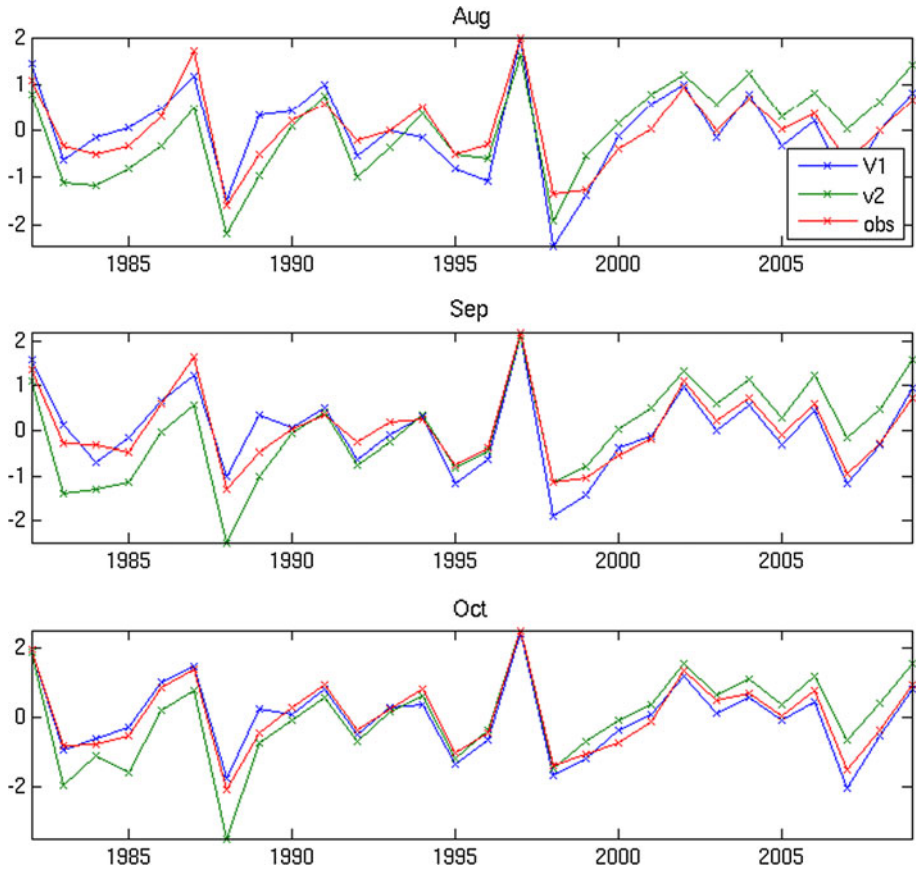

Fig. 8 Shortest-lead Nino3.4 SST anomaly predictions of CFSv1 (*blue*) and CFSv2 (*green*) and corresponding observations (*red*) for start times at beginning of (*top*) August, (*middle*) September, and (*bottom*) October over the 1982–2009 period

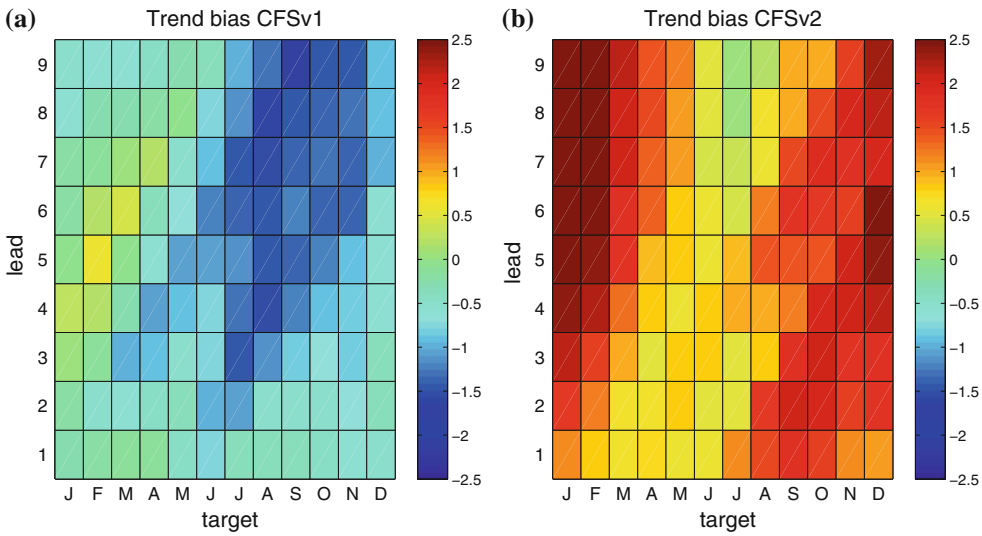

Fig. 9 Bias, relative to observations, in the slope of the linear trend fit over the 1982–2009 (°C per 27 year) period for Nino3.4 predictions of **a** CFSv1 and **b** CFSv2 as a function of target month and lead time

and higher probability. Similarly, the lower inset shows that zero-probability predictions for above normal SST that are issued more than 41 % of the time by CFSv1 are issued 33 % of the time by CFSv2, indicating a greater expressed forecast uncertainty.

The somewhat more reliable probabilistic predictions seen in CFSv2 than in CFSv1 are attributable to a combination of its generally higher skill (Figs. 1, 3) and its slightly less sharp, more conservative probabilities that better reflect the true level of uncertainty in the ocean–atmosphere system. This outcome is consistent with the greater inflation above unity of the standard deviation ratio of the ensemble mean forecasts in CFSv1 than CFSv2 noted above (although high model variance, per se, would

Table 7 Local and field significance evaluation for linear trend bias with respect to the observed trend in the Nino3.4 SST during 1982-2009

Model version	% Cells with negative trend bias (significantly negative)	% Cells with positive trend bias (significantly positive)	108-Cell field significance: average z-statistic	108-Cell field significance status (significant means $p < 0.05$)
CFSv1	86 (31)	14 (0)	−1.05	Significant $p < 0.04$
CFSv2	0 (0)	100 (82)	2.25	Significant $p \ll 0.001$

Here, model data are not corrected using the dual climatology approach. The $p = 0.05$ level (one-sided) is used to count individual cells with significant trend biases for the percentages shown in the first two columns, and a one-sided test is used for field significance shown in last column

also contribute), especially at medium to long lead times (left panels of Fig. 5). Aside from model improvement, one reason for the better probabilistic forecast performance of CFSv2 than CFSv1 is the larger ensemble size of CFSv2 than CFSv1 (24 vs. 15 members), given that smaller ensemble sizes are associated with larger sampling variability in the ensemble mean and the ensemble distribution leading to the probability assignments.

Elimination of the discontinuity in the climatology of the predictions slightly helps to remedy the inflated standard deviation ratio of CFSv2 (lower right panel of Fig. 5). To determine the effect of the correction on CFSv2 reliability, the analysis is repeated using dual climatologies for the tercile boundary definitions for the model prediction category. Results (Fig. 10c) indeed indicate a slope closer to unity, and the observed relative frequencies for forecasts of zero probability become <2 %, suggesting that now such sharply low probabilities are justified in the absence of the spurious change in the forecast climatology within the hindcast period. Similarly, forecasts with 100 % probability are met with correctly verifying observations in about 95 % of cases when using the dual climatologies, but only about 80 % (90 %) for the above (below) normal category without the climatology adjustment. All told, the correction appears to improve probabilistic reliability for CFSv2. However, the small effective sample size of forecasts and observations must be noted. While results are suggestive, and consistent with findings shown earlier for the deterministic verifications, they are not likely to be statistically significant on their own, and are presented for qualitative interest.

4 Discussion and conclusion

Given the time and resources invested toward improvement, one would expect higher predictive skill in CFSv2 than in CFSv1. Here we examine skill differences between CFSv1 and CFSv2 in predictions of the ENSO state, as represented by the Nino3.4 SST anomaly.

Initial examination shows that CFSv2 is better able to predict the ENSO state than CFSv1 at long lead through the northern spring predictability barrier, the time of year when there is most need for improvement. By contrast, CFSv2 appears to fall short of CFSv1 in predictions for northern late summer and autumn start times—times for which ENSO prediction is known to be least challenging. Combining all times of year and all lead times, CFSv2 fails to show net improvement over CFSv1. However, CFSv2 is found to have a significant discontinuity in initial condition climatology near 1999 associated with discontinuities in the oceanic part of the Reanalysis observations generated using the high resolution CFSv2 (the CFSR). The size and impact of this discontinuity is most prominent in the tropical Pacific, in the form of an step-like increase in ENSO-related SST and associated changes in other tropical Pacific conditions around 1999, as described in Kumar et al. (2012), Xue et al. (2013), and other recent studies. This discontinuity is spurious, as it is not reflected in the observations. Here, we examine the consequences of the discontinuity for the performance of model predictions of the Nino3.4 SST anomaly. In identifying and removing those components of the differences in specific skill metrics likely related to the discontinuity, we aim to assess performance differences related to true model improvement (or lack thereof).

The initial condition discontinuity acts to diminish CFSv2's net skill in ENSO prediction, and masks some aspects of its standing relative to CFSv1 in predicting Nino3.4 SST. This diminishing effect is most noticeable for northern autumn start times when skill is highest and when CFSv1 is already quite skillful. The impact of the discontinuity on skill is evaluated by comparing skill with and without correction of the discontinuity by using two separate climatologies from which to form anomalies. The main results are summarized in Table 8. Without the climatology correction, CFSv2 is still seen to outperform CFSv1 in terms of (1) standard deviation ratio with respect to the observations, and (2) probabilistic reliability, due to its lesser degree of amplitude inflation and probabilistic overconfidence, respectively, than seen in CFSv1. A third

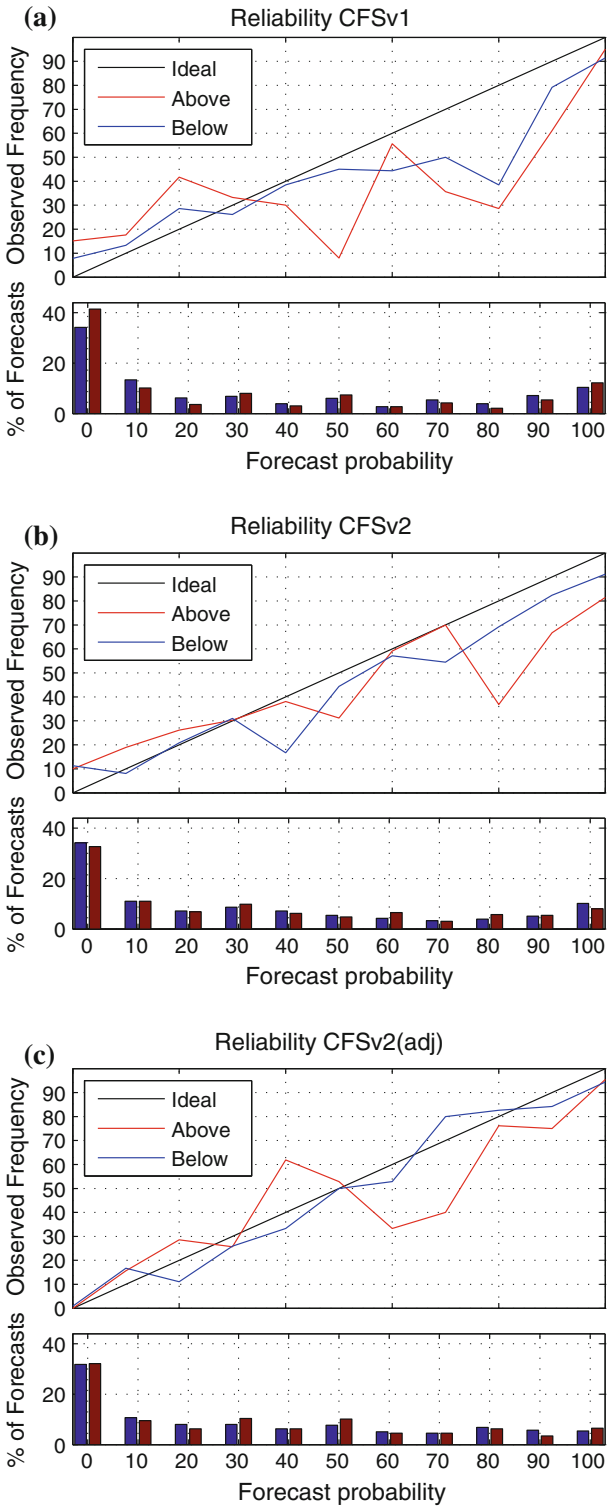

Fig. 10 Reliability diagrams for **a** uncorrected CFSv1, **b** uncorrected CFSv2, and **c** CFSv2 using dual prediction climatology predictions of Nino3.4 SST at 6-month lead time. *Red* (*blue*) *curve* indicates reliability for above (*below*) normal SST predictions. The *black diagonal* (45°) *line* represents the ideal perfect reliability. Probability bins are 10 % wide (e.g., 0.35–0.45), except for the *top* and *bottom* ones, which are 5 % wide. The *histograms* in the *insets below* the main *panel* show the frequency distribution for predictions among the probability *bins*

attribute in which CFSv2 outperforms CFSv1 without a climatology correction is lack of "target month slippage"—i.e., CFSv2 does not tend to verify better on target times earlier than those intended due to slowness in reproducing transitions in the ENSO state.

After discontinuity corrections, including correction of CFSv1's less severe discontinuity, performance of CFSv2 is found to exceed that of CFSv1 at most times of the year in anomaly correlation (although the difference is not statistically field significant) and RMSE (with a highly significant difference), the two most basic and commonly used deterministic skill metrics. After correction, improvement in performance of CFSv2 over CFSv1 is also more strongly field significant in standard deviation ratio with respect to the observations, as CFSv2 lacks the forecast amplitude inflation of CFSv1 to a greater extent. While not confirmed statistically, CFSv2 also appears further improved in probabilistic reliability (Fig. 10).

A constant bias, correctable with a single adjustment, does not degrade measures such as the anomaly correlation or the confidence-indicating slope of the reliability curves. A changing bias, by contrast, is equivalent to a nonsystematic error, uncorrectable unless the problem is identified (e.g., by inspecting Fig. 2) and treated with a combination of human intervention and automation in choosing the point of discontinuity and the correction parameters. The timing of the discontinuity in CFSv2 has been linked to the advent of ATOVS radiance measurements in late 1998, and the lesser discontinuity in CFSv1 to issues with XBT measurements before 1991. Knowledge of the likely causes justifies identification of the temporal break points in the hindcast time series, reducing concern that they are subjectively based.

CFSv2 is shown to have a larger upward trend in Nino3.4 SST than that observed, apart from the 1999 discontinuity. This appears despite the specification of realistic time-evolving CO_2 concentrations—an improvement over CFSv1, which had a fixed and outdated CO_2 concentration, and a possibly related negative trend bias with respect to observations. The positive trend bias in CFSv2, with currently unknown cause, may indicate potential for improvement in a future version of CFS.

In summary, based on Nino3.4 SST prediction skill results after adjustment for discontinuities in the climatologies of CFSv1 and CFSv2, we can conclude:

1. CFSv2 makes more skillful long-lead predictions than CFSv1 from early in the calendar year, through the northern spring predictability barrier. But its shorter lead forecasts through the barrier (e.g., from March start time) remain no more skillful than those of CFSv1 at short and medium lead times. For predictions that do not traverse the barrier, skills of the two model versions are comparable. Overall differences in

Table 8 Summary of general results of the study in terms of comparative performance of CFSv1 versus CFSv2 in Nino3.4 prediction before and after climatology correction for both model versions

Metric	Better performing model		Comments
	Without correction	With correction	
Correlation	CFSv1	CFSv2	Better medium/long-lead boreal summer predictions from CFSv2
RMSE	CFSv2	CFSv2**	Correction reduces CFSv2 RMSE dramatically
Stand. deviation ratio	CFSv2*	CFSv2*	CFSv2 has better SD ratio by larger margin after correction
Slippage	$CFSv2^{NT}$	NA	CFSv1 has 3-month slippage for long-lead predictions
Linear trend bias	CFSv1: negative* CFSv2: positive**	NA	Trend biases are apparent in addition to the discontinuities
Prob. reliability	$CFSv2^{NT}$	$CFSv2^{NT}$	CFSv2 has better reliability by larger margin after correction

The asterisk (two asterisks) denotes statistical significance of performance differences at the 5 % (1 %) level. "NA" means "not analyzed"; as superscript, NT means statistical significance was not tested

correlation skill between CFSv1 and CFSv2, while favoring CFSv2, are insufficient for statistical field significance over the 28-year hindcast period.

2. CFSv2 predictions have more realistic (i.e., lower) amplitude, and correspondingly more reliable probabilistic forecasts, than CFSv1, especially during seasons and leads when predictability is relatively low. This significant improvement in calibration, combined with the slight overall improvement in correlation, leads to a highly statistically significant overall improvement in RMSE.

Although the discontinuity has clearly discernible effects on CFSv2 predictions of ENSO-related SST, they are not large enough to materially degrade the model's predictions of climate across much of the globe, including those involving many of the ENSO-related climate teleconnections. Performance in climate predictions has been found significantly better than that of CFSv1 in many instances, including in the United States during winter when ENSO is a major governing factor (Peng et al. 2013), and in reproduction of the MJO (Weaver et al. 2011). The skill of CFSv2 is even found competitive with that of ECMWF system 4 for winter climate predictions over North America, despite its relative shortcomings in predictions of ENSO and the globally averaged tropical climate (Kim et al. 2012). A better CFSv2 than CFSv1 is expected on the basis of the factors shown in Table 1, including finer horizontal resolution, a more recent version of the GFS atmospheric model component, a more recent ocean model component, a larger ensemble size, more accurate (predicted) sea-ice, and evolving CO_2 concentration. Last but not least, CFSv2 is initialized from a more realistic Reanalysis—*except* for the 1999 discontinuity whose correction using the dual-climatology approach has been demonstrated necessary to recognize some critical aspects of the improved performance.

Acknowledgments The authors appreciate the thoughtful comments and suggestions of the anonymous reviewers. This work was funded by a grant/cooperative agreement from the National Oceanic and Atmospheric Administration (NOAA) (NA10OAR4310210), and also two grants from the MAPP Program of NOAA (NA12OAR4310091 and NA12OAR4310082). The views expressed are those of the authors and do not necessarily reflect the views of NOAA or its subagencies.

References

Barnston AG, Chelliah M, Goldenberg SB (1997) Documentation of a highly ENSO-related SST region in the equatorial Pacific. Atmos Ocean 35:367–383

Barnston AG, Tippett MK, L'Heureux ML, Li S, DeWitt DG (2012) Skill of real-time seasonal ENSO model predictions during 2002–11: is our capability increasing? Bull Am Meteor Soc 93:631–651

Behringer DW, Xue Y (2004) Evaluation of the global ocean data assimilation system at NCEP: the Pacific Ocean. Eighth symposium on integrated observing and assimilation systems for atmosphere, oceans, and land surface, AMS 84th Annual Meeting, Washington State Convention and Trade Center, Seattle, Washington, 11–15

Chelliah M, Ebisuzaki W, Weaver S, Kumar A (2011) Evaluating the tropospheric variability in National Centers for Environmental Prediction's Climate Forecast System Reanalysis. J Geophys Res (Atmos) 116 Art. No. D17107 doi:10.1029/2011JD015707

Chen WY (1982) Fluctuations in Northern Hemisphere 700 mb height field associated with the Southern Oscillation. Mon Weather Rev 110:808–823

Davis RE (1976) Predictability of sea surface temperature and sea level pressure anomalies over the North Pacific Ocean. J Phys Oceanogr 6:249–266

Deser C, Philips AS, Alexander MA (2010) Twentieth century tropical sea surface temperature trends revisited. Geophys Res Lett 37 doi:10.1029/2010GL043321

Ebisuzaki W, Zhang L (2011) Assessing the performance of the CFSR by an ensemble of analyses. Clim Dyn 37:2541–2550
Hayes WL (1973) Statistics for the social sciences. Rinehart and Winston, Holt 954
Hoerling MP, Kumar A (2002) Atmospheric response pattern associated with tropical forcing. J. Clim 15:2184–2203
Jin EK, Kinter JL (2009) Characteristics of tropical Pacific SST predictability in coupled GCM forecasts using the NCEP CFS. Clim Dyn 32:675–691
Jin EK et al (2008) Current status of ENSO prediction skill in coupled ocean–atmosphere models. Clim Dyn 31:647–664
Kanamitsu MW et al (2002) NCEP-DOE AMIP-II reanalysis (R-2). Bull Am Meterol Soc 83:1631–1643. doi:10.1175/BAMS-83-11-1631
Kim HM, Webster PJ, Curry JA (2012) Seasonal prediction skill of ECMWF System 4 and NCEP CFSv2 retrospective forecast for the Northern Hemisphere winter. Clim Dyn 39:2957–2973. doi:10.1007/s00382-012-1364-6
Kumar A, Bhaskar J, L'heureux M (2010) Are tropical SST trends changing the global teleconnection during La Nina? Geophys Res Lett 37:L12702. doi:10.1029/2010GL043394
Kumar A, Chen M, Zhang L, Wang W, Xue Y, Wen C, Marx L, Huang B (2012) An analysis of the non-stationarity in the bias of sea surface temperature forecasts for the NCEP climate forecast system (CFS) version 2. Mon Weather Rev 140:3003–3016
Levitus S, Antonov JI, Boyer TP, Locarnini RA, Garcia HE, Mishonov AV (2009) Global ocean heat content 1955–2008 in light of recently revealed instrumentation problems. Geophys Res Lett 36:L07608. doi:10.1029/2008GL037155
Livezey RE, Chen W-Y (1983) Field significance and its determination by Monte-Carlo techniques. Mon Weather Rev 111:46–59
Lyon B, DeWitt DG (2012) A recent and abrupt decline in the East African long rains. Geophys Res Lett 39:L02702. doi:10.1029/2011GL050337
Lyon B, Barnston AG, DeWitt DG (2013) Tropical Pacific forcing of a 1998–99 climate shift: observational analysis and climate model results for the boreal spring season. Clim Dyn 26 (in press)
Mason SJ, Goddard L (2001) Probabilistic precipitation anomalies associated with ENSO. Bull Am Meteorol Soc 82:619–638
Murphy AH (1973) A new vector partition of the probability score. J Appl Meteorol 12:595–600
Peng P, Barnston AG, Kumar A (2013) A comparison of skill between two versions of the NCEP climate forecast system (CFS) and CPC's operational short-lead seasonal outlooks. Weather Forecast 28:445–462
Reynolds RW, Rayner NA, Smith TM, Stokes DC, Wang W (2002) An improved in situ and satellite SST analysis for climate. J Clim 15:1609–1625
Ropelewski CF, Halpert MS (1987) Global and regional scale precipitation patterns associated with the El Niño Southern Oscillation. Mon Weather Rev 115:1606–1626
Saha S et al (2006) The NCEP climate forecast system. J Clim 19:3483–3517
Saha S et al (2010) The NCEP Climate Forecast System Reanalysis. Bull Am Meteorol Soc 91:1015–1057. doi:10.1175/2010BAMS3001.1
Saha S et al (2013) The NCEP Climate Forecast System Version 2. J Clim 26 (unpublished)
Tippett MK, Barnston AG, Li S (2012) Performance of recent multimodel ENSO forecasts. J Appl Meteorol Climatol 51:637–654
Van den Dool HM, Toth Z (1991) Why do forecasts for near normal often fail? Weather Forecast 6:76–85
Wang W, Xie P, Yo SH, Xue Y, Kumar A, Wu X (2011) An assessment of the surface climate in the NCEP Climate Forecast System Reanalysis. Clim Dyn 37:1601–1620. doi:10.1007/s00382-010-0935-7
Weaver SJ, Wang WQ, Chen MY, Kumar A (2011) Representation of MJO variability in the NCEP climate forecast system. J Clim 24:4676–4694
Wilks DS (2006) Statistical methods in the atmospheric sciences, 2nd edn. Academic Press, Oxford, p 648
Xue Y, Huang B, Hu Z-Z, Kumar A, Wen C, Behringer D, Nadiga S (2011) An assessment of oceanic variability in the NCEP Climate Forecast System Reanalysis. Clim Dyn 37:2511–2539. doi:10.1007/s00382-010-0954-4
Xue Y, Chen M, Kumar A, Hu Z-Z, Wang W (2013) Prediction skill and bias of tropical Pacific Sea surface temperatures in the NCEP Climate Forecast System version 2. J Clim 26. http://dx.doi.org/10.1175/JCLI-D-12-00600.1
Zhang L, Kumar AK, Wang W (2012) Influence of changes in observations on precipitation: a case study for the Climate Forecast System Reanalysis (CFSR). J Geophys Res Atmos 117:D08105. doi:10.1029/2011JD017347