Science and Technology Infusion Climate Bulletin
NOAA's National Weather Service
37th NOAA Annual Climate Diagnostics and Prediction Workshop
Fort Collins, CO, 22-25 October 2012

A Comparison of Skill of CFSv1 and CFSv2 Hindcasts of Nino3.4 SST

Anthony G. Barnston[1] and Michael K. Tippett[1, 2]

[1]*International Research Institute for Climate and Society,*
The Earth Institute at Columbia University, Lamont Campus, Palisades, NY 10964

[2]*Center of Excellence for Climate Change Research, Dept. of Meteorology,*
King Abdulaziz University, Jeddah, Saudi Arabia

1. Background

The first version of the NOAA/NCEP Climate Forecast System coupled model (CFSv1; Saha *et al.* 2006) was used operationally between 2004 and 2011. In 2011 it was supplanted by the second version, CFSv2 (Saha *et al.* 2013). Some basic characteristics of the two model versions are shown in Table 1. The CFSv2 carries several major improvements. Besides changes in the model dynamics and increases in forecast resolution and ensemble size, the CO_2 concentration in CFSv2 evolves realistically over time, while for CFSv1 the CO_2 value is fixed at the observed 1988 concentration. Another difference 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). It is stated in Saha *et al.* (2010) that the atmospheric analysis, and therefore the initial conditions, based on the CFSR is more realistic than for the R-2.

	CFSv1	**CFSv2**
Horizontal & Vertical Resolution	T62 (~2°), 64 levels	T126 (~1°), 64 levels
Atmospheric Model	GFS from 2003	GFS from 2009
Ocean Model	MOM3	MOM4
No. Ensemble Members / Month	15	24
Source of Initial Condition Data	NCEP/DOE Reanalysis	Climate Forecast Sys. Reanalysis (CFSR)
Sea Ice	Climatology	Predicted
Carbon Dioxide Concentration Setting	Fixed at 1988 level	Evolving with time

Table 1 Some basic specifications for CFSv1 and CFSv2

Given the improvements in CFSv2 compared with CFSv1, one would expect relatively better predictive skill in CFSv2. 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 hindcast integrations beginning that year—especially those in the tropical Pacific (Xue *et al.* 2011; Kumar *et al.* 2012). Here we compare the skill of predictions of Nino 3.4 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, or to the 1999 discontinuity in the initial conditions due to the CFSR.

2. Results

Here the skill results include verification measures for deterministic predictions, including trend analysis and forecast timing error analysis, and also reliability analysis for the probabilistic aspect of the predictions.

a. Anomaly correlation and RMSE

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

Correspondence to: Anthony G. Barnston, International Research Institute for Climate and Society, The Earth Institute at Columbia University, Lamont Campus, Palisades, NY 10964; E-mail: tonyb@iri.columbia.edu.

difference is found in forecasts for northern summer at medium and long lead times, where CFSv1 has relatively low skill (correlations of 0.5 or lower) while CFSv2 shows higher skill (0.6 to 0.7).

These forecasts are for target months beyond the northern spring ENSO predictability barrier that are made before that barrier—the condition known to present greatest predictive difficulty. However, another 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 beginning of the episode (*e.g.*, a forecast for February made in July). These "easier" predictions appear to be made better by CFSv1 than CFSv2. Why would this be the case for a model that outperforms its predecessor in the most difficult prediction conditions?

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.

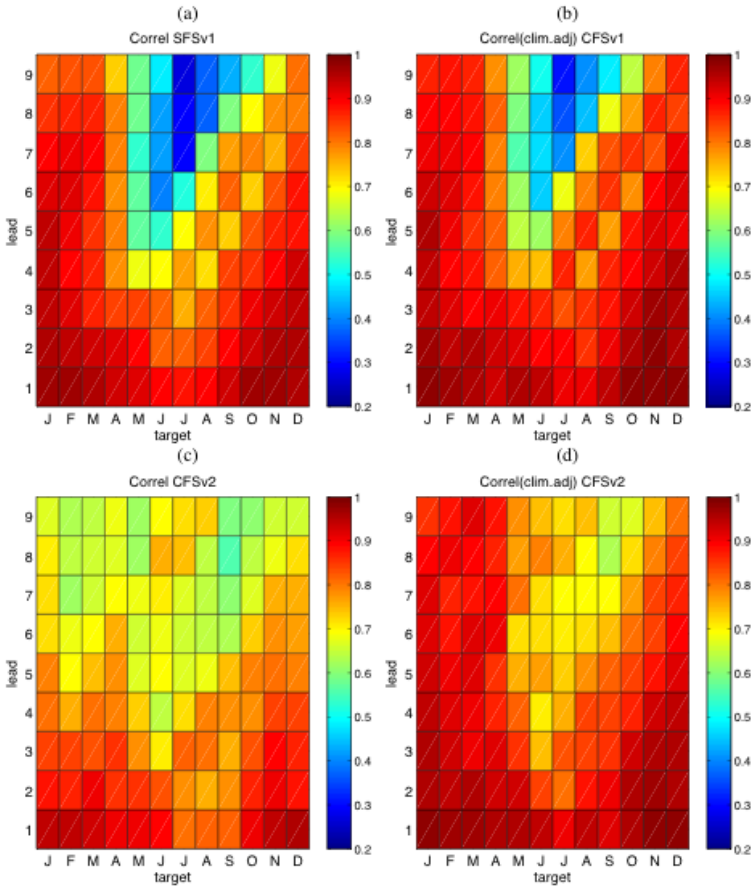

Fig. 1 Temporal correlation between (a) CFSv1 and (c) CFSv2 predictions of Nino3.4 SST and verifying observations over the 1982-2009 period. The target month is indicated on the horizontal axis, and lead time on the vertical axis. A lead time of 1 month implies a prediction made at the very beginning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

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 further here. The CFSv2 error discontinuity, on the other hand, is associated with 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 (Xue *et al.* 2011; Kumar *et al.* 2012). This SST change has been attributed to the introduction of the ATOVS[1] 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.* 2012; Deser *et al.* 2010; Kumar *et al.* 2010; Lyon and DeWitt 2012), and is therefore seen as artificial. Such a positive change in tropical Pacific SST behavior around 1999 would be important because the SST in that region, besides reflecting the ENSO state in its own right, would affect remote teleconnections to seasonal climate. A change in the climatology of tropical Pacific reanalyzed 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 in

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

several studies, noted most strongly in the general vicinity of the tropical Pacific (Wang *et al.* 2011; Chelliah *et al.* 2011; Ebisuzaki *et al.* 2011). It will be shown below that the signature of the 1999 discontinuity in the predictions of Nino3.4 SST appears in the shortest lead time, propagates to longer lead times, and exhibits some seasonal dependence.

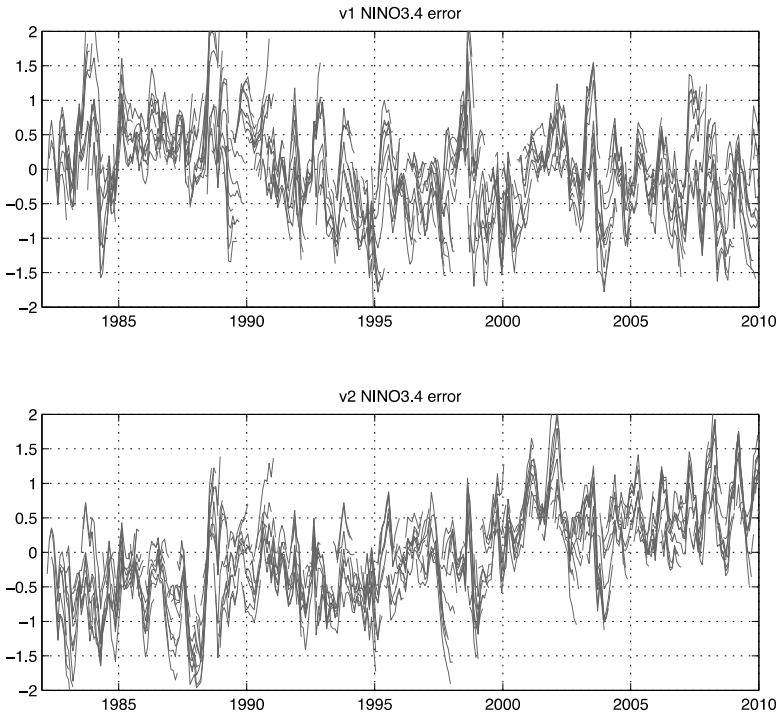

Fig. 2 Error (°C) in Nino3.4 SST predictions of CFSv1 (top) and CFSv2 (bottom) for start times (indicated on horizontal axis) over the course of the 1982-2009 period. Errors for predictions at all lead times are shown.

To free the evaluation of 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 adjustment (or correction) 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 than CFSv1. In CFSv2, higher correlations are seen in all seasons and leads, but most notably for predictions for late northern autumn and winter made during summer or later — forecasts considered least challenging but relatively lacking in skill compared with CFSv1 before the correction. A summary of the correlation differences between CFSv2 and CFSv1 before and after the discontinuity corrections for both models is 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 clear with or without the correction, but with the correction CFSv2 no longer presents a degradation for moderate and long lead predictions for northern winter made from earlier within the same ENSO cycle. It may be noted, however, that CFSv1 performed about as well for these predictions as CFSv2. Following the correction, then, the better performance of CFSv2 applies to most seasons and leads.

A similar skill comparison is conducted for RMSE using standardized anomalies[2], with results shown in Fig. 4. The results for RMSE differ noticeably in pattern to those of correlation because biases in both mean and in amplitude contribute to RMSE but not to correlation.

RMSE scores are reduced considerably with the dual climatology correction for both model versions, indicating the importance of the sub-period biases that can greatly exacerbate the squares of the largest errors in the direction of the bias. Comparing the RMSE for the corrected versions of the two model versions, it is seen again that the main difference is a substantial improvement in CFSv2 in the errors of predictions traversing the northern spring predictability barrier, particularly for late northern summer target months made early in the calendar year. Such predictions are for ENSO conditions generally not yet observed at the time of the forecast.

[2] 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.

b. Standard Deviation Ratio

Figure 5 shows the ratio of the interannual standard deviation of the model 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 standard deviation ratio would be no higher than unity throughout all seasons and leads, and lower to the extent that predictive skill is imperfect: Theoretically, it should be the square root of the fraction of observed variance explained by the predictions. While the correction results mainly in subtle changes in the ratios, 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 at its seasonal minimum. CFSv2 lacks this weakness and, following the bias correction, shows ratios fairly close to unity for many seasons and leads. In keeping with the expected lower skill expected for forecasts traversing the northern spring predictability barrier, ratios of less than unity are noted in CFSv2 for predictions for June to October made at medium and long leads.

c. Target month slippage

"Target month slippage" is a systematic error that occurs when predictions verify with higher skill for target months earlier or later than those intended (Tippet et al. 2012; Barnston et al. 2012), such as a 4-month lead prediction intended for July verifying better against observations of May or June. Typically slippage occurs with predictions late in reproducing observed changes, such as onsets or endings of ENSO episodes. Slippage cannot be

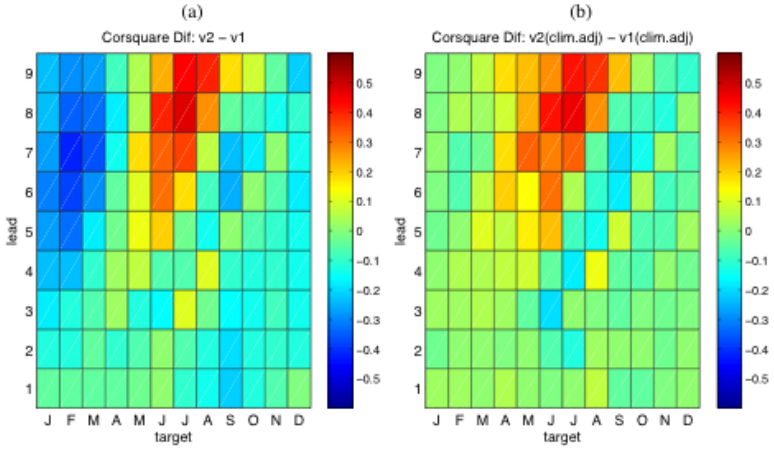

Fig. 3 Difference in squared correlation (of predictions vs. observations) of CFSv2 and CFSv1 without treatment for discontinuities and following treatment using dual climatologies for each model version (a and b, respectively). Negative sign is retained upon squaring. The target months and lead times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

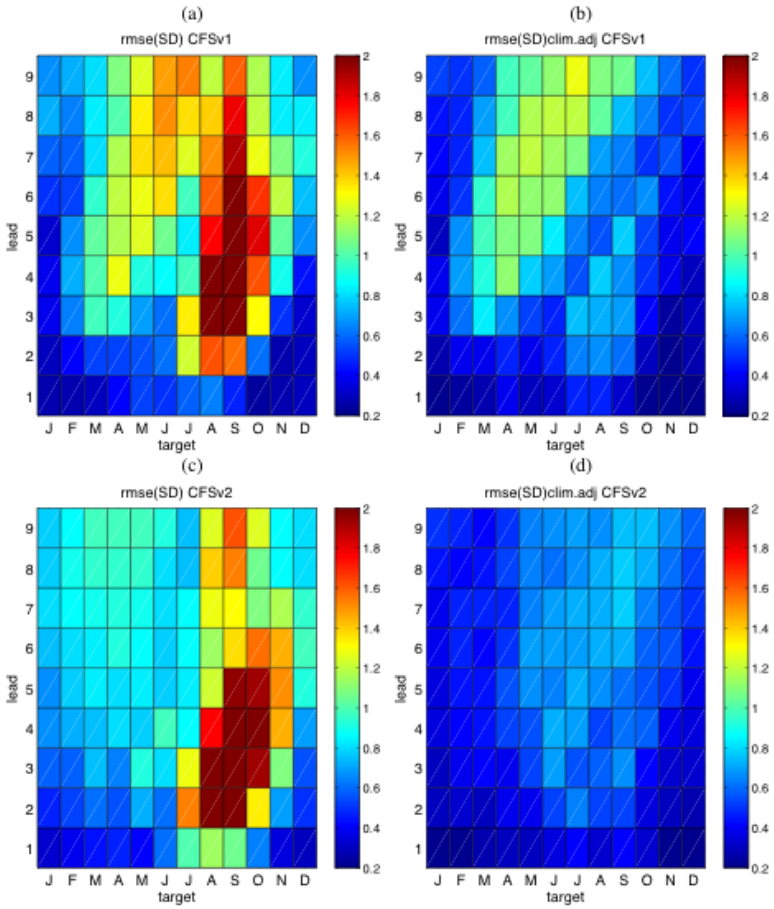

Fig. 4 Root mean squared error of predicted vs. observed standardized anomalies of (a) CFSv1 and (c) CFSv2 without treatment for discontinuities and following treatment using dual climatologies for each model version (b and d, respectively). In the absence of any skill, RMSE of 1.41 is expected. The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

diagnosed from the usual skill measures, which only compare forecasts with the verifying observations of the intended target time. Although slippage is a systematic temporal 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. Because CFSv1 is seen to underperform CFSv2 in such predictions crossing the northern spring predictability barrier, greater slippage might be expected in CFSv1 than CFSv2.

Slippage is shown in plots of skill as a function of the lag time between the measured target period and the intended one. To overcome the small sample issue, the diagnosis is made for all seasons together. To the extent that slippage is systematic, it can be corrected using statistical methods, such as multiple regression, that define optimum shifts of 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 an MSE-based skill metric. Figures 6 and 7 show slippage and skill results for CFSv1 and CFSv2, respectively, before and after the correction.

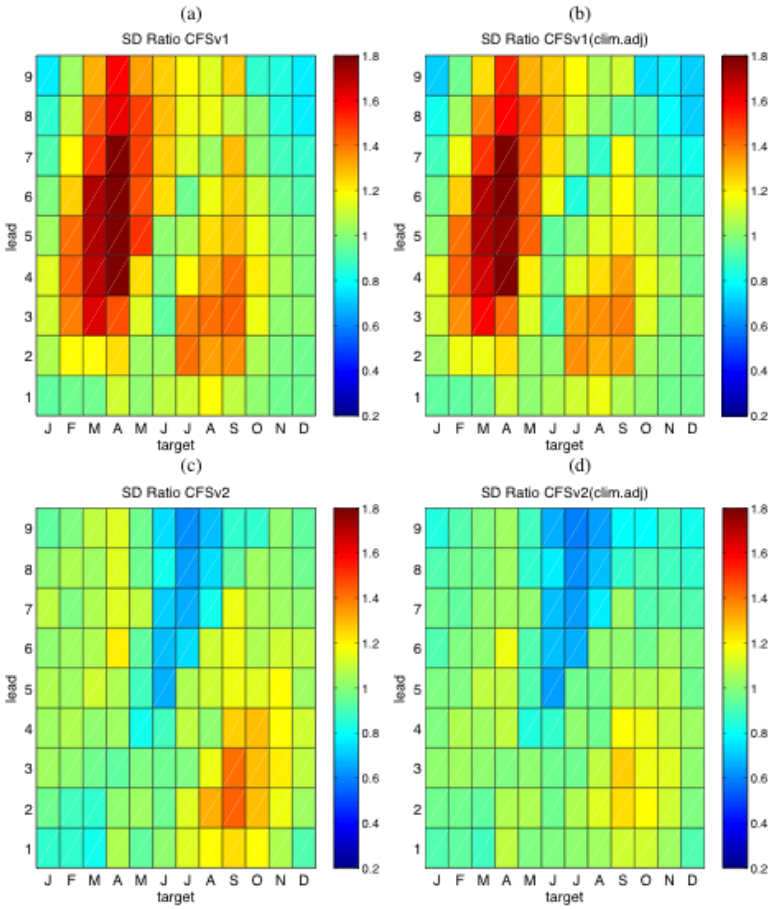

Fig. 5 Ratio of interannual standard deviation of predicted vs. observed anomalies of CFSv1 (a) and CFSv2 (c) without treatment for discontinuities and following treatment using dual climatologies for each model version (b and d, respectively). Ideally, the ratio is unity or less. The target months and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

Slippage is obvious in CFSv1 (top left panel of Fig. 6), and it increases with increasing lead times. The MSE-based skill score (bottom left panel) indicates sub-zero skill for long-lead CFSv1 forecasts for northern summer. After the statistical 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 correction does little to improve the already good performance.

d. Trend Bias

The time-conditional biases indicated in the CFSv1 and CFSv2 predictions discussed earlier (Fig. 2) create trend biases in the sense that a linear trend fit to the predictions exhibit slopes that do not appear in such a fit to the observations. Each model also exhibits more gradual trends within each of its sub-periods, particularly for start months around northern 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 the earlier discussion, 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 of positive biases to increase with time from 1999 forward.

At the earliest lead time, predictions are expected to be influenced heavily by the initial conditions. The systematic discrepancies between the short-lead predictions and the observations shown in Fig. 8 are thus indicative of biases in the SST initial conditions, and in this case 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 at each lead time. The CFSv2 positive trend biases for the shortest lead predictions of August, September and October are noted in the bottom row of cells. Figure 9 (right) shows that these northern autumn biases amplify as they propagate to predictions at later target months with increasing lead times.

The initial condition bias is thus seen to be 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. This relatively simple correction is sufficient to uncover evidence of the substantial general improvement in predictive skill of CFSv2 compared to CFSv1.

A reason for a remaining gradual positive trend in CFSv2 predictions relative to observations even after the discontinuity correction using dual climatologies is not obvious, but may reflect a problem of radiation balance in the model. This possibility may be an issue for consideration in the development of the future version of the CFS.

The trend bias in CFSv1 is negative for virtually all months and leads, mainly because of the discontinuity in 1991 but also to some degree because of a gradual trend 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

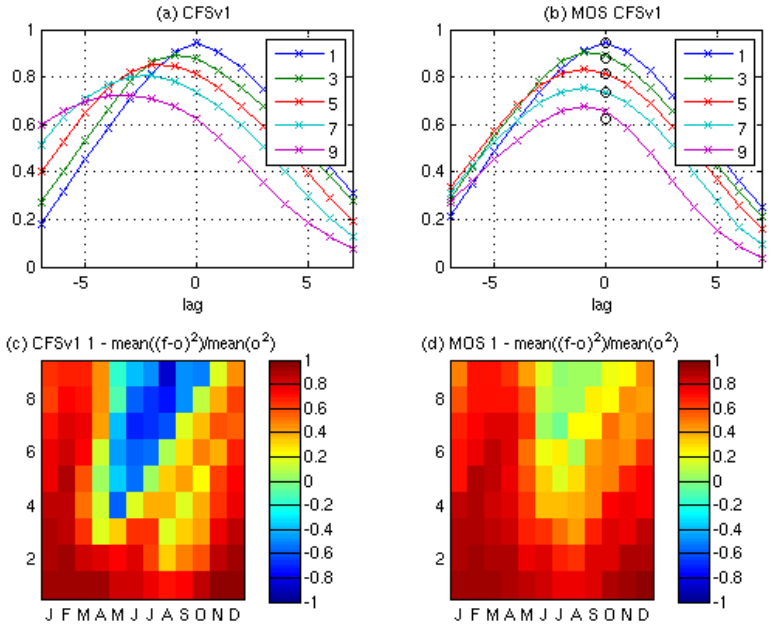

Fig. 6 Target period slippage, and its correction, in CFSv1: (top) Correlation between predictions and observations as a function of lag time between verified target month and intended target month, for leads of 1, 3, 5, 7 and 9 months before (left) and after (right) a MOS correction for slippage based 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 (MSE) skill score as a function of target month and lead time before (left) and after (right) the MOS correction.

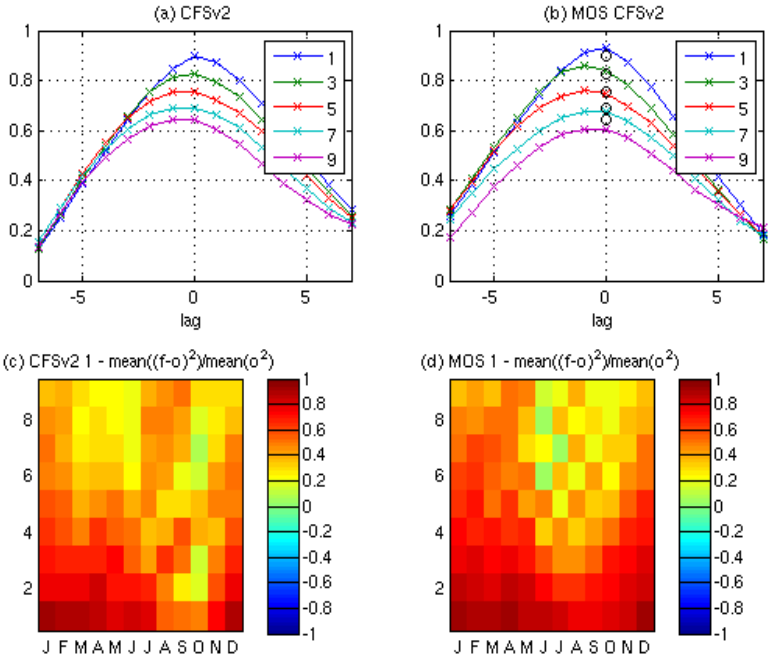

Fig. 7 As in Fig. 6, except for CFSv2 slippage and its correction.

a non-evolving CO_2 concentration setting, and this is one possible reason for the slowly declining Nino3.4 SST predictions relative to the observed SST.

e. Probabilistic reliability

We assess the reliability and sharpness of the probabilistic predictions of Nino3.4 SST from the two CFS versions using reliability analysis. 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, where the categories are defined using tercile cutoffs for the study period. The observations are categorized likewise. The three categories may be loosely representative of La Nina, neutral and El Nino conditions. Reliability analysis is carried out for the above and below normal forecast categories separately. We ignore the near-normal category, which has repeatedly been demonstrated to have weak performance.

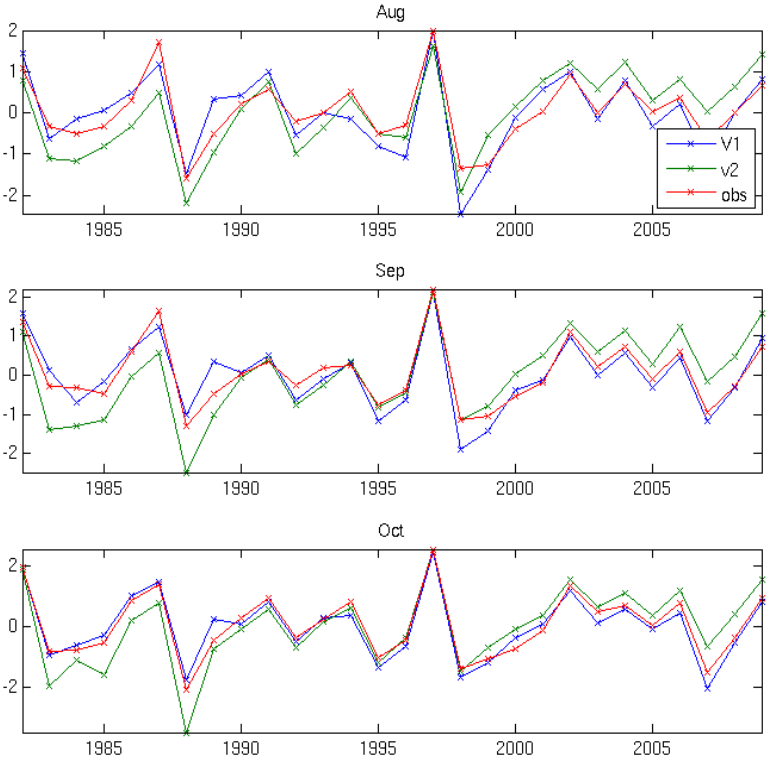

Fig. 8 Shortest-lead Nino3.4 SST anomaly predictions of CFSv1 (blue) and CFSv2 (green) and corresponding observations (red) for start times at beginning of August (top), September (middle) and October (bottom) over the 1982-2009 period.

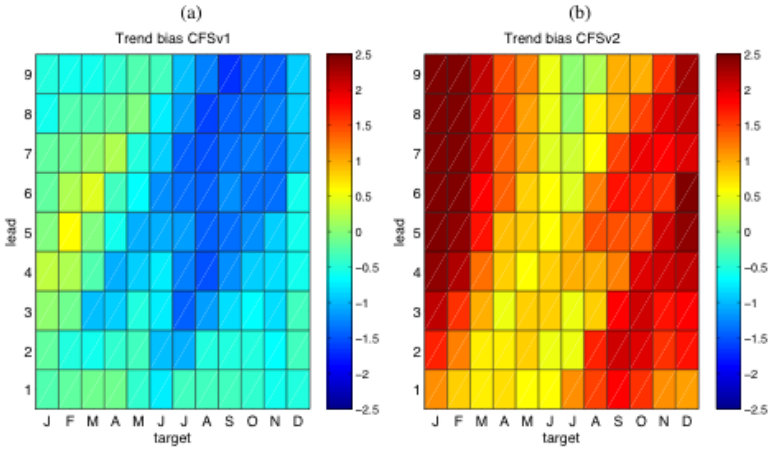

Fig. 9 Bias, relative to observations, in the slope of the linear trend fit over the 1982-2009 (°C per 28 yr) period for Nino3.4 predictions of (a) CFSv1 and (b) CFSv2 as a function of target month and lead time.

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. Perfect reliability would be achieved, for example, if for the 20 instances when the above normal Nino3.4 SST category is assigned a probability of 40%, the corresponding later observed anomalies were above normal category in 8 (40%) cases. Here we examine just the 6-month lead predictions, and combine all target months. We form eleven 10%-wide forecast probability bins. Then there are (28×12) = 336 predictions, resulting in an expected average of about 31 predictions per probability bin.

The reliability diagrams for the below and above normal categories are shown for the two CFS model versions, with uncorrected climatologies, in Fig. 10 as the red and green curves, respectively. For each category, forecasts are binned for forecast probability spanning from lowest to highest (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 plots insets below the main panel show the percentage of forecasts having probabilities in each bin.

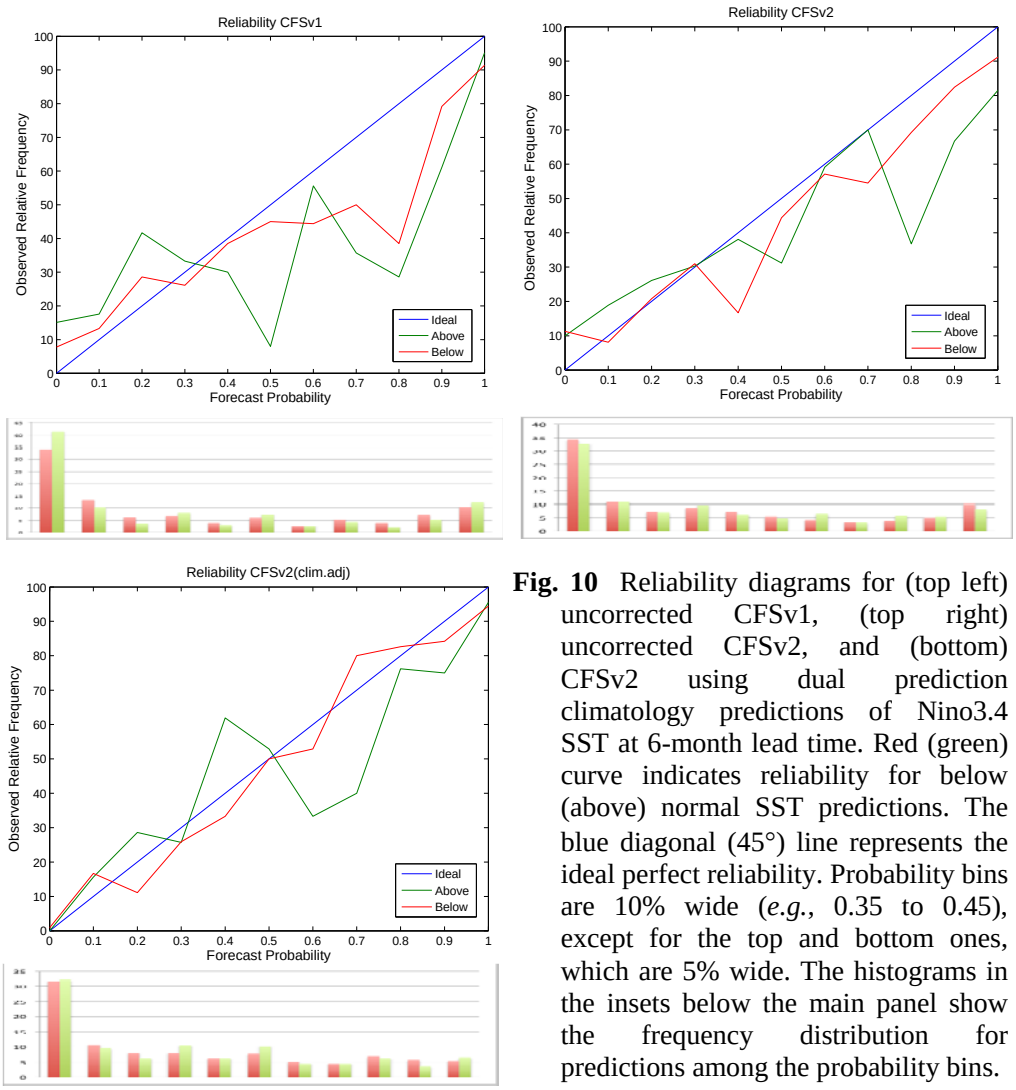

Fig. 10 Reliability diagrams for (top left) uncorrected CFSv1, (top right) uncorrected CFSv2, and (bottom) CFSv2 using dual prediction climatology predictions of Nino3.4 SST at 6-month lead time. Red (green) curve indicates reliability for below (above) normal SST predictions. The blue diagonal (45°) line represents the ideal perfect reliability. Probability bins are 10% wide (*e.g.*, 0.35 to 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.

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 related to the somewhat small sample sizes per bin. However, the average slope of both curves is seen to be somewhat less than unity. Thus, forecasts with very low (high) probabilities do not result in comparably low (high) frequencies of occurrence — *i.e.* the forecasts exhibit overconfidence, particularly 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 to 0.05) is by far the most frequently issued probability, followed by the highest bin (0.95 to 1.00) and the second lowest bin (0.05 to 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. 10, upper right), while roughy similar to that of CFSv1, shows milder overconfidence: the curves have slope closer to (but still less than) unity, with smaller deviations below the ideal reliability (45°) line for bins for 0.50 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 only 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 and 3) and its slightly less sharp, more conservative probabilities that better reflect the true level of uncertainty in the model's reproduction of the ocean-atmosphere system. This outcome is consistent with the greater inflation above unity of the standard deviation ratio of CFSv1 than CFSv2 noted above, especially at medium to long lead times (left panels of Fig. 5). Elimination of the discontinuities in the climatology of the predictions slightly helps to remedy the inflated standard deviation ratio of CFSv2 (lower right panel of Fig. 5), and a similar improvement would be expected

in the reliability analysis. To confirm this expectation, the CFSv2 analysis is applied using dual climatologies for the tercile boundary definitions for the model prediction category. Results (Fig. 10c) indicate an overall slope closer to unity than when using a single prediction climatology; and the observed relative frequencies associated with forecasts of zero probability are less than 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. Likewise, forecasts with 100% probability are met with correctly verifying observations in about 95% of cases for the dual climatologies, rather than only about 80% (90%) for the above (below) normal category without the climatology adjustment. All told, the adjustment results in an improvement in probabilitistic reliability for CFSv2—most noticeably for forecasts deviating most sharply from climatology. That the extreme probability forecasts are most able to be improved in reliability makes sense in view of the expected effect of an artificial mean shift in the climatological on forecasts probabilities that heavily define the reliability curve, both because they are issued frequently (in this example) and because they form the end points of the curve.

3. Conclusion

Given the large amount of time and resources used to achieve an improved CFSv2 compared with the earlier CFSv1, one would expect relatively better predictive skill in CFSv2. Here we examine the skill difference between CFSv1 to CFSv2 in predictions of the ENSO state, as represented by Nino3.4 SST anomaly.

CFSv2 is better able to predict the ENSO state than CFSv1 through the northern spring predictability barrier, the time of year when the need for better predictions is greatest. By contrast, on initial examination CFSv2 appears to fall short of CFSv1 in ENSO prediction skill for northern summer and autumn start times — times for which ENSO prediction is known to be least challenging and skill is highest. However, CFSv2 is found to be affected by a significant discontinuity in initial condition climatology near 1999 associated with a corresponding discontinuity in the high resolution Reanalysis observations generated using CFSv2 (the CFSR). The size and impact of this discontinuity turns out to be most prominent in the tropical Pacific region (Xue *et al.* 2011; Kumar *et al.* 2012). Here, focusing on the skill for Nino3.4 SST anomaly, we highlight differences in skill diagnostics that may be related to model improvement, or on the other hand caused by the discontinuity.

The initial condition discontinuity masks CFSv2's net predictive skill and its general superiority over CFSv1 in prediction Niño3.4 SST. This impediment is most noticeable for northern autumn start times when skill is highest, when CFSv1 already achieves a high skill level that is difficult to exceed. The skill impact of the discontinuity is evaluated by examining skill with versus without the benefit of correction of the discontinuity by defining two separate climatologies from which to form anomalies. After correcting for the 1999 discontinuity, performance of CFSv2 is found to equal or exceed that of CFSv1 more generally at nearly all times of the year in terms of anomaly correlation, RMSE, and interannual standard deviation ratio with respect to the observations. CFSv2 also exhibits better probabilistic reliability than CFSv1, mainly because of its lesser degree of probabilistic overconfidence, and the climatology correction still further increases this margin of superiority. Finally, CFSv2 largely lacks "target month slippage" compared with CFSv1— *i.e.*, it does not tend to verify better on target times earlier than those intended due to being slow to reproduce major transitions in the ENSO state.

Comparing verifications before and after the climatology correction, the measures seen to be most noticeably adversely affected by the uncorrected 1999 change are first the RMSE, and secondly the temporal anomaly correlation. The standard deviation ratio and probabilistic reliability analyses are noticeably, but less dramatically, affected. When one realizes that the problem is one of a changing calibration, it is easy to expect *all* verification measures to be degraded without a correction. A constant miscalibration is easily corrected, and the lack of a correction would not degrade measures such as the anomaly correlation or the slope of the reliability curves. However, a changing miscalibration becomes equivalent to a nonsystematic error unless the time series is examined by eye (*e.g.*, Fig. 2) and the problem identified and treated with a combination of human intervention and machine automation (*i.e.*, choosing the appropriate correction procedure).

CFSv2 is shown to have a larger upward trend in Nino3.4 SST than found in the observations, 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. This exaggerated positive trend may be related to a problem in the radiation budget, and indicates a potential area of improvement for the next improved version of CFS.

Although the discontinuity has clearly discernible effects on predictions of ENSO-related SST by CFSv2, they are not so large as to materially degrade the model's predictions of climate across much of the globe. In fact, performance in climate predictions has been found significantly better than that of CFSv1, including for example in the United States during winter when ENSO is a major governing factor (Peng *et al.* 2013) and reproduction of the MJO (Weaver *et al.* 2011). The skill of CFSv2 is found competitive with that of ECMWF system 4 for winter climate predictions over North America, despite relative shortcomings in predictions of ENSO and the globally averaged tropical climate (Kim *et al.* 2012).

References

Barcikowska, M., F. Feser, and H. von Storch, 2012: Usability of best track data in climate statistics in the western north pacific. *Mon. Wea. Rev.*, **140**, 2818-2830.

Barnston, A. G., M. K. Tippett, M. L. L'Heureux, S. Li, and D. G. DeWitt, 2012: Skill of real-time seasonal ENSO model predictions during 2002-11: Is our capability increasing? *Bull. Amer. Meteor. Soc.*, **93**, 631-651.

Behringer, D.W., and Y. Xue, 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., W. Ebisuzaki, S. Weaver, and A. Kumar, 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.

Deser, C., A. S. Philips, and M. A. Alexander, 2010: Twentieth century tropical sea surface temperature trends revisited. *Geophys. Res. Lett.*, **37**, doi:10.1029/2010GL043321.

Ebisuzaki, W., and L. Zhang, 2011: Assessing the performance of the CFSR by an ensemble of analyses. *Clim. Dyn*., **37**, 2541-2550.

Kim, H. M., P. J. Webster and J. A. Curry, 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., J. Bhaskar, and M. L'heureux, 2010: Are tropical SST trends changing the global teleconnection during La Nina? *Geophys. Res. Lett.*, **37**, L12702, doi:10.1029/2010GL043394.

—, M. Chen, L. Zhang, W. Wang, Y. Xue, C. Wen, L. Marx, and B. Huang, 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. Wea. Rev*., **140**, 3003-3016.

Lyon, B., and D. G. DeWitt, 2012: A recent and abrupt decline in the East African long rains. *Geophys. Res. Lett*., **39**, L02702, doi: 10.1029/2011GL050337.

Peng, P., A. G. Barnston, and A. Kumar, 2013: A Comparison of Skill between Two Versions of the NCEP Climate Forecast System (CFS) and CPC's Operational Short-Lead Seasonal Outlooks. *Weather and Forecasting*, **27**, in press.

Saha, S., and Coauthors, 2006: The NCEP Climate Forecast System. *J. Climate*, **19**, 3483–3517.

—, and Coauthors, 2010: The NCEP Climate Forecast System Reanalysis. *Bull. Amer. Meteor. Soc.*, **91,** 1015-1057. doi: 10.1175/2010BAMS3001.1

—, and Coauthors, 2013: The NCEP Climate Forecast System Version 2. *J. Climate*, **19**, submitted.

Tippett, M. K., A. G. Barnston, and S. Li, 2012: Performance of recent multimodel ENSO forecasts. *J. Appl. Meteor. Climatol.*, **9**, 637–654.

Wang, W., P. Xie, S. H. Yo, Y. Xue, A. Kumar, and X. Wu, 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, S. J., W. Q. Wang, M. Y. Chen, and A. Kumar, 2011: Representation of MJO variability in the NCEP Climate Forecast System. *J. Climate*, **24**, 4676-4694.

Xue, Y., B. Huang, Z.-Z. Hu, A. Kumar, C. Wen, D. Behringer, and S. Nadiga, 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.

Zhang, L., A. K Kumar, and W. Wang, 2012: Influence of changes in observations on reanalysis products: A Case Study for the CFSR. *J. Geophys. Res. - Atmosphere.* Conditionally accepted.