Clim Dyn (2013) 41:2785–2795
DOI 10.1007/s00382-013-1965-8

Improved reliability of ENSO hindcasts with multi-ocean analyses ensemble initialization

Jieshun Zhu · Bohua Huang · Magdalena A. Balmaseda · James L. Kinter III · Peitao Peng · Zeng-Zhen Hu · Lawrence Marx

Received: 13 March 2013 / Accepted: 4 October 2013 / Published online: 16 October 2013

Abstract Currently, ensemble seasonal forecasts using a single model with multiple perturbed initial conditions generally suffer from an "overconfidence" problem, i.e., the ensemble evolves such that the spread among members is small, compared to the magnitude of the mean error. This has motivated the use of a multi-model ensemble (MME), a technique that aims at sampling the structural uncertainty in the forecasting system. Here we investigate how the structural uncertainty in the ocean initial conditions impacts the reliability in seasonal forecasts, by using a new ensemble generation method to be referred to as the multiple-ocean analysis ensemble (MAE) initialization. In the MAE method, multiple ocean analyses are used to build an ensemble of ocean initial states, thus sampling structural uncertainties in oceanic initial conditions (OIC) originating from errors in the ocean model, the forcing flux, and the measurements, especially in areas and times of insufficient observations, as well as from the dependence on data assimilation methods. The merit of MAE initialization is demonstrated by the improved El Niño and the Southern Oscillation (ENSO) forecasting reliability. In particular, compared with the atmospheric perturbation or lagged ensemble approaches, the MAE initialization more effectively enhances ensemble dispersion in ENSO forecasting. A quantitative probabilistic measure of reliability also indicates that the MAE method performs better in forecasting all three (warm, neutral and cold) categories of ENSO events. In addition to improving seasonal forecasts, the MAE strategy may be used to identify the characteristics of the current structural uncertainty and as guidance for improving the observational network and assimilation strategy. Moreover, although the MAE method is not expected to totally correct the overconfidence of seasonal forecasts, our results demonstrate that OIC uncertainty is one of the major sources of forecast overconfidence, and suggest that the MAE is an essential component of an MME system.

Keywords ENSO prediction · Reliability · Multi-ocean analyses ensemble · CFSv2

1 Introduction

Even though the ability of dynamical models to predict El Niño and the Southern Oscillation (ENSO) has improved significantly over the past few decades (e.g., Latif et al. 1998; Jin et al. 2008; Wang et al. 2010; and references

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.

J. Zhu (✉) · B. Huang · J. L. Kinter III · L. Marx
Center for Ocean-Land-Atmosphere Studies, 270 Research Hall,
Mail Stop 6C5, George Mason University, 4400 University
Drive, Fairfax, VA 22030, USA
e-mail: jieshun@cola.iges.org

B. Huang · J. L. Kinter III
Department of Atmospheric, Oceanic, and Earth Sciences,
College of Science, George Mason University,
Fairfax, VA, USA

M. A. Balmaseda
European Centre for Medium-Range Weather Forecasts,
Reading, UK

P. Peng · Z.-Z. Hu
Climate Prediction Center, National Centers for Environmental
Prediction/NOAA, College Park, MD, USA

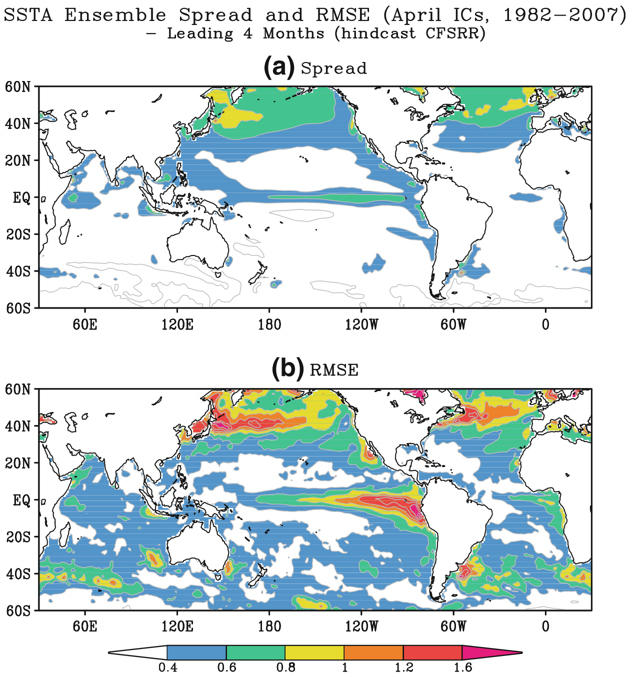
SSTA Ensemble Spread and RMSE (April ICs, 1982–2007)
– Leading 4 Months (hindcast CFSRR)
(a) Spread
(b) RMSE
60N
40N
20N
EQ
20S
40S
60S
60E
120E
180
120W
60W
0
0.4
0.6
0.8
1
1.2
1.6

Fig. 1 **a** Ensemble spread and **b** RMSE of SSTA at the 4-month lead from hindcast CFSRR with IC in April of 1982–2007. Contour interval is 0.2°C, with above 0.4°C *colored shading*

therein), ENSO prediction is still far from perfect, using both deterministic and probabilistic metrics. An example of the latter is that current ensemble seasonal forecasts at operational centers are generally found to have an apparent "overconfidence" problem, i.e., the ensemble perturbations have limited growth relative to the amplitude of mean error (Palmer et al. 2004; Vialard et al. 2005; Saha et al. 2006; Weisheimer et al. 2009; also seen in Fig. 1). As a result, events (e.g., warm ENSO events) occur more frequently in the ensemble forecasts than the fraction of times such events are observed (e.g., Weigel et al. 2009; Langford and Hendon 2013). This means that the low ensemble spread underestimates the forecast uncertainty and makes it less reliable. The lack of reliability may seriously affect subsequent applications of the forecasts.

Therefore, an important measure of ensemble forecasting is whether the resulting probabilities are reliable, i.e., the forecast probabilities match the observed frequencies (Johnson and Bowler 2009). Efforts have been made to enhance the reliability of seasonal forecasts in different ways. In general, the lack of reliability originates from an inadequate sampling of the uncertainty associated with the errors inherent in current forecast system. Seasonal sea surface temperature anomaly (SSTA) forecasts are primarily subject to three types of errors: (I) amplification of errors in ocean initial conditions (OIC), (II) errors due to the unpredictable nature of the synoptic atmospheric variability, and (III) coupled model errors. A good ensemble forecast system is necessary to sample the effect of all these error sources. Uncertainties of type (II) (i.e., atmospheric perturbations, AP) have been considered in current seasonal forecast systems, usually by using multiple atmospheric initial conditions. Type (III) errors can be sampled by employing the so-called stochastic physics (Vialard et al. 2005), or by adopting a multi-model ensemble approach (MME, see Palmer et al. 2004; Weisheimer et al. 2009; Kirtman and Min 2009). In addition, the reduction of the systematic errors by some empirical corrections may also improve ENSO predictions (Manganello and Huang 2009; Pan et al. 2011; Magnusson et al. 2012).

Comparatively less attention has been paid to type (I) errors in current initialization strategies, even though adequately sampling the OIC uncertainty is vital for ensemble seasonal forecasting (the MME approach implicitly includes the sampling of different ocean initial conditions, but this aspect is usually not highlighted in the literature and, as far as we know, its effectiveness has yet to be shown). Generally, single-model-based operational seasonal forecast systems use relatively simple procedures to produce perturbations in OIC. For instance, the Climate Forecast System, version 2 (CFSv2), of the National Centers for Environmental Predictions (NCEP) applies the traditional lagged ensemble (LE) approach to generate ensemble members in both atmospheric and oceanic initial states (Saha et al. 2013). In this approach, the ensemble is built by aggregating predictions from a succession of neighboring initial states. In contrast, operational climate predictions at the European Center for Medium-range Weather Forecasts (ECMWF) are initialized with five perturbed ocean states generated by random perturbations inherent in its ocean data assimilation analysis (Molteni et al. 2011; Balmaseda et al 2013). Neither method samples the structural uncertainty associated with the data assimilation system used in the production of the ocean analyses. More sophisticated techniques have been adopted from the successful practice in numerical weather prediction, such as the singular vector (Palmer et al. 1994) and breeding (Toth and Kalnay 1997) methods, in limited experimental cases (Yang et al. 2008) or with simplified forecast systems (e.g., the empirical singular vector by Kug et al. 2011). However, possibly due to their intrinsic limitations (Kug et al. 2010) for seasonal predictions and because the complexity of CGCMs, such methods have not been implemented in any CGCM-based operational ensemble seasonal forecast systems.

The adequacy of the above ensemble generation strategies in accounting for the OIC uncertainty has not been fully tested. In comparison with an atmospheric initial condition, the uncertainty of an OIC may be higher and more dependent upon geographical locations because the observational measurements are much fewer in number and are more tightly clustered geographically in the ocean. This source of uncertainty cannot be taken into account in the initialization approaches discussed above, because they are all based on their respective ocean analysis systems with their individual ocean model and assimilation techniques. The ocean analysis systems are generally based on different ocean models forced by different atmospheric fluxes, and apply different assimilation techniques to assimilate slightly different ocean datasets (see Table 1 in the "Appendix" as an example). These differences have resulted in substantial uncertainties in the estimated ocean states (Zhu et al. 2012a; Xue et al. 2012; also see Figs. 5, 6 in "Appendix").

In a recent study, Zhu et al. (2012b) found that there was a substantial difference in the ENSO prediction skill with different ocean analyses. Moreover, the grand ensemble mean of the predictions initialized from all available ocean analyses, referred to as the multiple-ocean analysis ensemble (MAE) initialization, gives prediction skill (or accuracy) at least as good as the best set of forecasts derived from an individual ocean analysis (Zhu et al. 2012b, 2013). It is known that deterministic measures of skill cannot provide the information about prediction uncertainties. In this study, we further explore the potential of utilizing multiple analysis initialization for probabilistic forecasting. Specifically, we examined the effect of the MAE initialization on ENSO forecasting reliability by analyzing groups of hindcasts generated using CFSv2. In addition to comparing with hindcasts considering AP only, we also compare MAE with the LE approach (i.e., hindcasts from NCEP CFS Reanalysis and Reforecast (CFSRR) Project using CFSv2). The paper is organized as follows. The CGCM, the experimental design and datasets are described in the next section. The results are presented in Sect. 3. A summary and discussion are given in Sect. 4.

2 Model, hindcast experiments and datasets

The coupled model used in this study is the NCEP CFSv2 (Saha et al. 2013). CFSv2 has been the operational forecast system for seasonal-to-interannual prediction at NCEP since March 2011, replacing its predecessor, CFSv1. As a national climate model, CFSv1 has been particularly successful in seasonal-to-interannual climate forecasting, both retrospectively and operationally (Saha et al. 2006). In CFSv2, the ocean model is the GFDL MOM version 4 (Griffies et al. 2004), which is configured for the global ocean with a horizontal grid of 0.5° × 0.5° poleward of 30°S and 30°N and meridional resolution increasing gradually to 0.25° between 10°S and 10°N. The vertical coordinate is geopotential (z-) with 40 levels (27 of them in the upper 400 m). The maximum depth is approximately 4.5 km. The atmospheric model is the global forecast system (GFS), which has horizontal resolution at T126 (105-km grid spacing, a coarser resolution than is used for the GFS operational weather forecast), and 64 vertical levels in a hybrid sigma-pressure coordinate. The oceanic and atmospheric components exchange surface momentum, heat and freshwater fluxes, as well as SST, every 30 minutes. More details about CFSv2 can be found in Saha et al. (2013).

The hindcasts initialized from multiple ocean analyses have been described in Zhu et al. (2012b), where all validations were based on the ensemble mean fields using deterministic metrics. This group of hindcasts starts from each April during 1979–2007, and lasts for 12 months. In the group of hindcasts, four ocean analyses from the NCEP and ECMWF were used as OIC—ECMWF COMBINE-NV (Balmaseda et al. 2010), ECMWF ORA-S3 (Balmaseda et al. 2008), NCEP Forecast System Reanalysis (CFSR) (Saha et al. 2010), and NCEP GODAS (Behringer 2005) (see the "Appendix" for more details about these ocean analyses). For each OIC, four atmospheric/land initial conditions (the atmospheric/land instantaneous states at 00Z of the first four days in April in the CFSR) were applied to represent the uncertainties in the atmospheric/land initial states as in the LE approach (see the "Appendix" for more details about the hindcast experiment design). Thus, for hindcasts with each OIC, AP is taken into account with four ensemble members generated. These hindcasts are referred to as hindcasts AP_cbn, AP_ora3, AP_cfsr, and AP_gds, corresponding to the above four OIC sources, respectively. The four sets of hindcasts with different OICs are further clustered together to generate a grand ensemble, which is referred to as hindcast MAE with a total of 16 ensemble members. In hindcast MAE, in addition to AP, the uncertainties in OIC, a more important factor affecting seasonal-interannual forecasting, are also sampled.

To validate the MAE method, we also analyzed the retrospective forecasts from NCEP CFSRR, where the LE approach is applied to generate ensembles. The CFSRR hindcasts were produced by NCEP using CFSv2, and cover predictions initialized from all calendar months during Jan 1982 to Dec 2010, with each run extending to around 9 months. For each year, 4 predictions were produced every 5 days beginning on January 1st with ocean and atmosphere initial conditions (ICs) from the NCEP CFSR (Saha et al. 2010). In this analysis, we used forecasts from 16 ICs in March and early April to build our ensemble predictions starting from each April during 1982–2007. Specifically, the 16 predictions are from ICs on Mar. 22, and 27, as well as Apr. 1 and 6 at 00Z, 06Z, 12Z, and 18Z. This group of hindcasts is referred to as hindcast CFSRR.

In addition, the predictions initialized from different days (i.e., Mar. 22, 27, Apr. 1 and 6) are also separately combined to form four subsets to be referred to as hindcast Mini-CFSRR, each with four ensemble members. In these cases, the ensemble perturbations mainly reside in atmospheric initial conditions (AIC), considering the longer memory of the ocean.

The predicted SSTA is derived by subtracting a lead time-dependent climatology from the total SST. The observation-based monthly SST analysis used for validation is from the optimum interpolation analysis, version 2 (OIv2) SST dataset (Reynolds et al. 2002), which has a resolution of 1.0° × 1.0°.

3 Results

To validate the effectiveness of MAE, we first examine the predicted SSTA spread in the tropical Pacific. In a reliable forecasting system, it is required that a given forecast member should have the same statistical properties as the truth. In another words, the true state can be considered as a member of the ensemble (Johnson and Bowler 2009). As a necessary condition for reliability, the standard deviation (spread) of the ensemble should be comparable to the root-mean-square error (RMSE) of the ensemble-mean SSTA forecast (Johnson and Bowler 2009). In practice, however, the SSTA ensemble spread is substantially smaller than the RMSE in all single-model ensembles (see Fig. 1 as an example), which means the predictions tend to be "overconfident" (Palmer et al. 2004; Vialard et al. 2005; Saha et al. 2006; Weisheimer et al. 2009).

Figure 2 shows the spatial distributions of ensemble spread versus ensemble-mean RMSE ratio for hindcast SSTA at lead times of 2, 5, and 8 months. For the four AP hindcasts (upper four rows of Fig. 2), the ratios are generally comparable: at the 2-month lead, all have two regions where the forecasts are overconfident (regions with minimum ratio)—the tropical mid-basin and far eastern basin; at the 5- and 8-month leads, the mid-basin center is less well-defined, but the minimum in the far eastern basin is still apparent. The increasing difference in the spread/RMSE ratio between the mid-basin and far eastern basin centers with increasing lead time implies that the former may be mainly related to the surface ocean processes with shorter time scales, while the latter is mostly attributable to the subsurface processes with longer time scales. In hindcast CFSRR, at all three lead times, the low ensemble spread/RMSE ratios are mostly confined to the far eastern basin, extending westward toward the mid-basin. It is clear that CFSRR has a higher ratio than AP hindcasts in the mid-basin, but the ratio is equivalent to the AP runs in the far eastern basin.

In hindcast MAE, combining ensembles from the four AP hindcasts, the ensemble spread/RMSE ratios are significantly increased. Although the ratio in all AP hindcasts is smaller than 0.6 over a large region of the tropical Pacific at all three lead times, in hindcast MAE it is larger than 0.7 over most of the tropical Pacific. This improvement is also apparent comparing hindcast MAE with hindcast CFSRR. Particularly, in the far eastern basin, very few points in hindcast MAE have a ratio less than 0.5, in contrast with smaller ratios in AP hindcasts and hindcast CFSRR. In the mid-basin at the 2-month lead, in contrast to the general characteristics described above, the hindcast MAE is slightly worse than the hindcast CFSRR, which will be discussed below. It is also interesting to notice that the ensemble spread/RMS ratio is larger than 1 for hindcast LE and also for MAE in the Intertropical Convergence Zone (ITCZ) and South Pacific Convergence Zone (SPCZ), which may be a reflection of the relatively low potential predictability of the forecast model in these regions.

Figure 3 shows the temporal evolution of ensemble spread versus ensemble-mean RMSE ratio for the hindcast NINO3.4 SSTA index (averaged over 5°S–5°N, 170°W–120°W). At all lead times, the AP and Mini-CFSRR hindcasts generally have comparable spread/RMSE ratios, mostly lower than 0.6 for 0–9 months lead time, which is lower than the desired value of 1. This indicates that atmospheric initial perturbations only cannot generate sufficient ensemble spread in the hindcasts. On the other hand, both the MAE (combined from the four AP hindcasts) and CFSRR (combined from the four Mini-CFSRR hindcasts) have clearly increased spread/RMSE ratios at all leading times, which demonstrate the improvement by perturbing OIC.

Moreover, there is a clear distinction between the MAE and CFSRR hindcasts. Comparing hindcast MAE with hindcasts CFSRR, we found that the former has a higher ratio than the latter for 0–9 months lead time as a whole (0.68 vs. 0.59). In particular, hindcast MAE generates significantly higher ensemble spread at long lead times (longer than 2 months), with the spread/RMSE ratio larger (smaller) than 0.7 in the MAE (CFSRR) at these leading months. At lead times shorter than 2 months, a slightly lower ensemble spread is achieved by the hindcast MAE. This may be related to the fact that different ODAs depart from each other more clearly below the surface than at the surface (Figs. 5, 6), and it takes a couple of months to for the SSTA spread to respond to the subsurface memory. Meanwhile, in the tropics the subsurface differences (Fig. 5b) among ODAs mainly reside in the off-equatorial regions (larger than 0.2 °C), which, as a result of propagating equatorial waves, further contributes to larger spreads at longer lead times. In addition, there is a concern about whether the improvement in hindcast MAE comes

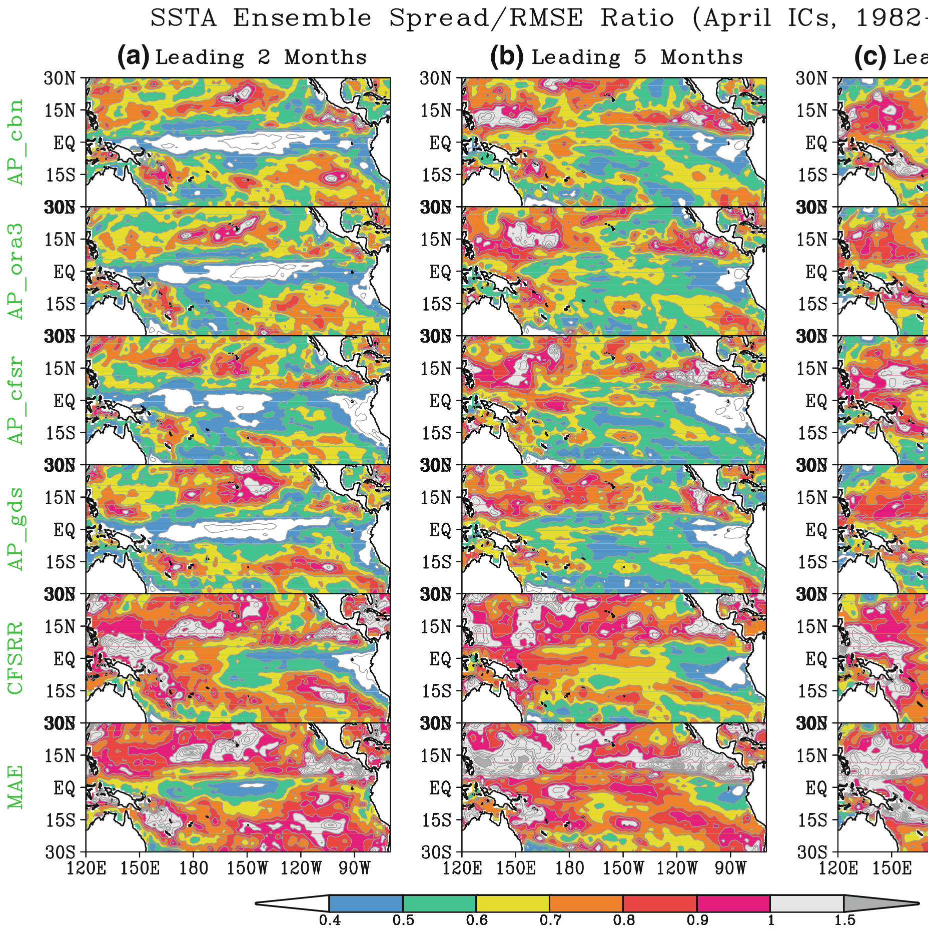

Fig. 2 Distribution of the ensemble spread-to-RMSE ratios for the predicted SST anomalies in the tropical Pacific at **a** 2-, **b** 5-, and **c** 8-month lead times with IC in April of 1982–2007. The results for hindcasts AP_cbn, AP_ora3, AP_cfsr, AP_gds, CFSRR and MAE are shown from the most upper row to the lowest row. Contour interval is 0.1, with above 0.4 *colored shading*

simply from the increased number of ensemble members. To examine this possibility, we computed the spread/RMSE ratio for four Mini-MAEs (Fig. 3b), each of which consists of four ensemble members with different OIC and randomly chosen atmospheric initial conditions. It is clear that the Mini-MAE generally produces a higher spread/RMSE ratio than the Mini-CFSRR or the AP hindcasts, confirming that the increase of spread/RMSE ratio in hindcast MAE is mainly due to including the uncertainty in different OICs, not enlarging the sample size.

It should be noted that the smaller spread in hindcast MAE in mid-basin at short lead times may be related to a weakness in the OIC generation, i.e., adopting monthly mean data for the initial conditions rather than using instantaneous fields, which are difficult to obtain. A set of test runs showed that using the monthly mean fields as OIC has little impact in deterministic terms (Zhu et al. 2012b). However, this choice may have an effect on the probabilistic metrics, especially at short forecast leads. In fact, high frequency features, which should enhance the uncertainty in the OIC, especially near the surface, have been greatly smoothed out by monthly averaging. For example, tropical instability waves (TIW) provide a potential source of the OIC uncertainty in this region. Previous studies have demonstrated that TIW can induce intensive air-sea feedback (Zhang and Busalacchi 2008). Apparently, the monthly averaged oceanic state weakens the TIW signal and its subsequent growth, consequently reducing the departures among different OICs in MAE on this time scale. On the other hand, these signals are included in hindcast CFSRR, because the more frequent instantaneous ocean analysis can detect the different temporal phases of TIW, introducing extra variance in the OIC (Wen et al. 2012). Thus, it is not surprising that hindcast MAE using monthly mean data as OIC produces slightly less SSTA variance at short lead times in the mid-tropical Pacific basin, where TIW is active near the surface. This suggests that the instantaneous fields should be used in the future MAE forecast systems when such OIC become available.

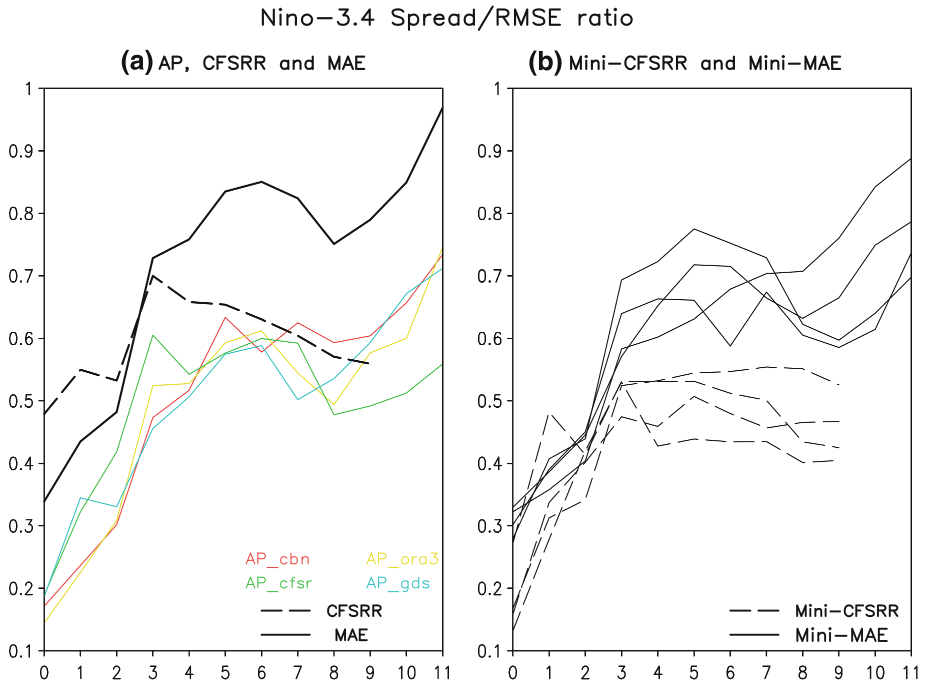

Fig. 3 Evolution of the ensemble spread-to-RMSE ratios for the predicted Niño-3.4 index during 1982–2007 with respect to lead months. In **a**, *colored curves* are for AP hindcasts with four different ocean analyses, and *solid* (*dashed*) *black curves* are for hindcast MAE (CFSRR). In **b**, *solid* (*dashed*) *black curves* are for four Mini-CFSRR (Mini-MAE) hindcasts, each with four ensemble members

We use the reliability diagram (Wilks 2006; Corti et al. 2012; Peng et al. 2012) to quantitatively examine the reliability of ENSO forecasts, which compares the forecast probabilities against the corresponding frequencies of observed occurrence. If a forecast system is perfectly reliable in probabilistic forecasts, the predicted probability of an event occurrence should be equal to the observed relative frequency, which is represented as a 1:1 diagonal line in the diagram. To accumulate a large enough sample of cases for this analysis, three measures are used in our calculations: 1) a contingency table is calculated for each grid point in the NINO3.4 area; 2) all forecasts during at 0–9 months lead time are used, which gives a bulk measure for all lead times; and 3) a relatively small number of probability bins is chosen: 0–20, 20–40, 40–60, 60–80 and 80–100%. We then assess the ability to predict three categories of ENSO events: warm (with SSTA larger than 0.43 °C), cold (SSTA less than −0.43 °C) and neutral (SSTA falling in between). The value 0.43 is chosen is because 43% of a standard deviation is the tertile threshold for normally distributed data, and the standard deviation of NINO3.4 index during the hindcast period is about 1 °C (choosing instead a 0.5 °C threshold does not change the results).

In Fig. 4, we show the reliability diagram for hindcasts MAE and CFSRR, which both have the same number of ensemble members (=16). In general, CFSv2 with both ensemble generation methods produces ENSO forecasts with relatively good reliability, even though the forecasts are still somewhat overconfident, as found in CFSv1 by Saha et al. (2006). The reliability lines in the hindcast MAE are closer to the 1:1 diagonal line for all warm, cold and neutral categories than the same for the hindcast CFSRR, as objectively shown by the differences in their respective slopes (the gray numbers in Fig. 4), despite a common “cold” forecast bias for the cold categories (blue curves in Fig. 4) in both hindcasts. In addition, the difference between MAE and CFSRR may be dependent upon the probability range. For instance, the reliabilities are nearly indistinguishable between the two sets of hindcasts when the predicted probability is below 0.8. On the other hand, CFSRR seems more likely to be overconfident in the high probability range (0.8–1.0) for all three categories of events. These results prove that MAE provides more reliable ENSO forecasts than CFSRR. The sharpness (three inset histograms in Fig. 4) is similar between hindcast MAE and hindcast CFSRR, with warm and cold (neutral) categories having high (intermediate) confidence.

The above analyses indicate that MAE can effectively reduce the “overconfidence” problem in single-model ensembles, suggesting that uncertainty in different OICs contributes to the reliability of the forecasts. However, ensemble spread is still lower than RMSE even though the MAE initialization is applied. This is possibly because another error source, the coupled model error, is not represented in our experiments. Therefore, to fully cover all error sources, stochastic physics or MME should be included, too. On the other hand, in the

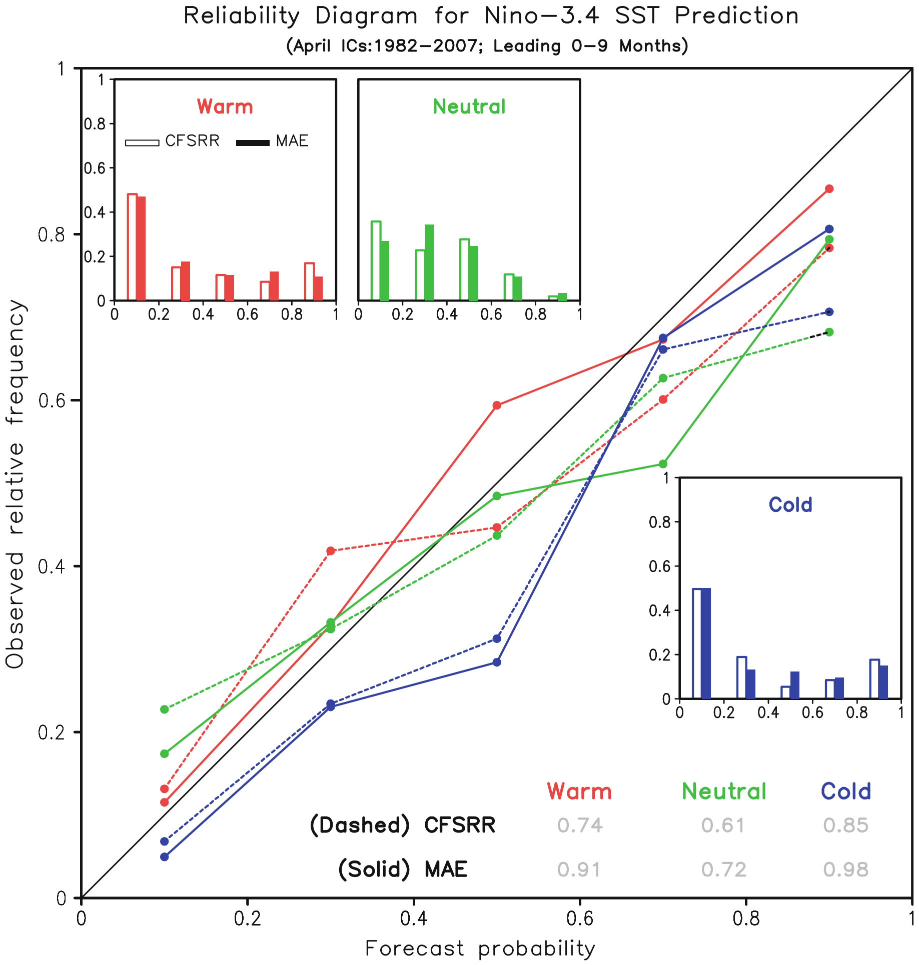

Fig. 4 Reliability diagram of forecast probabilities that predicted SSTs over the Niño-3.4 region fall in the upper (warm; *red curves*), middle (neutral; *green curves*), and lower (cold; *blue curves*) categories of the observed climatology for the leading 0–9 months with IC in April of 1982–2007. The *solid* (*dashed*) *curves* are for hindcast MAE (CFSRR). The y=x diagonal line (slope = 1.0) represents perfect reliability. The probabilities are binned as 0.2-wide intervals, e.g., 0–0.2 (plotted at 0.1). The *inset* histograms are the frequency distributions for these probability bins. *Red colors* correspond to forecasts for the upper (warm), *green* to the middle (neutral), and *blue* to the lower (cold) categories, with filled (outlined) bar for hindcast MAE (CFSRR). The *gray numbers* in the *right bottom* represent slopes of the indicated reliability lines by regression fit

current MME framework, each model component is commonly initialized from one single ocean analysis system, which is usually based on its own ocean model. Consequently, the uncertainty of OIC discussed in this study is also underestimated in the current MME. However, the MAE initialization can be easily applied in the MME framework.

4 Conclusion and discussion

This study presented a new ensemble generation method for seasonal forecasting, i.e., the multi-ocean analysis ensemble (MAE). This method is intended to address the apparent "overconfidence" problem in current ensemble seasonal forecast systems with single models, evidenced by the limited growth of ensemble perturbations with respect to the amplitude of the mean error. In this method, ocean initial conditions (OIC) are based on multiple ocean analyses, which can sample structural uncertainties in OIC originating from errors in the ocean model, forcing fluxes, the analysis method, and the assimilated ocean datasets.

In this study, the merit of MAE is assessed by examining ENSO forecast reliability. In particular, we compared the MAE method with methods that employ atmospheric perturbations (AP), and the lagged ensemble (LE) approach. The latter has been applied by operational climate prediction centers, such as NCEP. It is found that MAE can effectively enhance ensemble spread. The probabilistic reliability analysis indicates that the MAE method has better forecast reliability for all ENSO warm, neutral and cold categories. It is suggested that the MAE method is an easy but effective way to sample various kinds of uncertainties in OIC, and can be beneficial to seasonal forecasting as a potentially useful component in a multi-model ensemble (MME) framework. It is also suggested that, in the future, the MAE method should be applied using instantaneous OIC instead of monthly mean fields, when available.

As pointed by Vialard et al. (2005), an apparent drawback of the LE approach is that it introduces a delay in the forecast delivery date. For climate prediction, the long lag required to generate a large enough ensemble with sufficient oceanic perturbations aggravates this problem. For example, the ensembles in CFSRR are generated every 5 days, so that 15 days to construct a fairly minimal ensemble of 16 members. In addition to the delivery date issue, the incorporation of the ensemble members with a large time lag may also potentially degrade the value of more recent ensemble members, because the earlier members have larger model drift. On these aspects, the MAE initialization may have an advantage over the LE approach. On the other hand, LE has the potential of sampling the different phases of high-frequency phenomena such as TIW or the Madden-Julian Oscillation (MJO), which may have strong effect on ENSO prediction (Wang et al. 2011). In particular, intraseasonal variability may be better sampled in CFSRR than in the MAE. How this can affect the predictive skill can be tested by introducing the LE approach along with the MAE strategy in future studies.

In addition, some ocean analyses, like COMBINE-NV (Balmaseda et al. 2010) and ORA-S4 (Balmaseda et al. 2013), consist of an ensemble of OICs (5 members for instance). Let us call this method the ocean perturbation (OP) approach, where the same data assimilation system has been used to produce the OIC (by perturbing winds, observations or other aspects). The OP approach does not sample the “structural” uncertainty, while the MAE method does. It will be interesting to compare OP with MAE in future work, exploring how much of the uncertainty in the initial conditions sampled by MAE is “structural”, i.e., how does the reliability of MAE compare with the reliability obtained with multiple OIC from a single reanalysis (for example, the 5 ensemble members of COMBINE-NV). It will also be interesting to see a comparison between MAE and MME (for instance an ensemble of ECWMF seasonal forecast system 4 and CFSRR).

Acknowledgements Funding for this study was provided by grants from NSF (ATM-0830068), NOAA (NA09OAR4310058), and NASA (NNX09AN50G). The authors would like to thank Dr. J. Shukla for his guidance and support of this project. We thank ECMWF and NCEP for providing their ocean data assimilation analysis datasets, which made this project possible. The authors gratefully acknowledge NCEP for the CFSv2 model made available to COLA. Computing resources provided by NAS are also gratefully acknowledged.

Appendix

Four ocean analyses used in hindcast AP/MAE

In hindcast MAE, four different ocean analyses were used as OICs, with two from ECMWF and two from NCEP. They are the ECMWF COMBINE-NV (Balmaseda et al. 2010), the ECMWF Ocean Reanalysis System 3 (ORA-S3; Balmaseda et al. 2008), the NCEP Climate Forecast System Reanalysis (CFSR; Saha et al. 2010), and the NCEP Global Ocean Data Assimilation System (GODAS; Behringer 2005). The GODAS and CFSR (ORA-S3) ocean analyses have been used to initialize the operational seasonal predictions made by NCEP (ECMWF). COMBINE-NV, which is only slightly different from ECMWF Ocean Reanalysis System 4 (ORA-S4), has been used to initialize their decadal predictions at ECMWF. Table 1 briefly summarizes the major characteristics of these ocean analyses, including the models, resolutions, assimilation methods, and the assimilated data. To show the systematic differences among the four ocean analyses, Figs. 5 and 6 present the standard deviation of SST/HC300 systematic differences among them and the signal versus noise ratio maps, respectively. From Fig. 5, it can be seen that SST

Table 1 Brief summary of the used ocean analyses

System and institution (reference)	Model and resolution	Method	Input data
COMBINE-NV ECMWF, EU (Balmaseda et al. 2010)	NEMO 1×(0.3–1), 42 Levels	3D-VAR	EN3_v2a data set (including ocean station/CTD, XBT, Buoys, profilers inWOD05, GTSPP and ARGO)
ORA-S3, ECMWF, EU (Balmaseda et al. 2008)	HOPE 1×(0.3–1), 29 levels	3D OI with online bias correction	Altimeter (sea level anomalies and global trends), SST, T&S from XBT, CTD, Argo, TAO
CFSR, NCEP, USA (Saha et al. 2010)	NCEP CFS2 [MOM4 0.5×(0.25–0.5), 40 levels]	Partially Coupled Data Assimilation (3D-VAR for OM)	SST, T&S profiles from XBT, CTD, Argo, TAO
GODAS, NCEP, USA (Behringer 2005)	MOM3 1×(0.3–1), 40 levels	3D-VAR	SST, T profiles from XBT, CTD, Argo, TAO

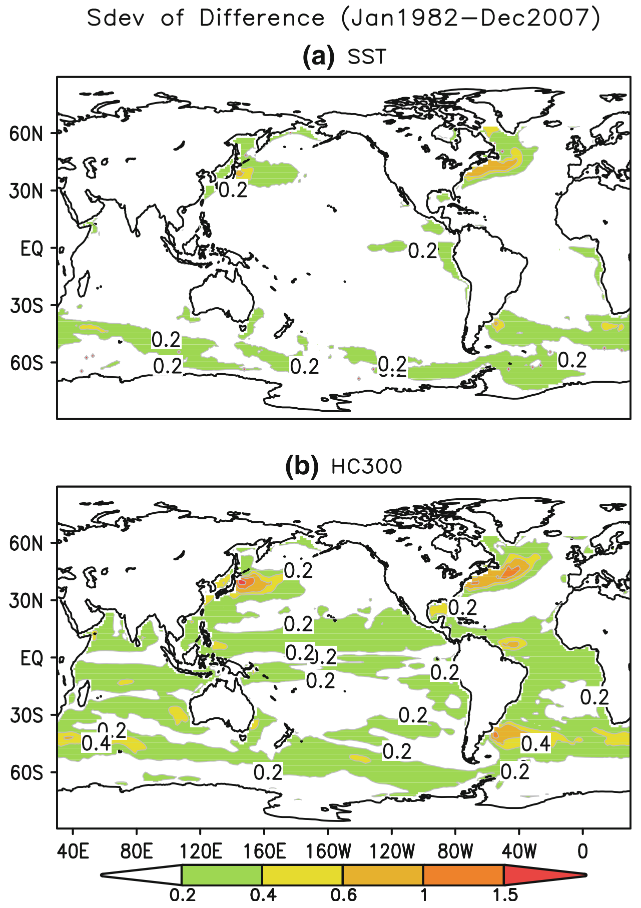

Fig. 5 The standard deviation of the systematic differences in **a** SST and **b** the upper 300 m mean ocean temperature (HC300) for 1982–2007 among the four ocean analyses used in the study (i.e., COMBINE-NV, ORA-S3, CFSR, GODAS). Contours 0.2, 0.4, 0.6, 1, 15 are shown. The systematic differences are calculated as the difference between four individual datasets and their ensemble mean. Unit: °C

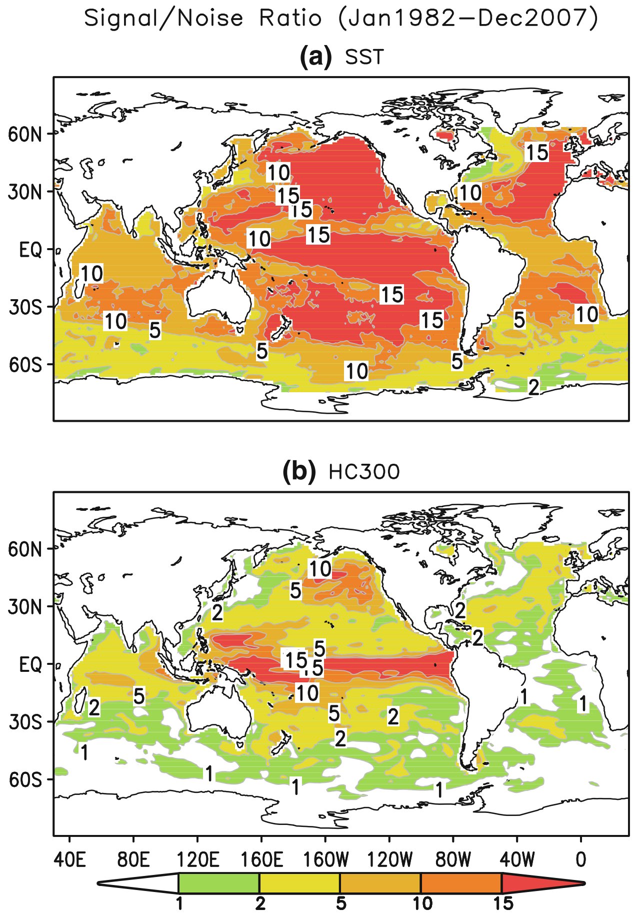

Fig. 6 The global maps of the signal versus noise ratio of **a** SST and **b** HC300 for 1982–2007 derived from the four ocean analyses (i.e., COMBINE-NV, ORA-S3, CFSR, GODAS). Contours 1, 2, 5, 10, 15 are shown. Here signal is defined as the interannual variance in the ensemble mean of four datasets; noise is defined as the variance of the difference between four individual datasets and their ensemble mean

shows little difference among them in the tropics expect for the far eastern coastal regions. There is larger difference among them in the subsurface as shown in Fig. 5b. Particularly, in the off-equatorial regions the difference in HC300 is larger than 0.2 °C, with less difference along the equator. Figure 6 also indicates that SSTs have relatively lower level of noises, while HC300s have larger level of noises, particularly over the off-equatorial regions.

Experimental design for hindcast AP/MAE

For hindcast AP_cbn, AP_ora3, AP_cfsr, and AP_gds, the atmosphere, land and sea ice initial states are specified in the same way, using the instantaneous fields from the CFSR. For each hindcast, four ensemble members are generated that differ in their atmosphere/land surface conditions, which are the instantaneous fields from 00Z of the first four days in April in the CFSR, respectively. For the OIC, to reduce the potentially negative effects of the mean biases in ocean analyses and the forecast model, and to make the predictions using OIC from different analyses comparable, we applied an anomaly ocean initialization strategy (e.g., Schneider et al. 1999) in these experiments. For this purpose, a monthly climatology for the CFSv2 ocean component was derived from the last 20 years of a 30-year simulation starting from CFSR state on November 1, 1980. The monthly anomalies of all variables from the ocean analyses are then calculated with respect to their own climatologies and superimposed on the CFSv2 monthly climatological states. The fields in March and April are averaged to represent the oceanic states at the start of April. Initializing the hindcasts using the monthly oceanic analyses is different from the operational practice of using an instantaneous analysis from the ocean data assimilation system. A set of test runs in Zhu et al. (2012b) showed that using the monthly fields as OIC has little impact on the deterministic forecasting skill (see Fig. S1 of the auxiliary material in Zhu et al. (2012b)).

References

Balmaseda M, Vidard A, Anderson D (2008) The ECMWF System 3 ocean analysis system. Mon. Wea. Rev. 136:3018–3034

Balmaseda M, Mogensen K, Molteni F, Weaver A (2010) The NEMOVAR-COMBINE ocean re-analysis. COMBINE technical report No. 1. p. 10 http://www.combine-project.eu/Technical-Reports.1668.0.html

Balmaseda M, Mogensen K, Weaver A (2013) Evaluation of the ECMWF ocean reanalysis ORAS4. Quart. J. Roy. Meteor. Soc. doi:10.1002/qj.2063

Behringer DW (2005) The global ocean data assimilation system (GODAS) at NCEP, 11th symposium on integrated observing and assimilation systems for the atmosphere, oceans, and land surface (IOAS-AOLS), San Antonio, TX, American Meteorological Society, 3.3

Corti S, Weisheimer A, Palmer TN, Doblas-Reyes FJ, Magnusson L (2012) Reliability of decadal predictions. Geophys Res Lett 39:L21712

Griffies SM, Harrison MJ, Pacanowski RC, Rosati A (2004) Technical guide to MOM4, GFDL ocean group technical report no. 5. NOAA/Geophysical Fluid Dynamics Laboratory. Available on-line at. http://www.gfdl.noaa.gov/*fms

Jin EK et al (2008) Current status of ENSO prediction skill in coupled ocean-atmosphere models. Clim Dyn 31:647–664

Johnson C, Bowler N (2009) On the reliability and calibration of ensemble forecasts. Mon. Wea. Rev. 137:1717–1720

Kirtman BP, Min D (2009) Multimodel Ensemble ENSO Prediction with CCSM and CFS. Mon. Wea. Rev. 137:2908–2930

Kug J-S, Ham Y-G, Kimoto M, Jin F-F, Kang I-S (2010) New approach on the optimal perturbation method for ensemble climate prediction. Clim Dyn. doi:10.1007/s00382-009-0664-y

Kug J-S, Ham Y-G, Lee EJ, Kang IS (2011) Empirical singular vector method for ensemble El Niño–Southern Oscillation prediction with a coupled general circulation model. J Geophys Res 116:C08029

Langford S, Hendon HH (2013) Improving Reliability of Coupled Model Forecasts of Australian Seasonal Rainfall. Mon. Wea. Rev. 141:728–741

Latif M, Anderson D, Barnett T, Cane M, Kleeman R, Leetmaa A, O'Brien J, Rosati A, Schneider E (1998) A review of the predictability and prediction of ENSO. J Geophys Res 103:14375–14393. doi:10.1029/97JC03413

Magnusson L, Balmaseda M, Corti S, Molteni F, Stockdale T (2012) Evaluation of forecast strategies for seasonal and decadal forecasts in presence of systematic model errors. Clim Dyn. doi:10.1007/s00382-012-1599-2

Manganello JV, Huang B (2009) The influence of systematic errors in the Southeast Pacific on ENSO variability and prediction in a coupled GCM. Clim Dyn 32:1015–1034. doi:10.1007/s00382-008-0407-5

Molteni F, Stockdale T, Balmaseda M, Balsamo G, Buizza R, Ferranti L, Magnusson L, Mogensen K, Palmer T, Vitart F (2011) The new ECMWF seasonal forecast system (System 4). ECMWF Technical Memorandum No. 656, p. 49

Palmer TN, Buizza R, Molteni E, Chen Y-Q, Corti S (1994) Singular vectors and the predictability of weather and climate. Philos Trans R Soc Lond 348:459–475

Palmer TN et al (2004) Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER). Bull. Amer. Meteor. Soc. 85:853–872

Pan X, Huang B, Shukla J (2011) The influence of mean climate on the equatorial Pacific seasonal cycle and ENSO: simulation and prediction experiments using CCSM3. Clim Dyn 37:325–341. doi:10.1007/s00382-010-0923-y

Peng P, Kumar A, Halpert MS, Barnston AG (2012) An analysis of CPC's operational 0.5-month lead seasonal outlooks. Wea Forecasting 27:898–917

Reynolds RW, Rayner NA, Smith TM, Stokes DC, Wang W (2002) An improved in situ and satellite SST analysis for climate. J. Climate 15:1609–1625

Saha S et al (2006) The NCEP Climate Forecast System. J. Climate 19:3483–3517

Saha S et al (2010) The NCEP climate forecast system reanalysis. Bull. Amer. Meteor. Soc. 91:1015–1057

Saha S et al (2013) The NCEP Climate Forecast System Version 2. Submitted to J, Climate

Schneider EK, Huang B, Zhu Z, DeWitt DG, Kinter JL III, Kirtman B, Shukla J (1999) Ocean data assimilation, initialization, and predictions of ENSO with a coupled GCM. Mon. Wea. Rev. 127:1187–1207

Toth Z, Kalnay E (1997) Ensemble Forecasting at NCEP: the breeding method. Mon. Wea. Rev. 125:3297–3319

Vialard J, Vitart F, Balmaseda MA, Stockdale TN, Anderson DLT (2005) An ensemble generation method for seasonal forecasting with an ocean-tmosphere coupled model. Mon. Wea. Rev. 133:441–453

Wang W, Chen M, Kumar A (2010) An assessment of the CFS real-time seasonal forecasts. Wea. Forecasting 25:950–969. doi:10.1175/2010WAF2222345.1

Wang W, Chen M, Kumar A, Xue Y (2011) How important is intraseasonal surface wind variability to real-time ENSO prediction? Geophys Res Lett 38:L13705. doi:10.1029/2011GL047684

Weigel AP, Liniger MA, Appenzeller C (2009) Seasonal ensemble forecasts: Are re-calibrated single models better than multimodels? Mon. Wea. Rev. 137:1460–1479

Weisheimer A et al (2009) ENSEMBLES: A new multi-model ensemble for seasonal-to-annual predictions—Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. Geophys Res Lett 36:L21711. doi:10.1029/2009GL040896

Wen C, Xue Y, Kumar A (2012) Ocean-Atmosphere Characteristics of Tropical Instability Waves Simulated in the NCEP Climate Forecast System Reanalysis. J. Climate 25:6409–6425

Wilks DS (2006) Statistical methods in the atmospheric sciences, 2nd edn. International Geophysics Series, vol 59. Academic Press, New York, p 627

Xue Y et al (2012) A Comparative Analysis of Upper Ocean Heat Content Variability from an Ensemble of Operational Ocean Reanalyses. J. Climate 25:6905–6929. doi:10.1175/JCLI-D-11-00542.1

Yang S-C, Kalnay E, Cai M, Rienecker MM (2008) Bred vectors and tropical pacific forecast errors in the NASA coupled general circulation model. Mon Wea. Rev. 136:1305–1326

Zhang R-H, Busalacchi AJ (2008) Rectified effects of tropical instability wave (TIW)-induced atmospheric wind feedback in the tropical Pacific. Geophys Res Lett 35:L05608. doi:10.1029/2007GL033028

Zhu J, Huang B, Balmaseda MA (2012a) An Ensemble Estimation of the Variability of Upper-ocean Heat Content over the Tropical

Atlantic Ocean with Multi-Ocean Reanalysis Products. Clim Dyn 39:1001–1020. doi:10.1007/s00382-011-1189-8

Zhu J, Huang B, Marx L, Kinter JL III, Balmaseda MA, Zhang R-H, Hu Z-Z (2012b) Ensemble ENSO hindcasts initialized from multiple ocean analyses. Geophys Res Lett 39. doi:10.1029/2012GL051503

Zhu J, Huang B, Hu Z-Z, Kinter JL III, Marx L (2013) Predicting U.S. Summer Precipitation using NCEP Climate Forecast System Version 2 initialized by Multiple Ocean Analyses. Clim Dyn. doi:10.1007/s00382-013-1785-x