ADVANCES IN ATMOSPHERIC SCIENCES, VOL. 26, NO. 2, 2009, 359–372

Ensemble Hindcasts of ENSO Events over the Past 120 Years Using a Large Number of Ensembles

ZHENG Fei[1] (郑 飞), ZHU Jiang*[2] (朱 江), WANG Hui[3] (王 慧), and Rong-Hua ZHANG[4]

[1] *International Center for Climate and Environment Science* (*ICCES*), *Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing* 100029

[2] *State Key Laboratory of Atmospheric Boundary Layer Physics and Atmospheric Chemistry* (*LAPC*), *Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing* 100029

[3] *National Meteorological Center, Beijing* 100081

[4] *Earth System Science Interdisciplinary Center* (*ESSIC*), *University of Maryland, College Park, Maryland, USA*

(Received 20 January 2008; revised 2 November 2008)

ABSTRACT

Based on an intermediate coupled model (ICM), a probabilistic ensemble prediction system (EPS) has been developed. The ensemble Kalman filter (EnKF) data assimilation approach is used for generating the initial ensemble conditions, and a linear, first-order Markov-Chain SST anomaly error model is embedded into the EPS to provide model-error perturbations. In this study, we perform ENSO retrospective forecasts over the 120 year period 1886–2005 using the EPS with 100 ensemble members and with initial conditions obtained by only assimilating historic SST anomaly observations.

By examining the retrospective ensemble forecasts and available observations, the verification results show that the skill of the ensemble mean of the EPS is greater than that of a single deterministic forecast using the same ICM, with a distinct improvement of both the correlation and root mean square (RMS) error between the ensemble-mean hindcast and the deterministic scheme over the 12-month prediction period. The RMS error of the ensemble mean is almost 0.2°C smaller than that of the deterministic forecast at a lead time of 12 months. The probabilistic skill of the EPS is also high with the predicted ensemble following the SST observations well, and the areas under the relative operating characteristic (ROC) curves for three different ENSO states (warm events, cold events, and neutral events) are all above 0.55 out to 12 months lead time.

However, both deterministic and probabilistic prediction skills of the EPS show an interdecadal variation. For the deterministic skill, there is high skill in the late 19th century and in the middle-late 20th century (which includes some artificial skill due to the model training period), and low skill during the period from 1906 to 1961. For probabilistic skill, for the three different ENSO states, there is still a similar interdecadal variation of ENSO probabilistic predictability during the period 1886–2005. There is high skill in the late 19th century from 1886 to 1905, and a decline to a minimum of skill around 1910–50s, beyond which skill rebounds and increases with time until the 2000s.

Key words: ENSO, ensemble prediction system, interdecadal predictability, hindcast

Citation: Zheng, F., J. Zhu, H. Wang, and R.-H. Zhang, 2009: Ensemble hindcasts of ENSO events over the past 120 years using a large number of ensembles. *Adv. Atmos. Sci.*, **26**(2), 359–372, doi: 10.1007/s00376-009-0359-7.

1. Introduction

Based on the intermediate coupled model (ICM) (Keenlyside and Kleeman, 2002; Zhang et al., 2005), a probabilistic EPS was developed. It has been demonstrated that this system can be improved for El Niño simulations and predictions through the use of the ensemble Kalman filter (EnKF; e.g., Evensen, 2003) data

*Corresponding author: ZHU Jiang, jzhu@mail.iap.ac.cn

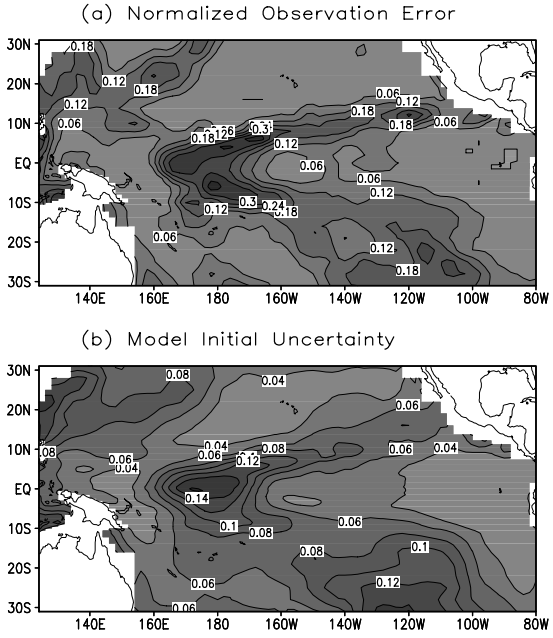

Fig. 1. Horizontal distributions of the normalized observation error (a) and the model initial uncertainty (b) of SST in January 1998. The model initial uncertainty is estimated from the EnKF analysis spread. The contour interval is 0.05°C in (a) and 0.02°C in (b).

assimilation approach for generating the initial ensemble conditions, as well as a linear, first-order Markov-Chain SST anomaly error model that was embedded into the ensemble prediction system (EPS) to provide model error perturbations (Zheng et al., 2006a). However, the model performance was only verified over a relatively short period with relative small number of events.

As pointed out by Chen et al. (2004), previous estimations of El Niño's predictability (e.g., Goswami and Shukla, 1991; Kirtman and Schopf, 1998; Latif et al., 1998) were mostly based on retrospective predictions for the last two or three decades (i.e., the hindcast period encompassed a relatively small number of events). With so few degrees of freedom and short hindcast periods, the statistical significance of those estimates is questionable. From available SST observations, Chen et al. (2004) used the Lamont ENSO prediction model to perform a retrospective forecast experiment of 148 years from 1856 to 2003, and found that ENSO predictability clearly had interdecadal variations. This was, to date, the first work that studied ENSO predictability by extending realistic forecasts to a period over 100 years. Also, Tang et al. (2008) compared ENSO predictabilities using three different models by performing 120-year retrospective forecasts, and confirmed the interdecadal variations in ENSO predictability were not model dependent.

However, the ENSO predictability in these models was only verified in the deterministic sense. Indeed, as considered in classic theories, ENSO should be viewed as a chaotic or irregular interannual fluctuation in the tropical Pacific (e.g., Tziperman et al., 1994). So we need to discuss the ENSO predictability in not only a deterministic sense but also in a probabilistic sense. With a realistic ENSO EPS, and newly-developed SST assimilation approaches (Zheng et al., 2006a), we recently completed a long-term retrospective ensemble forecast from 1886 to 2005 with 100 members, and analyzed the ENSO predictability and its variations in both a deterministic and probabilistic sense.

This paper is structured as follows: Section 2 describes the components of the EPS, and the historic SST data in detail. Section 3 examines the deterministic and probabilistic prediction skills of the EPS for the whole period from 1886 to 2005. In section 4, the interdecadal variations of the ensemble prediction skills in the ENSO EPS are examined in both the deterministic and probabilistic sense. A summary and discussion are given in section 5.

2. Ensemble prediction system components and dataset

2.1 *Basic deterministic model*

Our ensemble prediction system mainly contains three components. The EPS is firstly based on a deterministic model, and the basic intermediate coupled model was developed by Keenlyside and Kleeman (2002) and Zhang et al. (2003). Its dynamical component consists of both linear and non-linear components. The former was essentially a McCreary-type (McCreary, 1981) modal model, but was extended to include a horizontally-varying background stratification. In addition, ten baroclinic modes, along with a parameterization of the local Ekman-driven upwelling, were included. A SST anomaly model was embedded within this dynamical framework to simulate the evolution of the mixed-layer temperature anomalies. As demonstrated by Zhang et al. (2005), having a realistic parameterization for the temperature of the subsurface water entrained into the mixed-layer (T_e) is crucial to the performance of SST simulations in the equatorial Pacific. An empirical T_e model was constructed from historical data and was demonstrated to be effective in improving the SST simulations. The ocean model was coupled with a statistical atmospheric model, which specifically relates wind stress (τ) to SST anomaly fields. The two empirical models (the T_e model and the atmospheric model) were constructed based on the historic observations during the period 1963–96 (34 yr of data). All coupled-model components exchange

simulated anomaly fields. Information concerning the interactions between the atmosphere (τ) and the ocean (SST) was exchanged once a day.

2.2 *Initial ensemble condition*

Based on the ICM, a probabilistic EPS was developed by Zheng et al. (2006a). The initial ensemble conditions of the EPS were provided by the EnKF (e.g., Evensen, 2003, 2004) data assimilation approach through assimilating SST anomaly data into the model with 100 ensemble members (Zheng et al., 2006a). Figure 1 shows an example of horizontal distributions of the normalized observation error and the model initial uncertainty of SST at the initial time of January 1998. The distribution of the model uncertainty has the same shape as that of the normalized observation error. Thus, each initial ensemble member after assimilation represents an equally realistic initial condition. At the same time, the initial ensemble state variables are dynamically balanced within the model after a series of assimilation cycles. Thus, this ensemble initialization approach not only can generate accurate and dynamically consistent initial ensemble members, but also can provide reasonable surface initial stochastic uncertainties for the EPS by combining both background and observation errors during the assimilation cycles (Zheng and Zhu, 2008).

2.3 *Stochastic model-error perturbation*

As described by Zheng et al. (2006a), due to simulation deficiencies for coupled air-sea interactions and subsurface thermal effects in the SST anomaly model, a linear, first-order Markov stochastic model is embedded within the SST anomaly model of the ICM to represent the model uncertainties of forecasted SST anomaly fields. This perturbation method was verified to be capable of effectively simulating the time evolution of model uncertainties during the ensemble forecasting procedure (Zheng et al., 2007). Here, we make further refinements and extensions to the model error perturbation scheme by carefully analyzing the forecast errors (408 samples of the observation-minus-forecast values for 12-month lead time from 1963 to 1996, covering the same analysis period as the training period of the deterministic model) for the different lead times by an empirical orthogonal function (EOF) method, instead of the formulation used in Zheng et al. (2006a). After doing these, the time evolution of the model errors at different lead times can be represented as,

$$\begin{cases} Q_j^{(t)} = \sum\limits_{i=1}^{M} \lambda_i^{(t)} \times \Psi_{i,j} + \xi_j^{(t)} \\ \lambda_i^{(t)} = \alpha_{i,j} \times \lambda_i^{(t-1)} + \sqrt{1-\alpha_{i,j}^2} \times v_i^{(t)} \end{cases}, \tag{1}$$

where $\Psi_{i,j}$ represents the spatial pattern of the ith EOF mode for the (SST anomaly) model error Q at lead time of j month, and which is a constant horizontal distribution for each mode. $\lambda_i^{(t)}$ represents the random normalized time coefficient of the ith mode at time t, the coefficient $\alpha_{i,j}$ is the time correlation of the stochastic forcing for the ith mode at lead time of j month, $v_i^{(t)}$ is a random number of the ith mode at time t, with a mean equal to 0 and variance equal to 1, and the correlations between the random vector of each mode should be zero to allow the maintenance of the orthogonality of each mode. Therefore, this equation ensures that the variance in $\lambda_i^{(t)}$ is equal to 1 as long as the variance of $\lambda_i^{(t-1)}$ is also equal to 1. The subscripts i and j indicate the EOF mode number and the lead time respectively, the number M is the number of the EOF modes used in the stochastic model, and $\xi_j^{(t)}$ represents a residual random field for Q at time t that is obtained by taking out the first M EOF modes from the observation-minus-forecast values.

There are two advantages that should be addressed here for this simplified representation of the model errors. First, there is no longer any need to calculate the spatial correlation scales of the model errors as in Zheng et al. (2006a) at each grid point through perturbing the time coefficients with only the constant spatial patterns for each mode. Second, the temporal correlation coefficients α, for each mode in Eq. (1) for the stochastic model can be easily obtained by calculating the lagged correlations of the series of the time coefficients from EOF analysis results.

This model-error analysis was performed by comparing the SST anomalies' twelve-month observation-minus-forecast values. The model errors were computed from 408 samples over a 34-year period (coinciding with the model training period) starting in 1963 and extending until 1996, without considering the errors inherent within the initial conditions. The forecast initialization scheme was a nudging assimilation scheme, which was used to minimize the initial errors here (Zheng et al., 2006b). The details of the analysis process for estimating the model errors are as follows. Firstly, to obtain the approximate "perfect" initial fields, the observed SST anomaly data were nudged into the model at every time step and at each grid point, and this nudging process was started each month from December 1962 to November 1996 with a reasonable nudging intensity [i.e., 0.50 following Zheng et al. (2006b)] and 12-month nudging time length. Then, twelve-month forecasts were initialized from the nudging results each month during the 34-yr period from 1963 to 1996. Thirdly, twelve-month observation-minus-forecast values of the SST anomalies during this 34-yr period (408 samples) were ob-

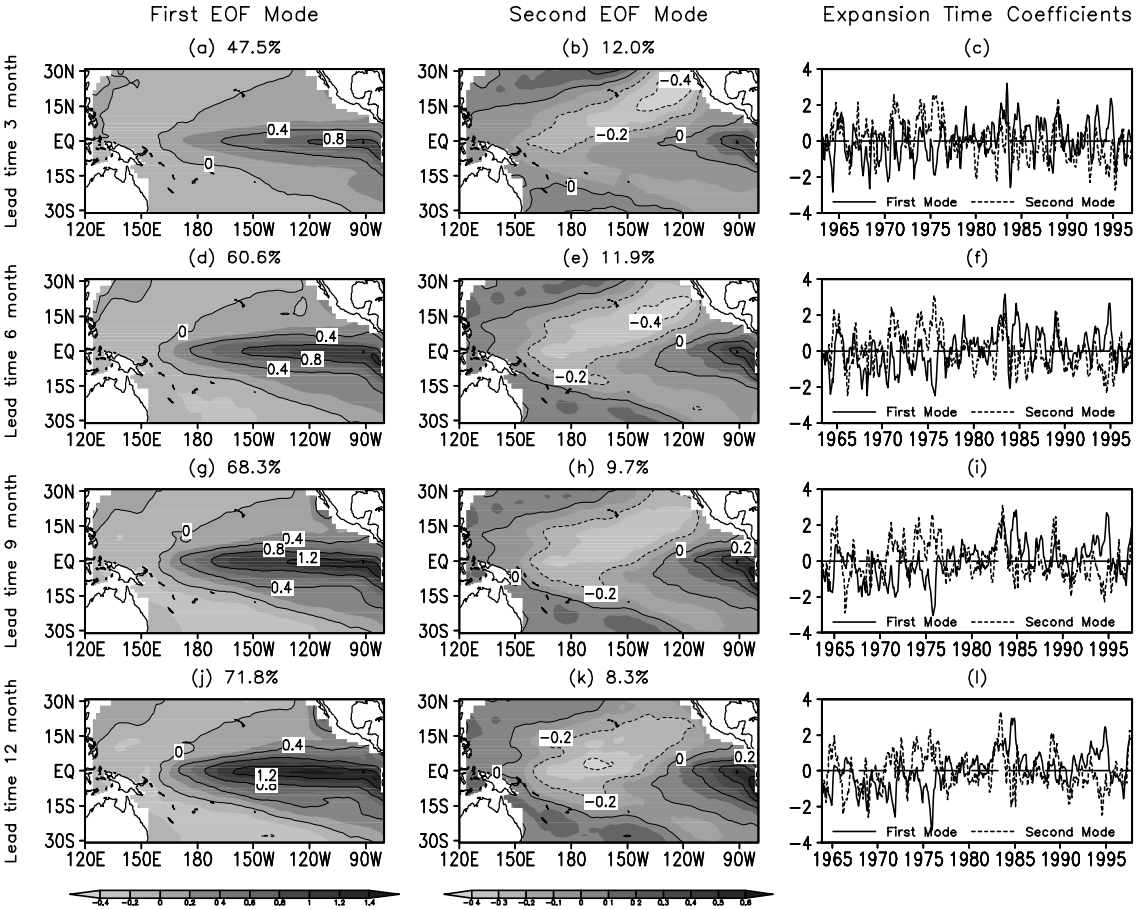

Fig. 2. Spatial patterns of the first mode (left column), second mode (middle column), and their associated normalized time coefficients (right column) for the SST anomaly model errors at 3-, 6-, 9-, and 12-month lead times. The contour interval is 0.2°C for the first mode and 0.1°C for the second mode.

tained as model errors. Finally, the properties of the model errors, such as spatial patterns and their associated temporal variations, were analyzed through the EOF method.

Figure 2 shows the spatial patterns [i.e., Ψ in Eq. (1)] of the first and second EOF modes for the SST anomaly model errors at lead times of 3, 6, 9, and 12 months, and the associated time series. The spatial structure indicates the regions, which are not predicted well by the model. For the first mode, model uncertainties are mainly located over the eastern equatorial Pacific, and extend into the central basin with longer lead times. In contrast to the first mode, the model uncertainties of the second mode are mainly located over the eastern costal regions and the central equatorial Pacific. And the proportion of the first mode in total covariance increases from 37.1% to 71.8% in the 12-month model-error analysis results, while the proportion of the second mode decreases to 8.3% at 12-month lead. These results indicate that the first several modes can explain and describe the variations of the model errors in the tropical Pacific, and the contributions of the first mode dominate the model-error simulations, especially at longer leads. The temporal correlation coefficients [i.e., α in Eq. (1)] for each mode were obtained by calculating the one-month lagged correlations for each EOF time-series. Table 1 presents the temporal correlation coefficients that are used in the stochastic model. The temporal correlation coefficients of each mode increase with increasing lead time, which indicates a decreased randomness in the expansion time coefficients, and the temporal correlation coefficient α of the first mode exceeds 0.95 at 12-month lead time. Thus, the variations of the major modes in the model-error model are allowed to be more random at short lead times, but with more stable and bias-correction like properties at longer lead times (e.g., Evensen, 2003).

After carefully building up a reasonable model-error model, we can use Eqs. (1) and (2) to provide a simple representation of a non-linear model, by embedding the above model-error system within the dy-

Table 1. Time-correlated coefficients of the stochastic model for the first ten modes from one-month to twelve-month lead times.

Lead time	EOF mode									
(months)	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
1	0.695	0.603	0.604	0.397	0.408	0.416	0.274	0.403	0.403	0.253
2	0.799	0.840	0.703	0.721	0.689	0.747	0.630	0.714	0.621	0.651
3	0.858	0.863	0.767	0.745	0.754	0.780	0.659	0.732	0.666	0.693
4	0.879	0.875	0.796	0.746	0.787	0.784	0.670	0.696	0.753	0.698
5	0.891	0.877	0.816	0.750	0.795	0.782	0.688	0.686	0.783	0.691
6	0.902	0.888	0.804	0.768	0.792	0.779	0.697	0.699	0.784	0.704
7	0.919	0.888	0.797	0.784	0.774	0.780	0.712	0.700	0.780	0.740
8	0.928	0.879	0.809	0.802	0.756	0.776	0.718	0.694	0.785	0.740
9	0.935	0.877	0.810	0.818	0.737	0.785	0.726	0.703	0.781	0.753
10	0.941	0.875	0.817	0.835	0.714	0.795	0.728	0.704	0.783	0.768
11	0.948	0.872	0.828	0.849	0.710	0.790	0.732	0.704	0.781	0.770
12	0.954	0.868	0.836	0.859	0.720	0.791	0.728	0.714	0.775	0.779

namical model to simulate the time evolutions of the model errors during the ensemble forecast process:

$$\psi_t = \boldsymbol{f}(\psi_{t-1}) + Q_t \tag{2}$$

where ψ_t represents the model state at time t, and $\boldsymbol{f}$ is the non-linear model operator. In order to achieve reasonable amplitudes, the first ten EOF modes were retained in simulating the random model errors, and the simulated random model errors of SST anomalies were generated and added into the physical model daily.

2.4 *Dataset*

The data used in this study is the monthly extended global SST (ERSST) dataset from 1854 to 2006 reconstructed by Smith and Reynolds (2004), with 2° horizontal resolution. Due to the relatively poor quality of the dataset prior to 1880, the observed SST anomalies before 1880 lack annual and seasonal variations (Smith et al., 2008), so the initial conditions can not trigger real annual oscillations and seasonal variations of the predicted signals (Tang et al., 2008). Thus we focus on the period from 1886 to 2005 in this study, and the data domain is configured as the tropical Pacific Ocean. A very important task in ENSO predictions is to optimize the oceanic initial conditions, and the assimilation of subsurface in-situ observations and satellite altimetry can significantly improve model skills (e.g., Tang and Hsieh, 2003; Zheng et al., 2007). However, the oceanic satellite altimetry and subsurface observation records are too short for our study. The only way solution is to only assimilate SST to initialize forecasts. Zheng et al. (2006a) used the EnKF data assimilation system to provide an initial condition ensemble for the ICM with 100 members. And this SST-only assimilation approach has been verified to be able to provide dynamically balanced initial fields and significantly improve El Niño predictions. In this study, only the observed monthly SST anomaly fields from Smith and Reynolds (2004) are assimilated into the ICM with the EnKF once a month. These observational data are also used for verifying the model predictions.

3. Retrospective forecast experiments

The retrospective forecast (or hindcast) experiments covering the period 1886–2005 are made and compared to available observations. A 12-month hindcast is initialized each month during this 120-yr period. For each initial month, an ensemble of 100 hindcasts is run, yielding a total of 144000 retrospective forecasts to be verified. Figure 3 directly shows the predicted ensemble mean of the Niño-3.4 (5°S–5°N, 120°–170°W) SST anomalies and the prediction spread at 6-month lead time from 1886 to 2005. The variability of the ensemble mean follows the Niño-3.4 observations quite well. Apart from a few exceptions the ensemble forecasts can encompass the observations. This indicates that the EPS is able to predict most of the warm and cold events that occurred in past 120 years at 6-month lead time, especially the relatively large El Niño and La Niña events. The skill of the hindcasts is examined from both a deterministic and a probabilistic perspective. The skill estimation in this section is based on the full hindcast period, 1886–2005, which corresponds to a total of 144000 members.

3.1 *Deterministic prediction skill*

Firstly, to check the deterministic predictability of the EPS for the large events, Fig. 4 shows long lead time deterministic retrospective forecast results for four of the largest warm episodes (as measured by the peak Niño-3.4 SST anomalies) of the past 120 years. In all cases, the EPS is able to predict the ob-

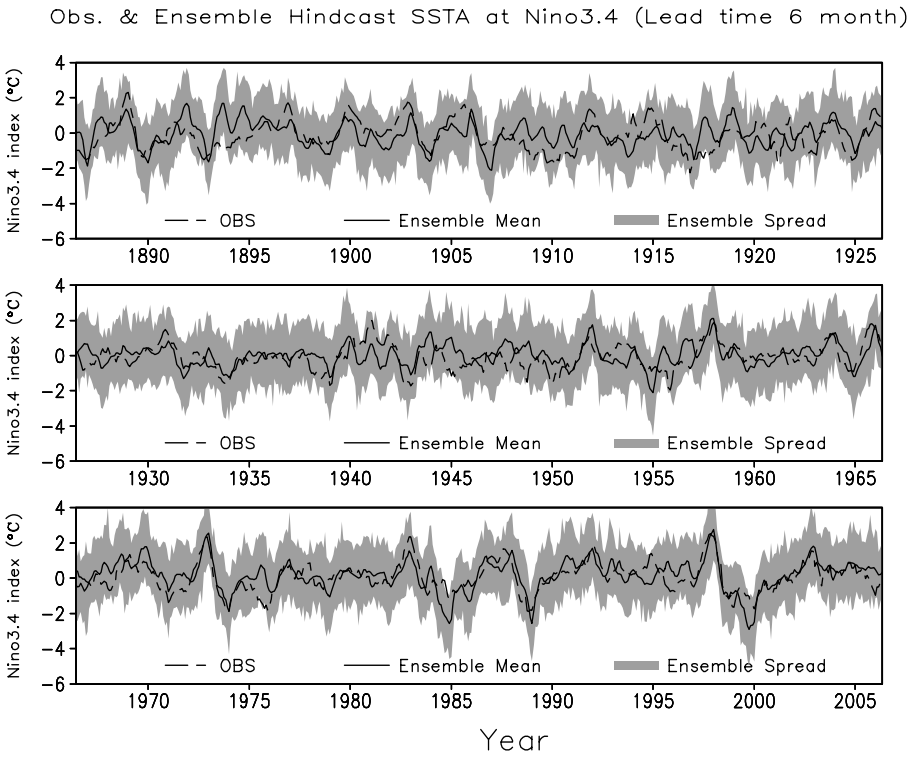

Fig. 3. Time series of observed and forecasted Niño-3.4 SST anomalies at 6-month lead time. The dashed line represents the observed SST anomalies, the solid line represents the ensemble mean, and the shaded area represents the prediction spread.

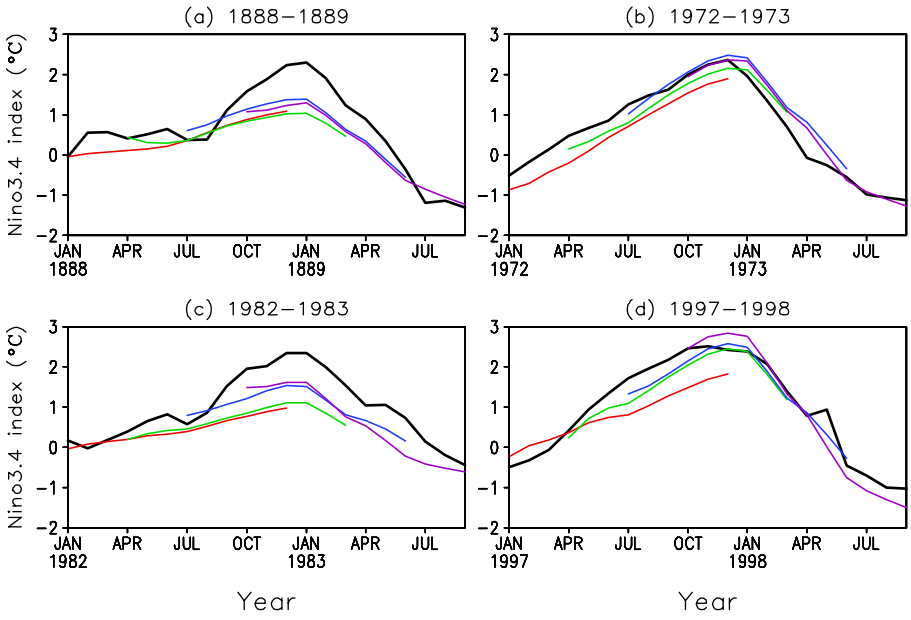

Fig. 4. Four of the largest El Niños since 1886. The thick black curves are observed Niño-3.4 SST anomalies, and the thin curves of red, green, blue and purple are ensemble mean predictions started respectively 12, 9, 6, and 3 months before the peak of each El Niño.

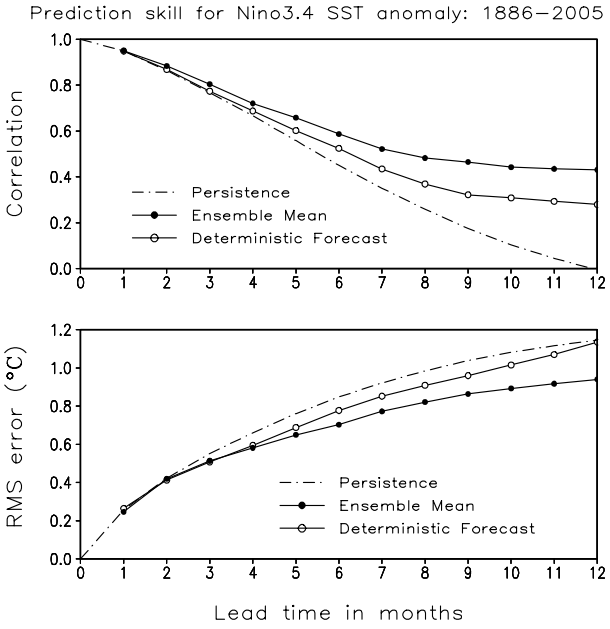

Fig. 5. Anomaly correlation (top) and RMS error (bottom) of the Niño-3.4 SST anomalies for the model ensemble mean hindcast (solid line with closed circle), the deterministic hindcast (solid line with open circle), and persistence (dot–dashed line) are shown as functions of lead time.

served strong El Niño events twelve months in advance, although some errors still exist in the forecasted onset and development, and in the magnitude of these events. The implication is that the signal components present in initial fields play a critical role in determining ENSO prediction skills (e.g., Peng and Kumar, 2005; Moore et al., 2006; Zheng et al., 2009).

Figure 5 shows the anomaly correlation and root mean square (RMS) error between observed and predicted average SST anomalies over the tropical Pacific Ocean Niño-3.4 region as a function of lead time. To compare with the original deterministic prediction skill, we also perform a prediction experiment whose initialization procedure is briefly described here, wherein the wind stress anomalies reconstructed from observed SST anomalies via a singular value decomposition (SVD) based model are used to integrate the ocean model over the whole forecast period to generate initial conditions for the dynamical component, and the SST anomaly model initial conditions are taken as the observed SST anomalies (Zhang et al., 2003). The skill scores for the ensemble mean hindcast are better than that of the original deterministic forecast scheme; both of hindcasts schemes have particularly high skill at short lead times and beat persistence for all lead times with a correlation coefficient of greater than 0.94 for the first month. Beyond 4-month lead time, there is a distinct difference of RMS errors between the ensemble mean hindcast and the original scheme. The RMS error of the ensemble mean remains smaller than 0.94°C over the 12-month prediction period, and is almost 0.2°C smaller than that of the original deterministic forecast scheme at a lead time of 12 months. Over the whole period, this improvement occurs because the advanced assimilation method can provide more dynamically consistent and accurate initial conditions than the original initialization method, and the ensemble mean can remove some unpredictable stochastic information.

3.2 *Probabilistic prediction skill*

We use Talagrand diagrams (also known as rank histograms) to evaluate whether the hindcast and the verifying observation are sampled from the same probability distribution (e.g., Talagrand et al., 1998; Hamill, 2001). The Talagrand diagrams are generated by ordering at each grid point the forecast values from each of the ensemble members from smallest to largest. For our full ensemble, with 100 members, this creates 101 intervals, and the value of the verifying observation then falls into one of the 101 categories. Figure 6 shows the Talagrand diagram for the SST anomalies over the Niño-3.4 region, and is a diagram of the frequencies as a function of the category index. For the SST anomalies, the distribution is flat, although the two extreme categories are somewhat higher than their adjacent categories. The 12-month lead hindcast is better in this respect than the 3-month lead hindcast, however. This indicates that the ensemble spread at longer lead is more reasonable. Also, there is a small shift of frequencies (i.e., the frequencies in the upper intervals are decreasing from shorter lead time to longer lead time, while the frequencies in the lower and middle intervals are increasing at the same time) of the verifying observation from the lower categories to the higher categories at all four lead times. The Talagrand diagrams indicate that the probability distribution of observations can be represented by the ensemble approach.

As described in section 2, our ensemble members are generated based on the hypothesis of a Gaussian distribution, but the standard normally distributed perturbations are processed at all model grids (not on regions), and thus we need to check whether probability distributions of the Niño-3.4 forecasted ensemble members accord with the Gaussian distribution. Figure 7 shows the normalized probability curve of the forecasted ensemble members over the Niño-3.4 region based on the entire 120-yr period. At different lead times, the forecasted ensemble members agree with the normal distribution quite well, and there are no

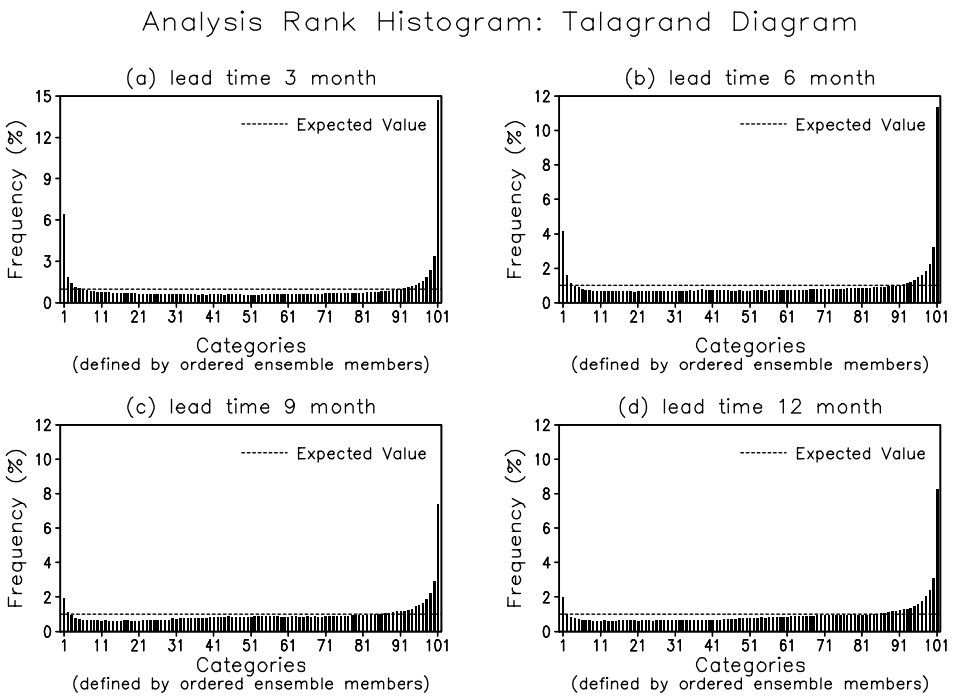

Fig. 6. Talagrand diagram for the full ensemble Niño-3.4 SST anomaly hindcast over the whole 120-year period: (a) 3-month lead time, (b) 6-month lead time, (c) 9-month lead time, and (d) 12-month lead time hindcasts. The dashed line marks the theoretical frequency for a perfectly reliable EPS.

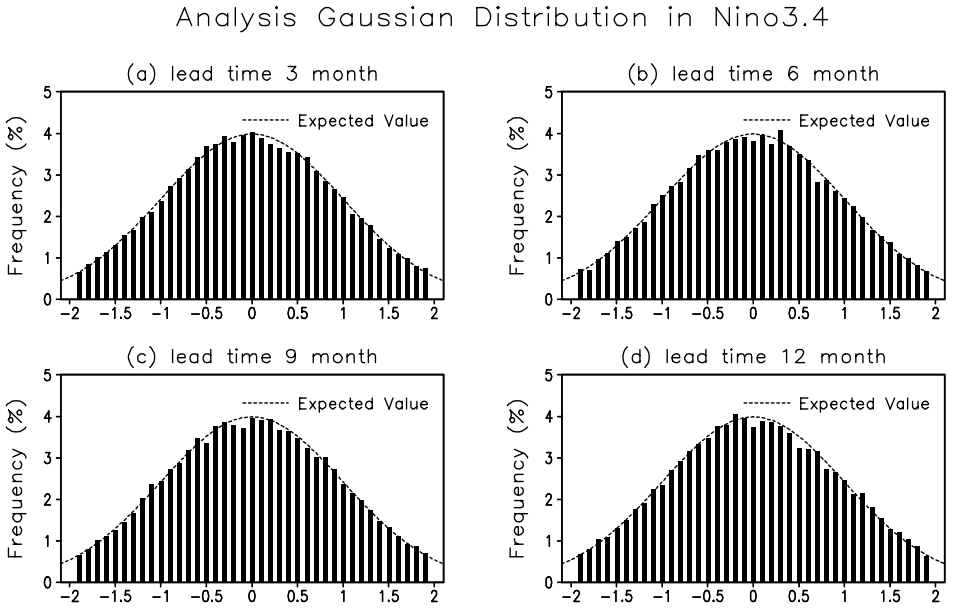

Fig. 7. Gaussian distribution diagram for the Niño-3.4 ensemble SST anomaly hindcast: (a) 3-month lead time, (b) 6-month lead time, (c) 9-month lead time, and (d) 12-month lead time. The dashed line marks a standard normal probability curve.

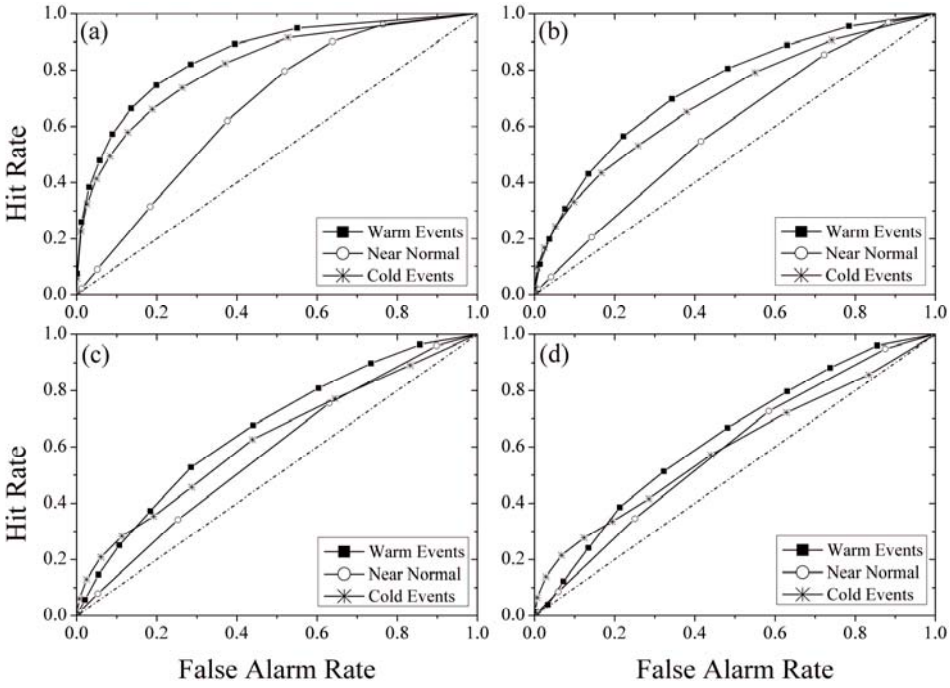

Fig. 8. Niño-3.4 ROC curves for a lead time of (a) 3 months, (b) 6 months, (c) 9 months, and (d) 12 months. Warm events (upper tercile) are denoted with closed squares, normal events (middle tercile) are denoted with open circles, and cold events (lower tercile) are denoted with asterisks.

double- or multi-modal peaks for the ensemble members. This indicates that the generation methods of the forecast ensemble members are reasonable, and the ensemble-mean forecast result is the most representative deterministic forecast, capable of illustrating the deterministic performance of the EPS.

To measure the probabilistic prediction skill more accurately, here we choose the method commonly referred to as relative operating characteristic (ROC; e.g., Mason and Graham, 1999) to measure the ensemble forecast performance by comparing the fraction of events that were properly forewarned (i.e., the hit rate) with the fraction of nonevents that occurred after a warning was issued (i.e., the false alarm rate). The ratios are determined from contingency tables and the events are predefined and expressed in binary terms. Given an ensemble of hindcasts, an ROC curve showing the different combinations of hit and false alarm rates given different forecast probabilities can be constructed. The ROC curve is useful for identifying optimum strategies for issuing warnings, by indicating the trade-off between false alarms and misses. Details and examples of the ROC calculation can be found in Mason and Graham (1999).

ROC curves for the Niño-3.4 hindcasts at lead times of 3, 6, 9, and 12 months are shown in Fig. 8. For all lead times, there are three curves representing three different event types: (i) warm events (upper tercile), (ii) cold events (lower tercile), and (iii) normal events (middle tercile), where both the retrospective forecasts and the observations have been normalized by their local standard deviation. An ideal probabilistic forecast system would have relatively large hit rates and small false alarm rates so that all the points on the ROC curve would cluster in the upper-left corner of the diagram (e.g., Kirtman, 2003). For a relatively poor forecast system, all the points of the ROC curve would lie very close to the dashed diagonal line indicating that the hit rate and the false alarm rate were nearly the same (i.e., no skill). Akin to previous studies (e.g., Kirtman, 2003; DeWitt, 2005), the EPS has relatively higher skill for the warm events and cold events, and it has relatively lower skill for the neutral events. For 3- and 6-month lead times, both warm and cold events are fairly well predicted. The false alarm rates are low and the hit rates are relatively high when the agreement among the ensemble members is relatively large. For a normal event forecast, the 3-month lead time also has some skill although smaller than for the extremes, whereas for 6-, 9-, and 12-month leads, the ROC curve lies close to the diagonal, indicating little skill. These results indicate that the EPS can capture and predict big SST anomaly signals or extreme events over the Niño-3.4 region in different sea-

sons quite well (Zhang et al., 2005), and the model is able to predict extreme events. At 9- and 12-month lead times, there is a considerable drop in skill. High confidence forecasts for warm and cold events are only marginally better than those for normal events, suggesting that a confident forecast for a warm or cold event at 12 months lead time is still not particularly useful. This is also appeared to be the case with the earlier studies (e.g., Barnston et al., 1999; Kirtman, 2003).

The ability to easily verify the hindcast skill of warm events and cold events separately is one of the advantages of the ROC calculation, and thus we further used the ROC area to verify the probabilistic skills of the EPS for the three different events. The ROC area is the area under the ROC curve, and a perfect forecast system would have a ROC area of 1 while a system with no ability to distinguish in advance between different events would have a score of 0.5. Figure 9 shows the ROC area of SST anomalies for warm events, normal events, and cold events over the Niño-3.4 region, as a function of lead time. Similar to the analysis results above for the 120-yr hindcast, the ROC areas for both the warm and cold events are clearly higher than that of the neutral events during the 12-month forecast period. This also indicates that a large (initial) signal can lead to a reliable prediction and high prediction skill, and that for small predicted signals, the evolution of predicted SST anomalies in our EPS might present a more chaotic evolution, which would degrade prediction skill and induce obvious decreases of predictability (e.g., Zheng et al., 2009).

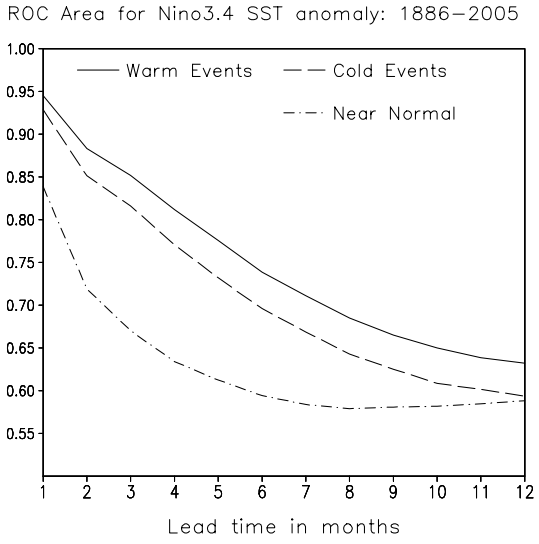

Fig. 9. ROC area of SST anomalies for warm events (solid line), normal events (dot-dashed line), and cold events (dashed line) over the Niño-3.4 region, shown as a function of lead time.

4. Variation of ENSO ensemble predictability

Similar to previous studies (e.g., Chen et al., 2004), Fig. 3 shows that the characteristics of the interannual variability obviously have changed with time. To examine the possible interdecadal variation of ENSO ensemble predictability, in this section, we calculate both deterministic and probabilistic prediction skills of 6 sub-periods of 20 years each.

4.1 *Deterministic predictability*

ENSO's deterministic predictability depends on the time period in which it is estimated (Balmaseda et al., 1995; Kirtman and Schopf, 1998; Chen et al., 2004). This is also evident in Fig. 10. For the six sub-periods of 20-year each, both anomaly correlation and RMS error vary over significant ranges, especially

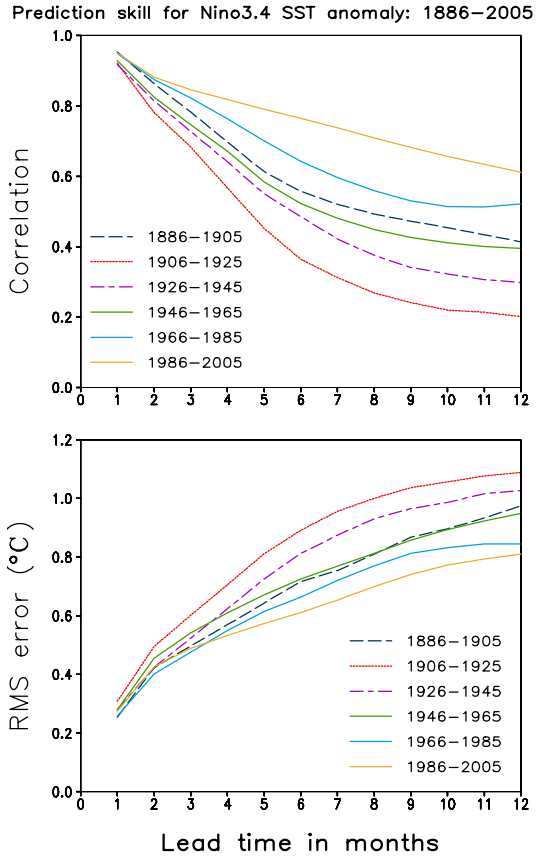

Fig. 10. Anomaly correlation (top) and RMS error (bottom) between the observed and the ensemble-mean predicted values of the Niño-3.4 index. These are shown as a function of lead time, for six consecutive 20-yr periods since 1886.

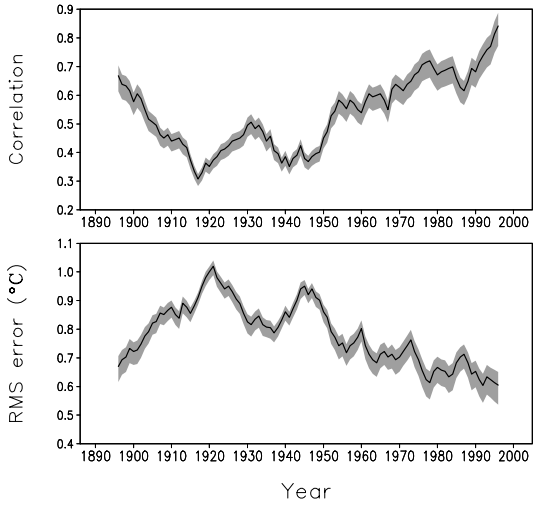

Fig. 11. The averaged correlation (top) and RMS error (bottom) between the observed and the predicted Niño-3.4 SST anomalies at 6-month lead time. The correlation and RMS error are computed at each running window of 20-yr period from 1886 to 2005. The shaded area represents the 95% confidence interval via bootstrap procedures.

at longer lead times. For example, high prediction skills appear in the late 19th century and the middle-late 20th century (i.e., 1886–1905, 1966–85, and 1986–2005), these periods are dominated by strong and regular ENSO events. The high scores for the 1966–85 and 1986–2005 periods might not be surprising because the model is trained using data from part of this period, and the high scores for the 1886–1905 period, which is free of artificial skill, indicate that the large El Niño and La Niña events can be highly predictable, even initialized with only SST anomaly data. But, the periods of 1906–25, 1926–45, and 1946–65 have relatively low prediction skills. The lower skill in these periods is consistent with there being fewer and smaller events to predict.

The consistent temporal variations of the deterministic prediction skills of the EPS are further displayed in Fig. 11, which shows the averaged correlation and RMS error at 6-month lead time measured by a running window of 20-yr from 1886 to 2005 (i.e., 1886–1905, 1887–1906, $\cdots$, 1986–2005). For example, the skill at 1896 was calculated using the samples from 1886–1905. The 20-yr window is shifted by one year for each time starting from 1886 to 2005. There is a striking interdecadal variation of ENSO deterministic predictability (in both the correlation and RMS error) over the past 120 years from 1886 to 2005 in the EPS. Generally, there is high predictability in the late 19th century and in the middle-late 20th century, and a low predictability from 1906 to 1951 (correlation is lower than 0.50).

A bootstrapped resampling procedure (Efron and Tibshirani, 1986) is also used to derive useful confidence limits for the skill scores in order to allow meaningful statistical conclusions to be drawn from these comparisons. The shaded area in Fig. 11 represents the 95% confidence interval computed using bootstrap procedures, and indicates the uncertainty of verification sampling. Considering the confidence interval, both the correlation and RMS error results have shown that verification sampling have smaller influence on the forecast skill scores than differences between the skills in the different decades. This might because the ensemble members match the Gaussian distribution quite well at different lead times (Fig. 7), and thus that the resampling process makes little adjustments on the distribution of the forecasted ensembles.

4.2 *Probabilistic predictability*

To verify the variations of the probabilistic predictability of the EPS, we examine the temporal changes of the ROC area for the three different event types. Figure 12 shows the ROC area in the Niño-3.4 region for warm events, cold events, and normal events in six consecutive 20-yr periods since 1886. Obviously, the ENSO probabilistic predictability for different events also depends on the time period. For warm and cold events, high probabilistic prediction skills still appear in the late 19th century (i.e., 1886–1905, when the skill for warm events is only a little higher than that in the early 20th century) and the middle-late 20th century (1966–85 and 1986–2005), with the highest skills for warm events in the sub-period 1966–85, and highest probabilistic prediction skills for cold events in sub-period 1986–2005.

In order to illustrate the interdecadal features of the probabilistic prediction skills more clearly, we further verify the consistent temporal variations of the probabilistic prediction skills of the EPS. Figure 13 shows the ROC area for the three different event types at 6-month lead time measured by a running window of 20-yr from 1886 to 2005. For the warm events, the highest skill appears in the late 20th century, and the lowest skill appears from 1910 to 1930. For the cold events, the highest skill also appears in the late 20th century, and the lowest skill emerges from 1920 to 1950. For the neutral events, the 20-yr averaged skill decreases from 1896 to 1910, and takes a linear increasing feature from 1910 to 1995. The uncertainty of verification sampling is also shown in Fig. 13 using the 95% confidence interval computed via bootstrap procedures. Considering the confidence interval, for the three different events, the ROC analysis results also show that verification sampling has smaller influence

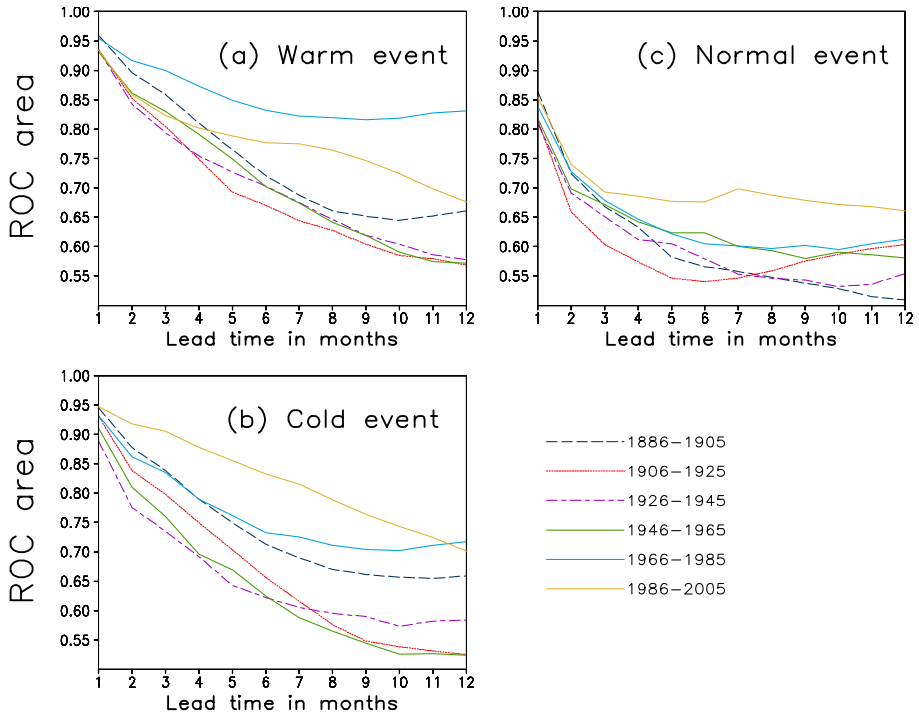

Fig. 12. ROC area in the Niño-3.4 region for (a) warm events (upper tercile), (b) cold events (middle tercile), and (c) normal events (lower tercile). These are shown as a function of lead time, for six consecutive 20-yr periods since 1886.

on the forecast skill scores than differences between the skills in the different decades. Compared to Fig. 11, the probabilistic verification uncertainties are larger than the deterministic verification uncertainties. However, in summary, for three different the event types, there are still obvious interdecadal variations of ENSO probabilistic predictability over the past 120 years from 1886 to 2005 in the EPS.

5. Discussions and conclusions

In this paper, long-term retrospective ensemble forecasts using 100 members covering the past 120 years are performed with an EPS. With the assimilation of only a historic SST dataset, the prediction skills of the EPS are verified in both a deterministic and probabilistic sense, and the EPS displays useful prediction skill. An interesting finding from the retrospective ensemble forecasts is that the EPS showed interdecadal variations in both deterministic and probabilistic prediction skills. Both deterministic and probabilistic prediction skills are high in the late 19th century from 1886 to 1905, and then decline with time, reaching a minimum around 1910–50, beyond which skill rebounds and increases with time from the 1960s onward. The EPS has relatively high prediction skill (but also including some artificial skill) from the 1960s onward, especially in the late 20th century from 1986 to 2005. These results are similar to previous studies (e.g., Chen et al., 2004; Tang et al., 2008), although there are still some differences in the prediction skills among different models [which is also shown in Tang et al. (2008)]. However, the trends of the interdecadal variations in different models appear comparable (i.e., higher predictability in the late 19th century and in the middle-late 20th century, and a lower predictability in early 20th century). These results all indicate that the interdecadal variability of ENSO (deterministic and probabilistic) predictability exists generally, and is not model dependent.

However, it should be noted that the theoretical framework discussed in this study is based on a relatively simple EPS, and one could argue that our analysis is not complete since we only use SST data. One serious question is whether or not this interdecadal variation in predictability discussed in this paper is due mainly to the differences of the quality of the data in different periods. For example, one can guess that the high prediction skill for the period from 1966 to 2005 is probably due to better data quality because of improvements of observation systems and the fact that the model was trained using the data from part of this period.

To explore this, we can examine the simulation skill

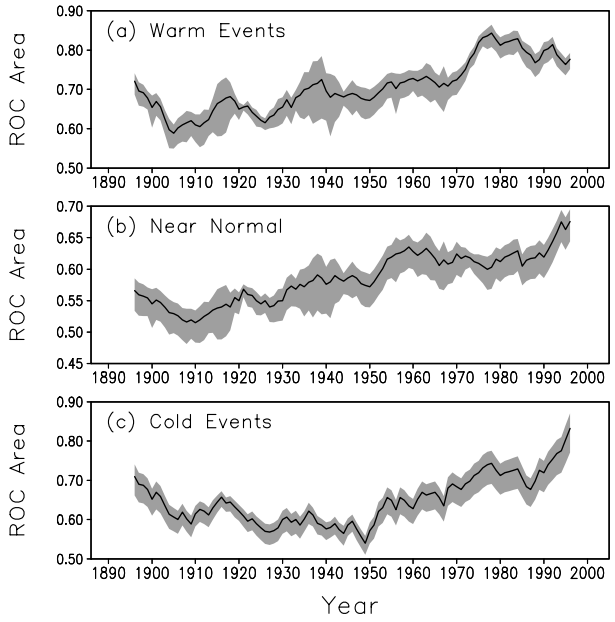

Fig. 13. The averaged ROC area of Niño-3.4 SST anomalies for (a) warm events, (b) normal events, and (c) cold events at 6-month lead time, respectively. The ROC areas are computed at each running window of 20-yr period from 1886 to 2005. The shaded area represents the 95% confidence interval via bootstrap procedures.

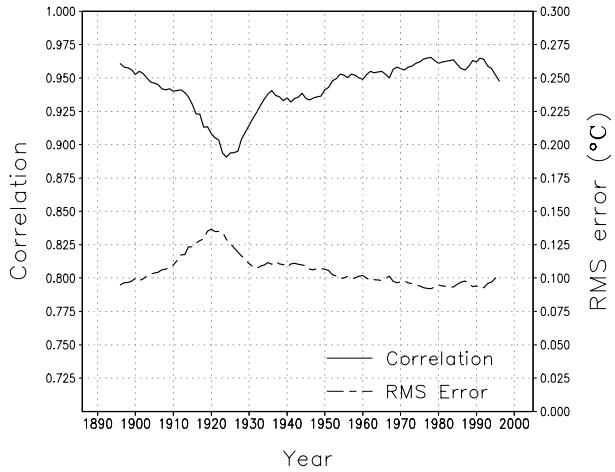

Fig. 14. The averaged correlation (solid line) and RMS error (dashed line) between the observed and the simulated Niño-3.4 SST anomalies forced by the reconstructed wind stress anomalies. The correlation and RMS error are computed at each running window of 20-yr period from 1886 to 2005.

of the model forced by the reconstructed wind stress from the atmospheric τ model. Thus, the quality of initial conditions of predictions and the model performance can be indicated by the simulation skill, with both inherent to the data quality. And the existence of an impact of the data quality on the model's simulation skill will be mostly felt through the quality of initial conditions, such as initial SST anomalies. Figure 14 shows the averaged correlation and RMS error between the observed and the simulated Niño-3.4 SST anomalies forced by the reconstructed wind stress anomalies at each running window of 20-yr period from 1886 to 2005, and indicates that the interdecadal difference of the simulation skill is not large in the model. The magnitude of variation is about 0.1 from maximum to minimum during the entire period for both correlation and RMS error (units: °C). A comparison between Figs. 11 and 14 reveals that the interdecadal variation in predictability does not agree with that in the simulation skill. Thus, interdecadal variation in predictability is not due to model performance associated with data quality. This is further suggested by the fact that noticeably higher prediction skill also occurs during the period from 1886 to 1905.

The results of the analyses in this paper motivate us to further investigate the possible reasons and sources of limited ENSO predictability in detail. These concerns in future works need to be addressed through more comprehensive analyses, and other possible sources (besides the ENSO signal) of controlling ENSO predictability also need to be further discussed, such as nonlinearity and stochastic noise. Nevertheless, this study is to date the first work to discuss both ENSO deterministic and probabilistic predictabilities using ensemble forecasts and long-term predictions. The results and conclusions found in this EPS might be helpful for the study of ENSO predictability.

Acknowledgements. The authors wish to thank the two anonymous reviewers for their very helpful comments and suggestions. This research is supported by the Chinese Academy of Science (Grant No. KZCX2-YW-202), National Basic Research Program of China (2006CB403600) and National Natural Science Foundation of China (Grant Nos. 40437017 and 40805033).

REFERENCES

Balmaseda, M. A., M. K. Davey, and D. L. T. Anderson, 1995: Decadal and seasonal dependence of ENSO prediction skill. *J. Climate*, **8**, 2705–2715.

Barnston, A. G., M. Glantz, and Y. He, 1999: Predictive skill of statistical and dynamical climate models in SST forecasts during the 1997–98 El Niño and the 1998 La Niña onset. *Bull. Amer. Meteor. Soc.*, **80**, 217–243.

Chen, D., M. A. Cane, A. Kaplan, S. E. Zebiak, and D. Huang, 2004: Predictability of El Niño in the past 148 years. *Nature*, **428**, 733–736.

DeWitt, D. G., 2005: Retrospective forecasts of interan-

nual sea surface temperature anomalies from 1982 to present using a directly coupled atmosphere-ocean general circulation model. *Mon. Wea. Rev.*, **133**, 2972–2995.

Efron, B., and R. Tibshirani, 1986: Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, **1**, 54–77.

Evensen, G., 2003: The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, **53**, 343–367.

Evensen, G., 2004: Sampling strategies and square root analysis schemes for the EnKF. *Ocean Dynamics*, **54**, 539–560.

Goswami, B. N., and J. Shukla, 1991: Predictability of a coupled ocean-atmosphere model. *J. Climate*, **4**, 3–22.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560.

Kirtman, B. P., and P. S. Schopf, 1998: Decadal variability in ENSO predictability and prediction. *J. Climate*, **11**, 2804–2822.

Kirtman, B. P., 2003: The COLA anomaly coupled model: Ensemble ENSO prediction. *Mon. Wea. Rev.*, **131**, 2324–2341.

Keenlyside, N., and R. Kleeman, 2002: On the annual cycle of the zonal currents in the equatorial Pacific. *J. Geophys. Res.*, **107**, doi: 10.1029/2000JC0007111.

Latif, M., and Coauthors, 1998: A review of the predictability and prediction of ENSO. *J. Geophys. Res.*, **103**, 14,375–14,393.

Mason, S. J., and N. E. Graham, 1999: Conditional probabilities, relative operating characteristics, and relative operating levels. *Wea. Forecasting*, **14**, 713–725.

McCreary, J. P., 1981: A linear stratified ocean model of the equatorial undercurrent. *Philosophical Transactions of the Royal Society* (*London*), **298**, 603–635.

Moore, A., and Coauthors, 2006: Optimal forcing patterns for coupled models of ENSO. *J. Climate*, **19**, 4683–4699.

Peng, P., and A. Kumar, 2005: A large ensemble analysis of the influence of tropical SSTs on seasonal atmospheric variability. *J. Climate*, **15**, 1068–1085.

Smith, T. M., and R. W. Reynolds, 2004: Improved extended reconstruction of SST (1854–1997). *J. Climate*, **17**, 2466–2477.

Smith, T. M., R. W. Reynolds, T. C. Peterson, and J. Lawrimore, 2008: Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880–2006). *J. Climate*, **21**, 2283–2296.

Talagrand, O., R. Vautard, and B. Strauss, 1998: Evaluation of probabilistic prediction systems. *Proc. Seminar on Predictability*, Reading, United Kingdom, ECMWF, 1–26.

Tang, Y., and W. W. Hsieh, 2003: ENSO simulation and predictions using a hybrid coupled model with data assimilation. *J. Meteor. Soc. Japan*, **81**, 1–19.

Tang, Y., Z. Deng, X. Zhou, and Y. Cheng, 2008: Interdecadal variation of ENSO predictability in multiple models. *J. Climate*, **21**, 4811–4833.

Tziperman, E., L. Stone, M. A. Cane, and H. Jarosh, 1994: El Niño chaos: Overlapping of resonances between the seasonal cycle and the Pacific ocean-atmosphere oscillator. *Science*, **264**, 72–74.

Zhang, R.-H., S. E. Zebiak, R. Kleeman, and N. Keenlyside, 2003: A new intermediate coupled model for El Niño simulation and prediction. *Geophys. Res. Lett.*, **30**(19), 2012, doi: 10.1029/2003GL018010.

Zhang, R.-H., S. E. Zebiak, R. Kleeman, and N. Keenlyside, 2005: Retrospective El Niño forecast using an improved intermediate coupled model. *Mon. Wea. Rev.*, **133**, 2777–2802.

Zheng, F., J. Zhu, R.-H. Zhang, and G.-Q. Zhou, 2006a: Ensemble hindcasts of SST anomalies in the tropical Pacific using an intermediate coupled model. *Geophys. Res. Lett.*, **33**, L19604, doi: 10.1029/2006GL026994.

Zheng, F., J. Zhu, R.-H. Zhang, and G.-Q. Zhou, 2006b: Improved ENSO forecasts by assimilating sea surface temperature observations into an intermediate coupled model. *Adv. Atmos. Sci.*, **23**(4), 615–624, doi: 10.1007/s00376-006-0615-z.

Zheng, F., 2007: Research on ENSO ensemble predictions. Ph. D. dissertation, Institute of Atmospheric Physics, Chinese Academy of Sciences, 159pp. (in Chinese)

Zheng, F., J. Zhu, and R.-H. Zhang, 2007: Impact of altimetry data on ENSO ensemble initializations and predictions. *Geophys. Res. Lett.*, **34**, L13611, doi: 10.1029/2007GL030451.

Zheng, F., and J. Zhu, 2008: Balanced multivariate model errors of an intermediate coupled model for ensemble Kalman filter data assimilation. *J. Geophys. Res.*, **113**, C07002, doi: 10.1029/2007JC004621.

Zheng, F., H. Wang, and J. Zhu, 2009: Impacts on ENSO ensemble prediction: Initial-error perturbations vs. model-error perturbations. *Chinese Science Bulletin*. (in press)