Ensemble Construction and Verification of the Probabilistic ENSO Prediction in the LDEO5 Model

YANJIE CHENG, YOUMIN TANG, AND PETER JACKSON

Environmental Science and Engineering, University of Northern British Columbia, Prince George, British Columbia, Canada

DAKE CHEN

Lamont-Doherty Earth Observatory of Columbia University, Palisades, New York, and State Key Laboratory of Satellite Ocean Environment Dynamics, Hangzhou, China

ZIWANG DENG

Environmental Science and Engineering, University of Northern British Columbia, Prince George, British Columbia, Canada

(Manuscript received 30 September 2009, in final form 19 May 2010)

ABSTRACT

El Niño–Southern Oscillation (ENSO) retrospective ensemble-based probabilistic predictions were performed for the period of 1856–2003 using the Lamont-Doherty Earth Observatory, version 5 (LDEO5), model. To obtain more reliable and skillful ENSO probabilistic predictions, first, four ensemble construction strategies were investigated: (i) the optimal initial perturbation with singular vector of sea surface temperature anomaly (SSTA), (ii) the realistic high-frequency anomalous winds, (iii) the stochastic optimal pattern of anomalous winds, and (iv) a combination of the first and the third strategy. Second, verifications were conducted to examine the reliability and resolution of the probabilistic forecasts provided by the four methods. Results suggest that reliability of ENSO probabilistic forecast is more sensitive to the choice of ensemble construction strategy than the resolution, and a reliable and skillful ENSO probabilistic prediction system may not necessarily have the best deterministic prediction skills. Among these ensemble construction methods, the fourth strategy produces the most reliable and skillful ENSO probabilistic prediction, benefiting from the joint contributions of the stochastic optimal winds and the singular vector of SSTA. In particular, the stochastic optimal winds play an important role in improving the ENSO probabilistic predictability for the LDEO5 model.

1. Introduction

The loss of ENSO predictability in a numerical model generally depends on uncertainties due to (i) errors in the initial conditions, (ii) model errors, and (iii) unexpected external stochastic noise (e.g., Moore and Kleeman 1998). These uncertainties develop during the forecast period as lead time increases, eventually rendering the forecast no better than climatology. As a response to the limitations imposed by these uncertainties, a more useful forecast strategy is to perform ensemble predictions and evaluate the uncertainties of the forecast system using probabilistic methods (Chen and Cane 2008).

Compared with a single forecast starting from the best initial conditions, an ensemble forecast has many advantages. First, ensemble averaging acts as a nonlinear filter; it removes less predictable parts and keeps more predictable features among the ensemble members (e.g., Leith 1974). A properly designed ensemble has higher skill than that of individual ensemble members in a statistical sense (Toth and Kalnay 1997). Second, ensemble prediction provides a practical tool for estimating the possible uncertainties in a forecast system. Ensemble forecasts can provide additional information, such as the probability distribution function (PDF) of forecast, ensemble-based potential skill measures (i.e., ensemble mean, ensemble spread, and ensemble ratio), and probabilistic skill measures, which are useful in decision

Corresponding author address: Dr. Youmin Tang, Environmental Science and Engineering, University of Northern British Columbia, 3333 University Way, Prince George, BC V2N 4Z9, Canada.
E-mail: ytang@unbc.ca

DOI: 10.1175/2010JCLI3453.1

making. It is shown that probability forecasts have greater potential economic value than corresponding single deterministic forecasts with uncertain accuracy (e.g., Palmer 2000).

To perform an ensemble-based ENSO probabilistic forecast, the crucial issue is to design a reliable and high-resolution ensemble prediction strategy that should include the major uncertainties of a forecast system. Many strategies have been used in the ensemble construction of weather forecasts and seasonal climate predictions. For example, some strategies are dynamically constrained methods such as the breeding vector (BV; Toth and Kalnay 1993), the ensemble transform (ET; an improved version of the BV; Bishop and Toth 1999; Wei et al. 2008), and the singular vector (SV; e.g., Lorenz 1965; Palmer 1993), which are used to optimally perturb the initial conditions for constructing ensemble forecasts. Other methods are also used to obtain the ''best'' initial conditions in ensemble constructions: the ensemble Kalman filter (EnKF; Evensen 1994, 2003), the ensemble transform Kalman filter (ETKF; Bishop et al. 2001; Wang and Bishop 2003), and the perturbed observation (PO; Houtekamer and Derome 1995). Using the three-parameter Lorenz (1963) model, Anderson (1997) found that random perturbations produce more skillful ensembles than BV and SV. Houtekamer and Derome (1995) found little difference in the quality of the ensemble mean forecasts between the BV, SV, and PO methods using a quasigeostrophic model. Hamill et al. (2000) compared BV, SV, and PO in a quasigeostrophic channel model. They found that the PO method is better than the BV and SV method. Descamps and Talagrand (2007) compared four strategies in the Lorenz (1963) model and a three-level atmospheric model and concluded that the relative performance, from best to worst, of these strategies was in the order EnKF $>$ ETKF $>$ BV $>$ SV.

Generally two kinds of strategies are used to produce optimal perturbations for ensemble-based ENSO predictability studies: (i) perturbation of the initial conditions and (ii) perturbations in the stochastic atmospheric noise through the whole forecast period. In addition, perturbation can be applied on model parameters for considering errors existing in physical/dynamical parameterizations, or a superensemble is constructed using multiple models (e.g., Kirtman and Min 2009). The first strategy was often used by SV analysis (e.g., Lorenz 1965; Chen et al. 1997; Xue et al. 1997a,b; Battisti 1988; Fan et al. 2000; Cai et al. 2003; Tang et al. 2006; Cheng et al. 2010a,b), whereas the second strategy was performed in the framework of the stochastic optimal theory (e.g., Kleeman and Moore 1997; Moore and Kleeman 1998, 1999; Tang et al. 2005). Significant progress has been made in using these optimal perturbations to study ENSO predictability as cited above. However, these previous studies mainly focused on the optimal error growth of ENSO deterministic predictions. The impact of perturbation construction on the ensemble probabilistic predictions has not been well addressed, especially using long-term retrospective ensemble predictions over periods as long as 100 yr. In this study, we will explore this issue using SV-based perturbation methods. So far, the SV method itself has not been well examined in the framework of ENSO ensemble probabilistic prediction. One reason is that the SV analysis needs a tangent linear model (TLM), which is often technically difficult. Another reason is the lack of a long-term forcing data for initializing predictions, so that previous retrospective predictions were limited to a short period of 20–40 yr, with a rather limited number of ENSO cycles. This may preclude statistically robust conclusions. Chen et al. (2004) used Kaplan SSTA reanalysis data and the Zebiak and Cane (ZC) model [i.e., the Lamont-Doherty Earth Observatory version 5 (LDEO5)] to perform a 148-yr hindcast experiment for the period of 1856–2003. They successfully predicted all of the prominent El Niño events during this period at lead times of up to 2 yr, with the SST being the only data used for model initialization. Tang et al. (2008a) further analyzed the interdecadal variation in ENSO prediction skill from 1881 to 2000 using multiple models. These retrospective ENSO predictions not only allow us to achieve a robust and stable study of statistical predictability of ENSO but also demonstrate that the long-term SSTA data are of good quality. Recently, a fully physically based TLM was constructed for the LDEO5 model, and singular vector analyses were performed for the 148-yr period from 1856 to 2003 in Cheng et al. (2010a). The long-term SVs obtained in Cheng et al. (2010a) makes it possible to construct ensemble predictions with the LDEO5 model, so that the shape of the forecast PDF that describes the prediction uncertainty can be estimated, and the probabilistic nature of ENSO predictability can be explored.

Another issue is the role of stochastic atmospheric noise in ensemble ENSO predictions. It has been well recognized that stochastic atmospheric forcing associated with synoptic-to-intraseasonal variability is critical in forming, developing, and maintaining ENSO cycles (e.g., Penland and Sardeshmukh 1995; Kleeman and Moore 1997; Eckert and Latif 1997; Blanke et al. 1997; Kirtman and Schopf 1998; Moore and Kleeman 1999; Thompson and Battisti 2000; Fluegel et al. 2004; Moore et al. 2006; Philip and van Oldenborgh 2009; Eisenman et al. 2005; Gebbie et al. 2007; Tziperman and Yu 2007; Zavala-Garay et al. 2005; Perez et al. 2005; Zhang and Busalacchi 2008). These previous studies addressed the

role of stochastic forcing in ENSO formation and development, namely, answering a central question on the ENSO mechanism, ''Is ENSO a nonlinear system or a linear system driven by stochastic forcing?'' For example, Zhang and Busalacchi (2008) suggested that tropical instability wave-induced wind feedback to the ocean can have a rectified effect on large-scale mean ocean state and interannual variability. The realistic stochastic forcing that was found to have important impact on ENSO mainly included synoptic-scale atmospheric processes and high-frequency variability such as westerly wind bursts and the Madden–Julian oscillation (MJO). There have been already some works to discuss the loss of ENSO predictability (deterministic skill) due to these stochastic forcing. However, the importance and significance of stochastic forcing on ENSO predictability have not been well addressed in the sense of probabilistic prediction skill. Thus, it is not very clear so far how the stochastic atmospheric noise impacts ENSO probabilistic predictions.

An important task associated with ensemble construction is to evaluate an ensemble-based probabilistic prediction system by probabilistic verification methods, from which the performance of the prediction system and the ensemble construction method can be quantitatively evaluated. Probabilistic verification is known as an important complement to deterministic verification, which provides a useful and quantitative way to measure uncertainty (Palmer 2000; Kirtman 2003). In contrast with the traditional prediction skill measures such as anomaly correlation R skill and root-mean-square error (RMSE) skill, the verification of an ensemble-based probabilistic forecast system focuses on measuring two properties, reliability and resolution, which are the two most important characteristics of a probabilistic forecast system (Toth et al. 2003). An introduction of these properties and the probabilistic verification methods will be described in section 4.

This study will introduce both initial condition uncertainty and additive stochastic atmospheric noise into the LDEO5 model and examine their impacts on ENSO probabilistic prediction. It is unrealistic to evaluate all ensemble construction methods available for ENSO probabilistic prediction, so we focus on evaluating four methods, chosen based on previous studies as referred to previously: (i) initial condition perturbation using the singular vector of SSTA (SV1_sst), (ii) realistic stochastic winds as a continuous external forcing during the forecast period (UV_realstoc), (iii) stochastic optimal winds (SO1_wind) as a continuous external forcing during the forecast period, (iv) a combination of the first method SV1_sst and the third method SO1_wind (SO1_wind+ SV1_sst). Several probabilistic verification methods are used to evaluate the reliability and resolution of ensemble-based probabilistic ENSO predictions, including the reliability diagram (RD) and the Brier skill score, the ranked probability score (RPS), and the ranked probability skill score (RPSS). Emphasis is placed on assessing which ensemble construction method provides more reliable and skillful probabilistic ENSO predictions.

This paper is structured as follows: section 2 briefly introduces the LDEO5 model and the metrics of ensemble prediction skill. Section 3 discusses the four ensemble construction methods used in this study. Section 4 gives the introduction of probabilistic forecast verification methods. Section 5 presents the ensemble prediction results followed by the conclusions and discussion in section 6.

2. Model and ensemble forecast

a. The LDEO5 model

The model used in this study is the ZC model (Zebiak and Cane 1987), which has been widely applied for ENSO simulation and prediction. LDEO5 is the latest version of the ZC model (Chen et al. 2004). The atmosphere dynamics follows Gill (1980) using steady-state, linear shallow-water equations. The circulation is forced by a heating anomaly that depends on the SST anomaly and moisture convergence. The ocean dynamics uses the reduced-gravity model, and ocean currents were generated by spinning up the model with monthly wind. The thermodynamics describe the SST anomaly and heat flux change. The model time step is 10 days. The spatial region is focused on the tropical Pacific Ocean (28.75°S–28.75°N, 124°E–80°W). The grid for ocean dynamics is 2° longitude × 0.5° latitude, and the grid for SST physics and the atmospheric model is 5.625° longitude × 2° latitude.

The SSTA dataset used in this study is a reconstructed analysis data by Kaplan et al. (1998) with the period from January 1856 to December 2003. It is the only an oceanic dataset available for initializing long-term retrospective ENSO prediction over 100 yr. With the initialization of the SSTA dataset, the LDEO5 model successfully predicted all of the prominent El Niño events during at lead times of up to 2 yr and achieved a good hindcast skill (e.g., Chen et al. 2004; Tang et al. 2008a). Note that in the coupled initialization procedure of the LDEO forecast system, assimilated SST data are not simply putting a constraint on the ocean model with SST observations; they translate to surface wind field and subsurface ocean memory.

There are two model output statistics (MOS) schemes to correct model bias in the LDEO5. One scheme is for SST, and the other is applied to thermocline depth and winds. Bias correction terms are given at each time step

(Chen et al. 2000). With the two statistical bias correction schemes, the imbalance among those model variables (e.g., SST, thermocline depth, and winds) due to SST assimilation or perturbation of initial SST in the framework of ensemble can be expected to adjust quickly during the prediction period.

b. Metrics for ensemble prediction deterministic skill

Several ensemble construction schemes are designed in this study, focusing on different aspects of uncertainties related to the predictability (i.e., the initial conditions and stochastically external forcing). These ensemble retrospective ENSO predictions were performed by perturbing SST or wind, or both, using a given method as described in section 3. The model is initialized by only the assimilation of SST every month for 1856–2003 from Chen et al. (2004), thus a total of 148 yr $\times$ 12 months yr^{-1} (=1776) forecast initial conditions were obtained. From each initial time, an ensemble forecast was performed with the ensemble size M of 100, and for a period of 24 months. Thus, there are a total of 1776 months $\times$ 100 members $\times$ 24 months lead-time (=4 262 400) forecasts for the ensemble experiment of a given ensemble construction method.

In this study, we use the error of the ensemble mean ($\mathrm{RMSE_{EM}}$) and ensemble spread to assess ensemble deterministic prediction skill, defined by

$$\mathrm{SPREAD}(i,t) = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}[T_i^p(m,t) - \mathrm{EM}(i,t)]^2}, \tag{1}$$

$$\mathrm{EM}(i,t) = \frac{1}{M}\sum_{m=1}^{M=100} T_i^p(m,t), \tag{2}$$

where EM is the ensemble mean, a function of initial time i and lead time t; M is the ensemble size (i.e., 100 here); and T is the index of Niño-3.4 SSTA, with the superscripts p and o denoting predictions (forecasts) and observations, respectively. Here N is the number of initial conditions used (N = 1776):

$$\mathrm{SPRD}(t) = \frac{1}{N}\sum_{i=1}^{i=N}\mathrm{SPREAD}(i,t), \tag{3}$$

where the SPRD in (3) is the averaged ensemble spread over all the initial times, it is a function of lead time t only.

3. Strategies of ensemble construction

a. Perturbation of initial condition with SV of SSTA

In Cheng et al. (2010a), SV analysis was performed for the period 1856–2003 using the LDEO5 model. The leading singular vectors (SV1s), representing the optimal growth pattern of initial perturbations/errors, were obtained by perturbing the constructed TLM of the LDEO5 model. It was found that the first singular vectors of SSTA are dominated by a west–east dipole spanning most of the equatorial Pacific, with one center located in the east and the other in the central Pacific. The SV1s are less sensitive to initial conditions (i.e., are independent of seasons and decades). Thus, we will use the 148-yr-averaged SV1 of SSTA (denoted by SV1_sst) to perturb all initial conditions. As found in Cheng et al. (2010a), the fastest perturbation growth rate occurs at a 9–12-month lead in the LEDO5 model. Correspondingly, the prediction RMSE skill varies slowly with lead time after 12-month leads (Chen et al. 2004; Chen and Cane 2008). This motivates us to choose the SV1_sst of the 12-month lead in the following discussion. Note that the ensemble construction by two or more SV patterns does not show higher resolution or reliability than that constructed from the SV1 alone (not shown); thus, only the SV1-based ensemble is used, so that we perturbed the initial model SST by the SV1_sst pattern. The construction of initial perturbation Y can be expressed by (4), where random numbers X were normalized, and α is a constant value controlling the perturbation magnitude, set to 0.25 here according to Karspeck et al. (2006):

$$Y = \mathrm{SV1\_sst} \times X \times \alpha. \tag{4}$$

b. Realistic stochastic winds

In this study, we use two methods to generate the stochastic wind perturbations: high-frequency (<90 days) realistic winds and stochastic optimal winds. The first of these, denoted by UV_realstoc, is our second ensemble construction strategy. A dataset of the atmospheric high-frequency components were first obtained by applying a 3-month high-pass filter to the National Centers for Environmental Prediction (NCEP) daily wind dataset from 1948 to 2000 (Deng and Tang 2008). This dataset, referred to as the noise dataset, realistically represents all possible temporal and spatial characteristics of atmospheric noise. Then, the atmospheric model (winds) is perturbed using the high-frequency winds, randomly drawn from the noise dataset, at each model time step (10 days).

c. Stochastic optimal perturbation

The spatial structure of initial perturbations has an important effect on the ensemble forecasts. The third method used for ensemble construction in this study is the stochastic optimal (SO) mode perturbation (Farrell and Ioannou 1993; Kleeman and Moore 1997; Moore

et al. 2006; Tang et al. 2005; Tang et al. 2008b). Instead the realistic high-frequency winds that might not generate optimal perturbation growth, we used the leading SO mode of winds (SO1_wind) to perturb the model through the entire forecast period. As discussed in Tang et al. (2005, 2008b), for white noise in time, the SOs are the eigenvectors of the operator S:

$$S = \int_0^{\tau} R^*(0,t)R(0,t)\,dt. \tag{5}$$

Here τ is the forecast interval of interest, set at 24 months in this study, $R(0, t)$ is the forward tangent propagator of the TLM that advances the state vector of the system from time 0 to time t, and $R^*(0, t)$ is the transpose of $R(0, t)$. A detailed description of the SO can be found in Moore et al. (2006), Tang et al. (2005), and Tang et al. (2008b). Specifically, at each initial time, the perturbation was held constant for a total of 30 days, as a continuous wind perturbation following Karspeck et al. (2006), and then a new temporally uncorrelated perturbation was applied. The perturbations were controlled by (4) but using SO1_wind instead of SV1_sst, where X is still a normalized random number; and $\alpha = 0.7$ equivalent to the RMSE of winds anomaly of 0.7 m s^{-1}, obtained using sensitivity experiments based on the first verification principle (6) described in section 4.

d. *Combination of stochastic optimal and initial SSTA perturbations*

The fourth ensemble construction method is denoted by SO1_wind+SV1_sst, including the SV1_sst perturbation at initial conditions and the SO1_wind during the whole forecast period. Thus, two key sources of uncertainties were included in the SO1_wind+SV1_sst method. Comparisons between the SO1_wind+SV1_sst method against the SV1_sst method and the SO1_wind method reflect relative importance of the uncertainty from the SO1_wind and from the SV1_sst in ensemble probabilistic predictions.

4. Verification principles of probabilistic forecasts

ENSO probabilistic forecasts are made for three categories in this study: La Niña, neutral, and El Niño. The category classification follows the definition (available online at http://portal.iri.columbia.edu/portal/server.pt?open=512&objID=945&PageID=0&cached=true&mode=2&userID=2) used by the International Research Institute for Climate Prediction (IRI) ENSO forecast system, where the LDEO5 model is one of the forecast models used routinely for ENSO probabilistic forecast. Specifically, three ENSO categories are defined by the observed Niño-3.4 SSTA binned at its climatological frequency (the overall sample average frequency $\overline{O}$, or named as the base rate) of 25%, 50%, and 25%, respectively, which approximately match the common historical ENSO events during 1950–2002.

It is necessary to mention key properties of a probabilistic system here. A probabilistic forecast system has two important attributes: (i) reliability—defined by statistical consistency between forecast probability (P_f, the proportion of ensemble members that indicate the occurrence of an event) and the corresponding observed frequencies P_o over the long time period (Toth et al. 2003). For example, the forecast system for precipitation is reliable if the proportion of occurrences of rain is close to P_o. However, reliability alone is not sufficient for a probabilistic forecast system. For example, a system always forecasting the climatological probability of the event is reliable but not useful because the system would not provide any forecast information beyond climatology. Thus another key property of a probabilistic system is also required: (ii) resolution—which measures the difference between observed frequencies P_o and climatological probability $\overline{O}$ (Murphy 1973). Compared to the base rate, a larger P_o indicates a higher resolution of the forecast system. Note P_o is obtained by compiling a set of cases for forecasts with P_f, P_o depends on P_f implicitly. To achieve a reliable and high "resolution" ensemble-based probabilistic ENSO forecast, several principles used to measure the two properties are applied to evaluate our ensemble construction methods as discussed next.

a. *Ensemble spread and error of ensemble mean*

If the observation is statistically indistinguishable from the ensemble members, then the error of the ensemble mean (i.e., $\mathrm{RMSE_{EM}}$) must close to the mean distance of the individual members from their mean (i.e., ensemble standard deviation or SPRD; Buizza 1997; Stephenson and Doblas-Reyes 2000; Toth et al. 2003). In addition, the $\mathrm{RMSE_{EM}}$ is comparable to the RMSE of the deterministic forecast ($\mathrm{RMSE_{CTL}}$), obtained from the unperturbed initial condition. However, when nonlinearity becomes pronounced with increased lead time, the ensemble prediction could be better than the control forecast (Toth and Kalnay 1997). Furthermore, the standard deviation of the observed SSTA distribution over a long time period indicates the upper limit of RMSE for ENSO climatological predictions. With the observed NIÑO-3.4 SSTA index for the period of 1856–2003, the standard deviation value is 0.71.

Thus, if an ensemble forecast system includes all possible uncertainties of the realistic ENSO system, over a long time period, the following relationship is valid:

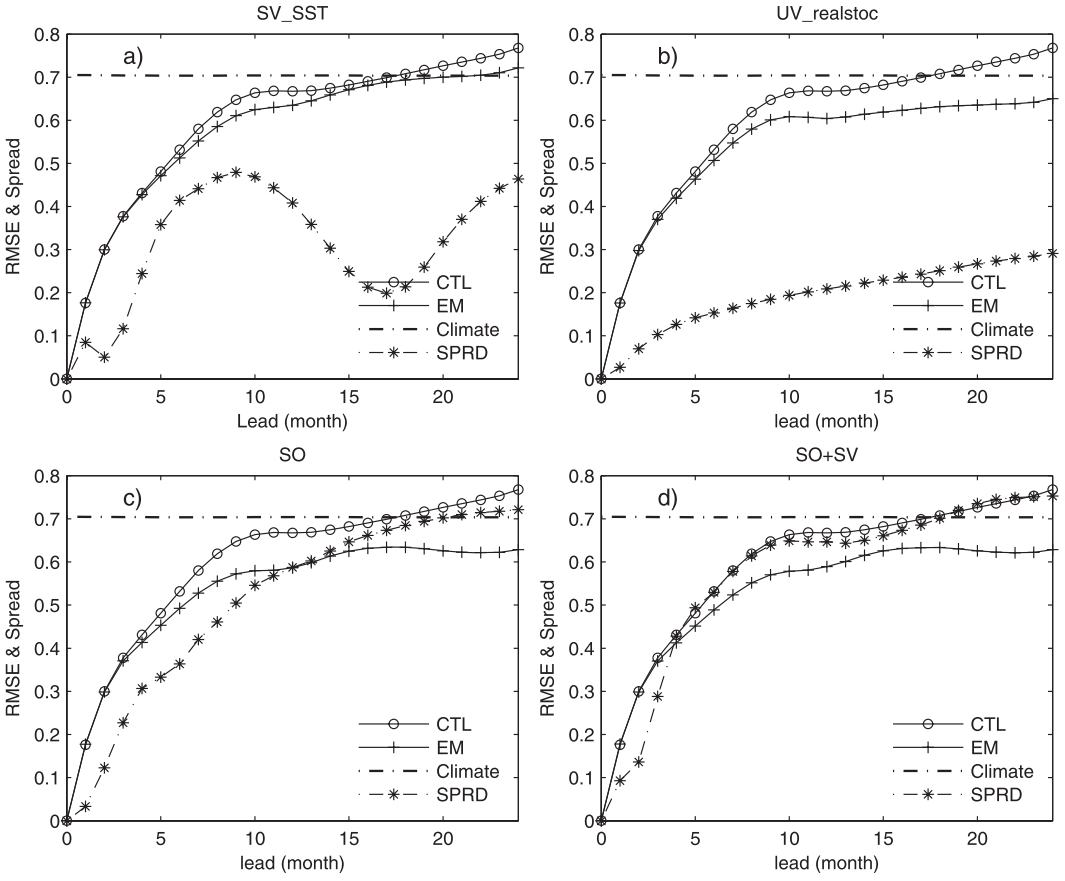

FIG. 1. RMSE of Niño-3.4 SSTA for the control run (CTL, circle), ensemble mean ($RMSE_{EM}$, plus), along with ensemble spread (SPRD, asterisk), and climatological standard deviation of SSTA (dashed–dotted line) as a function of lead time (month), averaged over 1856–2003. RMSE and SPRD for (a) SV1_sst method, (b) UV_realstoc method, (c) SO1_wind method, and (d) SO1_wind+SV1_sst method.

$$\text{SPRD} \approx \text{EME} \leq \text{RMSE}_{\text{CTL}} \leq 0.71. \quad (6)$$

Note the SPRD in (6) is a function of lead time t only. We will see in section 5b that (6) is useful in generating a reliable ensemble system.

b. Reliability diagram (RD)

The traditional reliability diagram (RD; Wilks 1995) is often used to evaluate the reliability of probability forecast, which examines the consistency of the observed relative frequency of event occurrence P_o and the forecast probabilities P_f. The P_o is calculated at a set of forecast probabilities from 0% to 100% in 10% intervals. The reliability diagram is a plot of P_o against P_f. If the forecast is perfectly reliable, P_o should be equal to P_f. The RD method is good at evaluating and calibrating the reliability of a two-category (yes–no) forecast. It also can be applied for a multicategory forecast by examining the reliability of individual categories separately. Also, one can evaluate the reliability by another method, the multicategory reliability diagram (MCRD) method (Hamill 1997).

c. The Brier score

The Brier score (BS; e.g., Wilks 1995) is a commonly used verification measure for assessing the accuracy of probability forecasts. It is the mean squared distance between the forecast probability and the observed frequency over the verification period:

$$\text{BS} = \frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2, \quad (7)$$

where N is the number of total verification samples (N = 1776 here), P_i is the forecast probability, and O_i has a value of 1 or 0 depending on whether the event occurred or not. Similar to the deterministic prediction skill RMSE, a smaller BS indicates a good forecast system.

The BS can be decomposed into three items: reliability (REL), resolution (RES), and uncertainty (UNC) as follows (e.g., Wilks 1995):

$$\mathrm{BS} = \underbrace{\left[\frac{1}{N}\sum_{k=1}^{K=10} n_k(P_{fk} - \overline{O}_k)^2\right]}_{\mathrm{REL}} - \underbrace{\left[\frac{1}{N}\sum_{k=1}^{K=10} n_k(\overline{O}_k - \overline{O})^2\right]}_{\mathrm{RES}} + \underbrace{[\overline{O}(1-\overline{O})]}_{\mathrm{UNC}}. \quad (8)$$

Over the verification period, the observed frequency of occurrence P_o can be partitioned into K bins ($K = 10$ in this study) according to the forecast probability P_f. Here P_{fk} is the averaged forecast probability at bin k and $\overline{O}_k$ is the corresponding observed frequency. The uncertainty term UNC and base rate $\overline{O}$ are obtained from the long-term observed data they are independent of the forecast system. For the cold, neutral, and warm ENSO categories, UNC is $^3/_{16}$, $^4/_{16}$, and $^3/_{16}$, respectively, according to the definition of the climatological frequency (base rate) from IRI as mentioned earlier. Here n_k is the number of the forecast and observation pairs located in an individual bin k. The first term reliability RES on the rhs of (8) is actually equal to the mean squared deviation of the reliability curve from the diagonal line in RD plot. A smaller reliability term REL indicates a better consistency between P_{fk} and $\overline{O}_k$, which results in a smaller BS and a more reliable ensemble system. The second term resolution RES is equivalent to the variance of observed distribution. RES measures the ability of a forecast system to discern different situations where the frequency of the occurrence of the event is different from the base rate s. Note that the RES term has a negative sign, but it is often used without the negative sign, as a positive-oriented measure. A good Brier score occurs at a large RES item and a small REL item, corresponding a high resolution and good reliability. The ideally perfect RES value equals to the uncertainty item UNC that gives the upper limit of the predictability of the probabilistic prediction system.

To compare the Brier score to that for a reference forecast system $\mathrm{BS}_{\mathrm{ref}}$, it is convenient to use the Brier skill score (BSS; e.g., Wilks 1995):

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}}. \quad (9)$$

If the climatological forecast is taken as reference prediction, $\mathrm{BS}_{\mathrm{ref}} = \mathrm{UNC} = s(1 - s)$. Here BSS is positively oriented. It has the range of $-\infty$ to 1. A negative BSS indicates that the forecast is less accurate than the climatology forecast. Here BSS is equal to 1 for a perfect system, and 0 for a system that performs like the climatology forecast.

From (8), (9) can be rewritten as

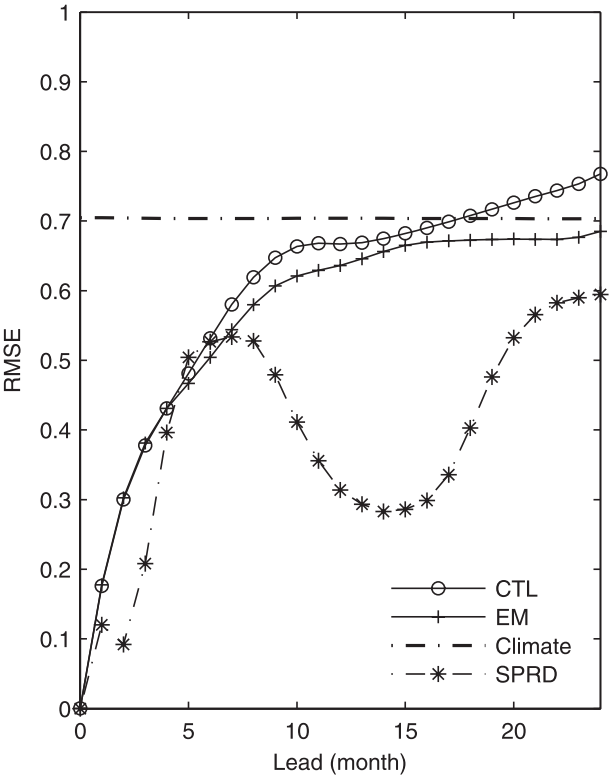

FIG. 2. As in Fig. 1, but for the first ensemble construction method SV1_sst, with a larger SSTA initial perturbation magnitude (1.5 times that in Fig. 1a).

$$\mathrm{BSS} = \frac{\mathrm{RES}}{\mathrm{UNC}} - \frac{\mathrm{REL}}{\mathrm{UNC}} = B_{\mathrm{res}} - B_{\mathrm{rel}}. \quad (10)$$

In (10), B_{rel} and B_{res} are named as the reliability term and resolution term of the BSS, respectively. The B_{rel} is negatively oriented and B_{res} is positively oriented, consistent with the signs of the RES and REL terms in the BS score. Both $B_{\mathrm{res}} = 1$ and $B_{\mathrm{rel}} = 0$ indicate a perfect forecast system.

d. The RPS score

The RPS (Epstein 1969; Murphy 1969, 1971) is another commonly used skill (resolution) measure for probabilistic forecasts, defined in terms of the squared differences between the cumulative probabilities in the forecast and observation vectors:

$$\mathrm{RPS}(t,i) = \sum_{l=1}^{3}\left[\sum_{k=1}^{l} p_k(t,i) - \sum_{k=1}^{l} O_k(t,i)\right]^2, \quad (11)$$

where P_k is the forecast probability assigned to the lth category and $O_k = 1$ when the observation falls into lth category and 0 otherwise. The RPSS (Wilks 1995) is

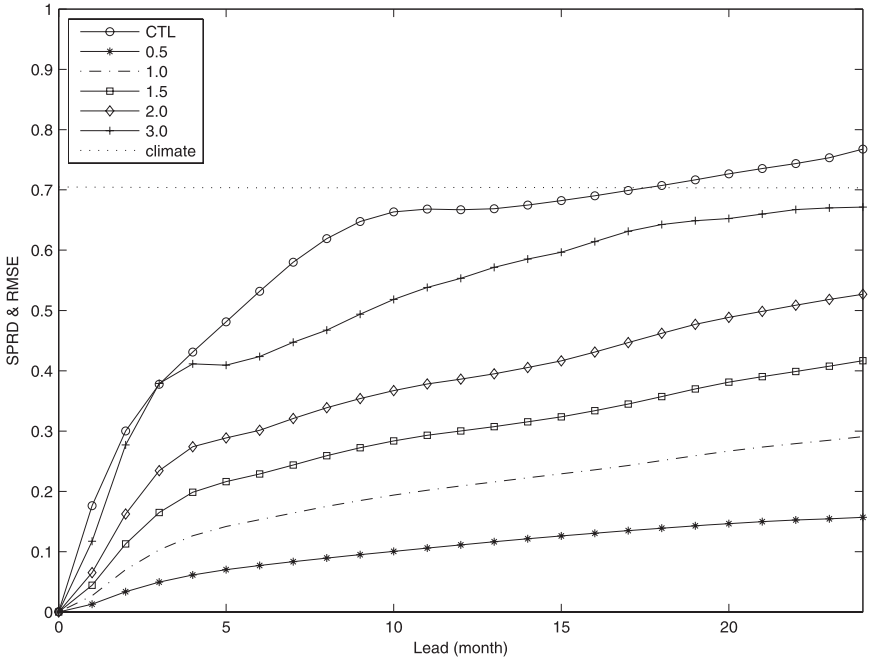

FIG. 3. A sensitivity study for the SPRD by adjusting the strength of stochastic winds in the second ensemble construction method. The perturbation magnitude varies from 0.5 to 3.0 times of that of NCEP high-frequency winds.

defined using the RPS and a reference forecast defined to have zero skill. Here, the climatological forecast is used as the reference forecast:

$$\mathrm{RPSS} = 1 - \frac{\mathrm{RPS}}{\mathrm{RPS}_{\mathrm{clim}}}, \tag{12}$$

where RPS/RPSS scores are functions of lead time t and initial time i.

In summary, a good ensemble-based probabilistic forecast system should have the following: (i) an ensemble spread (SPRD) should be close to the $\mathrm{RMSE}_{\mathrm{EM}}$, and the RMSE of the control deterministic forecast, as given in (6); (ii) probabilistic forecasts must be reliable, as measured by the reliability diagram and the reliability term of the BSS (i.e., B_{rel}); and (iii) a skillful probabilistic forecast system should have good resolution measured by the resolution term of the BSS score, (i.e., B_{res}). In addition, a good probabilistic forecast should have a small RPS and a large BSS/RPSS score.

5. Results

a. Ensemble spread

We begin by first examining whether ensemble prediction experiments can satisfy the first principle in (6). As discussed in section 4a, the first principle offers a measure to judge that whether an ensemble construction can include sufficient uncertainties of the model. The SPRDs of four ensemble experiments are compared against the RMSE of the control run ($\mathrm{RMSE}_{\mathrm{CTL}}$) and the RMSE of the ensemble mean ($\mathrm{RMSE}_{\mathrm{EM}}$; Figs. 1a–d). In Fig. 1a, although the $\mathrm{RMSE}_{\mathrm{EM}}$ for the SV1_sst method is close to the $\mathrm{RMSE}_{\mathrm{CTL}}$ and the standard deviation of the climatological forecast (0.71; the blue dashed–dotted line), the ensemble SPRD underestimates the model uncertainty significantly. Of note is the decrease in ensemble SPRD at lead times of 10–17 months, suggesting a limitation of using linear SV theory in ensemble construction over long lead times. We explored the evolution of the model error growth in Cheng et al. (2010b). It was found that the error growth reaches its maximum around the lead times of 9–12 months and is controlled by the underlying dynamical processes [i.e., linear and linearized nonlinear heating processes (horizontal and vertical advection/mixing)]. The linear/nonlinear heating has an offsetting effect and opposite contribution to the total error growth. A strong offsetting effect can be observed in SV1_SST at the leading time of around 15–17 months. This explains why the decrease in ensemble spread happens after the lead times of around 10 months and the ensemble spread rebounds and increases after the lead time of around 17 months. Certainly, a large SPRD could be obtained by increasing the perturbation magnitude α of (4). Figure 2 shows a larger SPRD occurring as the α increases 1.5 times of that in Fig. 1a. However,

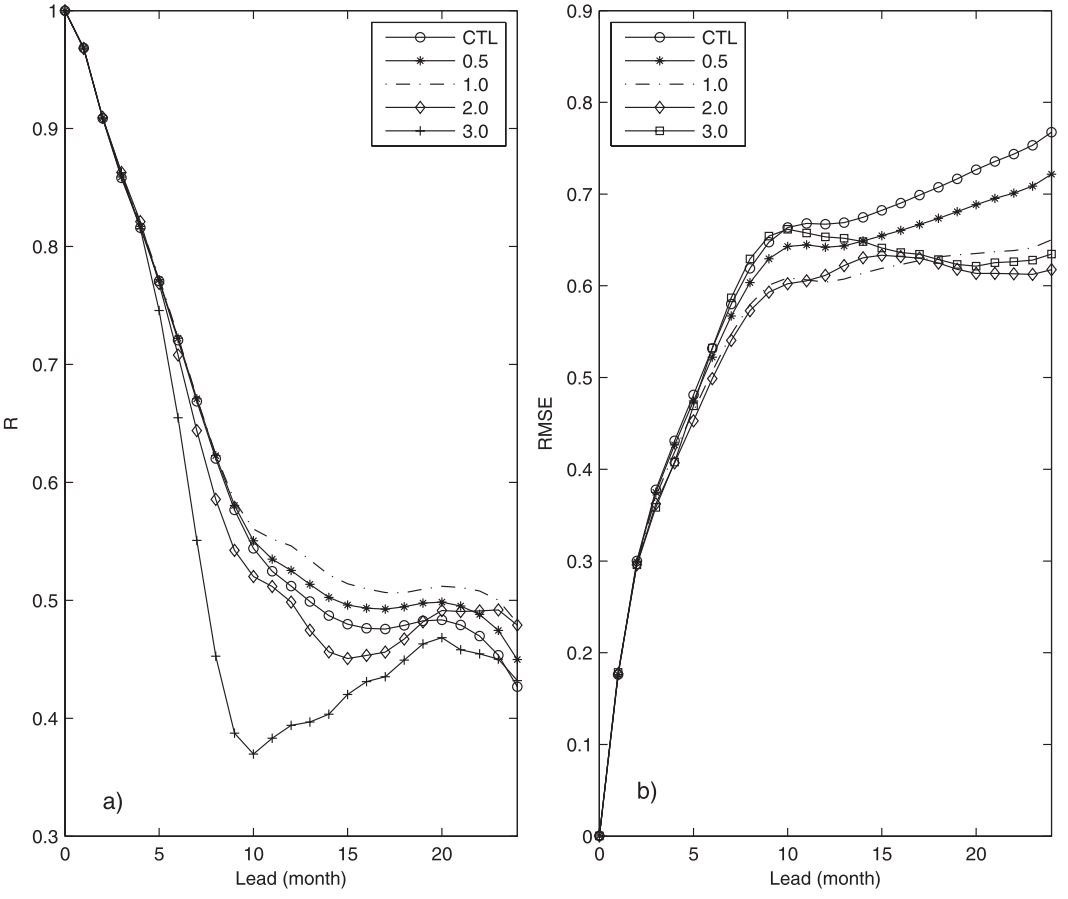

FIG. 4. (a) Anomaly correlation skill R from the control run (circle) and ensemble mean forecast with different level of stochastic wind perturbation (0.5, 1, 2, and 3 times) than the realistic NCEP winds. (b) As in (a), but for RMSE.

Fig. 2 still violates (6) after a lead time of 6 months. Thus, the SV1_sst method might not be a good ensemble construction strategy for the long lead time ENSO ensemble predictions.

In the second ensemble construction method (UV_realstoc), the high-frequency realistic stochastic winds are used during the forecast period. The current LDEO5 model is free of atmospheric random forcing; thus, using realistic stochastic winds might be able to potentially improve ENSO predictability. Unfortunately, the UV_realstoc method also underestimates the model error/uncertainty, showing a small SPRD far away from the $\mathrm{RMSE_{CTL}}$ and the standard deviation of the climatological forecast in Fig. 1b. Increase in the magnitude of external forcing can produce a large spread as shown in Fig. 3; however, artificial adjustment of the strength of stochastic winds results in unrealistic stochastic forcing. For example, the spread is close to the $\mathrm{RMSE_{CTL}}$ when the perturbation magnitude is increased to 3 times the original NCEP winds in Fig. 3. This is in agreement with the result from the LDOE4 model in Karspeck et al. (2006), where a sufficient spread could not be obtained until using an unrealistic strong wind forcing, with a standard deviation of 10 m $\mathrm{s^{-1}}$. Figure 4 shows that if the stochastic winds are unrealistically large (e.g., a strong wind perturbation 3.0 times as large as the original NCEP winds), the anomaly correlation R and RMSE

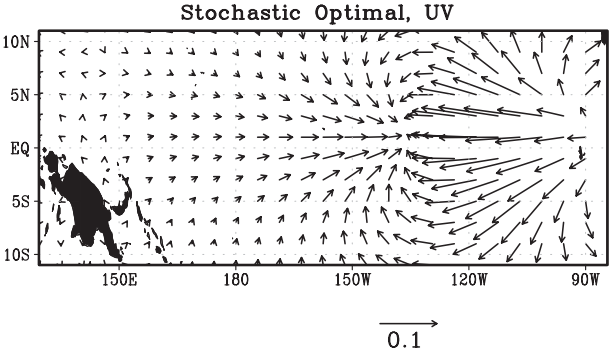

FIG. 5. The 148-yr-averaged leading mode of the stochastic optimal (SO) winds (m $\mathrm{s^{-1}}$). This mode explains the 30%–40% of the original variance of S in (5).

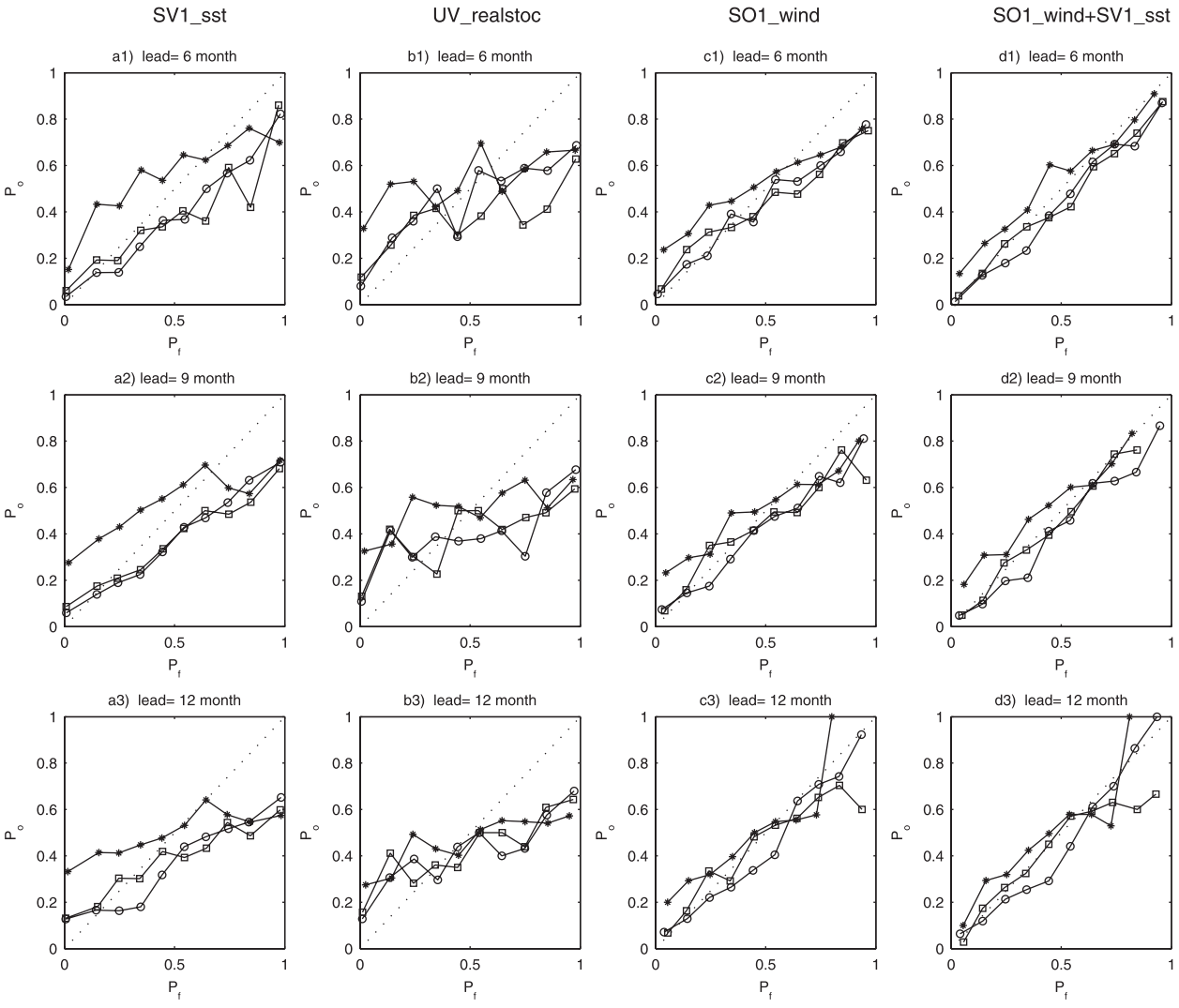

FIG. 6. (a1)–(a3) The RD for the first ensemble construction method: SV1_sst, at lead times of 6, 9, and 12 months. In each plot, three reliability curves represent three ENSO categories: warm (circle), neutral (asterisk), and cold events (square). These are calculated based on 100-member ensemble hindcasts for all months over the 1856–2003. For the second method: (b1)–(b3) UV_realstoc, (c1)–(c3) SO1_wind, and (d1)–(d3) SO1_wind+SV1_sst.

degrade in spite of a large SPRD. An unrealistic strong wind perturbation may bias the model system and produce a large dynamical imbalance. Thus, the second strategy fails to construct a good ensemble forecast either.

In summary, both the SV1_sst and the UV_realstoc methods cannot introduce sufficient uncertainties that we expects for a good ensemble construction. For the SV1_sst method, large differences between the SPRD and $\mathrm{RMSE}_{\mathrm{CTL}}$ at longer lead times suggest that the perturbation introduced at the initial SSTA cannot effectively persist through the forecast period because of dispersion. For the UV_realstoc method, uncertainty estimated from the high-frequent components of NCEP winds cannot produce sufficient prediction uncertainties or errors due to the random nature of the perturbation spatial structure. As mentioned in section 3c and in the introduction section, the spatial structure of stochastic wind perturbation is important in ensemble construction.

In the third experiment, we used the stochastic optimal mode to construct the ensemble prediction for the period from 1856 to 2003, as discussed in Kleeman and Moore (1997) and in section 3c. To achieve this, we first calculated the leading SO mode of winds (denoted by SO1_wind) for each calendar month for the optimal period of 24 months over the 148-yr period. It was found that the spatial pattern of the SO1_wind is not sensitive to initial conditions; thus, the average SO1 wind pattern over all initial conditions, as shown in Fig. 5, was used

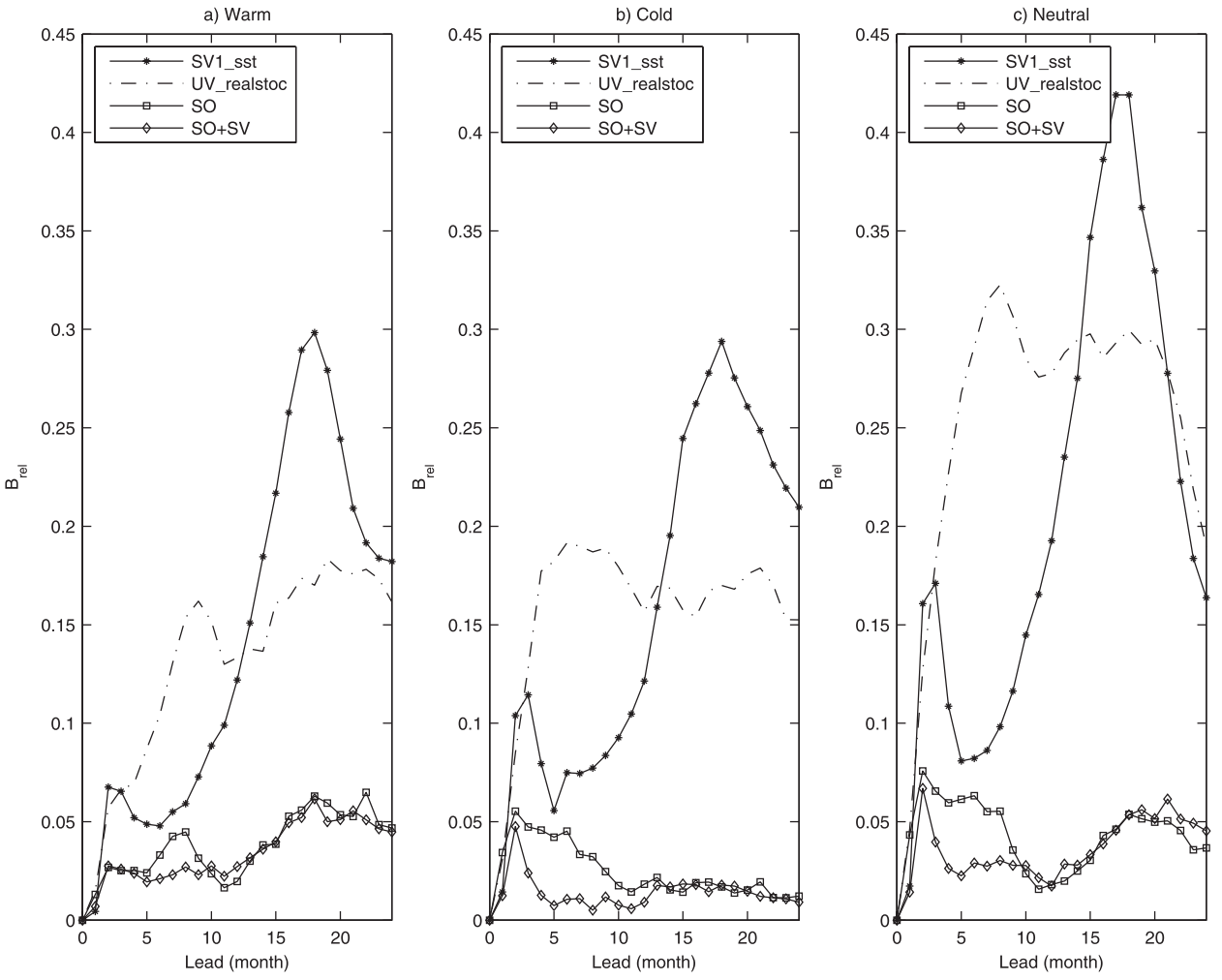

FIG. 7. The reliability term B_{rel} of the BSS as a function of lead time for the four ensemble prediction experiments at (a) warm, (b) cold, and (c) neutral ENSO categories as functions of lead time (month). SV1_sst (asterisk); UV_realstoc (dashed–dotted); SO1_wind (square); and SO1_wind+SV1_sst (diamond). Note that B_{rel} is negatively oriented.

for the ensemble construction. Similar to the SV1 of SSTA, the SO1_wind contributes to about 30%–40% of the total variance. There is a strong convergence region of winds centered at 140°W and a divergence at the cold tongue region of the eastern tropical Pacific. That such a structure is favorable for perturbation growth is probably inherent in ENSO dynamics. For example, this pattern generates corresponding downwelling and upwelling in the eastern tropical Pacific, and induces warm eastward-propagating Kelvin waves and cold westward-propagating Rossby waves, which in turn impacts on ENSO variability according to the delayed oscillator theory (Suarez and Schopf 1988). Figure 1c shows the SPRD variation as a function of lead time, generated by the SO1_wind method. As can be seen, the RMSE_{EM} and SPRD from this method are closer to the RMSE_{CTL} and the standard deviation of ENSO climatological prediction (i.e., 0.71) than the first two methods, satisfying the first principle (6). Comparison of Fig. 1b with Fig. 1c suggests the importance of the spatial structure of wind perturbations in ensemble construction. Note that the perturbation magnitude used here is much smaller than that of UV_realstoc (0.7 vs 2.5 m s^{-1}).

The fourth perturbation ensemble construction method (SO1_wind+SV1_sst) is to combine the SV1_{SST} and the SO1_wind perturbations. In terms of the first principle in (6), the ensemble spread produced by this method is the best, as shown in Fig. 1d. Compared with the SO1_wind and the SV1_sst method, the SPRD from SO1_wind+SV1_sst method is the closest to the RMSE_{EM} and RMSE_{CTL}, showing the important effect of both the SV1_{SST} and the SO1_wind perturbation on the ensemble spread. Especially, SO1_wind likely dominates the ensemble spread of long lead times.

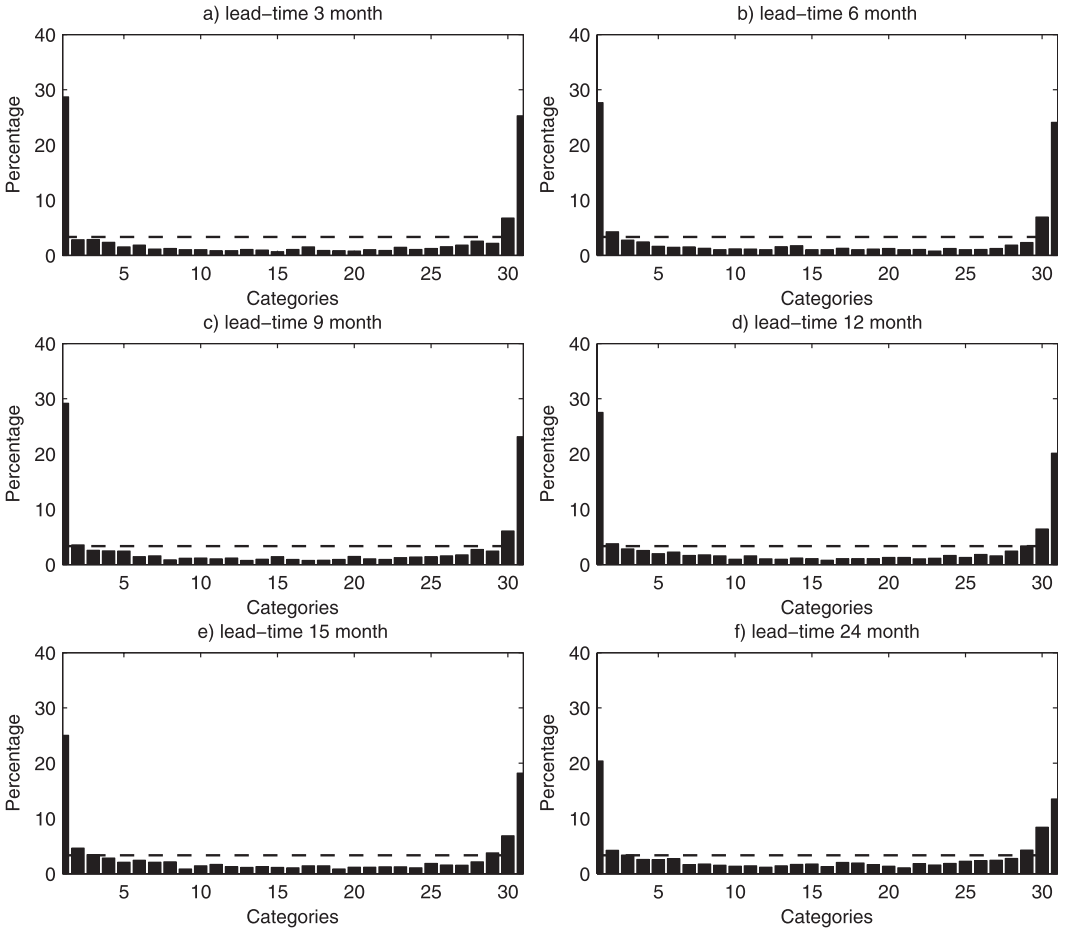

FIG. 8. Analysis rank histograms for a small SPRD case (i.e., the second ensemble construction method UV_realstoc using the original high-frequency winds) at different lead times of 3, 6, 9, 12, 15, and 24 months. The perfect percentage is 3% (dashed line). The rank of the verification is tallied 500 times.

b. Reliability

The second principle, ''reliability,'' is examined by the RD. The forecasted/observed SSTA are grouped into three categories representing the cold, neutral, and warm ENSO states, as defined at the beginning of section 4. In each ENSO category, a RD curve is made by using the forecast probability P_f at 11 intervals of 0%, 10%, . . . , 100% against the corresponding observed relative frequency P_o over the 148 yr. The diagonal line in an RD diagram indicates a perfect reliable system (i.e., $P_f = P_o$).

The RD diagrams are shown in Fig. 6 for the four ensemble construction methods and at three different lead times: 6, 9, and 12 months. The RD curves from the first two ensemble construction methods cross the diagonal line from the upper left to bottom right showing poor reliability and overconfidence(Figs. 6a,b). These features are probably due to the smaller SPRDs of the first two methods. For the last two ensemble construction methods, their reliability is greatly improved as shown in Figs. 6c,d where the RD curves oscillate around the diagonal lines, especially for the fourth method SO1_wind+SV1_sst.

Reliability can be further quantified using the reliability component of the BSS (B_{rel}; Wilks 1995). Figure 7 shows the reliability scores for four ensemble experiments for the three ENSO categories. Again, the two SO-based ensemble construction methods provide more reliable results (i.e., smaller reliability scores) than the other two methods over all lead times and in all categories, especially at long lead times. Thus, both the RD analysis and the reliability score B_{rel} demonstrate the importance of the stochastic optimal winds in the ensemble construction.

Next, we will use the verification rank histogram to explore the role of ensemble SPRD on reliability. The rank histogram diagram, also called Talagrand diagram (e.g., Anderson 1996; Talagrand et al. 1997), is another

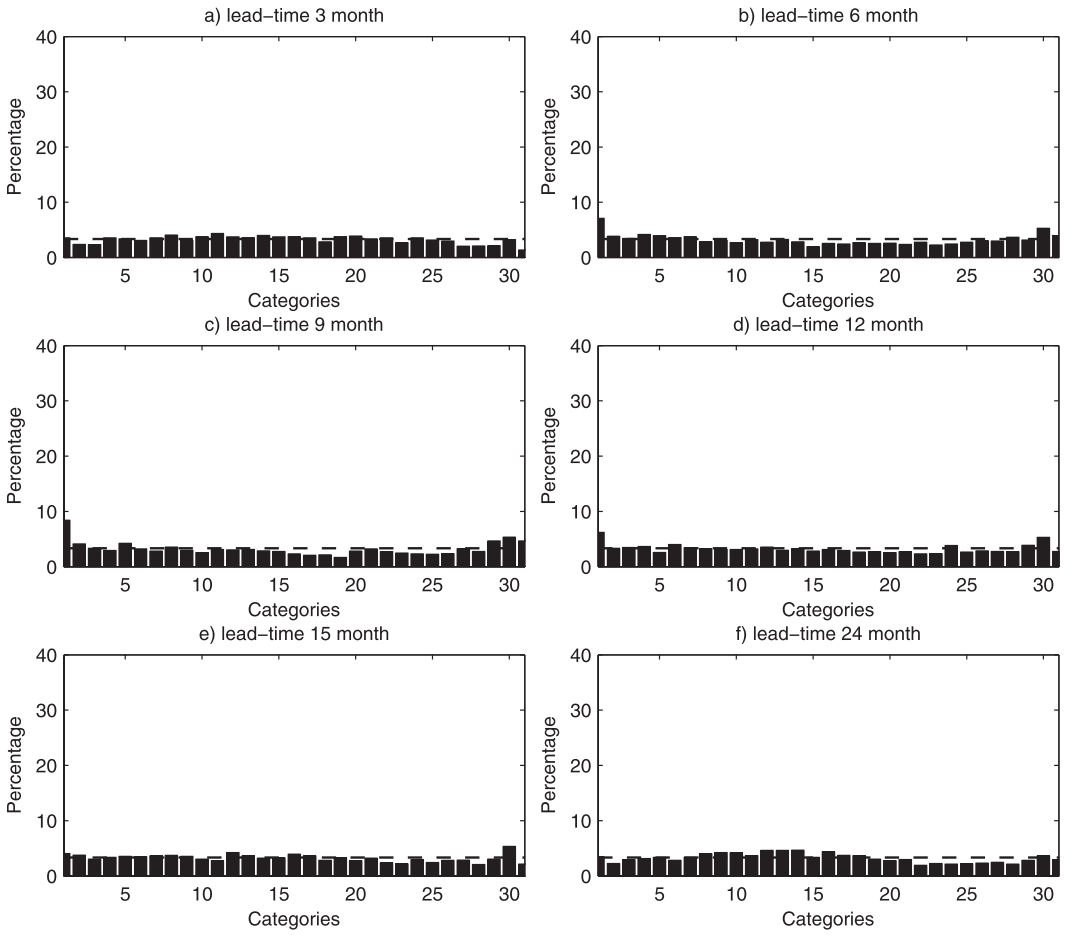

FIG. 9. As in Fig. 8, but for a good/sufficient SPRD case, derived by a strong wind perturbation (3 times of the original high-frequency NCEP winds).

way to present the reliability of an ensemble forecast system. The underlying basis of the rank histogram diagram is that the observation and ensemble members in a reliable ensemble system are subject to an identical probability distribution. To make a rank histogram plot, first, at a given lead time, ensemble forecasts with M members will be ranked in an order from the smallest to the largest, which will provide $M + 1$ categories including two open-ended categories. Observations falling in the two open-ended categories represent those observations that cannot be resolved by the forecast system. Finally, over a long time period, the frequency of observation at each category will be obtained.

A reliable system would be equally likely to contain the observed value in a rank histogram (Toth et al. 2003). For a small SPRD case, observations fall more frequently on the first and the last categories and rarely show in the middle categories, which results in a U-type distribution in a rank histogram. Figure 8 shows the rank histogram of the ensemble predictions by the second construction method that has the smallest SPRD. As can be seen, Fig. 8 shows a U-like type at all lead times, where the perfect percentage value is 3%, but there are 20%–30% observations fall in the first or the last category, whereas few samples fall in the middle categories. Figure 8 was obtained using a bootstrap strategy to ensure a robust histogram structure: 1) for each lead time, 30 members out of 100 ensemble predictions are randomly chosen to calculate the rank; 2) the process in 1) was repeated 500 times and the averaged percentage values of the 500 times are shown in Fig. 8.[1] For a relatively large SPRD case [e.g., the SPRD satisfying (6) by using an unrealistic strong wind perturbations in the second construction method], the rank histogram displays a homogenous distribution, as shown in Fig. 9, that is, the frequency

[1] We also tried different sample sizes in bootstrap experiments from 30, 40, 50, etc., and found that all results are similar to Fig. 8, which is actually consistent with the histogram using 101 categories.

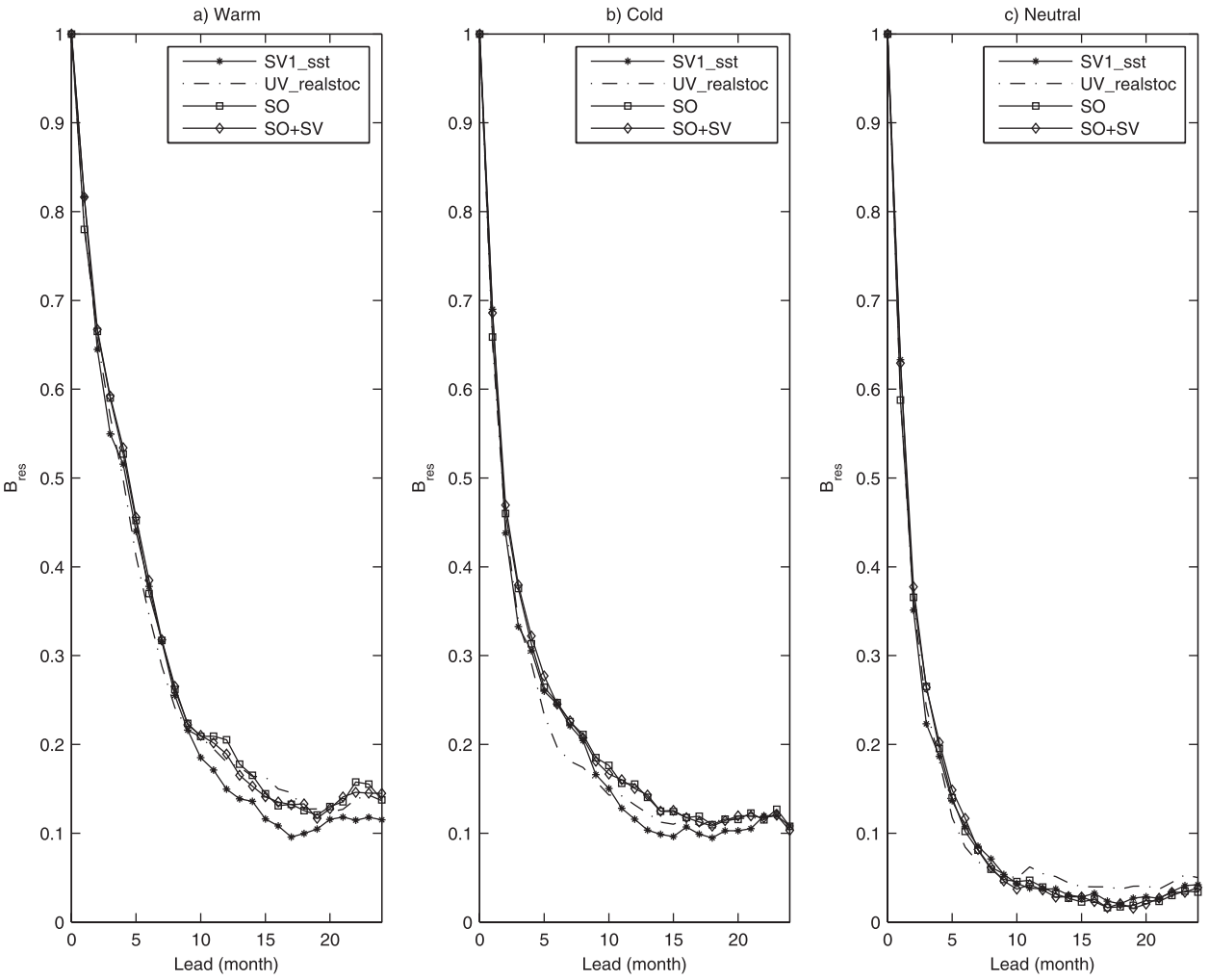

FIG. 10. As in Fig. 7, but for the resolution term B_{res} of the BSS, where B_{res} is positively oriented.

distribution is around the perfect percentage line, indicating the consistency of forecast and observed distributions (a good reliability). Thus, Figs. 8 and 9 suggest that an ensemble system with good ensemble SPRD can result in a reliable probabilistic forecast system.

c. Resolution

To examine the resolution of the four ensemble construction methods, we will analyze the resolution item B_{res} of the Brier score in (8). Figure 10 displays the B_{res} for the warm, cold, and neutral ENSO states as a function of lead time. Two common features can be seen: (i) The B_{res} scores for the warm and cold ENSO events are greater than those of the neutral ENSO state for a given lead time, and resolution drops faster at the neutral ENSO state than the others, indicating that El Niño and La Niña events are more predictable than neutral events. This signal-dependent characteristic of ENSO predictability is in agreement with many studies (e.g., Chen and Cane 2008; Tang et al. 2008a). (ii) Compared with the large differences of reliability terms among four methods in Fig. 7, resolution terms for the four methods only show slight differences, although their SPRD are visibly different in Fig. 1. This implies that ensemble SPRD is more related to reliability than resolution. In other words, the reliability of ENSO probabilistic forecast is more sensitive to choice of ensemble construction strategy than the resolution.

d. Overall probabilistic skill

The overall performance of the four ensemble construction methods is evaluated by BSS score and RPS/RPSS score, as defined by (10)–(12). The BSS measures the overall probabilistic skill, contributed by reliability and resolution scores for each ENSO category. The RPS and RPSS are accumulated skill scores for three ENSO categories. Figure 11 presents BSS for four ensemble

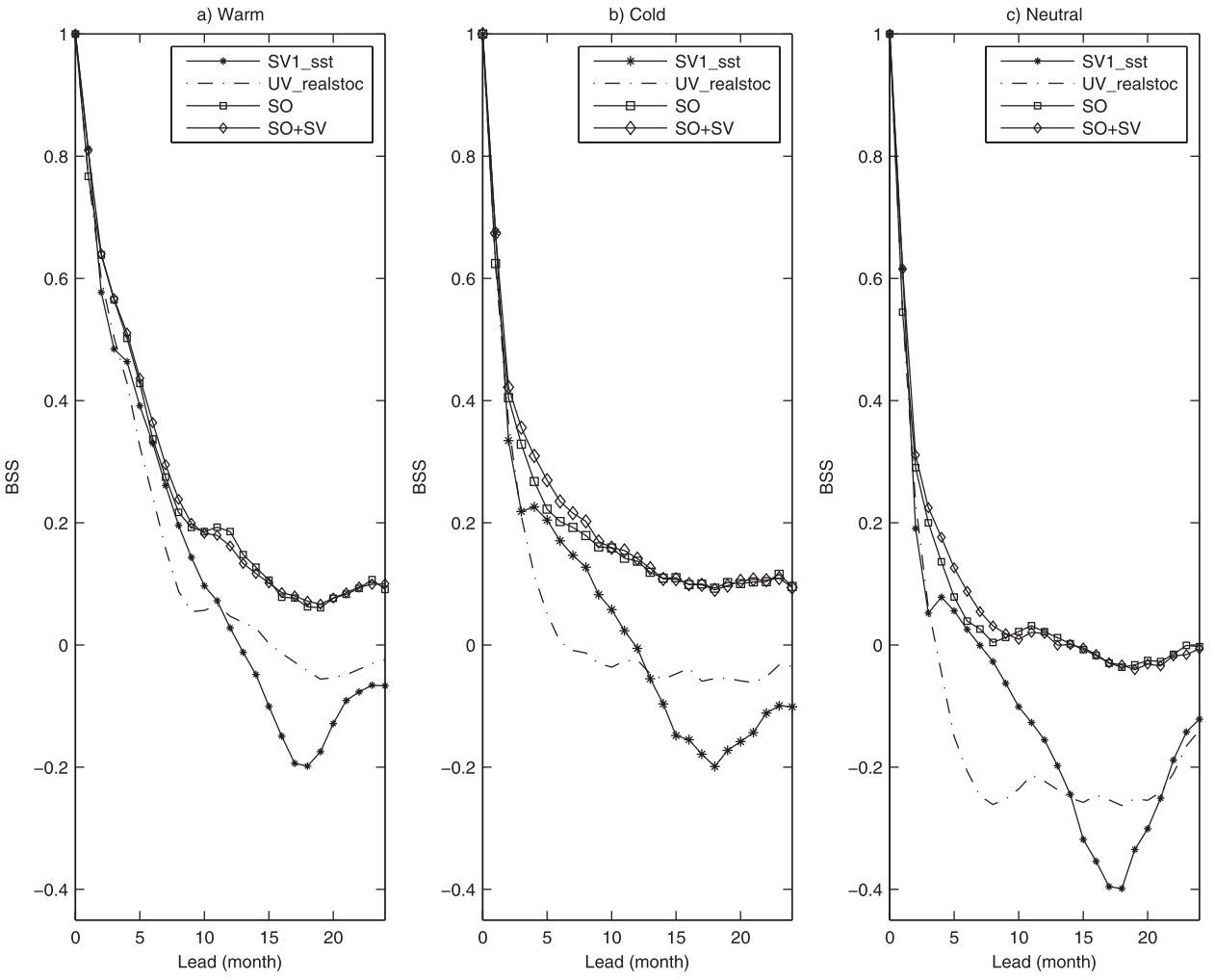

FIG. 11. As in Fig. 7, but for the BSS, which is positively oriented.

methods at cold, neutral, and warm ENSO states. As seen, the SO1_wind and SO1_wind+SV1_sst methods provide better BSS skill scores than SV1_sst and UV_realstoc methods, indicating the important role of stochastic optimal winds in improving probabilistic skill. The larger BSS scores in the last two methods mainly benefit from the better reliability terms in Fig. 7 because four experiments have resolution terms similar to those shown in Fig. 10. Figure 11 also indicates the upper limit of ENSO predictability of the LDEO5 model using BSS score. Warm and cold ENSO events are predictable for lead times of 2 yr or longer, whereas the neutral ENSO state reaches its lowest predictability at the lead time of 10 months (i.e., at the lead time longer than 10 months, the BSS is negative, indicating the system has no skill at longer lead times).

The RPS and RPSS scores measure the distance between the probability of the forecast and observation similar to the RMSE value, but in a probabilistic sense. From the definition of RPS, the range of the RPS score is between 0 (the perfect forecast) and 1. The RPSS score is zero or positive if the forecast skill equals to or exceeds that of the climatological probabilities, whereas a negative RPSS represents that the forecast skill is worse than climatology (e.g., Mason 2004). A smaller RPS or a larger RPSS score indicates higher predictability. To compare the skill of the four ensemble methods, individual RPS and RPSS scores were calculated over 148 yr for lead times from 0 to 24 months. The averaged RPS and RPSS score over the 148 yr are given in Fig. 12. The SO1_wind+SV1_sst method has the smallest RPS score and largest RPSS score, providing a more skillful forecast than other methods. It is worth noting that the RPSS scores shown in Fig. 12 are averaged over 3 ENSO categories for lead times of 0–24 months over the 148 yr; thus, although the averaged RPSS scores have negative

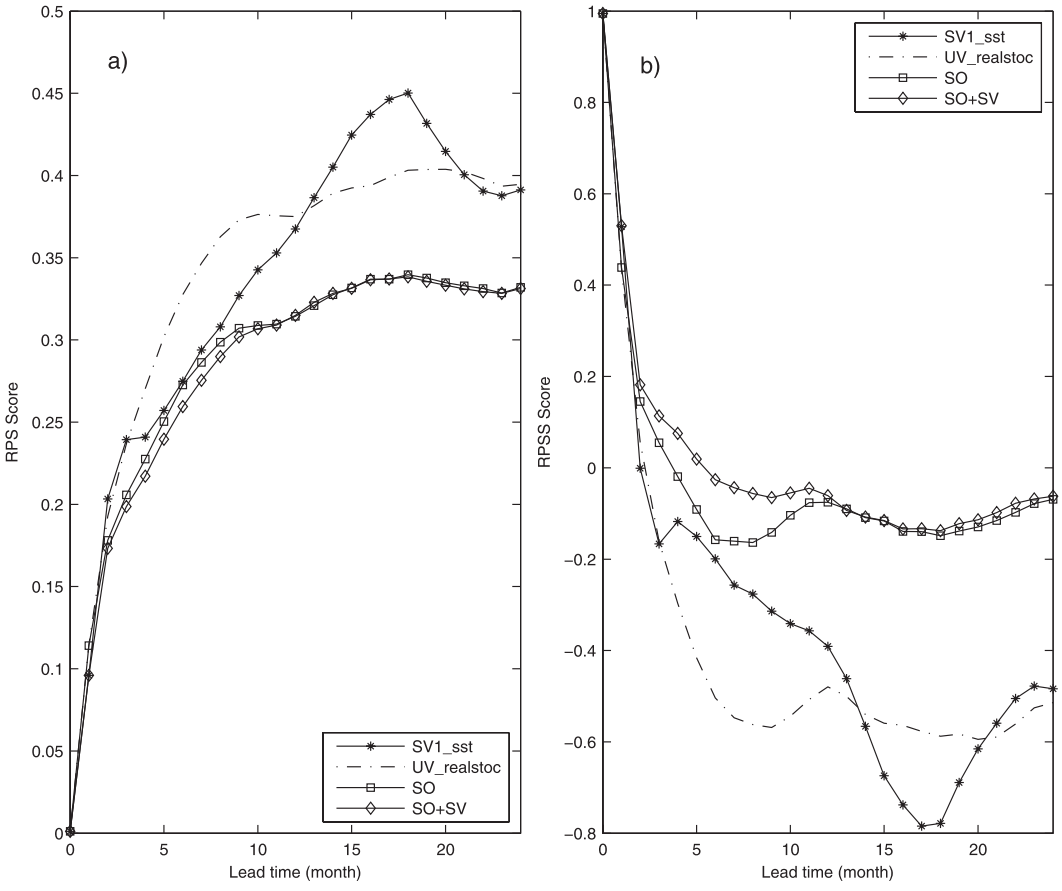

FIG. 12. (a) The RPS as a function of lead time for four ensemble construction methods. (b) As in (a), but for the RPSS.

values at lead times >5 months, for individual forecasts of warm or cool events the skill score can have positive RPSS scores.

In summary, Figs. 11 and 12 indicate that the fourth ensemble construction method SO1_wind+SV1_sst is superior to the other three, providing the highest probabilistic prediction skill. Also, the third method SO1_wind has relatively higher prediction skill than the first and the second methods. Thus, we have demonstrated that the stochastic optimal winds play important roles in constructing ensemble prediction in the LDOE5 model.

6. Conclusions and discussion

Skillful ENSO predictions will assist in the management of natural resources and the environment. Significant progress has been made in ENSO prediction over the past few decades (e.g., Latif et al. 1998; Goddard et al. 2001). Currently there are a few ENSO prediction models issuing routine predictions (e.g., IRI online at http://portal.iri.columbia.edu), including statistical models, intermediate complexity dynamical models, hybrid coupled models, and fully coupled general circulation models. However, some important issues still remain and are challenging to the ENSO and seasonal climate prediction community. One specific issue is the measures of the uncertainties in ENSO prediction.

An ideal approach to deal with prediction uncertainty is to issue probabilistic prediction, which has been widely applied in weather forecasting. Compared with weather probabilistic forecasting, ENSO probabilistic prediction has not been well addressed. Probabilistic predictions are typically generated by ensemble prediction methods. Thus, an interesting question is the following: which ensemble construction method can lead to the best ENSO probabilistic model? In this study, we explored four typical ensemble construction methods through the LDOE5 model. A long-term retrospective ensemble prediction was carried out for the past 148 yr (1856–2003) for each ensemble construction method. The performance of probabilistic prediction is measured using several probabilistic verification methods (e.g., the spread principle, reliability diagram, RPS and RPSS, and BSS). The reliability, resolution, and amplitude of the

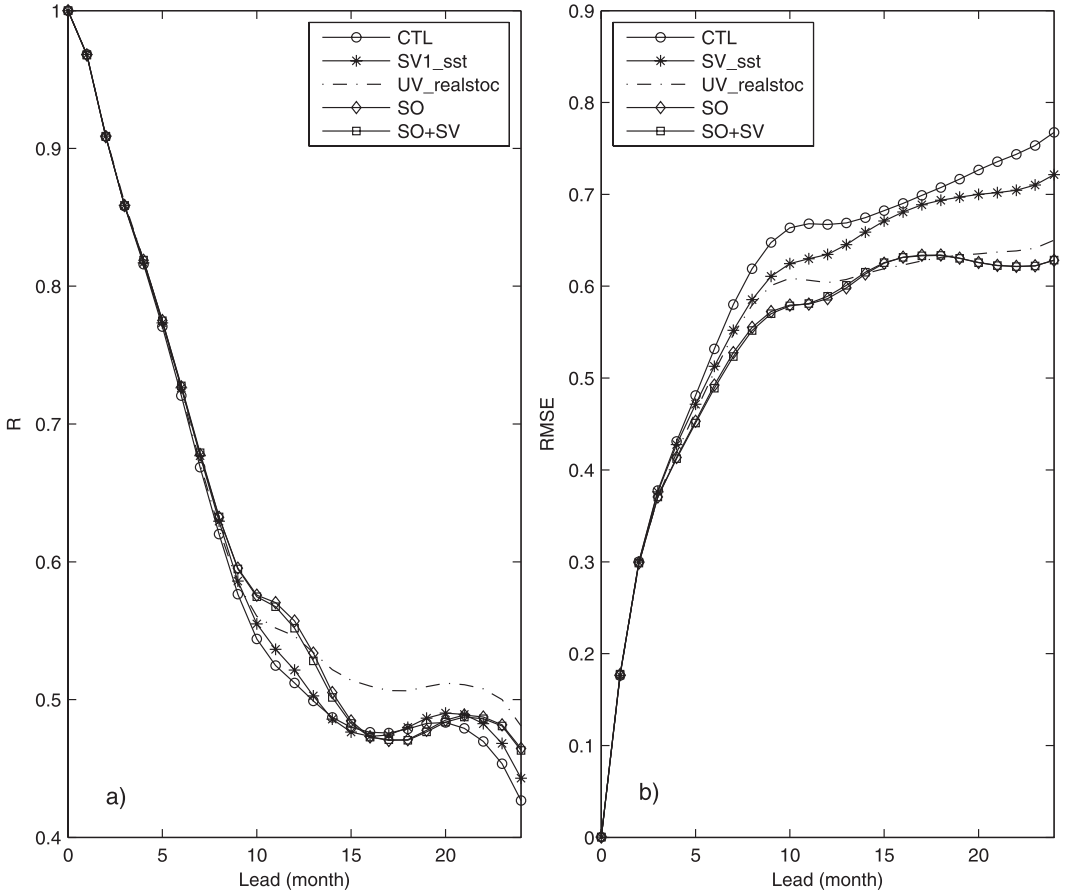

FIG. 13. (a) The anomaly correlation R as a function of lead time for four ensemble construction methods. (b) As in (a), but for the RMSE.

ensemble spread were considered as the key principles to evaluate the performance of ensemble construction methods.

It was found that the SV1_sst ensemble construction method and the realistic stochastic winds method UV_realstoc failed to generate reliable probabilistic predictions because they characterize insufficient uncertainties in the model, resulting in small ensemble spreads. Meanwhile, their lower reliability and poor BSS skill are revealed by several probabilistic methods. The small spreads in both of the SV1_sst and UV_realstoc methods are probably due to the limitation of linear SV theory at the longer lead times and to the random nature of spatial structure in the high-frequency realistic winds, respectively. To overcome the small spread issue, stochastic optimal perturbation of winds were applied over the whole forecast period in the last two SO-based methods. After removing the spread issue, the two SO-based methods exhibit good reliability in probabilistic measures.

Among four ensemble construction methods, the overall probabilistic skills measured by BSS and RPSS indicate that the SO1_wind+SV1_sst ensemble construction method is superior to the other three. Also, the third method SO1_wind has a higher BSS score than the first and the second methods, suggesting the stochastic optimal winds play important roles in constructing ensemble prediction in the LDOE5 model. The skillful perturbation method (large BSS or RPSS score) mainly benefits from the good reliability contributed by the stochastic optimal winds. However, the differences of resolution scores are subtle among the four ensemble construction methods in spite of large differences in the reliability scores existing. This indicates that the reliability score is much more sensitive to the ensemble construction method than resolution score, suggesting that the merits of a good ENSO probabilistic prediction system are mainly reflected in the reliability score. Basically, a good ensemble SPRD helps to achieve a good reliability score, thereby bringing a higher overall probabilistic skill (i.e., BSS/RPSS).

One interesting finding in this study is the great importance of stochastic forcing on ENSO probabilistic

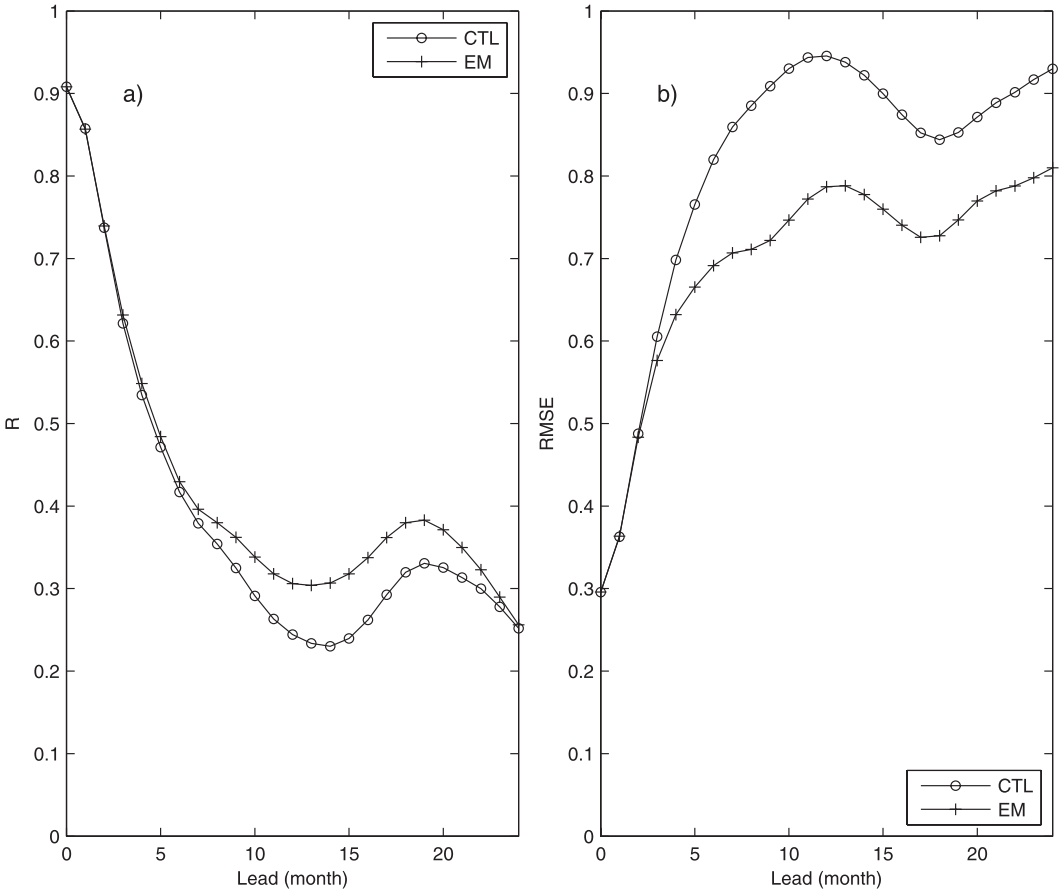

FIG. 14. Correlation skill *R* and RMSE skill for the control run and the ensemble prediction experiment with SV1_sst construction method, as functions of lead times (without using two MOS schemes).

prediction. Generally, there are two kinds of sources that limit ENSO predictability: the chaotic behavior of the nonlinear dynamics of the coupled system (e.g., Jin et al. 1994; Chen et al. 2004); and the stochastic nature of the coupled system characterized by weather noise and other high-frequency variations, such as westerly wind bursts and the Madden–Julian oscillation (e.g., Penland and Sardeshmukh 1995; Kleeman and Moore 1997; Moore et al. 2006; Gebbie et al. 2007). It is still not clear which source plays the more dominant role. Thus, the importance of stochastic forcing on ENSO probabilistic prediction provides insight into this central question challenging the ENSO community.

Different from older versions of the ZC model and many models, the LDEO5 version can effectively remove model biases through two MOSs (model output statistics) schemes: one for SSTA and the other one for other variables (Chen et al. 2000; Chen et al. 2004). Both MOS schemes take effect at each time step during the whole forecast period, and well consider uncertainties in model parameters. Thus, the advantage of ensemble methods developed in this study might have been weakened by the MOS schemes. For example, the ensemble predictions do not show improvement in correlation skill than the control run experiment at short lead times in Fig. 13a (however, the RMSE skill of all ensemble mean predictions are actually still better than control prediction for lead time longer than 4 months as shown in Fig. 13b). It is expected that our proposed methods such as stochastic optimal modes might be more useful and powerful for these models, which do not have bias correction schemes. To confirm this point, we performed an ensemble prediction perturbed by SV1_sst and a control prediction, both using the ZC model without MOS schemes. The results were shown in Fig. 14, indicating that both correlation and RMSE skill are improved at lead times of 5 months and beyond. Since SV1_sst perturbation method only considers the initial condition perturbation, the results suggests that ensemble-mean prediction skill could be significantly improved only by the initial condition perturbation. Figure 14 also suggests that including MOS schemes is probably a primary

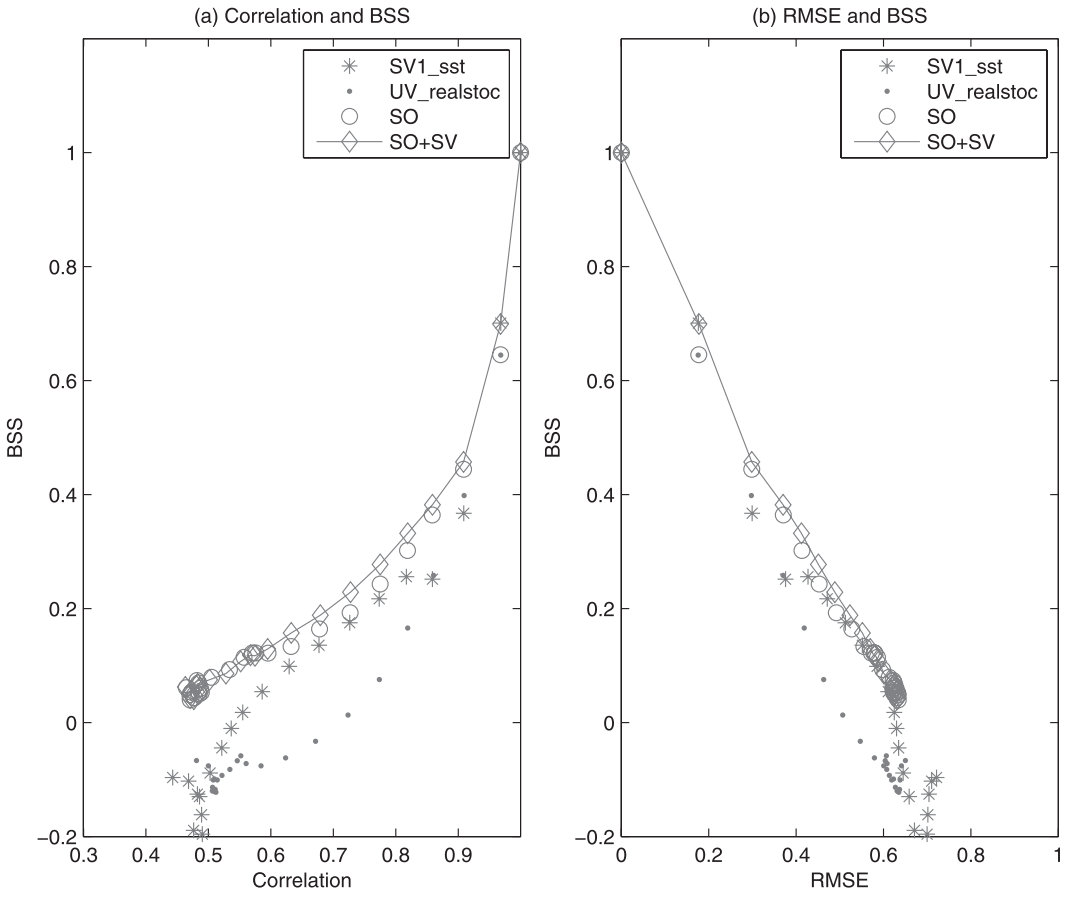

FIG. 15. Scatterplots of deterministic prediction skill and probabilistic prediction skill as functions of lead times for four ensemble construction methods. (a) Correlation and BSS at each lead time. (b) RMSE and BSS. The curve and solid line indicate the relationships from the most skillful ensemble construction method: SO1_wind+SV1_sst.

reason why our perturbation methods have less significant improvement in ensemble mean prediction skill for short lead times.

Another interesting result found in this study is that the probabilistic and deterministic skills are not always consistent with each other. For example, the second method UV_realstoc provides the best correlation skill correlation skill as shown in Fig. 13a but has much worse reliability score (Fig. 7). In other words, a reliable and skillful ENSO probabilistic prediction system might not necessarily have better deterministic skills than a poor reliable system. A typical case happens when an unrealistically large perturbation puts on the wind (e.g., 3 times as much as the high-frequent realistic wind). As such, a dynamical imbalance occurs, degrading the deterministic prediction skills although a good reliability can be obtained. For example, SO1_wind and SO1_wind+SV1_sst lead to similar RMSE skills but they produce differences in BSS, RPSS, and B_{rel} scores at short lead times. In other words, a reliable and skillful ENSO probabilistic prediction system might not necessarily have better deterministic skills than a poor reliable system. The possible reasons for the inconsistency between probabilistic skill measure and deterministic measure include the following. (i) Reliability/SPRD and RMSE have different meaning in concept. Reliability evaluates the consistency between forecast and observation distribution for events dependent on the classification of events (categories) whereas RMSE does not fully include the category information. (ii) SPRD is different from RMSE unless the model is perfect. It is possible that a large RMSE could have a small SPRD in an imperfect ensemble system. Because of the possible inconsistency between deterministic and probabilistic skill measure, it seems inappropriate to evaluate a probabilistic prediction system using deterministic skill measures.

It is interesting to discuss the general relationship between deterministic prediction skill and probabilistic skill for ensemble predications. In Wang et al. (2009),

a nonlinear relationship was found between correlation skill and probabilistic skill (i.e., BSS and ROC core) in seasonal prediction for winter precipitation predictions using the multiple model ensemble method (super-ensemble). They found that a correlation skill of 0.6 corresponds to a BSS of 0.1. Shown in Fig. 15 is the scatterplot of correlation skill against BSS for all ensemble construction methods, where correlation R, RMSE, and BSS are overall prediction skill averaged over the 148 yr, and they are functions of lead time. In Fig. 15a, we found similar results in ENSO ensemble predictions, namely, that there is a nonlinear relationship between BSS and correlation skill, especially in the most skillful ensemble construction method (SO1_wind+SV1_sst). A correlation skill of 0.6 corresponds to a BSS of 0.13, which is very close to the result in Wang et al. (2009). In addition, the relationship between BSS and RMSE skill is a linear relationship (Fig. 15b).

Several cautions should be borne in mind. First, we only investigated four ensemble construction methods. Based on a recent study of Ham et al. (2009), Zheng et al. (2009) suggested that the ENKF data assimilation approach is a good ensemble construction method that can provide reliable and high-resolution ensemble predictions. Thus, further comparisons of the SO-based methods with other methods such as ENKF and ET methods are expected. Second, we only perturbed two variables (i.e., the SSTA and anomalous winds); other variables could also have important impacts on ENSO predictability. For example, Karspeck et al. (2006) suggested that thermocline depth H1 or subsurface temperature Te might have large impacts on error growth and predictability in the LDEO4 model. However, because SSTA is the only initial conditions used in the LDEO5 model, choosing the errors and uncertainties from SV1_sst at the initial time and using SO1_wind to represent external atmospheric wind noise seems to be a reasonable way of perturbing the LDEO5 model. Nevertheless, a good ensemble construction strategy found in this study provides a reliable and skillful ENSO probabilistic prediction, offering a fundamental tool for the further study of ENSO predictability.

Acknowledgments. This work is supported by Canadian Foundation for Climate and Atmospheric Sciences (CFCAS) Research Grant GR-7027. Y. Cheng is also supported by the Natural Sciences and Engineering Research Council (NSERC) Scholarship PGS D2-362539-2008. D. Chen is supported by research grants from National Basic Research Program (Grant 2007CB816005) and National Science Foundation of China (Grant 40730843). We thank two anonymous reviewers for their constructive comments and suggestions.

REFERENCES

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate,* **9,** 1518–1530.

——, 1997: Impact of dynamical constraints on the selection of initial conditions for ensemble predictions: Low-order perfect model results. *Mon. Wea. Rev.,* **125,** 2969–2983.

Battisti, D. S., 1988: The dynamics and thermodynamics of a warming event in a coupled tropical atmosphere–ocean model. *J. Atmos. Sci.,* **45,** 2889–2919.

Bishop, C. H., and Z. Toth, 1999: Ensemble transformation and adaptive observations. *J. Atmos. Sci.,* **56,** 1748–1765.

——, B. J. Etherton, and S. Majumdar, 2001: Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Mon. Wea. Rev.,* **129,** 420–436.

Blanke, B., J. D. Neelin, and D. Gutzler, 1997: Estimating the effect of stochastic wind stress forcing on ENSO irregularity. *J. Climate,* **10,** 1473–1486.

Buizza, R., 1997: Potential forecast skill of ensemble prediction and spread and skill distribution of the ECMWF ensemble prediction system. *Mon. Wea. Rev.,* **125,** 99–119.

Cai, M., E. Kalnay, and Z. Toth, 2003: Bred vectors of the Zebiak–Cane model and their potential application to ENSO predictions. *J. Climate,* **16,** 40–56.

Chen, D., and M. A. Cane, 2008: El Niño prediction and predictability. *J. Comput. Phys.,* **227,** 3625–3640.

——, ——, S. E. Zebiak, R. Canizares, and A. Kaplan, 2000: Bias correction of an ocean–atmosphere coupled model. *Geophys. Res. Lett.,* **27,** 2585–2588.

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

Chen, Y. Q., D. S. Battisti, R. N. Palmer, J. Barsugli, and E. Sarachik, 1997: A study of the predictability of tropical Pacific SST in a coupled atmosphere–ocean model using singular vector analysis. *Mon. Wea. Rev.,* **125,** 831–845.

Cheng, Y., Y. Tang, X. Zhou, P. Jackson, and D. Chen, 2010a: Further analysis of singular vector and ENSO predictability in the Lamont model—Part I: Singular vector and the control factors. *Climate Dyn.,* **35,** 807–826, doi:10.1007/s00382-009-0595-7.

——, ——, P. Jackson, D. Chen, X. Zhou, and Z. Deng, 2010b: Further analysis of singular vector and ENSO predictability from 1856-2003—Part II: Singular value and predictability. *Climate Dyn.,* **35,** 827–840, doi:10.1007/s00382-009-0728-z.

Deng, Z., and Y. Tang, 2008: The retrospective prediction of ENSO from 1881-2000 by a hybrid coupled model—(II) Interdecadal and decadal variations in predictability. *Climate Dyn.,* **12,** 415–428, doi:10.1007/s00382-008-0398-2.

Descamps, L., and O. Talagrand, 2007: On some aspects of the definition of initial conditions fore ensemble prediction. *Mon. Wea. Rev.,* **135,** 3260–3272.

Eckert, C., and M. Latif, 1997: Predictability of a stochastically forced hybrid coupled model of the tropical Pacific ocean–atmosphere system. *J. Climate,* **10,** 1488–1504.

Eisenman, I., L. S. Yu, and E. Tziperman, 2005: Westerly wind bursts: ENSO's tail rather than the dog? *J. Climate,* **18,** 5224–5238.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.,* **8,** 985–987.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.,* **99,** 43–62.

——, 2003: The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dyn.,* **53,** 343–367.

Fan, Y., M. R. Allen, D. L. T. Anderson, and M. A. Balmaseda, 2000: How predictability depends on the nature of uncertainty in initial conditions in a coupled model of ENSO. *J. Climate,* **13,** 3298–3313.

Farrell, B. F., and P. J. Ioannou, 1993: Stochastic dynamics of baroclinic waves. *J. Atmos. Sci.,* **50,** 4044–4057.

Fluegel, M., P. Chang, and C. Penland, 2004: The role of stochastic forcing in modulating ENSO predictability. *J. Climate,* **17,** 3125–3140.

Gebbie, G., I. Eisenman, A. Wittenberg, and E. Tziperman, 2007: Modulation of westerly wind bursts by sea surface temperature: A semistochastic feedback for ENSO. *J. Atmos. Sci.,* **64,** 3281–3295.

Gill, A. E., 1980: Some simple solutions for heat-induced tropical circulation. *Quart. J. Roy. Meteor. Soc.,* **106,** 447–462.

Goddard, L., S. J. Mason, S. E. Zebiak, C. F. Ropelewski, R. Basher, and M. A. Cane, 2001: Current approaches to climate prediction. *Int. J. Climatol.,* **21,** 1111–1152.

Ham, Y. G., J. S. Kug, and I. S. Kang, 2009: Optimal initial perturbations for El Niño ensemble prediction with ensemble Kalman filter. *Climate Dyn.,* **33,** 959–973, doi:10.1007/s00382-009-0582-z.

Hamill, T. M., 1997: Reliability diagrams for multicategory probabilistic forecasts. *Wea. Forecasting,* **12,** 736–741.

——, C. Snyder, and R. E. Morss, 2000: A comparison of probabilistic forecasts from bred, singular-vector, and perturbed observation ensembles. *Mon. Wea. Rev.,* **128,** 1835–1851.

Houtekamer, P. L., and J. Derome, 1995: Methods for ensemble prediction. *Mon. Wea. Rev.,* **123,** 2181–2196.

Jin, F.-F., J. D. Neelin, and M. Ghil, 1994: El Niño on the devil's staircase—Annual subharmonic steps to chaos. *Science,* **264,** 70–72.

Kaplan, A., M. A. Cane, Y. Kushnir, A. C. Clement, M. B. Blumenthal, and B. Rajagopalan, 1998: Analysis of global sea surface temperature 1856–1991. *J. Geophys. Res.,* **103** (C9), 18 567–18 589.

Karspeck, A. R., A. Kaplan, and M. A. Cane, 2006: Predictability loss in an intermediate ENSO model due to initial error and atmospheric noise. *J. Climate,* **19,** 3572–3588.

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

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

——, and D. Min, 2009: Multimodel ensemble ENSO prediction with CCSM and CFS. *Mon. Wea. Rev.,* **137,** 2908–2930.

Kleeman, R., and A. M. Moore, 1997: A theory for the limitation of ENSO predictability due to stochastic atmospheric transients. *J. Atmos. Sci.,* **54,** 753–767.

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

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.,* **102,** 409–418.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.,* **20,** 130–141.

——, 1965: A study of the predictability of a 28-variable atmospheric model. *Tellus,* **17,** 321–333.

Mason, S. J., 2004: On using ''climatology'' as a reference strategy in the Brier and ranked probability skill scores. *Mon. Wea. Rev.,* **132,** 1891–1895.

Moore, A. M., and R. Kleeman, 1998: Skill assessment for ENSO using ensemble prediction. *Quart. J. Roy. Meteor. Soc.,* **124,** 557–584.

——, and ——, 1999: Stochastic forcing of ENSO by the intraseasonal oscillation. *J. Climate,* **12,** 1199–1220.

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

Murphy, A. H., 1969: On the ranked probability skill score. *J. Appl. Meteor.,* **8,** 988–989.

——, 1971: A note on the ranked probability skill score. *J. Appl. Meteor.,* **10,** 155–156.

——, 1973: A new vector partition of the probability score. *J. Appl. Meteor.,* **12,** 595–600.

Palmer, T. N., 1993: Extended-range atmospheric prediction and the Lorenz model. *Bull. Amer. Meteor. Soc.,* **74,** 49–66.

——, 2000: Predicting uncertainty in forecasts of weather and climate. *Rep. Prog. Phys.,* **63,** 71–116.

Penland, C., and P. D. Sardeshmukh, 1995: The optimal growth of tropical sea surface temperature anomalies. *J. Climate,* **8,** 1999–2024.

Perez, C. L., A. M. Moore, J. Zavaly-Garay, and R. Kleeman, 2005: A comparison of the influence of additive and multiplicative stochastic forcing on a coupled model of ENSO. *J. Climate,* **18,** 5066–5085.

Philip, S. Y., and G. J. van Oldenborgh, 2009: Atmospheric properties of ENSO: Models versus observations. *Climate Dyn.,* **34,** 1073–1091, doi:10.1007/s00382-009-0579-7.

Stephenson, D. B., and F. J. Doblas-Reyes, 2000: Statistical methods for interpreting Monte Carlo forecasts. *Tellus,* **52A,** 300–322.

Suarez, M. J., and P. S. Schopf, 1988: A delayed action oscillator for ENSO. *J. Atmos. Sci.,* **45,** 3283–3287.

Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. *Proc. ECMWF Workshop on Predictability,* Reading, United Kingdom, ECMWF, 1–25. [Available from ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, United Kingdom.]

Tang, Y., R. Kleeman, and A. Moore, 2005: On the reliability of ENSO dynamical predictions. *J. Atmos. Sci.,* **62,** 1770–1791.

——, ——, and S. Miller, 2006: ENSO predictability of a fully coupled GCM model using singular vector analysis. *J. Climate,* **19,** 3361–3377.

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

——, R. Kleeman, and A. Moore, 2008b: Comparison of information-based measures of forecast uncertainty in ensemble ENSO prediction. *J. Climate,* **21,** 230–247.

Thompson, C. J., and D. S. Battisti, 2000: A linear stochastic dynamical model of ENSO. Part I: Model development. *J. Climate,* **13,** 2818–2883.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.,* **74,** 2317–2330.

——, and ——, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.,* **125,** 3297–3319.

——, O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. *Forecast Verification: A Practitioner's Guide in Atmospheric Science,* I. T. Jolliffe and D. B. Stephenson, Eds., John Wiley & Sons Ltd., 137–163.

Tziperman, E., and L. Yu, 2007: Quantifying the dependence of westerly wind bursts on the large-scale tropical Pacific SST. *J. Climate,* **20,** 2760–2768.

Wang, B., and Coauthors, 2009: Advance and prospectus of seasonal prediction: Assessment of the APCC/CliPAS 14-model ensemble retrospective seasonal prediction (1980–2004). *Climate Dyn.,* **33,** 93–117, doi:10.1007/s00382-008-0460-0.

Wang, X., and C. Bishop, 2003: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.,* **60,** 1140–1158.

Wei, M., Z. Toth, R. Wobus, and Y. Zhu, 2008: Initial perturbations based on the ensemble transform (ET) technique in the NCEP global ensemble forecast systems. *Tellus,* **60A,** 62–79.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*. International Geophysics Series, Vol. 59, Academic Press, 467 pp.

Xue, Y., M. A. Cane, and S. E. Zebiak, 1997a: Predictability of a coupled model of ENSO using singular vector analysis. Part I: Optimal growth in seasonal background and ENSO cycles. *Mon. Wea. Rev.,* **125,** 2043–2056.

——, ——, and ——, 1997b: Predictability of a coupled model of ENSO using singular vector analysis. Part II: Optimal growth and forecast skill. *Mon. Wea. Rev.,* **125,** 2057–2073.

Zavala-Garay, J., C. Zhang, A. M. Moore, and R. Kleeman, 2005: The linear response of ENSO to the Madden–Julian Oscillation. *J. Climate,* **18,** 2441–2459.

Zebiak, S. E., and M. A. Cane, 1987: A model El Niño–Southern Oscillation. *Mon. Wea. Rev.,* **115,** 2262–2278.

Zhang, R. H., and A. J. Busalacchi, 2008: Rectified effects of tropical instability wave (TIW)-induced atmospheric wind feedback in the tropical Pacific. *Geophys. Res. Lett.,* **35,** L05608, doi:101029/2007GL033028.

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,** 359–372.