Consolidation of Multimodel Forecasts by Ridge Regression: Application to Pacific Sea Surface Temperature

MALAQUIAS PEÑA

SAIC, Environmental Modeling Center, NCEP/NOAA, Camp Springs, Maryland

HUUG VAN DEN DOOL

Climate Prediction Center/NCEP/NOAA, Camp Springs, Maryland

(Manuscript received 7 September 2007, in final form 14 March 2008)

ABSTRACT

The performance of ridge regression methods for consolidation of multiple seasonal ensemble prediction systems is analyzed. The methods are applied to predict SST in the tropical Pacific based on ensembles from the Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER) models, plus two of NCEP's operational models. Strategies to increase the ratio of the effective sample size of the training data to the number of coefficients to be fitted are proposed and tested. These strategies include objective selection of a smaller subset of models, pooling of information from neighboring grid points, and consolidating all ensemble members rather than each model's ensemble average. In all variations of the ridge regression consolidation methods tested, increased effective sample size produces more stable weights and more skillful predictions on independent data. While the scores may not increase significantly as the effective sampling size is increased, the benefit is seen in terms of consistent improvements over the simple equal weight ensemble average. In the western tropical Pacific, most consolidation methods tested outperform the simple equal weight ensemble average; in other regions they have similar skill as measured by both the anomaly correlation and the relative operating curve. The main obstacles to progress are a short period of data and a lack of independent information among models.

1. Introduction

Forecasts arising from a combination of multiple models of similar skill generally outperform forecasts from individual models. This is true both for forecasts produced by the same model but with perturbed initial states and for forecasts produced by models that differ in numerics or physics or both (Clemen and Murphy 1986), which may be run at different institutions. Efforts to make the best single forecast out of a number of forecast inputs, a consolidation forecast, have resulted in different types of combination approaches. The best forecast minimizes the average of an error metric (e.g., the root-mean-square error) over a series of past events.

Numerous studies (e.g., Doblas-Reyes et al. 2000; Kharin and Zwiers 2002, hereafter KZ02; Peng et al. 2002; Hagedorn et al. 2005) have shown that simple (equal weights) multimodel averaging of ensemble members (MMA) can produce forecasts consistently more accurate and more probabilistically reliable than forecasts from any single participating model. Other more sophisticated consolidation methods have been developed (Krishnamurti et al. 1999, 2000) to further improve forecast skill; however, whether these methods can outperform MMA is still a matter of debate, the main obstacle being a lack of sufficiently long datasets of retrospective forecasts (KZ02; DelSole 2007). Independently on whether these methods can improve upon MMA, given that more and more prediction systems are becoming available to forecasters and other users, an objective procedure is necessary to deal with the information overload. For this, optimal weights should be determined for all input models taking into consideration their individual past performance and collinearity among models.

Judging the success of a consolidation is difficult be-

Corresponding author address: Malaquias Peña, NOAA/National Centers for Environmental Prediction, 5200 Auth Rd., Room 807, Camp Springs, MD 20746.
E-mail: malaquias.pena.mendez@noaa.gov

DOI: 10.1175/2008JCLI2226.1

cause so many diverse issues play a role. To organize the discussion we devote a paragraph each to the topics of hindcasts, overfit, and collinearity.

a. Value of hindcasts

The value of a long, consistent retrospective forecast (hindcast) database for each of the models can hardly be overestimated. First, this information serves for the removal of systematic errors (SE) and other types of forecast calibration, even for a model in its own right (Hamill et al. 2004, 2006). Since all dynamical models drift toward their own climatology, assessing SE is especially crucial in long-range forecast systems (e.g., Stockdale 1997). Second, hindcasts help in the optimization of weights by giving a degree of credibility to a particular model depending on its past performance. The success of a consolidation method will depend on its ability to learn from the often small sample of past situations.

b. Overfit

The stunning lack of hindcast data vis-à-vis the number of coefficients that needs to be fitted in optimization procedures is one of the major problems for consolidation. When the length of the training dataset is too short compared to the number of input forecast models, overfitting occurs and optimization procedures fail to be successful with independent data. Several approaches to reduce or avoid overfitting have been proposed. The simple "regression-improved ensemble mean" method described in KZ02 reduces the number of regression coefficients to just one by linearly regressing only the MMA against the observations rather than all individual models simultaneously. This simple approach was the most successful among the more sophisticated consolidation methods KZ02 presented, a very telling result. Other approaches to reduce overfitting include accumulating statistics from neighboring grid points or across a region and pooling leads and neighboring start times. An extreme case of pooling is to use information from all the grid points in the region of analysis, thus producing space independent weights as in Van den Dool and Rukhovets (1994) and Peng et al. (2002)—this adds stability at the expense of any spatial dependence in the weights. Robertson et al. (2004) average across subsamples to improve the Bayesian methodology of Rajagopalan et al. (2002).

c. Collinearity

Consolidation of a large number of model forecasts can also lead to problems when at least one of the models is not entirely independent from the rest. That is, forecasts from one of the input models can be specified to within a certain small error by a linear combination of forecasts from the other models. When the set of forecasts are collinear the covariance matrix is ill conditioned (or nearly so) and regression produces a spurious and unstable solution, even with plentiful data. Approaches to reduce the shortcomings in forecast performance due to collinearity include truncating the singular value set of the covariance matrix of forecasts (e.g., Derome et al. 2001; Yun et al. 2003) and "regularization" methods, particularly ridging (Tikhonov 1963). The latter method has been applied to various problems in climate (Meisner 1979; Crone at al. 1996; Krakauer et al. 2004), medium-range forecast consolidation (Van den Dool and Rukhovets 1994), and the constructed analog technique (Van den Dool 1994; Van den Dool et al. 2003) used for seasonal SST and soil moisture prediction, respectively. This study extends this approach to consolidate both dynamical and statistical forecast models for the prediction of the tropical Pacific SST.

The purpose of this study is to compare different types of consolidation methods, particularly ridging methods, with emphases on 1) a cross-validation procedure that is more stringent than the 1-year-out procedure, CV-1, but also reduces the degeneracy effect (Barnston and Van den Dool 1993) that produces unrepresentative skill estimates when only one year is held out in a regression procedure; 2) a two-stage procedure to remove models with negative weights (it is the authors' belief that input models bring either some or no skill to the consolidation, in the latter case the no-skill model should not be included at all or weights should be reconfigured); and 3) an increase of the effective sample size by gathering information from the grid points neighboring the various regions of influence, as well as from individual ensemble members rather than ensemble means only.

The article is organized as follows. Section 2 describes the data used in the study. Sections 3 and 4 outline the theoretic foundation of consolidation methods. Section 5 describes strategies to increase the effective sampling size. Section 6 describes the evaluation methodology. Section 7 discusses the results for deterministic skill assessment. Section 8 discusses the results for the probabilistic assessment. Section 9 summarizes the results.

2. DEMETER PLUS data

This study is based on *monthly means* of ensemble forecasts from a suite of eight dynamical and one

TABLE 1. Some information on the DEMETER-PLUS models.

Acronym	Full name	Layout	Period
D1, D2, . . . , D7	DEMETER Models*	Ensemble members: 9 Leads: 0–5 months Initial months: Feb, May, Aug, Nov	1980–2001
CFS	NCEP Climate Forecast System	Ensemble members: 15 Leads: 0–8 months Initial months: Jan to Dec	1981–2006
CA	CPC constructed analog	Ensemble members: 12 Leads: −3–12 Initial months: Jan to Dec	1956–2006

* Institutions developing these models: the European Centre for Medium-Range Weather Forecasts, Max Planck Institute, Météo-France, Met Office, Instituto Nazionale de Geofisica e Vulcanology, Laboratoire d'Oceanographie Dynamique et de Climatologie, and the European Centre for Research and Advanced Training in Scientific Computation.

empirical seasonal prediction systems. Seven of the dynamical prediction systems are coupled ocean–atmosphere models from the Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER) project (Palmer et al. 2004) denoted D1, D2, . . . , D7, and the remaining two—the dynamical climate forecast system (CFS; Saha et al. 2006) and the empirical constructed analog (CA; Van den Dool 1994; Van den Dool 2006, chapters 7.2 and 7.4)—are in-house models currently used for operational short-term climate predictions at the National Centers for Environmental Prediction (NCEP). Details of these models are summarized in Table 1.

The more extensive datasets of CFS and CA were reduced to make them consistent with the four starts a year (February, May, August, and November) and leads 0–5 months in the DEMETER models. The DEMETER and CA forecasts initialized in 1980 are not used here because the CFS starts only in 1981. The DEMETER models start from the first of each month, all at the same time. The ensemble members in the CFS hindcast are not initialized at the same time but in three groups of five days centered at days 11 and 21 of the previous month and at the first of the lead 0 month (see Saha et al. 2006 for details.) This ensemble generation procedure results in having older members (those that were initialized earlier) with generally larger errors than in the more recent model runs. Thus, it is possible that the CFS would require an optimization procedure of its own (not pursued) to account for this effect. Analysis of the errors in the central Pacific SST indicates (Peña and Toth 2008) that this difference in skill between CFS members is relevant in the 0–2-month lead times. To ameliorate this problem only the 10 more recent CFS members (centered at days 21 and 1) are considered here to calculate the ensemble mean and assume that the 10 members are "equal."

The period of analysis is 1981–2001, when all the forecast outputs coincide. The monthly Reynolds optimum interpolation SST v2 (OIv2; Reynolds et al. 2002) is used as the verification field for the months November 1981 and on. The NCEP Global Ocean Data Assimilation System (GODAS; Behringer 2007) is used to complete the time series in early 1981. The forecast and observed fields are given on a 2.5° × 2.5° latitude–longitude grid.

3. Unconstrained consolidation methods

A consolidation forecast is expressed here as a linear combination of K participating predictions, $\zeta_i, i = 1, . . , K$, each multiplied by a correspondent coefficient or weight α_i. Thus, the consolidation of forecasts verifying at time t is related to the verification (or target) o as

$$\boldsymbol{\zeta}^{\mathrm{T}}\boldsymbol{\alpha} = o + \varepsilon, \tag{1}$$

where $\zeta^{\mathrm{T}} = (\zeta_i) = (\zeta_1, \zeta_2, \ldots, \zeta_K)$ is the $1 \times K$ row vector containing the participating forecasts, $\boldsymbol{\alpha} = (\alpha_1\ \alpha_2 \ldots \alpha_K)^{\mathrm{T}}$ is the $K \times 1$ column vector containing the weights, T denotes the transpose, and ε, a random error, is the residual term. In this study, ζ and o are anomalies with respect to an observed (external) climate. The obvious question is, what should $\boldsymbol{\alpha}$ be?

To find $\boldsymbol{\alpha}$, a training period of N time points (e.g., number of years) is selected from the hindcast data. Equation (1) is then generalized for this set by defining a matrix whose columns contain the training time points of the input forecast models,

$$\mathbf{Z} = (\zeta_{t,i}), t = 1, \ldots, N, i = 1, \ldots, K \tag{2}$$

and column vectors $\mathbf{o} = (o_t) = [o_1\ o_2 \ldots o_N]^{\mathrm{T}}$ and $\boldsymbol{\varepsilon} = (\varepsilon_t) = [\varepsilon_1\ \varepsilon_2\ \varepsilon_N]^{\mathrm{T}}$, which contain the corresponding verification and random error, respectively. Thus, consolidation methods will involve optimizing $\boldsymbol{\alpha}$ during the training period:

$$\mathbf{Z}\boldsymbol{\alpha} = \mathbf{o} + \boldsymbol{\varepsilon} \tag{3}$$

and applying $\boldsymbol{\alpha}$ to the testing period. This is repeated for each lead and initial month available in the hindcast.

a. Multimodel average (MMA)

A simple approach to generate a multimodel consolidation is by averaging with equal weights ($\alpha_i = 1/K$, $i = 1, K$). This approach generally outperforms individual model skill according to deterministic (KZ02; Peng et al. 2002; DelSole 2007) and probabilistic measures (Doblas-Reyes et al. 2000; Hagedorn et al. 2005). MMA is used in this study as a benchmark for the more sophisticated consolidation methods. One of the advantages of MMA is that it obviously does not require training data to optimize the weights; however, it still needs a long enough hindcast dataset to identify and remove SE, an aspect requiring cross validation.

b. Skill-based weights

In this category, weights are computed from past performance (on a training dataset) of individual models. A method that weights according to past skills of individual models, referred to as COR, is the following:

$$\alpha_i = \frac{1}{f}\frac{\sum_{t=1}^{N}\zeta_{t,i}o_t}{\sum_{t=1}^{N}\zeta_{t,i}^2} = \frac{b_i}{fa_{i,i}}, \tag{4}$$

where $f = \Sigma_{i=1}^{K}(b_i/a_{i,i})$ is a factor that makes $\Sigma_i\alpha_i = 1$, $\zeta_{t,i}$ is the ith column of $\mathbf{Z}$, b_i is the covariance between model i and observations, and $a_{i,i}$ is the variance of model i. Note that all α_i are positive as long as each method has skill (positive correlation). Moreover, for this method any $\alpha_i < 0$ are set to zero. Equation (4) is as if each model is regressed individually and independently against the observations. The α_i are regression coefficients, but they are very closely related to the anomaly correlations (AC) as: $\alpha_i = \mathrm{AC} \times \sigma_o\, a_{ii}^{-1/2} f^{-1}$, where σ_o is the standard deviation of observations.

The rationale to use the COR method is that it is the objective method nearest to what forecasters at the Climate Prediction Center (CPC) do subjectively, which is to combine the real-time forecasts by methods A, B, and C with matching maps of estimates of a priori skill (calculated over a 25 or 50 yr hindcast period) so as to decide which methods should be trusted most. They assign weights accordingly. The objective COR method does not protect against "double counting," that is, imagine a perverse situation in which method A and B always produce identically the same forecast (and thus have the same a priori skill). COR would give A and B the same weight, which is unfair relative to an independent method C. A smart enough living forecaster may have noticed this (he or she would ignore either A or B). But an objective method needs to study the correlation among the methods, which is a prominent feature in the ridging approaches described in the following section.

c. Unconstrained regression

The general problem of consolidation consists of finding a vector of weights $\boldsymbol{\alpha}$ that minimizes the sum of squared errors (SSE) given by the following expression:

$$\mathrm{SSE} = (\mathbf{Z}\boldsymbol{\alpha} - \mathbf{o})^{\mathrm{T}}(\mathbf{Z}\boldsymbol{\alpha} - \mathbf{o}). \tag{5}$$

Then $\partial/\partial\alpha(\mathrm{SSE}) = 0$ leads to $\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\boldsymbol{\alpha} = \mathbf{Z}^{\mathrm{T}}\mathbf{o}$; so the weights are formally given by

$$\mathbf{a} = \mathbf{A}^{-1}\mathbf{b}, \tag{6}$$

where $\mathbf{A} = \mathbf{Z}^{\mathrm{T}}\mathbf{Z}$ is the covariance matrix, $\mathbf{b} = \mathbf{Z}^{\mathrm{T}}\mathbf{o}$ and the superscript -1 denotes the inverse operation. Equation (6) is the solution for the ordinary (unconstrained) linear regression (UR). Equation (6) reduces to Eq. (4) if all off-diagonal elements in $\mathbf{A}$ are set equal to zero. Thus, the COR method defined in the previous subsection is a particular case of UR when the different models are uncorrelated or treated as uncorrelated.

It has been recognized that UR is unsatisfactory in many applications when collinearity exists in the data or the training data is too short for K coefficients to be fitted accurately (KZ02). UR may fit a sample dataset very well but give bad results when applied to independent data. In addition, the UR method sometimes produces (large) negative coefficients, implying that the consolidation takes the opposite of what the corresponding model forecast indicates. Several techniques have been introduced to reduce collinearity by reducing the degrees of freedom in the data. A common technique (Golub and Van Loan 1980) is to decompose $\mathbf{A}$ into its principal components, retain the components associated with the largest eigenvalues, and remove those associated with eigenvalues close to zero, considered noise.

Although in a situation with collinearity negative weights are mathematically possible and not necessarily wrong in certain academic problems, the authors reject this possibility for the following reasons: 1) The inputs are forecasts with presumably positive skill (if not, they should not participate). To forecast cold weather in some area because model A predicts warm weather and model A needs to be turned into its opposite does not

sound credible to a user. 2) Many consolidation procedures require that the sum of the weights is unity. This becomes problematic if weights are allowed to be negative. 3) The construction of the probability density function (PDF) is difficult to imagine with negative weights.

4. Constrained consolidation methods

a. Ridge regression (RID)

When matrix $\mathbf{A}$ in (6) is ill conditioned or nearly so—that is, the difference between the largest and smallest singular values of $\mathbf{A}$ is too large—the weights become very sensitive to small perturbations in $\mathbf{Z}$. One approach to reduce this problem is through regularization procedures, one of which is ridging (Tikhonov 1963). Ridging is a multiple linear regression with an additional penalty term to constrain the size of the squared weights in the minimization of SSE (5):

$$\mathbf{J} = (\mathbf{Z}\boldsymbol{\alpha} - \mathbf{o})^{\mathrm{T}}(\mathbf{Z}\boldsymbol{\alpha} - \mathbf{o}) + \lambda\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha}. \tag{7}$$

Minimization of $\mathbf{J}$ leads to

$$\boldsymbol{\alpha} = (\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{b}, \tag{8}$$

where $\mathbf{I}$ is the identity matrix, and λ, the regularization (or ridging) parameter, indicates the relative weight of the penalty term. Similarities between the ridging and Bayesian approaches for determining the weights have been discussed by Hsiang (1976) and DelSole (2007). In the Bayesian view, (8) represents the posterior mean probability of $\boldsymbol{\alpha}$, based on a normal a priori parameter distribution with mean zero and variance matrix $(\varepsilon^2/\lambda)\mathbf{I}$, where $\varepsilon^2\mathbf{I}$ is the matrix variance of the regression residual, assumed to be normal with a mean zero.

In the approach adopted here, λ is constrained to be the minimum value that makes the coefficients α_i stable to perturbations in $\mathbf{Z}$. Values of λ are increased from zero to (usually no more than) 0.50 at intervals of 0.05 until α_i is judged to be stable and nonnegative on the training data. This method is referred to as RID.

In contrast to the methods described in section 3, collinearity among models is now taken into account, but to the extent collinearity causes instability its effect is reduced by increases in λ. This can be seen by performing a singular value decomposition of $\mathbf{A}$ and $\mathbf{A} + \lambda\mathbf{I}$ and computing in each case the condition number. The condition numbers are, respectively,

$$\frac{\lambda_A^{(\max)}}{\lambda_A^{(\min)}} \quad \text{and} \quad \frac{\lambda_A^{(\max)} + \lambda}{\lambda_A^{(\min)} + \lambda},$$

where $\lambda_A^{(\max)}$, $\lambda_A^{(\min)}$ are the largest and smallest eigenvalues of $\mathbf{A}$, respectively. Collinearity is identified when the condition number is too large, usually when the smallest eigenvalue is close to zero. Note that a positive λ reduces the problem.

b. Multimodel mean constraint

The function in (7) can be modified depending on which characteristic of the solution one is trying to constrain. DelSole (2007) proposed a constraint that penalizes the amount the coefficients $\boldsymbol{\alpha}$ depart from $1/K$, the multimodel average. In this case, the weights become

$$\alpha = (\mathbf{A} + \lambda\mathbf{I})^{-1}\left(\mathbf{b} + \frac{\lambda}{K}\mathbf{1}\right), \tag{9}$$

where $\mathbf{1}$ is a column vector of size K with all elements equal to one. This method is referred to as RIM. Note that for large λ all α_i tend to $1/K$, and for small λ Eq. (9) tends to Eq. (8).

c. Weighted mean constraint

It is very reasonable to constrain the weights for their departures from the skill-based weights, as given in (4). An ad hoc formula developed here for the $\boldsymbol{\alpha}$ is

$$\boldsymbol{\alpha} = (\mathbf{A} + \lambda\mathbf{I})^{-1}(\mathbf{b} + \lambda\boldsymbol{\alpha}^{\mathrm{COR}}), \tag{10}$$

where $\alpha_i^{\mathrm{COR}} = 1/f(b_i/a_{i,i})$ as in (4) are the regression coefficients from the COR method. This method is referred to as RIW.

For very large λ, coefficients from RIW converge to those from COR method

$$\alpha_i \to \alpha_i^{\mathrm{COR}}. \tag{11}$$

In both RIM and RIW, $\Sigma_{i=1}^{K}\alpha_i \to 1$ for large λ.

In situations where the collinearity is too difficult to deal with, COR may be the least damaging strategy, and close to what is used, so far, in practice at CPC. The rationale for the use RIW is that it emerges starting with the COR solution. In that limit the off-diagonal elements (the collinearity or redundancy) are completely ignored. When lowering λ the collinearity is taken into account increasingly, all the way to the point where the solution becomes too sensitive. This should not be interpreted as the diagonal elements being more accurately known than the off-diagonal elements. Even if (very high) collinearity is accurately calculated from plentiful data the solution may be too sensitive. For instance, in the limit of the perverse situation (two identical methods are entered) it is not possible to deter-

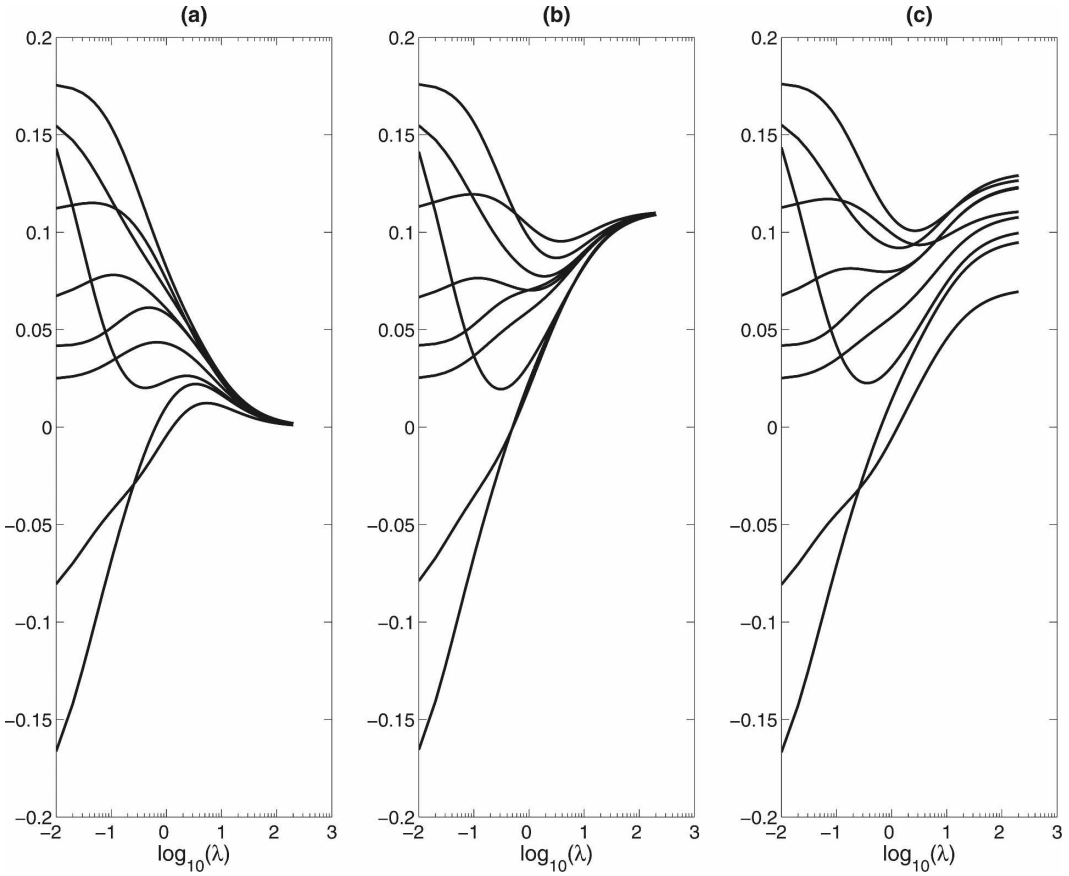

FIG. 1. Useful weights for ensemble averages corresponding to the 9 models in the DEMETER+PLUS data as a function of ridging amount (λ in $\log_{10}$ scale) for (a) ridging, RID, (b) ridging with departure-from-equal-weight penalty, RIM, and (c) ridging with departure-from-skill-based-weights penalty, RIW.

mine weights using unconstrained regression, no matter how much data are used.

d. Asymptotic behavior of weights and cutoff values of the ridging parameter

To illustrate the concepts of the ridging methods described so far, Fig. 1 is a plot of the nine regression coefficients corresponding to each of the model's weights for a lead 1 SST forecast for initial month February at a grid point in the western Pacific (12.5°S, 150°E), using all 21 yr. The figure shows by example how the weights change as the ridging amount increases from practically zero to 200. The zero ridging is equivalent to the regular (unconstrained) regression, and from the figure it is clear that in this case some of the models have a negative weight and that there is a large difference in weights among models for small λ. In the left panel, regular ridging gradually reduces the amplitude of all the coefficients. For very large ridging the coefficients tend to zero, which means that the RID consolidation converges to climatology. On the other hand, note that for RIM in the center panel the coefficients approach $1/K$, where $K = 9$ is the number of models; whereas for RIW in the right panel the coefficients approach a skill-weighted mean given by (11). In the RIW case a model with a negative weight would eventually have to be removed. Note that, in terms of weights, there is a large difference between zero ridging and a tiny amount of ridging, and keep in mind that literally zero ridging does not exist on a computer using "real" numbers. Note that for modest λ the RID, RIM, and RIW solutions are not necessarily all that different. The difference is more apparent in asymptotic behavior.

Most of the results in this study have λ smaller than 0.5 since the regression coefficients reach stability in this range. To illustrate this point, Fig. 2 shows the sensitivity of α_i to sampling for the same grid point in Fig. 1. The sensitivity is measured by the standard deviation of each weight across the 21 samples generated under the cross-validation procedure used in this study

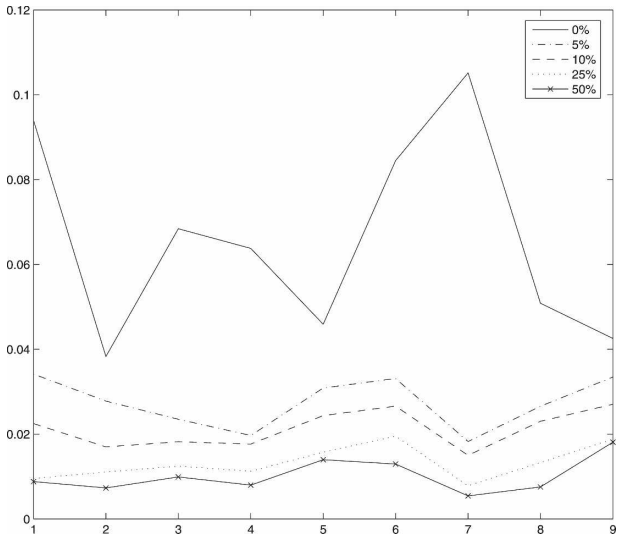

FIG. 2. Std dev of each of the useful weights (α_i, $i = 1, \ldots, 9$) across the 21 samples generated by the cross-validation procedure used in this study for different values (%) of the ridging parameter λ using the RID method.

(described in section 6) as a function of λ. Here, the weights are obtained using the RID method and are stable whenever their standard deviations become small. Thus, the figure indicates that the weights become quickly less sensitive to sampling as λ increases. With 25% of ridging the standard deviation of each weight is around 0.02 (or lower than for UR by a factor of 4) for this particular point.

Table 2 summarizes the consolidation methods tested. Although KZ02 have already shown that UR will fail on datasets as short as 20–30 yr, results based on this method are included for the sake of completeness and to make the point that skill can be very different with dependent and independent data.

5. Increasing the effective sampling size

a. Selection-combination (double pass) strategies

With less than 21 yr of hindcasts in cross-validation mode, it is necessary to find a way to increase the ratio of the training data points to the number of weights to be found. A common approach to deal with overfitting is to remove or explicitly set to zero the weights of some "bad" or highly redundant models. Reducing the number of coefficients improves the estimates of the covariance matrix and potentially increases prediction accuracy (Robertson et al. 2004). An approach for model removal is used here by performing ridge regression twice. First, to detect bad or redundant models (their weights will be negative after the first pass). Then, carry out ridging again only for the models whose weights are positive. Another approach is to set to zero the weights of the models whose AC is negative before entering them into the optimization procedure. Such approach is used for the COR and RIW methods to ensure the removal of negative weights.

b. Mixing data from neighboring grid points

Van den Dool and Rukhovets (1994) and Peng et al. (2002) increased the sampling size by using information from all grid points in the domain of analysis. In this case the weights do not change geographically and so do not allow for flexibility in cases where the ranking by model skill changes from one region to another (assuming there are enough data to make that determination). They report that in this case weights become more stable. The effective sample size can also be increased by mixing neighboring leads of the forecasts, or mixing forecasts issued close to the initial month. This last approach is not applicable to the data in this study since the initial forecast months are separated by three months, a serious limitation of DEMETER.

c. Mixing information from individual members

Another strategy to stabilize the weights, as per sample size increases, is by using the forecasts of single ensemble members. The difference among single members of the same model is the initial conditions, which can be small departures superimposed on the initial condition of reference, as in the DEMETER models and some of the CA members, or can be initial conditions corresponding to previous days or weeks, as in the CFS (and some CA members). The models used in this study have at least nine ensemble members each (Table 1). It is a reasonable practice to use only the ensemble mean of each model to carry out consolidation. If one were to consider a consolidation of all the individual members, it would imply finding solutions to 81 or more unknown coefficients. The strategy adopted here consists of regressing individual ensemble members onto the same observations and constraining the weights to be the same within each model; therefore, there will be, as before, nine unknowns to be determined but with 9 times the original sample size. Suppose that M is the minimum number of available ensemble members, then $\mathbf{Z}$ is augmented as

$$\mathbf{Z} = \zeta_{M\times \mathrm{ti}}, \tag{12}$$

where the number of rows of $\mathbf{Z}$ is M times larger than in (2). This requires duplicating M times the verification

TABLE 2. Summary of consolidation techniques and corresponding weights.

Acronym	Method	Weight
MMA	Multimodel ensemble average	$\alpha = K^{-1}\,\mathbf{1}$, where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)$, K = number of participating models and $\mathbf{1}$ is a column vector of size K and all elements equal to 1
COR	Correlation	$\alpha_i = \frac{1}{f}\frac{\sum_{t=1}^{N}\zeta_{t,i}o_t}{\sum_{t=1}^{N}\zeta_{t,i}^2} = \frac{b_i}{fa_{i,i}}$, where $f = \sum_{i=1}^{K}\frac{b_i}{a_{i,i}}$, ζ_{ti} is the training time series forecast of ith model, b_i is the covariance function between model i and observations, and $a_{i,i}$ is the variance of model i
UR	Unconstrained regression	$\boldsymbol{\alpha} = \mathbf{A}^{-1}\mathbf{b}$, where $\mathbf{A} = \mathbf{Z}^{\mathrm{T}}\mathbf{Z}$, $\mathbf{b} = \mathbf{Z}^{\mathrm{T}}\mathbf{o}$ and $\mathbf{Z} = (\zeta_{ti})$, $t = 1, \ldots, N$, $i = 1, \ldots, K$
RID	Ridging	$\boldsymbol{\alpha} = (\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{b}$, λ is such that $\sum_{i+1}^{K}\alpha_i^2$ is small and $\alpha_i \geq -0.01$, $i = 1, \ldots, K$
RI2	Double-pass ridging	First pass is regular RID; then set to zero any $\alpha_i < 0$, $i = 1, \ldots, K$; then carry out a second RID
RIM	Ridging with MMA constraint	$\boldsymbol{\alpha} + (\mathbf{A}+\lambda\mathbf{I})^{-1}\left(\mathbf{b} + \frac{\lambda}{K}\mathbf{1}\right)$
RIW	Ridging with weighted mean constraint	$\boldsymbol{\alpha} = (\mathbf{A} + \lambda\mathbf{I})^{-1}(\mathbf{b} + \lambda\alpha^{\mathrm{COR}})$, where $\alpha_i^{\mathrm{COR}} + \frac{1}{f}\left(\frac{b_i}{a_{ii}}\right)$ are the COR regression coefficients

vector **o**. Figure 3, in comparison to Fig. 1, shows the effect of this procedure on the weights. Most weights become smaller in magnitude for small λ—for very large λ Figs. 1 and 3 converge again by construction. Taking all ensemble members into account (rather than just ensemble means) can be looked upon as either an increase in the sample size or ridging in a natural way. In theory, for $\lambda = 0$, there should be no difference, whether one uses the ensemble mean or all members (with equal weight within each model), at least in terms of the resulting lhs of Eq. (1). The difference between the two figures when $\lambda = 0$ solely reflects the larger variance of individual members compared to the variance of the ensemble mean—the larger weights in Fig. 1 compensate for the smaller variance.

6. Evaluation methodology

a. Cross-validation (CV) procedure

Cross validation is necessary to establish the skill level to be expected with independent data. However, CV has problems of its own. To avoid both artificial skill and degeneracy, the consolidation has been evaluated using a 3-years-out cross-validation scheme designated as CV-3R. Of the 3 yr, one is the test element and two are chosen at random (hence R) without repetition. Barnston and Van den Dool (1993) describe the difficulties in obtaining a representation of real skill *for regression methods* when using a cross validation in which a very small portion of the sample is excluded from the development part. Leaving out the three successive values still has significant problems, especially when trends are present.

The training dataset is used both to compute SE and to carry out optimization for the weights. Clearly, the cross validation will more strongly affect the sophisticated consolidation methods than MMA, since the consolidation methods have both weights and a SE to be cross validated while MMA has only the latter.

b. Systematic error correction and anomaly expression

The assessment is carried out on anomaly fields after an estimate of SE is removed for each model individually. SE is computed as the time average, denoted by an overbar, of the forecast minus observation, $\mathrm{SE} = \overline{\boldsymbol{\zeta} - \mathbf{o}}$, for each initial month and lead in the training period. SE correction is applied with the sign reversed to the test forecast, and anomalies are formed by subtracting $\mathbf{o}_{\mathrm{clim}}$, the observed time average over the full data period. Here, $\mathbf{o}_{\mathrm{clim}}$ remains the same (regardless which years are withheld) and is considered "external," as if it were an observed climatology entirely outside the 1981–2001 period, hence the notation CV-3RE. This practice also helps to mitigate a degeneracy in skill estimates due to problems with the observed climatological mean (Van den Dool 1987).

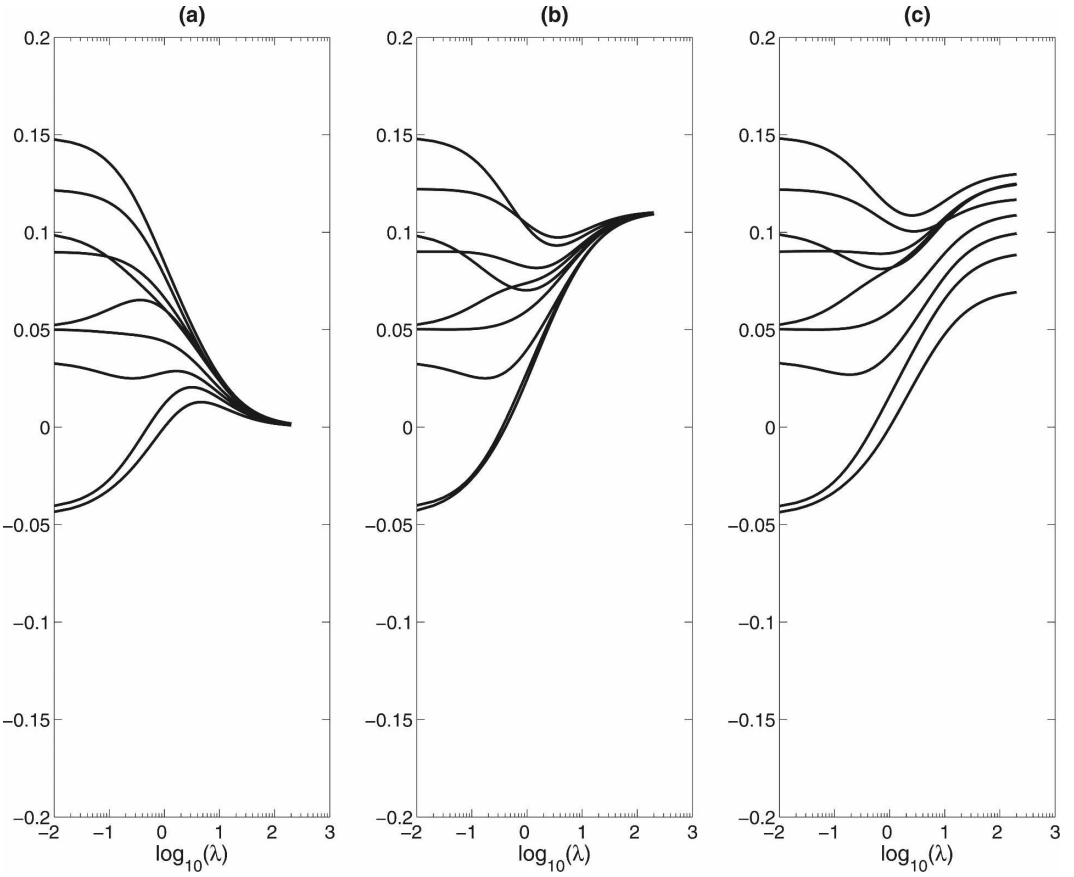

FIG. 3. Same as Fig. 1 but using all ensemble members, rather than the ensemble means of each model, to optimize the weights.

c. *Evaluation measures*

The AC between forecasts and analysis is used as a metric to measure deterministic forecast performance. AC is computed both at each grid point, to provide a regional assessment, and for the entire domain of analysis, usually called pattern anomaly correlation. In addition, the area below the relative operating characteristic curve (ROC; Mason and Graham 1999) is used to assess the ability to anticipate correctly the occurrence or nonoccurrence that SST anomalies will fall in the upper, middle, and lower terciles defined by the observed SST during the training period. The approach to construct the 3-category probability density function for the ROC application will be described in section 8.

7. Deterministic skill assessment of tropical Pacific SST

a. *Pattern correlation*

The performance of SE corrected forecasts of monthly mean SST in the deep tropical Pacific (12.5°S–12.5°N, 140°E–82.5°W) is given in Fig. 4. The height of the bars designates the average over all leads and initial months of the pattern anomaly correlation of each of the 9 models (D1 to CA) and for each of the consolidation methods (MMA to UR) described in sections 3 and 4. In Fig. 4 weights are computed gridpoint-wise using only the ensemble means of each participating model, as is commonly done. For each pair of bars, the one on the left corresponds to correlations using the full data, whereas the one on the right corresponds to correlations in a cross-validation mode, CV-3RE. Figure 4 shows that the AC of individual models goes down several points after the CV-3RE procedure. This is because the estimate of the SE based on 18 yr has nonnegligible error bars.[1]

It is clear that for the dependent case (all 21 yr, no cross validation), all of the sophisticated consolidation

[1] Reduction in skill in CA is not real. By construction CA has no SE (in the mean) so subjecting it to a SE procedure results in taking a random difference over 18 yr and applying it with the opposite sign to the test element. This lowers the skill estimate.

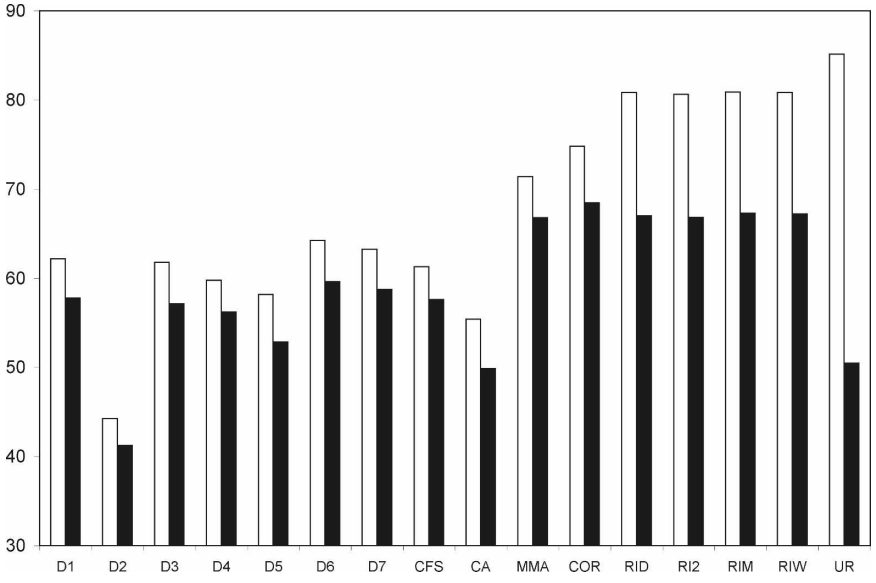

FIG. 4. Anomaly pattern correlation of systematic error corrected monthly SST over the tropical Pacific domain, averaged for all leads and initial months based on the 21 yr of data in the hindcasts (empty bars) and after 3 yr random cross validation (dark bars). The consolidation is done gridpoint-wise, which can be improved upon by increasing effective sample size (see Fig. 5).

methods outperform MMA but this may just reflect the overfitting problem of some of the methods. In particular, UR method shows a large artificial skill for the dependent case. UR yields much lower correlation in the cross-validation mode, even worse than most individual input models. To a lesser extent, the other consolidation methods also produce artificial skill in the dependent case, but the ridging method can be credited for limiting the damage. Nevertheless, as has been established earlier (e.g., KZ02; Peng et al. 2002), the figure shows the difficulty for sophisticated consolidation methods to outperform the simple MMA when the hindcasts are not long enough or models do not bring enough independent information to the consolidation, or both.

Next is the assessment of the performance of consolidation methods after increasing the effective sample size. In Fig. 4 the full sample size was only 21 (at each point in space). To maximize the limited information available in the hindcasts several strategies have been described in the introduction. Here, the impact of both spatial pooling of information and use of individual ensemble members is explored. Figure 5, left panel, shows the average of the anomaly correlation for all leads and initial months for the domain in the tropical Pacific for the following 4 strategies/situations using the ensemble average of participating models:

1) Grid point by grid point (consistent with results in Fig. 4);
2) Box 3×3 that includes the point of analysis plus the 8 closest grid points;
3) Box 9×9 with the 80 closest neighboring grid points;
4) All the grid points in the domain.

Strategies 1–4 are marked in Fig. 5 along the abscissa. Note that the verification domain and procedure is always the same. These same 4 strategies are repeated in the right panel of Fig. 5 but now using all nine ensemble members of each model. The solid horizontal line is the AC average of the MMA methods used as a reference.

Except for the COR methods, as the information from the closest neighboring grid points is added (second entry in left panel), the skill improves and it becomes more apparent that sophisticated methods outperform MMA, although not by huge margins in this SST application. However, including information from grid points farther away from the point of analysis (third entry in left panel) is no longer an improvement. For the case where all grid points are included (case number 4) and thereby the weights are space independent (fourth entry in left panel), all consolidation methods are generally higher than in entry 1. When the ensemble members of each model are used in the optimization procedure, almost all the methods outperform MMA by a noticeable margin.

A striking feature in both panels of Fig. 5 is that the COR method gets its highest score when the weights are optimized using the analyzed grid point; using

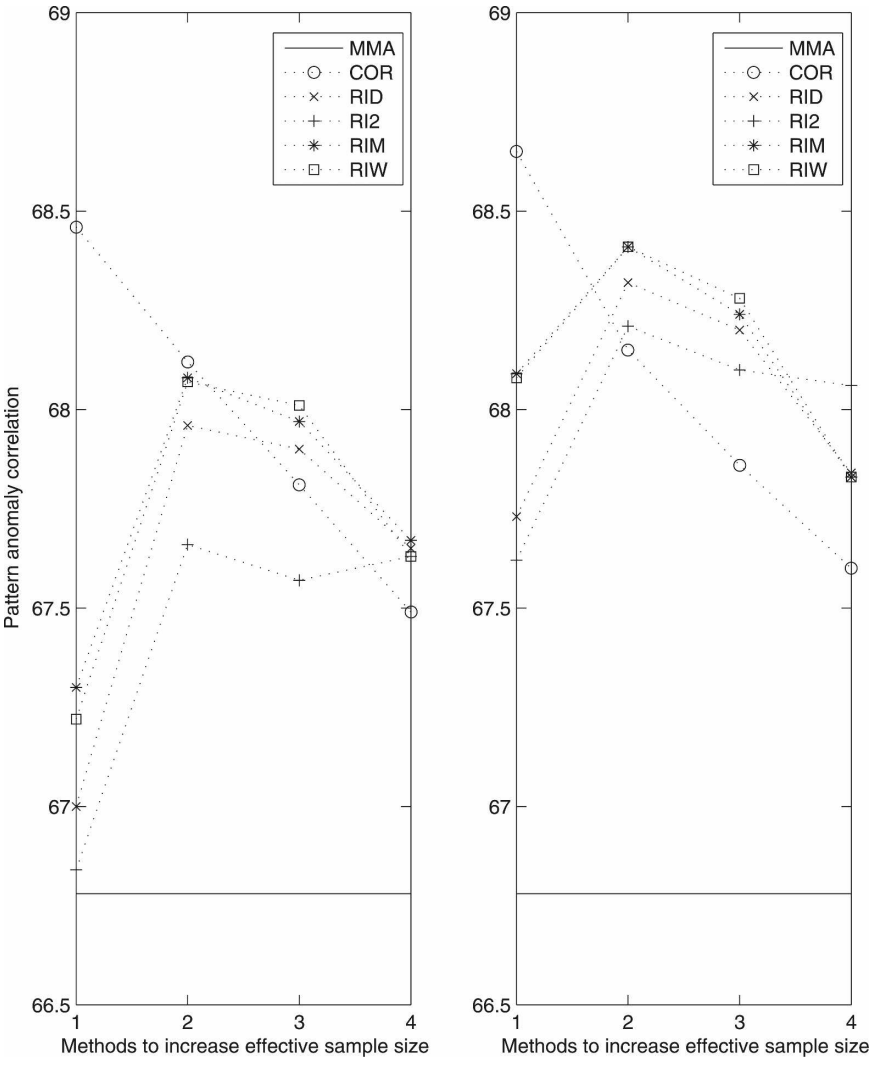

FIG. 5. Average anomaly pattern correlation for the consolidation methods for all leads and initial months. Along the abscissa are the four different effective sampling size strategies.

neighboring grid points deteriorate its skill. COR improves when using all ensemble members. The ridge regression methods, on the other hand are only marginally better than MMA when they use only the ensemble average and data at each grid point to optimize the weights (first entry in left panel). As it will be discussed further in section 9, the choice of λ in the ridging methods were not optimized to have the best skill but rather to have stable solutions for the weights. It is found that for this particular application λs around 75% have skills larger than COR.

b. Consistency

A second aspect assessed in the performance of consolidation methods is how *frequent* these methods outperform MMA. Previous studies (e.g., Hagedorn et al. 2005) have shown that MMA forecasts consistently outperform those from individual ensemble models. The impact of increasing the sample size on consistency is analyzed here using pattern correlation scores over the study domain. Table 3 shows the percentage of cases in which consolidation methods outperform MMA for each of the 6 leads (0–5) and initial months (4 initial months). In the first row, the sampling strategy "ensemble mean: 1×1" is the traditional case in which the weights are optimized using the ensemble means and individual grid points. In this case, the ridging methods outperform MMA around 50% of the cases. The second row shows the percentage of cases for the sampling strategy where the weights are optimized by pooling information from all ensemble members and the closest neighboring grid points (consistent with case 2; right

TABLE 3. Frequency of cases (%) in which ridging consolidation methods outperform MMA for three sampling strategies.

Sampling strategy	RID	RI2	RIM	RIW
Ensemble mean: 1 × 1	50.0	41.7	54.2	54.2
All members: 3 × 3	83.3	75.0	91.7	91.7
All members: All grid points*	91.7	83.3	91.7	91.7

* For a domain 30°S–30°N in the tropical Pacific.

panel of Fig. 5). Here RIM and RIW outperform MMA in more than 90% of the cases. An even further increase is obtained by pooling more information from beyond the study domain (third row). Thus, consistent improvement with respect to the MMA appears to increase when the effective sampling size is large. The table suggests that while the scores may not increase sensationally when increasing the effective sampling size, the benefit is seen in terms of consistent improvements due to having more stable weights (Fig. 3).

c. Gridpoint-wise anomaly correlation

Figure 6 shows the temporal anomaly correlation for all initial months averaged over all leads. The upper panel shows the AC based on MMA consolidation while the lower shows the difference of AC between RID and MMA. In the central equatorial Pacific, the MMA shows correlation values above 0.9 and above 0.7 in most of the tropical Pacific. This is consistent with previous studies (e.g., Stephenson et al. 2005; Hagedorn et al. 2005; Saha et al. 2006), indicating that after SE correction, most current seasonal prediction systems tend to be highly skillful in predicting the evolution of the SST in the central equatorial Pacific. In the lower panel, it is apparent that the improvements of the RID method over MMA are mostly confined to the western Pacific. This result is consistent with Stephenson et al. (2005) who showed a Bayesian approach that outperforms the MMA for seasonal predictions of SST in the equatorial Pacific. Their results show that a major contribution from this gain in skill is the increased reliability of calibrated probability forecasts.

An alternative way in understanding why sophisticated consolidation methods improve in the western Pacific over MMA is that there is apparently a large discrepancy in skill among input models. This allows weighted average methods to screen out models that do not perform well enough. In contrast, in the central and eastern Pacific where all the models perform well after bias correction, there is less chance that the consolidation methods can discriminate between some of the models. An illustration of this is given in Fig. 7, where the AC across the equatorial Pacific is plotted as a function of longitude for the ensemble mean of participating models (upper panels) for leads 0–3, and their corresponding consolidations based on MMA and RID (lower panels). In the western Pacific, the ridging method is able to screen out the less skillful methods

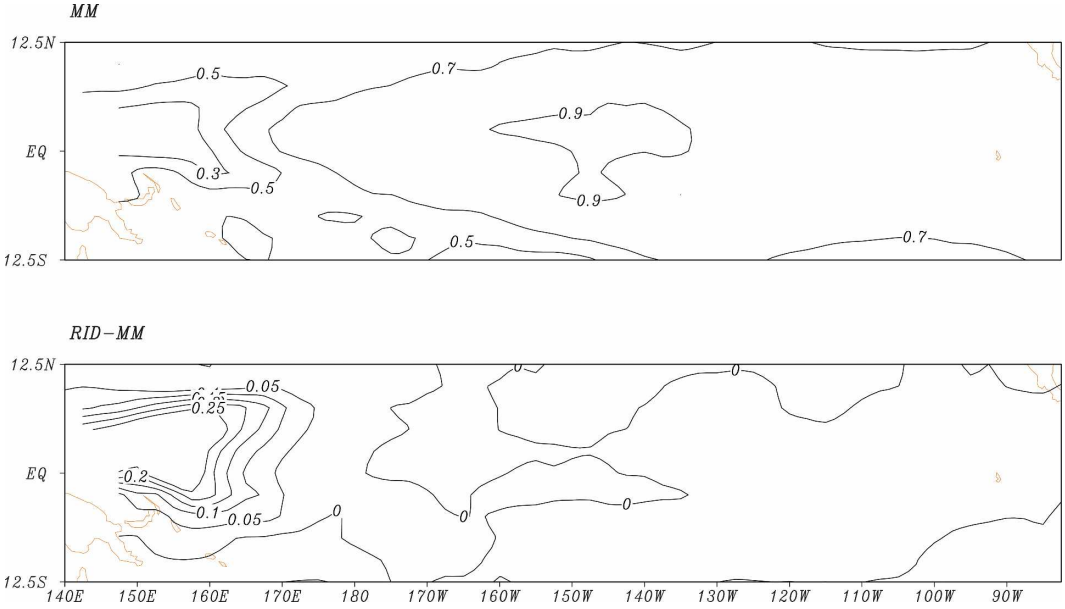

FIG. 6. Grid point by grid point anomaly correlation average over all leads and initial months available of (top) MMA and (bottom) the difference RID − MMA; contour intervals (CIs) 0.2 and 0.05, respectively. Weights were optimized using information from all grid points in the domain and 9 ensemble member of each model.

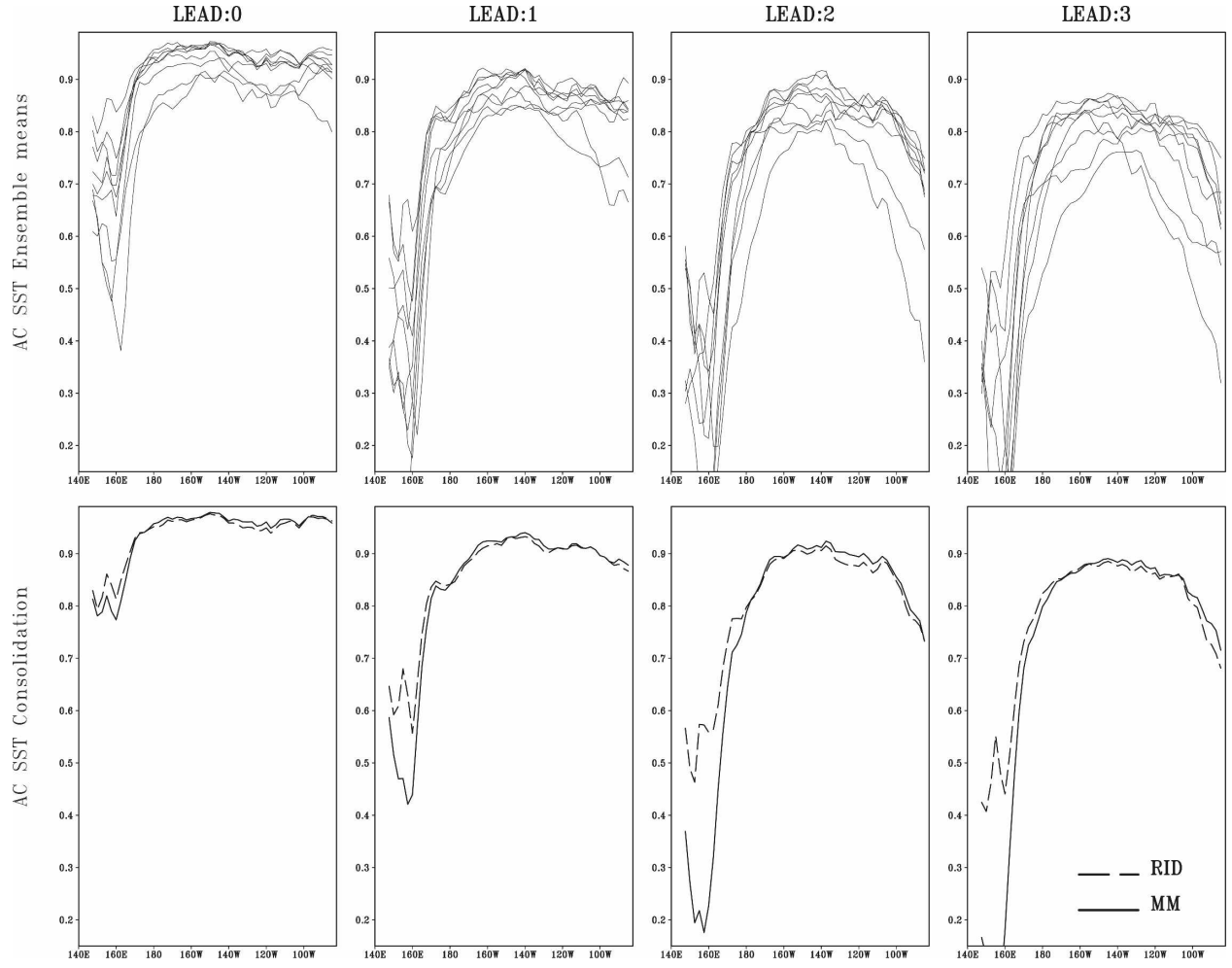

FIG. 7. (top) Anomaly correlation of ensemble means of each participating model along the equator in the Pacific. (bottom) Anomaly correlation of MMA (continuous curve) and RID (dashed curve).

and outdo MMA, especially as the lead increases, whereas in central and eastern Pacific the skill is similar.

8. Probabilistic skill assessment

a. Probability density function

In this section, the procedure to construct a probability density function from optimal weights is described. A common approach to extract probability information from ensemble forecasts is by counting the fraction of ensemble members that falls into predefined categories. Thus, for a given lead time, initial month, and category m, the probability can be expressed as $P_m = K_m/K$, where K_m is the number of ensemble members that are predicted to fall in category m and K is the total number of ensemble members. This is a straightforward procedure when ensemble members have equal weights. A number of other approaches, even for equal weights, are available in the literature, such as the kernel Gaussian (D. Unger 2007, personal communication) which constructs a Gaussian PDF centered at each member forecast with a width proportional to both the historical mean square error and current spread. The Bayesian method (Rajagopalan et al. 2002; Stephenson et al. 2005; Krzysztofowicz and Evans 2008; Luo et al. 2007) and Bayesian model average (Raftery et al. 2005) provide an a posteriori PDF directly.

A challenge arises when the weights of ensemble members are not equal, as is the case for weights derived from several consolidation methods discussed in this paper. As shown in previous sections, optimal weights from sophisticated consolidation methods reflect how well a model performed in a given training period. In other words, a large weight assigned to a particular model indicates a higher likelihood that its

forecast will verify. One way to quantitatively convey this likelihood is by increasing the relative frequency of that particular forecast with respect to the others. Units of the relative frequency, referred to as "stacks," can be defined so as to ensure that in a given ensemble forecast the sum of all relative frequencies is equal to unity. This way, the model's optimized weights determine the number of "stacks" given to its corresponding forecast. This study applies this approach to assess the probabilistic performance of RID and MMA.

The tercile categories are defined by the observed climatology and the limits are determined assuming a Gaussian distribution for the observations. The category limits are thus computed as 0.4308 times the observed standard deviation for each grid point. The same procedure as in the deterministic evaluation was followed, namely, that the forecasts are cross validated (CV-3R), but here the weights of each consolidation method are normalized. The problem is simplified by using each of these categories as a binary event, which means, for example, that the "above normal" is actually "above normal" versus "non–above normal."

An illustration of the approach described is given in Fig. 8. The figure displays a particular multimodel ensemble (81 members) forecast as a set of small vertical bars overlying the observed climatology depicted by a Gaussian distribution, along with the category limits. The value of each forecast is given by its position in the abscissa axis. The thick tall vertical line is the verifying observation. The numbers indicate the fraction of members that fall in each category. This is the probability that the forecast will verify on each tercile according to the issued ensemble forecast. The probability of the three categories adds up to one. In the upper panel, assuming equal weights (the MMA method, for example), there is 67/81 chance that the observation will verify in the upper tercile. In the bottom panel, the height of the small bars is determined by the weights obtained under the RID method for the full data. Since the approach uses the weights to determine the frequency (height of the bars) of forecasts only, the position of the small bars in the bottom panel aligns with those in the upper panel. In this case, the models with higher weights (taller bars) fall in the upper tercile giving a probability of 76.3/81 = 0.942.

b. ROC curves

The ability of consolidation methods to anticipate correctly the occurrence or nonoccurrence that SST anomalies will verify within tercile categories is assessed based on ROC. Figure 9 shows the "area below curve" (ABC) gridpoint-wise. The ABC ranges from 0 to 1, with values above 0.5 indicating that the model is more skillful than climatology and values equal to 1 indicating perfect skill. The figure shows the average of all months and leads 3–5 only, to indicate the more distinct features among the two methods (MMA and RID). For leads 0–2 (not shown) the ABC is closer to 1, but the difference between MMA and RID is negligible. The left panel of Fig. 9 shows that MMA is highly skillful in predicting whether the observation will fall in the upper or lower terciles, particularly in the central Pacific. Skill is lower for the middle class, as reported before by Van den Dool and Toth (1991) and Kharin and Zwiers (2003). In the right panels, positive values indicate the regions where RID are better than MMA. For the upper and lower terciles the improvement is limited to the western Pacific consistent with Fig. 6. There is a tendency for RID to outperform MMA in larger regions in the lower tercile than in the other tercile classes. The performance, averaged over the full study domain is, however, very similar for both RID and MMA. This conclusion differs from Fig. 5, which shows all RID variations to be consistently better (although not by much) than MMA by deterministic measures. An additional Brier score (BS) evaluation indicates RID has slightly lower BS in the western Pacific, mainly because of slightly improved reliability.

9. Discussion and conclusions

This study assesses the performance of seven consolidation methods of multimodel ensemble forecast systems to predict monthly SST in the deep tropical Pacific. The consolidation methods tested vary on degree of sophistication from a simple multimodel average method, MMA, which is used here as benchmark, to ridging regression approaches that take into account past performance and collinearity among models. The evaluation was made in a stringent cross-validation mode, the CV-3RE.

For the sophisticated consolidation methods, the length of the training period should be large enough for the optimization procedure to correctly estimate the covariance matrix **A**, from which the regression coefficients are computed. When the number of participating models is large and the hindcasts are short, the process leads to overfitting. Earlier consolidation attempts in meteorology appear to be working on a grid point by grid point basis and with ensemble averages. This works well with dependent data or under insufficient cross validation but more stringent cross-validation procedures, such as the one introduced here, show that with a short sampling the sophisticated consolidation approaches are only marginally better than the simple multimodel ensemble average. Furthermore, when the

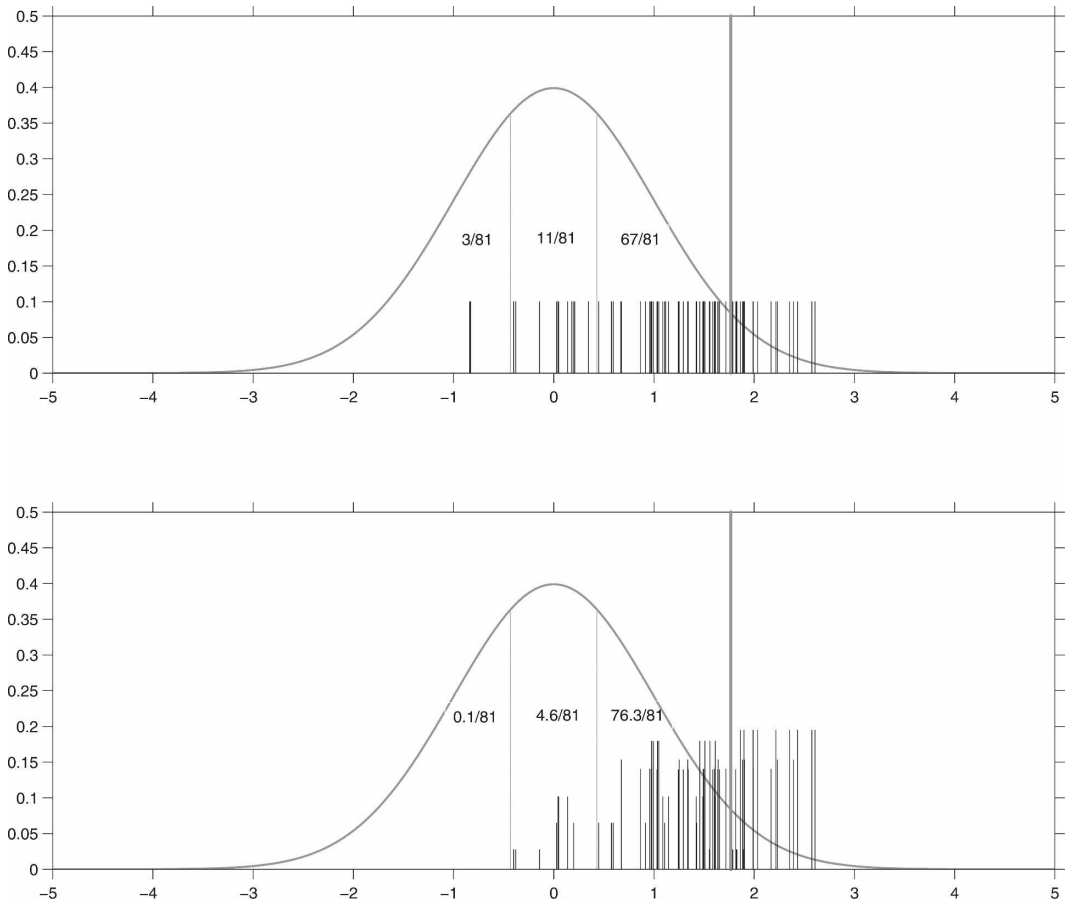

FIG. 8. Illustration of the 3-category PDF construction based on an 81-member multimodel ensemble forecast for a particular grid point, lead, and initial month, along with its corresponding verifying observation (thick tall line) and climatology (Gaussian curve). (top) The probability (fraction number of cases) on each category when the weights are equal, like MMA. (bottom) The probability for the same case but when the weights are optimized using the full data and RID method. Scales on the *y* axis are for the Gaussian distribution. Values of 0.1 in the height of the small bars correspond to a count of 1 out of 81 cases.

number of input models is large, even if there is sufficient data for training, collinearity could produce an ill-conditioned covariance matrix preventing the inverse to be known with sufficient accuracy.

In view of these problems, this study assesses the benefits of increasing the effective sampling size for consolidation methods that deal with collinearity among models, in particular the use of the ridging regression to determine optimal weights. While there are many variations of the ridging methodology, only three (RID, RIM, and RIW) were analyzed in more detail. Skills of these methods were compared with those from the ensemble mean forecast of individual models, the multimodel average method, the skill-weighted methods and the unconstrained regression.

Three approaches to increase the sampling size relative to the number of coefficients to be fitted were tested: reducing the set of input models, mixing information from neighboring grid points, and using information from individual ensemble members. To reduce the set of input models in the first approach, a double-pass consolidation technique was used. After the first pass, models with negative weight were assigned zero weight before performing the second pass. This was not always a successful approach to select and remove bad models because in some cases, the reduced set still produced negative weighs after the second pass, but now for other models. Another, more effective procedure was to remove upfront, models with negative AC. The latter approach was used for the COR and RIW methods. Mixing information from both neighboring grid points, and individual ensemble members had a large positive impact on the stability of the weights and produced some skill improvements on all the sophisticated consolidation methods. For the ridging regression methods, it was found that the benefit in skill by mixing

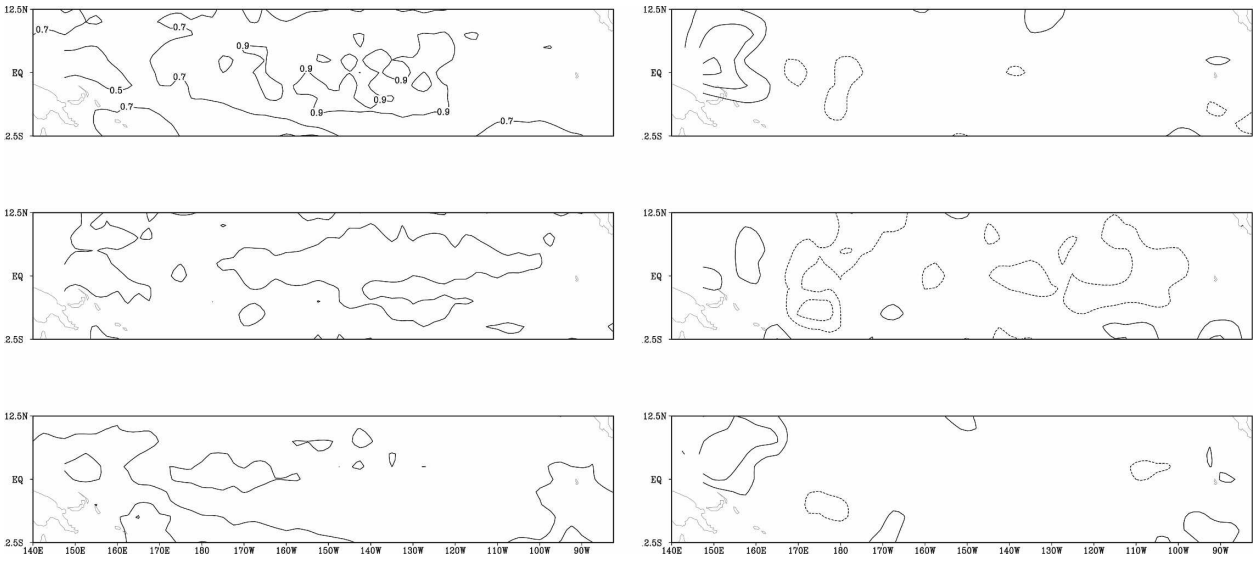

FIG. 9. ABC scores for equal weights, (left) MMA, and (right) difference RID minus MMA for the three categories: (top) upper, (middle) middle, and (bottom) lower tercile. Scores denote average of all initial months available and leads 3–5. CIs are 0.2 for the left panels and 0.02 without the zero contours and with spatial smoothing for right panels.

grid points reached a maximum when the closest neighboring grid points were included. Overall, the COR method outperformed the rest of the consolidation methods when no mixing of gridpoint information was used. The ridging methods performed similarly and reached a maximum in skill when the effective sampling size was increased with information of the closest neighbor grid point. Using individual ensemble members, rather than the ensemble mean, to optimize the weights benefited the overall skill.

The results in this study corroborate that for the tropical Pacific SST, the skill of MMA is, on average, higher than that for any particular model ensemble average. The results also corroborate that, given a short sampling size, other sophisticated consolidation methods fail to show a large improvement over the MMA. However, the study indicates that increasing the effective sampling size does produce more stable weights and more skillful (and much more consistently so) predictions of SST. It is found that sophisticated consolidation methods tend to outperform MMA largely in the western equatorial Pacific and only marginally so in the rest of the ocean basin, which is consistent with results using the Bayesian approach (Stephenson et al. 2005).

The ridging method has been presented here largely in terms of obtaining a stable solution, see especially Fig. 2. Obviously an unstable solution is not good because it will not hold up in independent data. A specific choice of λ to optimize skill was not sought. The skill optimizing approach was tried by DelSole (2007), who sought to find λ (at each grid point) to optimize skill but failed to improve skill that way when it was tested under two layers of cross validation. Nevertheless, the reader must wonder what happens to skill when λ is varied (even if the reason for the variation is stability or asymptotic behavior, or both). Figure 10 shows the do-

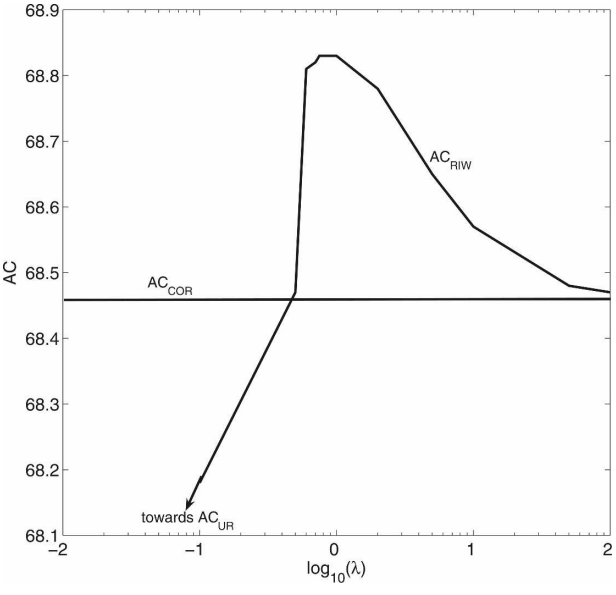

FIG. 10. Anomaly correlation average for all leads and initial months available for RIW and COR (horizontal line) methods as a function of ridging amount (λ). The values are after CV-3R cross validation for the case when weights are optimized using the ensemble mean of each model and without mixing of grid points (corresponding to Fig. 4 and the first entry in left panel of Fig. 5).

main aggregated AC performance of the RIW method (under 3CV-RE) for all leads and starting times combined as a function of λ. For very large λ one obtains the same skill as the COR method, while for negligibly small λ one gets the skill of UR. Based on this graph one gets the impression that skill has a rather flat maximum in the wide range λ = 0.6–2.0, that is, ridging 60%–200%. The quoted λ values indicate surprisingly strong ridging and, in general, much more than is needed for stability alone. It thus appears that the collinearity cannot be exploited in full, at least in the problem at hand. Still, optimized RIW is better than COR, so even for a 200% ridging some of the collinearity is exploited beneficially.

An approach to form a 3-class PDF from the optimized weights was used in which the weights determine the "stacks" associated with each model leaving the forecast values unchanged (i.e., no regression applied). The ROC was used to measure the ability of the methods to anticipate the occurrence or nonoccurrence that the target observation will fall in each tercile category. Results indicate that RID outperform MMA in the western Pacific, consistent with the deterministic evaluation of AC. Analysis of the Brier score (not shown) indicates that RID marginally improves over MMA in the same region. Such improvement is attributed to the more reliable forecasts of RID compared to MMA (not shown). The methods have the lowest skill to predict the middle tercile, in agreement with Van den Dool and Toth (1991). A more efficient treatment may be needed to adjust the PDF generated in association with the ridge regression methods. However, accurate corrections of the SE of the standard deviation and other higher moments of the PDF will be hard to obtain given the shortness of the hindcasts.

Judging the success of a consolidation is difficult because so many diverse issues get mixed in (this also makes claims in the literature hard to accept as universally applicable). One of them is the cross-validation procedure, which requires a study on its own. Given the shortness of the hindcast data in some models, estimates of the SE had large variations depending on which years are left out, and this hurts the performance of some of the models. The verification metric, and implicitly the control forecast, used has an influence on what is considered "success." A final difficulty in judging success is the application itself. If consolidation method A is better than method B, one should not rule out that this finding is application dependent. The prediction of SST has features of appreciable skill and high collinearity, but application to, say, European 2-m temperature prediction (which has lower skill, and less collinearity) may lead to different conclusions about the advantages of a certain variation of ridge regression or any other consolidation method.

Acknowledgments. This work was supported as a pilot project by the Climate Test Bed through NOAA's Climate Program Office. The authors thank Suranjana Saha for providing an organized set of forecast data from the different centers for this study. The authors also thank A. Johansson and A. Vintzileos who provided important corrections and comments to a preliminary draft of the paper, and Tim DelSole for an insightful discussion. Thanks also to two anonymous referees whose comments improved the paper.

REFERENCES

Barnston, A. G., and H. M. Van den Dool, 1993: A degeneracy in cross-validated skill in regression-based forecasts. *J. Climate,* **6,** 963–977.

Behringer, D., 2007: The Global Ocean Data Assimilation System (GODAS) at NCEP. Preprints, *11th Symp. on Integrated Observing and Assimilation Systems for the Atmosphere, Oceans, and Land Surface (IOAS-AOLS),* San Antonio, TX, Amer. Meteor. Soc., 3.3.

Clemen, R. T., and A. H. Murphy, 1986: Objective and subjective precipitation probability forecasts: Some methods for improving forecast quality. *Wea. Forecasting,* **1,** 213–218.

Crone, L. J., L. M. Mcmillin, and D. S. Crosby, 1996: Constrained regression in satellite meteorology. *J. Appl. Meteor.,* **35,** 2023–2035.

DelSole, T., 2007: A Bayesian framework for multimodel regression. *J. Climate,* **20,** 2810–2826.

Derome, J., G. Brunet, A. Plante, N. Gagnon, G. J. Boer, F. W. Zwiers, S. Lambert, and H. Ritchie, 2001: Seasonal predictions based on two dynamical models. *Atmos.–Ocean,* **39,** 485–501.

Doblas-Reyes, F. J., M. Déqué, and J.-P. Piedelevre, 2000: Multimodel spread and probabilistic forecasts in PROVOST. *Quart. J. Roy. Meteor. Soc.,* **126,** 2069–2087.

Golub, G. H., and C. F. Van Loan, 1980: An analysis of the total least square problem. *SIAM J. Numer. Anal.,* **17,** 883–893.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting. Part I: Basic concept. *Tellus,* **57A,** 219–233.

Hamill, T. M., J. S. Whitaker, and X. Wei, 2004: Ensemble reforecasting: Improving medium range forecast skill using retrospective forecasts. *Mon. Wea. Rev.,* **132,** 1434–1447.

——, ——, and S. L. Mullen, 2006: Reforecasts: An important data set for improving weather predictions. *Bull. Amer. Meteor. Soc.,* **87,** 33–46.

Hsiang, T. C., 1976: A Bayesian view on ridge regression. *Statistician,* **24,** 267–268.

Kharin, V. V., and F. W. Zwiers, 2002: Climate predictions with multimodel ensembles. *J. Climate,* **15,** 793–799.

——, and ——, 2003: Improved seasonal probability forecasts. *J. Climate,* **16,** 1684–1701.

Krakauer, N. Y., T. Schneider, J. T. Randerson, and S. C. Olsen, 2004: Using generalized cross-validation to select parameters in inversions for regional carbon fluxes. *Geophys. Res. Lett.,* **31,** L19108, doi:10.1029/2004GL020323.

Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachio-

chi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran, 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. *Science,* **285,** 1548–1550.

——, ——, Z. Zhang, T. LaRow, D. Bachiochi, E. Williford, S. Gadgil, and S. Surendran, 2000: Multimodel ensemble forecasts for weather and seasonal climate. *J. Climate,* **13,** 4196–4216.

Krzysztofowicz, R., and W. B. Evans, 2008: Probabilistic forecasts from the National Digital Forecast Database. *Wea. Forecasting,* **23,** 270–289.

Luo, L., E. F. Wood, and M. Pan, 2007: Bayesian merging of multiple climate model forecasts for seasonal hydrological predictions. *J. Geophys. Res.,* **112,** D10102, doi:10.1029/2006JD007655.

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

Meisner, B. N., 1979: Ridge regression-time extrapolation applied to Hawaiian rainfall normals. *J. Appl. Meteor.,* **18,** 904–912.

Palmer, T. N., and Coauthors, 2004: Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). *Bull. Amer. Meteor. Soc.,* **85,** 853–872.

Peña, M., and Z. Toth, 2008: Forecast error dynamics in a coupled ocean–atmosphere prediction system. Preprints, *19th Conf. on Probability and Statistics,* New Orleans, LA, Amer. Meteor. Soc., 10.3.

Peng, P., A. Kumar, H. Van den Dool, and A. G. Barnston, 2002: An analysis of multimodel ensemble prediction for seasonal climate anomalies. *J. Geophys. Res.,* **107,** 4710, doi:10.1029/2002JD002712.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.,* **133,** 1155–1174.

Rajagopalan, B., U. Lall, and S. E. Zebiak, 2002: Categorical climate forecasts through regularization and optimal combination of multiple GCM ensembles. *Mon. Wea. Rev.,* **130,** 1790–1811.

Reynolds, R. W., N. A. Rayner, T. M. Smith, D. C. Stokes, and W. Wang, 2002: An improved in situ and satellite SST analysis for climate. *J. Climate,* **15,** 1609–1625.

Robertson, A. W., U. Lall, S. E. Zebiak, and L. Goddard, 2004: Improved combination of multiple atmospheric GCM ensembles for seasonal prediction. *Mon. Wea. Rev.,* **132,** 2732–2744.

Saha, S., and Coauthors, 2006: The NCEP Climate Prediction System. *J. Climate,* **19,** 3487–3517.

Stephenson, D. B., C. A. S. Coelho, F. J. Doblas-Reyes, and M. Balmaseda, 2005: Forecast assimilation: A unified framework for the combination of multi-model weather and climate predictions. *Tellus,* **57A,** 253–264.

Stockdale, T. N., 1997: Coupled ocean–atmosphere forecasts in the presence of climate drift. *Mon. Wea. Rev.,* **125,** 809–818.

Tikhonov, A., 1963: Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.,* **4,** 651–667.

Van den Dool, H. M., 1987: A bias in skill in forecasts based on analogues and antilogues. *J. Climate Appl. Meteor.,* **26,** 1278–1281.

——, 1994: Searching for analogues, how long must we wait? *Tellus,* **46A,** 314–324.

——, 2006: *Empirical Methods in Short-Term Climate Prediction.* Oxford University Press, 240 pp.

——, and Z. Toth, 1991: Why do forecasts for near-normal fail to succeed? *Wea. Forecasting,* **6,** 76–85.

——, and L. Rukhovets, 1994: On the weights for an ensemble-averaged 6–10 day forecast. *Wea. Forecasting,* **9,** 457–465.

——, H. J. Huang, and Y. Fan, 2003: Performance and analysis of the constructed analogue method applied to U.S. soil moisture over 1981–2001. *J. Geophys. Res.,* **108,** 8617, doi:10.1029/2002JD003114.

Yun, W. T., L. Stefanova, and T. N. Krishnamurti, 2003: Improvement of the multimodel superensemble technique for seasonal forecasts. *J. Climate,* **16,** 3834–3840.