ENVIRONMETRICS
Environmetrics 2004; **15**: 233–249 (DOI: 10.1002/env.638)

Prediction of sea surface temperature from the global historical climatology network data

Samuel S. P. Shen[1,*,†], Alan N. Basist[2], Guilong Li[3], Claude Williams[2] and Thomas R. Karl[2]

[1]*Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB T6G 2G1, Canada*
[2]*National Climatic Data Center, Asheville, NC 28801, U.S.A.*
[3]*Atmospheric Science Division, Meteorological Service of Canada-Ontario Region, 4905 Dufferin Street, Downsview, ON M3H 5T4, Canada*

SUMMARY

This article describes a spatial prediction method that predicts the monthly sea surface temperature (SST) anomaly field from the land only data. The land data are from the Global Historical Climatology Network (GHCN). The prediction period is 1880–1999 and the prediction ocean domain extends from 60°S to 60°N with a spatial resolution $5^\circ \times 5^\circ$. The prediction method is a regression over the basis of empirical orthogonal functions (EOFs). The EOFs are computed from the following data sets: (a) the Climate Prediction Center's optimally interpolated sea surface temperature (OI/SST) data (1982–1999); (b) the National Climatic Data Center's blended product of land-surface air temperature (1992–1999) produced from combining the Special Satellite Microwave Imager and GHCN; and (c) the National Centers for Environmental Prediction/National Center for Atmospheric Research Reanalysis data (1982–1999). The optimal prediction method minimizes the first-M-mode mean square error between the true and predicted anomalies over both land and ocean. In the optimization process, the data errors of the GHCN boxes are used, and their contribution to the prediction error is taken into account. The area-averaged root mean square error of prediction is calculated. Numerical experiments demonstrate that this EOF prediction method can accurately recover the global SST anomalies during some circulation patterns and add value to the SST bias correction in the early history of SST observations and the validation of general circulation models. Our results show that (i) the land only data can accurately predict the SST anomaly in the El Nino months when the temperature anomaly structure has very large correlation scales, and (ii) the predictions for La Nina, neutral, or transient months require more EOF modes because of the presence of the small scale structures in the anomaly field. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: climate data reconstruction; empirical orthogonal function (EOF); sea surface temperature (SST); GHCN data; data error analysis; mean square error (MSE)

1. INTRODUCTION

Historically, the surface air temperature observations with thermometers near the land surface were more accurate than the SST observations, but provided limited coverage over the globe. In order to

*Correspondence to: S. S. P. Shen, Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB T6G 2G1 Canada.
†E-mail: shen@ualberta.ca

Published online 6 February 2004

Received 25 February 2003
Accepted 30 June 2003

obtain global coverage of surface temperature, one has to spatially interpolate the historical in situ observations from scattered sites to the places which were data-void. The past interpolations were basically done in the way that uses the SST observations to interpolate a complete SST field and uses the land surface air temperature (LSAT) to interpolate a LSAT field. This interpolation is usually called reconstruction (Smith *et al*., 1998). However, the sparse coverage and large bias in the earlier historical SST observations imply large errors in the traditional interpolation or reconstruction results (Rayner *et al*., 1996). It is worthwhile, both mathematically and physically, to extrapolate the more accurate LSAT observations to ocean domains to infer the SST field by using the fact that the monthly SST and LSAT anomalies are inflexibly connected through general circulations of atmosphere, particularly during the strong SST anomaly months. This extrapolation will help us with the SST bias correction and validation of general circulation models before the World War II. The purpose of this article is to provide a method that predicts the SST field using the monthly LSAT data from the Global Historical Climatology Network (GHCN). Our predicted SST field will be of $5^\circ \times 5^\circ$ resolution and our prediction region will be limited to the latitude band of (60°S, 60°N).

The method of our spatial prediction is to use EOFs as basis functions over both land and ocean and to calculate the EOF coefficients. Our method is similar to the interpolation scheme of Kim *et al*. (1996), who used spherical harmonics as basis functions and computed the coefficients of the spherical harmonics for surface temperature. In Kim *et al*. (1996), the observational data were considered errorless. Folland *et al*. (2001) and Shen *et al*. (1998) showed that when an optimal method is used, the observational error might be larger than the sampling error. Thus, the errors of the observational data are considered in this article.

Our method minimizes the MSE between the true field and the predicted LSAT and SST over both land and ocean and differs from the reconstruction of Smith *et al*. (1996, 1998), which minimizes the error only over the area where observed data existed. The monthly SST reconstruction in the tropical Pacific ocean made by Meyers *et al*. (1999) used a method similar to that of Smith *et al*. (1998). Both studies paid attention to the number of EOF modes supported by data, and hence their number of modes used for prediction varied from month to month. Rayner *et al*. (1996) also used a variable number of modes to fill up data gaps and generated a complete Global Ice and Sea Surface Temperature field.

Kaplan *et al*. (1997) proposed an SST reconstruction method that minimizes a penalty function. The function is the observational projection error variance inversely weighted by the observational error covariance, plus the model error variance inversely weighted by the model error covariance. A Kalman filter was used to solve the minimization problem. Since a Markov climate model was introduced, the reconstructed field was supposed to be dynamically consistent. However, the Kalman recursive scheme requires both the covariance matrices of the model errors and observational errors. When assuming the spatial independence of the error fields, the covariance matrices become diagonal and its elements are the error variances. Kaplan *et al*. (1997) assessed the observational error variances using the α^2/n approach, where n is the number of intrabox measurements for the monthly data. The model error covariance matrix was determined by a string of simplified assumptions of the Markov model. Our observational error estimate is also done in the α^2/n way, but using the data of the Special Sensor Microwave Imager (SSMI) satellite and the multiple-station observations in a grid box.

When the modeling part of Kaplan *et al*. (1997) is dropped, their method becomes the weighted, linear multivariate regression. The weight matrix can be different. For example, Shriver and O'Brien (1995) adopted a diagonal matrix whose elements are the squares of the numbers of observations in the SST grid boxes. Further, when the observational error variance becomes uniform, the method becomes a projection. Our method is essentially a projection, but observational error variances are considered.

However, our method is limited by the stationarity assumption like most other EOF approaches. The assumption is that the majority of the variability and teleconnections in the climate circulation reflected in the EOFs can be effectively identified from the data record of the recent period. Clearly, this assumption ought to be scrutinized in every EOF application. This paper has checked the compatibility of the data for computing EOFs. In the conclusion section, we will also discuss how additional data can further improve prediction accuracy. Despite this limitation, due to the consideration of observational errors, our method of using land stations to infer the temperature field over the ocean is a step toward the ultimate prediction of the global temperature field from limited high quality data sources over the last 120 years. These high quality data may include both LSAT observations and some accurately observed SST.

The data and procedures for computing EOFs with area-factor are described in Section 2. The EOF prediction method and the error estimation are presented in Section 3. Examples of the prediction are given in Section 4. Conclusions and discussion are contained in Section 5.

2. DATA AND PROCEDURES FOR COMPUTING EOFs WITH AREA-FACTOR

Let $T(\hat{\mathbf{r}}, t)$ be the true, hence error-free, temperature anomaly field over the working domain Ω, i.e. the (60°S, 60°N) latitude band, where $\hat{\mathbf{r}}$ is the position vector and t indicates time. Its corresponding observed temperature is $\tilde{T}(\hat{\mathbf{r}}_j, t)$, which is equal to the error-free $T(\hat{\mathbf{r}}_j, t)$ plus the random observational error E_j for a given location $\hat{\mathbf{r}}_j$ and a time t. Our reconstruction uses the GHCN data and thus our data are described using the GHCN grid boxes (Peterson and Vose, 1997). Let $\mathcal{G}$ denote the GHCN observation network. Then

$$\tilde{T}(\hat{\mathbf{r}}_j, t) = T(\hat{\mathbf{r}}_j, t) + E_j \tag{1}$$

is the observed temperature anomaly data at the grid box $\hat{\mathbf{r}}_j \in \mathcal{G}$, where the position vector $\hat{\mathbf{r}}_j$ is the center of the grid box $\hat{\mathbf{r}}_j$. Hereafter, $\hat{\mathbf{r}}_j$ is used to denote either the grid box or its centroid. The random measurement error over a grid box is E_j, which includes sampling error, instrumental error, reading error, and other random erroneous influences, and $T(\hat{\mathbf{r}}_j, t)$ is the true value of the temperature anomaly field at the box $\hat{\mathbf{r}}_j$. The position vector $\hat{\mathbf{r}}_j$ runs through the centers of all the $5^\circ \times 5^\circ$ GHCN data boxes for different values of index j. It is impossible to compute the exact error E_j, but some error statistics, such as the error variance, can be estimated. Fortunately, our reconstruction needs only these statistics.

The systematic errors are assumed to have been removed from the raw data after the NCDC's homogenization procedure, and hence the remaining random error is uncorrelated with either the anomaly field or the errors at other locations:

$$\langle E_i T \rangle = 0, \quad \langle E_i E_j \rangle = 0 \quad \text{when} \quad i \neq j \tag{2}$$

where $\langle \cdot \rangle$ denotes the ensemble average.

The covariance function is defined by

$$\rho(\hat{\mathbf{r}}, \hat{\mathbf{r}}') = \langle T(\hat{\mathbf{r}}, t) T(\hat{\mathbf{r}}', t) \rangle \tag{3}$$

The continuous EOFs $\psi_n(\hat{\mathbf{r}})$ are the eigenfunctions of the kernel $\rho(\hat{\mathbf{r}}, \hat{\mathbf{r}}')$ over the domain Ω.

The working domain Ω is divided into regular J grid boxes of $5^\circ \times 5^\circ$ resolution, where $J = 1728$. The symmetric, discrete eigenvalue problem with area-factor can be formed following the method described in North *et al*. (1982) and Shen *et al*. (1998):

$$\sum_{j=1}^{J} \left[\sqrt{A_i}\rho(\hat{\mathbf{r}}_i, \hat{\mathbf{r}}_j)\sqrt{A_j} \right] \left\{ \psi_n(\hat{\mathbf{r}}_j)\sqrt{A_j} \right\} = \lambda_n \left\{ \psi_n(\hat{\mathbf{r}}_i)\sqrt{A_i} \right\}, \quad i = 1, 2, \ldots, J \tag{4}$$

Here

$$A_j = R^2 \times \left(\frac{5}{180}\pi \right) \times \left(\frac{5}{180}\pi \right) \cos \phi_j \tag{5}$$

is the area of the grid box $\hat{\mathbf{r}}_j$, where ϕ_j is the latitude of the center of the box j and R is the radius of Earth, which is approximately 6376 km. The discrete EOFs are the eigenvectors of the eigenvalue problem

$$[\mathbf{v}_j]_{j=1}^{J} = \left[\psi_n(\hat{\mathbf{r}}_j)\sqrt{A_j} \right]_{j=1}^{J} \tag{6}$$

The magnitude of each eigenvector is always normalized to one. The variance of the nth mode is given by the eigenvalue $\lambda_n, n = 1, 2, \ldots$.

Three sources of monthly data are used for the EOF computing: (i) the blended SSMI/GHCN product (1992–1999) (Basist *et al*., 1998; Peterson *et al*., 2000); (ii) the OI/SST data (1982–1999) (Reynolds and Smith, 1994); and (iii) the Reanalysis 2-m temperature data (1982–1999). The Reanalysis data in 1982–1999 are used only to fill in the missing data in the blended product over the land in 1982–1999. Since the blended product was available in 1992–1999, the land data in 1982–1991 are all from the Reanalysis. Because OI/SST covers all the ocean boxes, the Reanalysis data over the ocean are not used. The Reanalysis data in 1949–1981 are not used over either ocean or land in our EOF computing.

The anomalies of the three data sets are relevant to the Reanalysis 1961–1990 climatology calculated in the following way. For the 1992–1999 blended SSMI/GHCN product, the climatology is given by

$$\text{Blended}_{\text{clim}} = \text{Blended}_{92\text{–}99\ \text{mean}} - [\text{Reanalysis}_{92\text{–}99\ \text{mean}} - \text{Reanalysis}_{61\text{–}90\ \text{clim}}]$$

The anomalies of the blended product are hence relative to this climatology.

The anomalies of the OI/SST are computed based upon the same scheme. Namely, the OI/SST climatology is the 1982–1999 mean off-setting by the difference between the 1982–1999 Reanalysis mean and the 1961–1990 Reanalysis mean. Rigorously speaking, the off-setting amount should be calculated from the sea surface temperature rather than the 2-m air temperature over the ocean. However, since the 1961–1990 Reanalysis 2-m data are used as the overall climatology on both land and ocean, we thus employed the difference of 2-m temperature as a substitute for the corresponding SST.

The anomalies of the Reanalysis are relevant to its own 1961–1990 climatology. These anomalies have too large variances in the high-latitude and too small variances over the tropics. For instance, even after the temperature anomalies are averaged over 5-degree boxes, the anomaly value in a box can still reach as high as 25°C. Although some large anomalies have been observed in the high-latitude

areas, the observed values are several degrees lower and have never exceeded 18°C. The large variance of the Reanalysis data in high latitude may be due to the fact that the Reanalysis model cannot handle sea ice well. The model produces some unrealistically large fluctuations in high latitude, particularly in the regions where land, water and sea ice intersect. Because of the problems of too large fluctuations in high latitude and over-smoothness in the lower latitude, the Reanalysis data are not used as an independent piece of information; rather, they are used only as the auxiliary data to the blended SSMI/GHCN and OI/SST data. This decision was made after numerous numerical experiments with and without the inclusion of the 1949–1981 Reanalysis data.

Finally, the compatibility of the three data sets needs to be checked since the blended data, OI/SST, and Reanalysis were derived based upon different principles. We checked the compatibility by plotting the completed data blended from three sources for EOF computing. That no unusually large gradients appeared on the plots implies the desired compatibility.

3. PREDICTION FROM THE GHCN DATA WITH ERRORS AND PREDICTION MSE

The true temperature anomaly is expanded in terms of normalized, covariance-based, EOFs $\psi_m(\hat{\mathbf{r}})$ as follows:

$$T(\hat{\mathbf{r}}, t) = \sum_{m=1}^{\infty} T_m(t)\psi_m(\hat{\mathbf{r}}) \tag{7}$$

The EOF coefficients T_m are given by the integral

$$T_m(t) = \int_{\Omega} T(\hat{\mathbf{r}}, t)\psi_m(\hat{\mathbf{r}})d\Omega \tag{8}$$

The spectral approach of prediction is to use the GHCN data to estimate these EOF coefficients.

The integration (8) is possible only when accurate observed data are available everywhere. However, this is certainly not true. The observed data are only on scattered sites or are pre-processed to the scattered $5^\circ \times 5^\circ$ boxes, and the data are not perfectly accurate but have errors. Thus the above integral has to be computed via a discretization with a weight assigned to every data box, at which the observed data with errors are used to make the numerical integration. The discrete integration of the expression (8) is

$$\hat{T}_m(t) = \sum_{j=1}^{N} \tilde{T}(\hat{\mathbf{r}}_j, t)\psi_m(\hat{\mathbf{r}}_j)w_j^{(m)} \tag{9}$$

where N, less than the total number of boxes J in the working domain Ω, is the total number of GHCN boxes with data, $\tilde{T}(\hat{\mathbf{r}}_j, t)$ are the observed values in GHCN boxes (averaged from the station observations and including the observational errors as expressed in (1)), and $w_j^{(m)}(t)$ are the weights for the observed data, which vary according to mode number m and month t. The number of data boxes N also varies according to time. $\hat{T}(\hat{\mathbf{r}}, t) = \sum_m \hat{T}_m(t)\psi_m(\hat{\mathbf{r}})$ is called the 'reconstructed' field if $\hat{\mathbf{r}}$ is over the entire Ω, but the 'predicted' field (or called extrapolation) if $\hat{\mathbf{r}}$ is over the ocean. When $\hat{\mathbf{r}}$ is over the

land, $\hat{T}(\hat{\mathbf{r}}_j, t)$ is an interpolation. In general, the extrapolation and interpolation are all spatial predictions. Thus, the 'reconstructed field' refers to the joint field of the predicted SST and the interpolated LSAT over the land.

The weights are determined by the condition of minimum mean square error between the true field and the reconstructed field over the entire (60°S, 60°N) band. A normalization condition is imposed to the weights, i.e. the sum of the weights should be equal to the area of the integration domain

$$\sum_{j=1}^{N} w_j^{(m)} = A \tag{10}$$

This constraint makes the mean square error take a local minimum (i.e. a restricted minimum) rather than the global minimum without constraint. Adoption of this constraint is based upon computational reasons. When we find optimal weights by minimizing the mean square error, the weights may not always be positive and their magnitudes can have a large variance. In this case, leaving out one data box may have a great impact on the prediction result if this box has a very large positive or negative weight. Exclusion or inclusion of this box can result in an enormously large departure from the normal condition of climate, which is physically untrue. When this constraint is imposed, an unbiased prediction is ensured and an unrealistic departure from normal will not happen.

One can also justify this normalization condition directly from the condition of unbiased estimation of $\langle T_m \rangle$. This unbiased condition is

$$\langle T_m \rangle = \int_\Omega \langle T(\hat{\mathbf{r}}, t) \rangle \psi_m(\hat{\mathbf{r}}) d\Omega \approx \langle \hat{T}_m(t) \rangle = \sum_{j=1}^{N} \langle \tilde{T}(\hat{\mathbf{r}}_j, t) \rangle \psi_m(\hat{\mathbf{r}}_j) w_j^{(m)} \tag{11}$$

Equation (1) implies that $\langle \tilde{T}(\hat{\mathbf{r}}_j, t) \rangle = \langle T(\hat{\mathbf{r}}_j, t) \rangle$. Apparently, the normalization constraint becomes a necessary condition when both the mean temperature field and the EOFs have small spatial variations, namely when $\langle T(\hat{\mathbf{r}}, t) \rangle$ and $\psi_m(\hat{\mathbf{r}})$ are close to contants. However, when the spatial variance of the mean temperature field is large, the normalization constraint may not be necessary to ensure the unbiased estimate of higher order EOF coefficients.

With the estimated EOF coefficients, the reconstructed field over the entire Ω is given by

$$\hat{T}(\hat{\mathbf{r}}, t) = \sum_{m=1}^{M} \hat{T}_m(t) \psi_m(\hat{\mathbf{r}}) \tag{12}$$

where M is the mode truncation number and is usually determined by the criterion of explaining a certain percentage of variance with the first M modes. The more modes the data can support, the smaller is the residual MSE E_R^2 defined in (17) below. However, including the modes which are not supported by observations can only add noise to the results (Smith *et al*., 1998).

The field of the mean square error of the reconstruction over Ω is

$$E^2(\hat{\mathbf{r}}) = \langle (T(\hat{\mathbf{r}}, t) - \hat{T}(\hat{\mathbf{r}}, t))^2 \rangle \tag{13}$$

This formula can be expressed in terms of EOFs as follows:

$$E^2(\hat{\mathbf{r}}) = E_M^2(\hat{\mathbf{r}}) + E_R^2(\hat{\mathbf{r}}) \tag{14}$$

where

$$E_M^2 = \sum_{m=1}^{M} \epsilon_{(m)}^2 \psi_m^2(\hat{\mathbf{r}}) \tag{15}$$

is the MSE computed from the first M modes with

$$\epsilon_{(m)}^2 = \langle (T_m(t) - \hat{\boldsymbol{T}}_m(t))^2 \rangle \tag{16}$$

and the residual $E_R^2(\hat{\mathbf{r}})$ is the higher modes contribution to E^2,

$$E_R^2(\hat{\mathbf{r}}) = \sum_{m=M+1}^{\infty} \lambda_m \psi_m^2(\hat{\mathbf{r}}) \tag{17}$$

This term is the truncation error and is determined by the truncation order M and the properties of the eigenvalues and eigenfunctions, and it is not directly related to the weights of the data boxes. However, an indirect relation exists since the truncation order M is related to the density and distribution of the GHCN data boxes. Usually, a large number of data boxes support a larger number of modes, which means a larger M. The truncation order M, of course, also depends on properties of the anomaly field. However, if a prediction is successful after a validation check or a comparison with a field close to truth, then the truncation error should be reasonably small and independent of the sampling errors. Namely, if

$$T(\hat{\mathbf{r}}, t) = \sum_{m=1}^{M} \hat{\boldsymbol{T}}_m(t) \psi_m(\hat{\mathbf{r}}) + T_R(\hat{\mathbf{r}}, t) \tag{18}$$

then

$$\langle (T_m(t) - \hat{\boldsymbol{T}}_m(t)) T_R(\hat{\mathbf{r}}, t) \rangle = 0 \tag{19}$$

If this ensemble mean is far away from zero, then the prediction is necessarily not a good approximation to the original field.

Equation (15) is an approximation based upon the lack-of-correlation assumption that

$$\langle (T_m(t) - \hat{\boldsymbol{T}}_m(t)) \, (T_n(t) - \hat{\boldsymbol{T}}_n(t)) \rangle = 0$$

if $m \neq n$. Here $\hat{\boldsymbol{T}}_m(t)$ is computed from the GHCN data and optimal weights. In order for this lack-of-correlation assumption to hold, the GHCN sampling must be able to support sufficiently many EOF modes that can reflect the teleconnections between land and ocean. Unfortunately, these assumptions cannot be verified a priori with the short length of the data stream. They can only be tested by examining the prediction results.

The total MSE of the error field is the integral of $E^2(\hat{\mathbf{r}})$,

$$\epsilon^2 = \frac{1}{A}\int_{\Omega} E^2(\hat{\mathbf{r}})d\Omega \tag{20}$$

where A is the total area of the (60°S, 60°N) latitude band. The overall accuracy of the prediction is measured by this total MSE. The optimization minimizes the error resulting from the first M modes by finding the optimal weights for all the data boxes. This apparently is not the best optimization criterion since only the minimum MSE over the ocean region is what one wants to know. However, an optimal algorithm for minimizing the MSE over the ocean domain only has not yet been worked out.

With the EOF approach, the MSE formula above for the first M modes can be written as a sum of the MSE at each mode,

$$\epsilon^2 = \sum_{m=1}^{M} \epsilon_{(m)}^2 \tag{21}$$

where the MSE for mode m is defined by (16). Since each term in the sum of (21) is non-negative, the minimization of ϵ^2 becomes the minimization of each term, i.e. min $\epsilon_{(m)}^2$. In terms of EOFs, $\epsilon_{(m)}^2$ can be written as

$$\epsilon_{(m)}^2 = \sum_{n=1}^{\infty} \lambda_n \left[\delta_{mn} - \sum_{j=1}^{N} \psi_n(\hat{\mathbf{r}}_j)\psi_m(\hat{\mathbf{r}}_j)w_j^{(m)}\right]^2 + \sum_{j=1}^{N} \langle E_j^2\rangle \left(\psi_m(\hat{\mathbf{r}}_j)w_j^{(m)}\right)^2 \tag{22}$$

where δ_{mn} is the Kroneker delta, which is equal to one if $m = n$ and 0 otherwise. This formula bears some similarity with (16) and (22) of Kim *et al*. (1996) and (31) of Shen *et al*. (1994) without accounting for observational errors, and (27) of Shen *et al*. (1998) taking observational errors into account. The first term of the above formula is basically the numerical integration error of

$$\int_{\Omega} \psi_m(\hat{\mathbf{r}})\psi_n(\hat{\mathbf{r}})d\Omega \tag{23}$$

The second term is the observational error weighted by EOFs and optimal weights.

To minimize the MSE $\epsilon_{(m)}^2$ with constraint (10) for the weights, a Lagrange function is defined as

$$J_m[\mathbf{w}] = \epsilon_{(m)}^2(\mathbf{w}) + 2\Lambda_m\left(\sum_{j=1}^{N} w_j^{(m)} - A\right) \tag{24}$$

where Λ_m is the Lagrange multiplier, whose unit is the square of the temperature unit. The critical point for the Lagrange function is determined by

$$\frac{\partial J_m}{\partial w_j^{(m)}} = 0 \quad \text{and} \quad \frac{\partial J_m}{\partial \Lambda_m} = 0 \tag{25}$$

which leads to a set of $N+1$ linear equations for $w_j^{(m)}, j = 1, 2, \ldots, N$, and Λ_m:

$$\sum_{j=1}^{N} \rho(\hat{\mathbf{r}}_i, \hat{\mathbf{r}}_j)\psi_m(\hat{\mathbf{r}}_i)\psi_m(\hat{\mathbf{r}}_j)w_j^{(m)} + \langle E_i^2\rangle \psi_m^2(\hat{\mathbf{r}}_i)w_i^{(m)} + \Lambda_m = \lambda_m \psi_m^2(\hat{\mathbf{r}}_i) \tag{26}$$

$$\sum_{j=1}^{N} w_j^{(m)} = A \tag{27}$$

Here, $[\rho(\hat{\mathbf{r}}_i, \hat{\mathbf{r}}_j)], i,j = 1, 2, \ldots, N$, is a sub-matrix of the covariance matrix $[\rho(\hat{\mathbf{r}}_i, \hat{\mathbf{r}}_j)], i,j = 1, 2, \ldots, J$, without area-factor and is computed from the OI/SST data, blended SSMI/GHCN data and Reanalysis data. Please note that these data are not the observed GHCN box data $(\tilde{\boldsymbol{T}}(\hat{\mathbf{r}}_j, t), i = 1, 2, \ldots, N)$, where $N \ll J$. Thus, the above equation implies that the weights are determined by the covariance structure and the errors of the observed data, not the observed data themselves.

With $\epsilon_{(m)}^2(t), m = 1, 2, \ldots, M$, and $\lambda_n, \psi_n(\hat{\mathbf{r}}), n = 1, 2, \ldots, Y$, one can compute $E^2(\hat{\mathbf{r}}, t)$ by using the formulas (14)–(16). Finally, the field of the root mean square error (RMSE) for the first M-mode is

$$E_M(\hat{\mathbf{r}}, t) = \sqrt{E_M^2(\hat{\mathbf{r}}, t)} \tag{28}$$

Since the residual error is not included, this result may be regarded as the lower bound of the prediction error.

The optimal weights $w_j^{(m)}$ are a function of time and are obtained by solving the optimal-weight equations (26) and (27) for every month (since N changes from month to month), and for every mode m until the truncation mode M. Having obtained $w_j^{(m)}(t)$, $j = 1, 2, \ldots, N(t)$, $m = 1, 2, \ldots, M$, and $t = 1, 2, \ldots, Y$, one can finally produce the prediction using (9) and (12).

However, (26) requires the error variance $\langle E_j^2\rangle$ for all the GHCN data used in the prediction; yet, the error variance for the monthly data on grid boxes has never been determined, although Jones *et al*. (1997) estimated the grid box error for decadal data. Fortunately, the optimal weights are insensitive to the exact size of the error variance as tested in Shen *et al*. (1998) for optimal averaging of the tropical Pacific SST. That paper used the approximate error $(\langle E_j^2\rangle = 0.09[^\circ\mathrm{C}]^2)$ estimated by Parker *et al*. (1994) and Reynolds and Smith (1994). Our objective is to predict the SST from the GHCN land data and hence the errors required should be for the GHCN data, and we cannot use the oceanic value from Parker *et al*. (1994). A method similar to that of Williams *et al*. (2000) and Kaplan *et al*. (1997) is used to assess the error variance of the GHCN data, i.e. to compare the differences between the SSMI satellite and in situ observations. Assume that the error field over a grid box is white noise. Suppose that a grid box has n stations and SSMI data are also available. The mean square difference can be found between the data from the satellite and each station. The average of the mean square difference among the n stations is taken as an approximation of the MSE of one-station observation. The MSE of the n-station observation may be approximated by the one-station MSE scaled down by n. The one-station MSE may be approximated in many different ways, such as using the station nearest to the centroid of the grid box as the most representative station, or averaging the mean square differences of all individual stations. Since the absolute truth of the temperature over a box is unknown, one can only use a reference to assess the mean square difference. SSMI satellite data may be used as the reference to the GHCN data, or one can use the blended product as the reference. The latter under-estimates the GHCN error since the blended product has already used the GHCN data. However, this under-estimate

has some advantage in computing during the North Hemisphere winter. In the winter, much of the middle and high latitude areas are snow-covered and SSMI data are not available. Hence SSMI cannot be used as a reference. The blended product, using only the in situ observations now, can still give a relative value of the temperature variance over a grid box. Of course, this estimate must be limited to the boxes with two or more stations. For the grid box with only one station, its error has to be approximated by the error from the nearest grid box on the same latitude. With this consideration, we estimate the error variance for the grid box $\hat{\mathbf{r}}_i$ by the following formula:

$$\langle E_i^2 \rangle = \frac{1}{n}\left(\frac{1}{\alpha}\sum_{t=1}^{\alpha}[Blended(t) - Stn_{ic}(t)]^2\right) \tag{29}$$

where α is the number of months of data (January 1992–June 1999), $Stn_{ic}(t)$ is the station nearest to the centroid in the grid box $\hat{\mathbf{r}}_i$, at time t, *Blended(t)* is the blended data for the grid box at time t, and n is the total number of stations in the grid box. The above formula is applied to every grid box with $n \geq 2$.

This is certainly not the ideal way to compute the error variance of the GHCN data, yet it can provide a reasonable error approximation for (26) to calculate our optimal weights. Further discussion on the error calculation is given at the end of Section 5.

The above approximation formula and resulting error variances are applied to the GHCN data to predict the SST anomaly field in the latitude band (60°S, 60°N) in the data period of 1880-current. The prediction credibility is assessed using both temporal and spatial correlations, and prediction examples are discussed for El Nino, La Nina, neutral, data-dense and data-sparse months. These results are compared with two reference data sets: the prediction using Jones' land data and the OI/SST.

4. NUMERICAL RESULTS OF PREDICTION

The (60°S, 60°N) latitude band includes $J = 1728$ five-degree grid boxes. The number of GHCN data boxes over land, including islands, varies from the $N = 120$ in 1880 to 288 in 1930, to the maximum (close to 500) in the period from the 1950s to the mid-1980s. Due to the time lag (around 10 years) of data reporting from various countries, the number of the GHCN data boxes dropped to around 300 in the 1990s. Of course, this number will be back to nearly 500 in the future. These land GHCN data are used to predict the SST. To establish an impression of the spatial distribution of the GHCN data, the coverage of the GHCN data for three different months is displayed in Figure 1.

To examine the prediction accuracy, we used the January 1982–December 1998 OI/SST as the truth and calculated the temporal correlation between the prediction and OI/SST for each grid box. The distribution of the correlation is shown in Figures 2(a) and (b) for January and July monthly SST anomalies, respectively. In this prediction, 12 EOF modes are used. For both January and July, the correlation is positive everywhere. In many areas the correlations have reached 0.8. For January, the maximum correlation areas are Tropical and Northern Atlantic, Northern Pacific, Southern Pacific and Indian Ocean. The correlation over the eastern Tropical Pacific is low. This may be due to the large variance of the SST in this region and the prediction may not be accurate during the transient months between El Nino and La Nina. The July prediction has consistently higher correlation than the January prediction, particularly in the Southern Hemisphere. Again, the Atlantic prediction appears more reliable than the Pacific prediction. Different from the January prediction, July's Tropical Pacific prediction yields a remarkably high correlation around 0.8. In general, it may be concluded that the

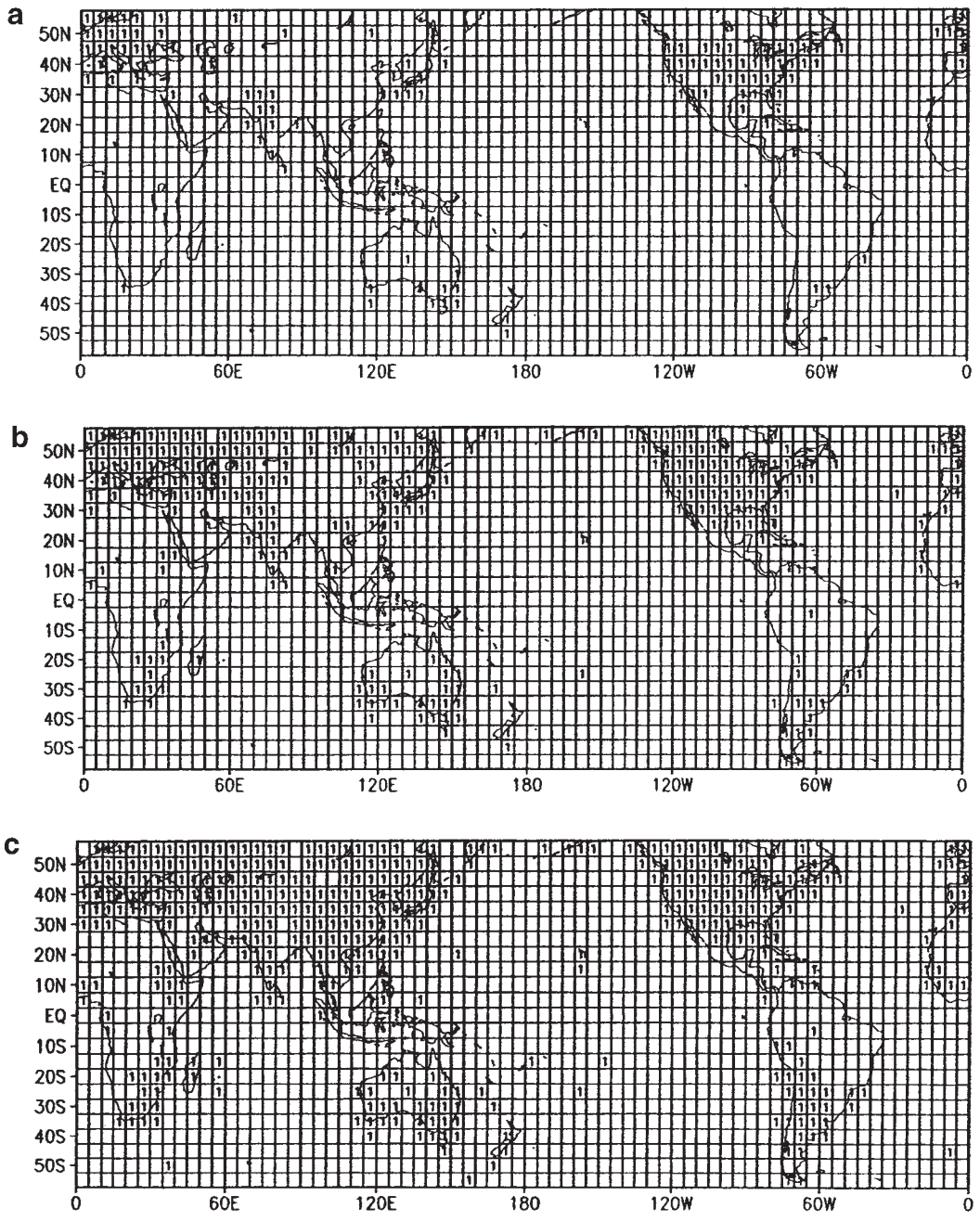

Figure 1. Representation of the global coverage of the GHCN data boxes: (a) January 1880, (b) January 1930 and (c) January 1983. The boxes with '1' indicate the existence of observed data for the month

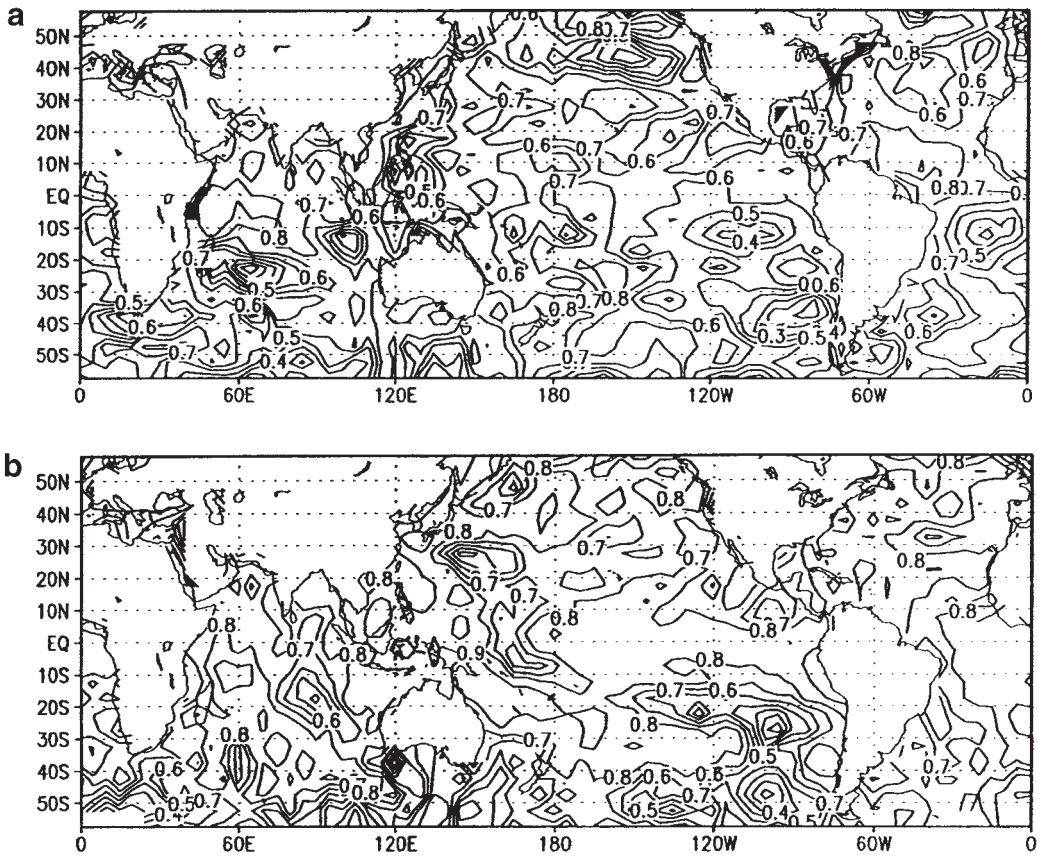

Figure 2. Temporal correlation from 1982 to 1998 between the GHCN predicted SST and the OI/SST for every grid box: (a) January and (b) July

July prediction is more accurate than January from the correlation point of view. The correlation reflects mainly the phase, and the prediction and the 'truth' are likely to have the same sign but they may differ greatly in magnitude.

To examine the success of the prediction of a large spatial signal, such as ENSO, one still needs to investigate some individual predictions and validate with the OI 'truth' or get some confidence of the results from predictions using different data sets. In this article, the different data set is the monthly land data of Jones (Jones *et al*., 1997). Let us examine January 1992. This month had moderate El Nino conditions. Two analyses are compared: the prediction and the OI/SST. The pattern correlation between the two fields is 0.82. The prediction explains 65% of the variance of the 'truth'. This is considered very good since 12 modes explain only 86% of the total variance in the EOF computing.

The next example is January of 1983, which had a major El Nino. Again, the OI/SST (Figure 3(a)) is used to validate the prediction. The largest SST anomaly occurred in the east equatorial Pacific, where a value exceeded 4°C. The prediction from the GHCN data correctly captured the location, magnitude and structure of this anomaly over the ocean (Figure 3(b)). Figure 3(c) shows a prediction, based on Jones' data set and the same EOFs, and it demonstrates slightly smaller values. All three fields correctly place a large area of negative anomalies in the southeastern Pacific, and the amplitude

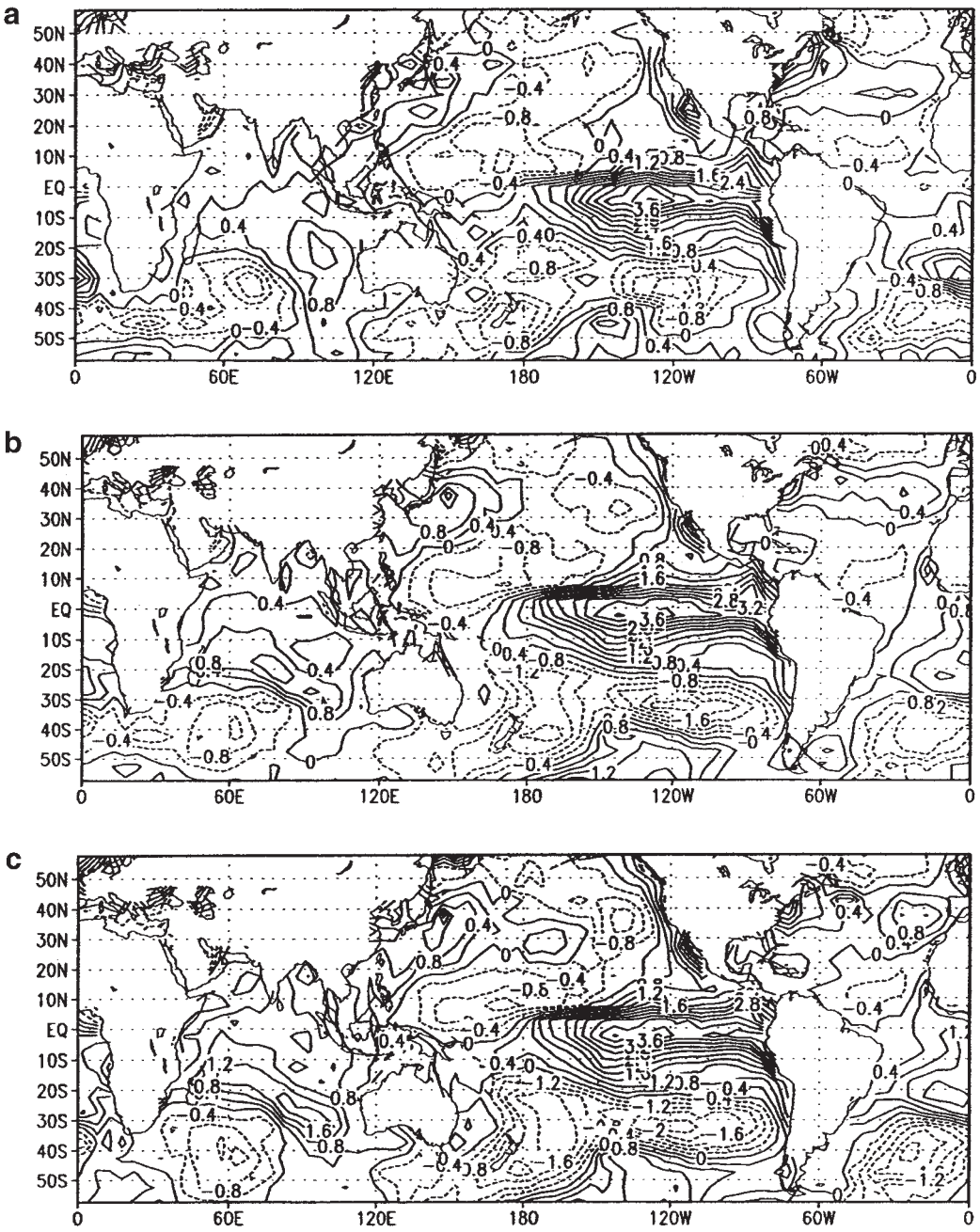

Figure 3. Prediction of the surface temperature anomalies of January 1983 (a strong El Nino month): (a) the OI/SST, (b) the prediction from GHCN data and (c) the prediction from Jones data

is similar for all of them. The mutual correlations among the three fields are high: between GHCH prediction and OI is 0.86, between Jones prediction and OI is 0.84, and between GHCN and Jones predictions is 0.96. The extremely high correlation between the predictions of GHCN and Jones data is due to the fact that the two data sets used basically the same raw station data with some different aggregation procedures. This high correlation also indicates the insensitivity of the prediction result to the data error. Thus, although it is difficult to obtain an accurate estimate of the error variance for data, a reasonable approximation of the data error can still generate the results of the same quality.

The predictions for moderate La Nina conditions like June 1988 are not as good as those for strong El Nino conditions, but can still be reliable and get the positive and negative phases right in most areas (figures are not shown). In the case of June 1988, the pattern correlations are 0.77 for GHCN prediction and OI/SST, 0.76 for Jones prediction and OI/SST and 0.96 for GHCN and Jones predictions.

The predictions for neutral conditions (i.e. the transient conditions between El Nino and La Nina) are less satisfactory. June 1995 was in such a condition. The agreement among the three data sets (OI/SST, GHCN prediction and Jones prediction) is not good (Figures are not shown), because of many short-scale spatial structures presented in the transient months. The pattern correlations are 0.70 between GHCN prediction and OI/SSI, 0.79 between Jones prediction and OI/SST, and 0.87 between GHCN and Jones predictions. To resolve the small-scale structure, one has to use high order EOFs. However, because the observations are over land and few islands, neither GHCN nor Jones' data can support many modes. In addition, because of the limited number of years of data used in the EOF computing, the higher order modes contain much noise. Bringing higher modes into prediction can introduce more noise. The prediction using only the lower modes is generally smoother and has less amplitude than that of the observations in neutral months.

The above predictions, although validated by the OI/SST data, were built upon the EOF framework that was partly made of the OI/SST. Thus, it is necessary to validate the prediction for the months not within the OI/SST period. We chose a month in the earlier period of the 120 years of GHCN data for another validation of prediction: July 1884, a summer month following the volcanic explosion of Krakatoa in Indonesia on 27 August 1883 (Press and Siever, 1982, p. 365). The volcanic plume caused much cooler than average global temperatures. Figures 4(a) and (b) show the predictions using the GHCN and Jones' data, respectively. Figure 4(a) shows that negative anomalies occur over the Indian Ocean, the Western Pacific and most of the Atlantic. The negative anomalies were most likely caused by the cold air, which in turn was caused by the high reflexibility of volcanic dust. Consequently, the atmosphere and ocean surface absorbed less than normal solar radiation. The Jones' data prediction (Figure 4b) yields similar results. The similarities between the two predicted fields increase our confidence about the SST prediction, although no validation data are available for confirmation.

Figure 5 shows the variation of the area-averaged error with respect to time and season. The error is computed by

$$E_{ave}(t) = \frac{1}{4\pi \sin 60^\circ} \sum_{i=1}^{1728} E_M(\hat{\mathbf{r}}_i, t) \left(\frac{5}{180}\right)^2 \sin \phi_i \tag{30}$$

where 1728 is the total number of boxes in the (60°S, 60°N) band, $E_M(\hat{\mathbf{r}}_i, t)$ is computed by (28), and ϕ_i is the latitude of the centroid of the box $\hat{\mathbf{r}}_i$. In the calculation, the mode truncation number is $M = 12$. The error decreases with respect to time until 1990 when the reported GHCN observations were dramatically reduced. The reduction was due to a time lag of international data aggregation. The four curves in the figure show the seasonality of the averaged errors: larger in northern hemisphere winters and smaller in summers.

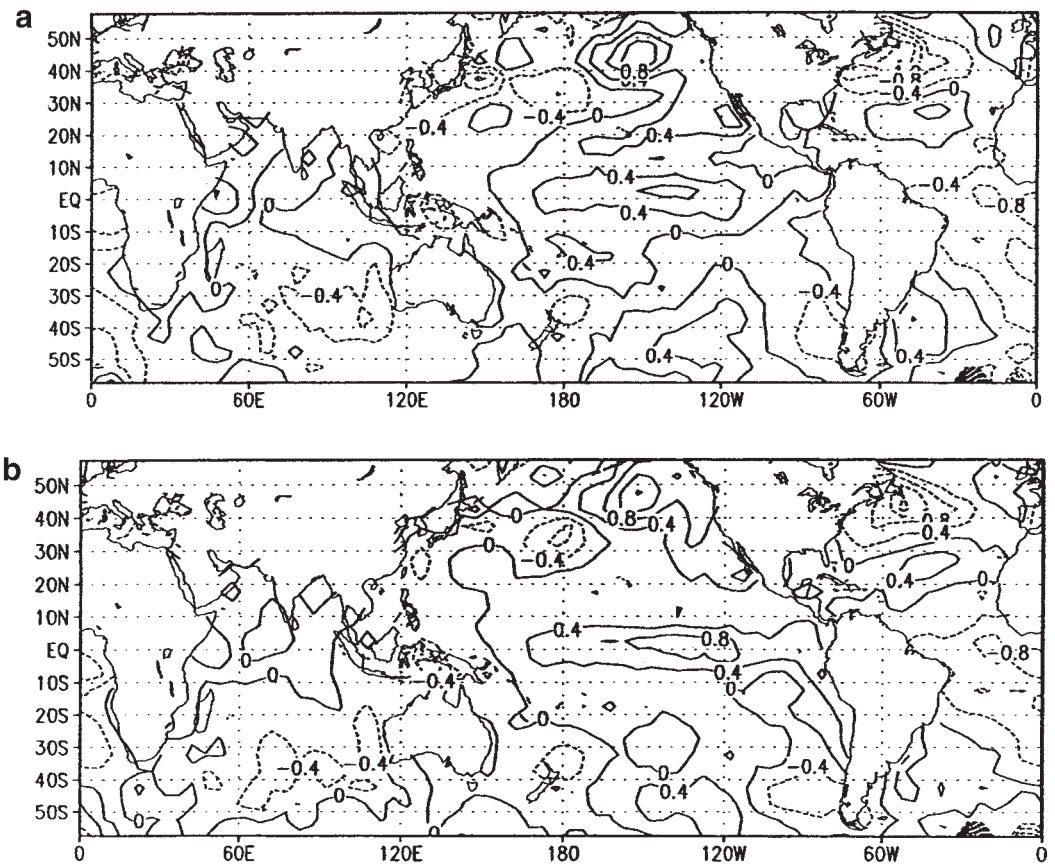

Figure 4. Prediction of the surface temperature anomalies of July 1884 (a data sparse month and the 'year without summer'): (a) the prediction from GHCN data and (b) the prediction from Jones data

5. CONCLUSIONS AND DISCUSSION

We have extended the spectral method of Kim *et al*. (1996) from the spherical harmonics basis to the EOF basis and provided a method to predict the monthly SST anomaly field from the land GHCN data. The study period is 1880–1999 and the study domain is the latitude band (60°S, 60°N) with $5° \times 5°$ spatial resolution. The OI/SST data (1982–1999), blended product (1992–1999) and Reanalysis data (1982–1999) have been used to calculate the spatial structure, i.e. EOFs. The optimization minimizes the first-M-mode mean square error between the true and predicted anomalies. In the optimization process, the data errors of the GHCN boxes have been used, and their contribution to the prediction error has been taken into account. Numerical experiments have shown that this EOF prediction method can recover the SST anomalies of the El Nino events with a remarkable accuracy considering that the prediction is anchored on land data only. We conclude that (i) the land only data can accurately predict the SST anomaly in the El Nino months when the temperature anomaly structure has very large correlation scales, and (ii) the predictions for La Nina, neutral, or transient months require more EOFs

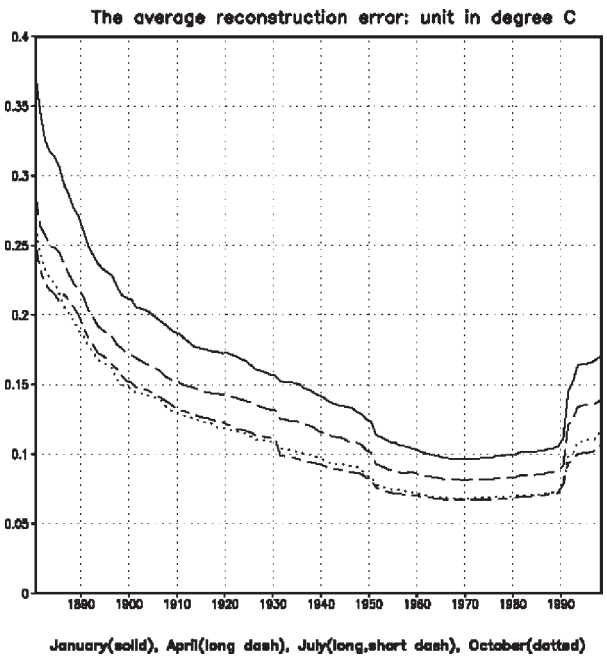

Figure 5. The area-averaged RMSE $E_{ave}(t)$ (see (24)) varies as a function of time for January, April, July and October

modes because of the presence of the small scale structures in the anomaly field. The success of our prediction is mainly the consequence of the strong tele-connection reflected in EOFs.

The prediction is less satisfactory or unreliable during the transition periods between El Nino and La Nina, when the teleconnections between the land and ocean surface are weak or out of phase. Learning from these results, we will attempt to improve the prediction by introducing some critical SST observations, but only a small subset of high quality SST observations will be used. Another data set is the Microwave Sounding Unit (MSU) satellite data for atmosphere temperature that will be calibrated by the in situ land-surface air temperature data to provide the coverage of the snow-covered areas in high latitudes. This data source will reduce the usage of the Reanalysis data in the EOF computing and make the EOF patterns closer to truth, since the Reanalysis data have too large variances in high latitudes and too small variances in equatorial areas.

The error estimate for the GHCN data can be improved when the satellite data coverage can be extended to a longer period. One improvement is to use the average MSE for all the stations and the improved satellite data as the reference, i.e.

$$\langle E_i^2 \rangle = \frac{1}{n^2} \sum_{j=1}^{n} \frac{1}{\alpha} \sum_{t=1}^{\alpha} [Stn_j(t) - Sat(t)]^2 \tag{31}$$

where α is the number of months of data available, $Stn_j(t)$ is the jth station in the grid box $\hat{\mathbf{r}}_i$ at time t, $Sat(t)$ is the satellite data for the grid box at time t, and n is the total number of stations in the grid box. This can be an important improvement when the coverage of the satellite data is extended to higher latitude areas.

The residual error $E_R^2(\hat{\mathbf{r}})$ defined in (14) is never estimated in our article. This error is likely large enough to be not negligible. If the data streams for computing EOFs are sufficiently long, say over 50 years, one might approximate this error directly by (17), in which the upper limit of the summation ∞ is replaced by 50. However, the mixture of higher modes may produce much computational noise and hence lead to an over-estimate of the residual error. This interesting problem is worth further investigation.

ACKNOWLEDGEMENTS

We appreciate our stimulating discussions with Tom Peterson and Jay Lawrimore. Francis Zwiers helped us to improve both the presentation and scientific quality of the article. Basist thanks the NOAA Office of Global Programs for a research grant. Shen thanks the National Climatic Data Center/NOAA for hosting him as a Visiting Scientist when most of this work was done. He also thanks the US National Research Council for the Associateship award, MITACS (Mathematics of Information Technology and Complex Systems) for a research grant, the NOAA Office of Global Programs for a research grant, and the Chinese Academy of Sciences for an Overseas Assessor's research grant and for the Well-Known Overseas Chinese Scholar award.

REFERENCES

Basist AN, Peterson NC, Peterson TC, Williams CN. 1998. Using the special sensor mi-crowave/imager to monitor land surface temperature, wetness, and snow cover. *J. Appl. Meteor.* **37**: 888–911.

Folland CK, Rayner NA, Brown SJ, Smith TM, Shen SSP, Parker DE, Macadam I, Jones PD, Jones RN, Nicholls N, Sexton DMH. 2001. Global temperature change and its uncertainties since 1861. *Geophys. Res. Lett.* **28**: 2621–2624.

Jones PD, Osborn TJ, Briffa KR. 1997. Estimating sampling errors in large-scale temperature averages. *J. Climate* **10**: 2548–2568.

Kaplan A, Kushnir Y, Cane MA, Blumenthal MB. 1997. Reduced space optimal analysis for historical datasets: 13 years of Atlantic sea surface temperatures. *J. Geophys. Res.* **102**: 27, 835–27, 860.

Kim K-Y, North GR, Shen SSP. 1996. Optimal estimation of spherical harmonic components from a sample with spatially nonuniform covariance statistics. *J. Climate* **9**: 635–645.

Meyers SD, O'Brien JJ. 1999. Prediction of monthly SST in the Tropical Pacific Ocean during 1868–1993 using adaptive climate basis functions. *Mon. Wea. Rev.* **127**: 1599–1612.

North GR, Bell TL, Cahalan RF, Moeng FJ. 1982. Sampling errors in the estimation of empirical orthogonal functions. *Mon. Wea. Rev.* **110**: 699–706.

Parker DE, Jones PD, Folland CK, Bevan A. 1994. Interdecadal changes of surface temperature since the late nineteenth century. *J. Geophys. Res.* **99**: 14,373–14,399.

Peterson TC, Vose RS. 1997. An overview of the global historical climatology network temperature database. *Bull. Am. Meteor. Soc.* **78**: 2837–2849.

Peterson TC, Basist AN, Williams C, Grody N. 2000. A blended satellite/in situ near-global surface temperature data set. *Bull. Am. Meteor. Soc.* **81**: 2157–2164.

Press F, Siever R. 1982. *Earth*, 3rd edition, W.H. Freeman and Co., p. 365.

Rayner NA, Horton EB, Parker DE, Folland CK, Hackett RB. 1996. Version 2.2 of the Global Sea-Ice and Sea Surface Temperature data set, 1903–1994. *Tech. Rep. 74*, Hadley Climate Center, 39 pp.

Reynolds RW, Smith TM. 1994. Improved global sea surface temperature analysis using optimum interpolation. *J. Climate* **7**: 929–948.

Shen SSP, North GR, Kim K-Y. 1994. Spectral approach to optimal estimation of the global average temperature. *J. Climate* **7**: 1999–2007.

Shen SSP, Smith TM, Ropelewski CF, Livezey RE. 1998. An optimal regional averaging method with error estimates and a test using tropical Pacific SST Data. *J. Climate* **11**: 2340–2350.

Shriver JF, O'Brien JJ. 1995. Low-frequency variability of equatorial Pacific ocean using a new pseudostress dataset: 1930–1989. *J. Climate* **8**: 2762–2786.

Smith TM, Reynolds RW, Livezey RE, Stokes DC. 1996. Prediction of historical sea surface temperatures using empirical orthogonal functions. *J. Climate* **9**: 1403–1420.

Smith TM, Livezey RE, Shen SSP. 1998. An improved method for interpolating sparse and irregularly distributed data onto a regular grid. *J. Climate* **11**: 1717–1729.

Williams CN, Basist AN, Peterson TC, Grody N. 2000. Calibration and validation of land surface temperature anomalies derived from the SSM/I. *Bull. Am. Meteor. Soc.* **81**: 2141–2156.

