Optimal Estimation of the Climatological Mean

BALACHANDRUDU NARAPUSETTY AND TIMOTHY DELSOLE

George Mason University, Fairfax, Virginia, and Center for Ocean–Land–Atmosphere Studies, Calverton, Maryland

MICHAEL K. TIPPETT

International Research Institute for Climate and Society, Palisades, New York

(Manuscript received 21 November 2008, in final form 23 March 2009)

ABSTRACT

This paper shows theoretically and with examples that climatological means derived from spectral methods predict independent data with less error than climatological means derived from simple averaging. Herein, ‘‘spectral methods’’ indicates a least squares fit to a sum of a small number of sines and cosines that are periodic on annual or diurnal periods, and ‘‘simple averaging’’ refers to mean averages computed while holding the phase of the annual or diurnal cycle constant. The fact that spectral methods are superior to simple averaging can be understood as a straightforward consequence of overfitting, provided that one recognizes that simple averaging is a special case of the spectral method. To illustrate these results, the two methods are compared in the context of estimating the climatological mean of sea surface temperature (SST). Cross-validation experiments indicate that about four harmonics of the annual cycle are adequate, which requires estimation of nine independent parameters. In contrast, simple averaging of daily SST requires estimation of 366 parameters—one for each day of the year, which is a factor of 40 more parameters. Consistent with the greater number of parameters, simple averaging poorly predicts samples that were not included in the estimation of the climatological mean, compared to the spectral method. In addition to being more accurate, the spectral method also accommodates leap years and missing data simply, results in a greater degree of data compression, and automatically produces smooth time series.

1. Introduction

Traditionally, the climatology is defined as the distribution of a random draw from the system conditioned on the phase of the annual or diurnal cycle. This definition is appropriate for cyclostationary systems—that is, systems whose statistics are invariant with respect to a shift in time equal to an integral multiple of a specified period—but may not be appropriate for systems exhibiting trends or other secular variations. As such, the climatological distribution depends on time, or more precisely on the phase of the annual or diurnal cycle. In any case, the climatological distribution is not known and therefore must be estimated from finite samples. There are at least two approaches to estimating the mean of the climatological distribution. The first, which we call simple averaging, is to average the state with respect to fixed phase of the annual and diurnal cycle. For instance, the climatological mean for 1 January would be estimated as the average of all 1 January states in a historical record. In practice, the resulting daily averages are smoothed to remove day-to-day variations, with the precise smoothing operation being essentially arbitrary. The second approach, which we call the spectral method, is to fit the time series to a sum of sines and cosines that are periodic on specified time scales. This approach also involves an essential arbitrariness, namely the number H of periodic functions used to fit the time series.

To decide which method of estimating climatological means is better, and to specify the arbitrary elements within each method, a criterion for ranking the quality of climatological means needs to be defined. Perhaps the most satisfactory criterion is how well the climatological mean produced by these methods predicts independent samples—that is, samples that were not used to derive the climatological mean itself. This criterion is justified

Corresponding author address: Balachandrudu Narapusetty, Center for Ocean–Land–Atmosphere Studies, 4041 Powder Mill Rd., Suite 302, Calverton, MD 20705.
E-mail: bala@cola.iges.org

DOI: 10.1175/2009JCLI2944.1

by the following standard fact: the parameter μ that comes as close as possible to a random process Y—in the sense that it minimizes the mean square residual

$$\epsilon = \langle (Y - \mu)^2 \rangle, \tag{1}$$

where brackets $\langle \cdot \rangle$ denote the expectation with respect to the climatological distribution—is the mean of Y. That is, the minimum ϵ is obtained when $\mu = \langle Y \rangle$. The parameter μ will vary with time if the climatological distribution varies with time, as is the case when annual and diurnal cycles are present. This result provides a basis for ranking different estimates of the climatological mean: the estimate that gives the smallest value of ϵ *in independent data* is considered more accurate than other estimates. The restriction to independent data is key. In many practical situations, we are not interested in defining a climatological mean for the available sample, but rather in defining the climatological mean for various other independent analyses. For instance, the climatological mean often is used to define ''anomalies,'' but if the climatological mean is wrong then the ''anomaly'' and climatological mean become mixed.

The relation between simple averaging and the spectral method is clarified considerably by a proof, given in the appendix, that simple averaging is a special case of the spectral method. For example, in the case of estimating daily climatologies, the two methods are identical if the spectral method contains the gravest 182 annual harmonics. Thus, the question of whether the spectral method is ''better'' than simple averaging reduces to the question of what is the ''best'' value of H, since either method can be recovered by suitable choice of H. If simple averaging is supplemented with a smoothing procedure, the smoothing can be interpreted as shrinking the spectral amplitudes at high frequencies. As the smoothing shrinks high-frequency components without significantly altering the low-frequency components, the method increasingly resembles the spectral method.

Given the above relation between the two methods, we can anticipate that the spectral method yields better forecasts of independent data than simple averaging when the length of the observational record is modest. To see this, consider estimating the climatological mean of daily data. In simple averaging, the daily climatology requires estimating at least 365 parameters (i.e., the sample mean for each day). In the spectral method, the daily climatology requires estimating $2H + 1$ parameters, where H is the number of annual harmonics; the parameters are the amplitudes of the sine and cosine function for each harmonic, plus the constant term. If H is small, say less than 10, then clearly the spectral method involves estimating many fewer parameters than simple averaging. A basic precept in statistical theory is that when the number of parameters estimated in a model becomes large relative to the set of observations, overfitting becomes a problem. Specifically, if an overfitted model is used to predict independent data, then the resulting prediction error ϵ tends to be larger than predictions generated by a model with fewer parameters. These considerations suggest that if H is small, the spectral method should produce smaller mean square error ϵ in independent data than simple averaging.

Previous studies have also constructed climatologies using sums of sines and cosines (Barnett 1981; Straus 1983; Epstein 1988), often specifying H arbitrarily; Barnett (1981) and Straus (1983) used the annual and semiannual harmonics. Epstein (1988) chose a value of H by estimating the spectral coefficients of each year independently and then testing the hypothesis that the coefficients vanish, using the year-to-year variability as a measure of the uncertainty in the coefficients. However, the use of a single year at a time to estimate spectral coefficients results in loss of accuracy. Epstein (1988) also included spectral estimation in the spatial dimension as well as in the temporal dimension. Because the climatological mean is anticipated to be dominated by large scales owing to the large-scale nature of the solar insolation, this approach can be argued to be superior to applying the spectral method to each grid point individually and independently. However, while spectral estimation in the spatial dimension is straightforward for a domain with periodic boundary conditions, which for the globe would involve spherical harmonics, it is inconvenient for nonperiodic domains, such as oceanic basins bounded by realistic coasts. In this study, we simply focus on the time domain.

The purpose of this paper is to demonstrate that simple climatological averages produce less accurate predictions of independent data than climatological means estimated from the spectral method. This fact will be demonstrated by using cross-validation techniques to estimate the prediction error in independent datasets. The results are easily interpreted as a consequence of overfitting. The spectral method also accommodates leap years and missing data very simply, in contrast to traditional averaging. In addition, the spectral method results in a greater degree of data compression than simple averaging, in the sense that it describes the annual cycle with many fewer parameters. Another obvious benefit of the spectral method is that it automatically produces smooth time series, in contrast to simple averaging, and gives optimal estimates in a least squares sense. The methodology can be extended to estimate climatological mean of any variable irrespective of its spatial boundary conditions. We will demonstrate the above claims using SST estimates from the Advanced Very High Resolution Radiometer (AVHRR) infrared satellite (Reynolds et al. 2007). We will also give summaries of the annual cycle of SST, which are of interest in their own right.

2. Estimation of the climatological mean

The climatological mean of y is defined as a time-dependent function $y_c(t)$, dependent only on the phase of the annual or diurnal cycle, that minimizes

$$\epsilon = \langle [y(t) - y_c(t)]^2 \rangle, \tag{2}$$

where the brackets define an expectation with respect to the climatological distribution. Let the N-dimensional vector $\mathbf{y}$ contain a realization of the random process y at times t_i, $i = 1, 2, \ldots, N$. One estimate of the climatological mean is obtained by averaging over times that occur at the same phase of the annual or diurnal cycle. This method will be called simple averaging, and the resulting estimate will be denoted $y_{\mathrm{SA}}(t_i)$.

An alternative approach is to assume the climatological mean has the functional form

$$y_{\mathrm{SM}}(t_i) = a_0 + \sum_{j=1}^{H} [a_j \cos(\omega_j t_i) + b_j \sin(\omega_j t_i)] \qquad i = 1, 2, \ldots, N, \tag{3}$$

where $\omega_j = 2\pi j/P$, P is the period, and H is the truncation parameter, and then to determine the parameters a_j and b_j that minimize the mean square difference between $y_{\mathrm{SM}}(t_i)$ and $y(t_i)$. To find the optimal parameters, we write the climatological mean in the form

$$\mathbf{y}_{\mathrm{SM}} = \mathbf{X}\mathbf{z}, \tag{4}$$

where $\mathbf{X}$ is an $N \times (2H + 1)$ matrix of sinusoidal values and $\mathbf{z}$ is a $(2H + 1)$-dimensional vector of amplitudes. That is,

$$\mathbf{X} = \begin{pmatrix} 1 & \cos(\omega_1 t_1) & \ldots & \cos(\omega_H t_1) & \sin(\omega_1 t_1) & \ldots & \sin(\omega_H t_1) \\ 1 & \cos(\omega_1 t_2) & \ldots & \cos(\omega_H t_2) & \sin(\omega_1 t_2) & \ldots & \sin(\omega_H t_2) \\ \vdots & \vdots & \ldots & \vdots & \vdots & \ldots & \vdots \\ 1 & \cos(\omega_1 t_N) & \ldots & \cos(\omega_H t_N) & \sin(\omega_1 t_N) & \ldots & \sin(\omega_H t_N) \end{pmatrix} \quad \mathbf{z} = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_H \\ b_1 \\ \vdots \\ b_H \end{pmatrix}. \tag{5}$$

The sum square residual (SSR) between $\mathbf{y}_{\mathrm{SM}}(t_i)$ and $y(t_i)$ is

$$\mathrm{SSR} = \sum_i \left(y_i - \sum_j X_{ij} z_j \right)^2 = (\mathbf{y} - \mathbf{X}\mathbf{z})^{\mathrm{T}} (\mathbf{y} - \mathbf{X}\mathbf{z}), \tag{6}$$

where superscript T denotes the matrix transpose. The coefficient vector $\mathbf{z}$ that minimizes the sum square residual (SSR) is a standard regression problem with solution

$$\mathbf{z}_{\mathrm{LS}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}. \tag{7}$$

The annual cycle for the spectral method is then defined as

$$y_A^{\mathrm{SM}}(t_i) = a_0^{\mathrm{LS}} + \sum_{j=1}^{H} [a_j^{\mathrm{LS}} \cos(\omega_j t_i) + b_j^{\mathrm{LS}} \sin(\omega_j t_i)], \tag{8}$$

where a_j^{LS} and b_j^{LS} are the least squares estimates derived from (7). If the time series is complete (i.e., no missing data), equally spaced in time, and has length N equal to an integral multiple of the period P, then the sines and cosines are orthogonal and the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is diagonal and easily inverted. In this case, the regression parameters $\mathbf{z}$ can be determined individually by simple projection methods. However, if the time series is partly missing, not equally spaced in time, or does not span an integral multiple of P, then the sines and cosines are not orthogonal and the full matrix inverse needs to be computed. This inverse, however, presents no difficulty with modern computers.

To handle leap years, one can simply set $P = 365.25$ days, in which case the sines and cosines exactly repeat themselves every 1461 days, as needed to account for leap years. Note that leap years are difficult to accommodate with simple averaging because 29 February has one fourth as much data for estimation purposes. In addition, missing data are handled automatically in spectral methods because these methods do not require t_i to be equally spaced or even to occur repeatedly at precisely the same phase of the period.

Regression theory provides confidence intervals for the regression parameters, but these intervals are difficult to apply in geophysical contexts because they are derived from the assumption that the residuals are independent. Many climate variables tend to be autocorrelated even after the climatological mean has been subtracted. However, the residuals might be approximately independent if we fit the data to an autoregressive integrated moving average (ARIMA) model (Box et al. 1994), but this entails further model fitting, which typically is not justified if only the climatology is desired.

We prove in the appendix that simple averaging is a special case of the spectral method, provided the time steps are equally spaced, the number of time steps is an integral multiple of the period P, and $H = [(P/2)$, where (J) denotes the largest integer less than or equal to J. Therefore, the question of which method produces the better estimate reduces to the question of whether the spectral method gives better estimates for $H = (P/2)$ compared to other values of H. We can also anticipate the connection between the two methods in cases in which simple averaging is supplemented by smoothing. Specifically, if the smoothing were accomplished by convolution of the time series with a smoothing window, then the convolution theorem implies that the Fourier transform of the smoothed time series would equal the product of the Fourier transforms of the smoothing window and the original time series. To the extent that the smoothing damps high-frequency components while leaving low-frequency components unmodified, the spectral decomposition of the smoothed time series will tend toward the spectral method with suitable values of H.

To estimate the prediction error of the climatological mean in independent datasets, we adopt the following standard approach. Our goal is to determine the dependence of the prediction error on the number of predictors and the number of samples used to develop the model. Consider a random process $\mathbf{y}$ generated by the model

$$\mathbf{y} = \mathbf{X}\mathbf{z} + \mathbf{e}, \tag{9}$$

where $\mathbf{X}$ is a generic specified $M \times K$ matrix of predictors (M is the sample size and K is the number of predictors) and $\mathbf{e}$ is Gaussian random noise with zero mean and covariance matrix $\sigma^2\mathbf{I}$. In what follows, the matrix $\mathbf{X}$ is general, not necessarily the same as in (5). The least squares estimate of $\mathbf{z}$ is obtained from (7). Now suppose an independent realization of $\mathbf{y}$ is obtained from the model

$$\mathbf{y}_I = \mathbf{X}\mathbf{z} + \mathbf{e}_I, \tag{10}$$

where $\mathbf{e}_I$ is an M-dimensional vector that is independent of $\mathbf{e}$. The predicted value of $\mathbf{y}_I$ is $\mathbf{X}\mathbf{z}_{\mathrm{LS}}$. The prediction error is then

$$\mathbf{y}_I - \mathbf{X}\mathbf{z}_{\mathrm{LS}} = (\mathbf{X}\mathbf{z} + \mathbf{e}_I) - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{z} + \mathbf{e}) \tag{11}$$

$$= \mathbf{e}_I - \mathbf{H}\mathbf{e}, \tag{12}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$. Hence, the mean square prediction error is

$$\frac{1}{M}\langle(\mathbf{y}_I - \mathbf{X}\mathbf{z}_{\mathrm{LS}})^{\mathrm{T}}(\mathbf{y}_I - \mathbf{X}\mathbf{z}_{\mathrm{LS}})\rangle$$

$$= \frac{1}{M}\langle(\mathbf{e}_I - \mathbf{H}\mathbf{e})^{\mathrm{T}}(\mathbf{e}_I - \mathbf{H}\mathbf{e})\rangle \tag{13}$$

$$= \frac{1}{M}\langle\mathbf{e}^{\mathrm{T}}\mathbf{H}\mathbf{e}\rangle + \frac{1}{M}\langle\mathbf{e}_I^{\mathrm{T}}\mathbf{e}_I\rangle \tag{14}$$

$$= \frac{1}{M}\mathrm{tr}[\mathbf{H}\langle\mathbf{e}\mathbf{e}^{\mathrm{T}}\rangle] + \frac{1}{M}\langle\mathbf{e}_I^{\mathrm{T}}\mathbf{e}_I\rangle \tag{15}$$

$$= \frac{1}{M}\sigma^2\mathrm{tr}(\mathbf{H}) + \frac{1}{M}\sigma^2 M \tag{16}$$

$$= \sigma^2\frac{M + K}{M}, \tag{17}$$

where we have used the fact that $\mathbf{H}^2 = \mathbf{H}$ and $\mathrm{tr} - (\mathbf{H}) = K$. The result [(17)] cannot be evaluated directly because σ^2 is unknown. As is well known, an unbiased estimate of σ^2 is

$$\sigma_{\mathrm{LS}}^2 = \frac{\mathrm{SSR}}{M - K}. \tag{18}$$

Substituting this value into (17) gives the estimated prediction error

$$\mathrm{FPE} = \sigma_{\mathrm{LS}}^2\left(1 + \frac{K}{M}\right), \tag{19}$$

where FPE is the "final prediction error" introduced by Akaike (1969).

Let us now evaluate FPE for the two methods. For the spectral method, $K = 2H + 1$ and $M = N$, which gives

$$\mathrm{FPE} = \sigma_{\mathrm{LS}}^2\left(1 + \frac{2H + 1}{N}\right) \quad \text{for} \quad \text{spectral method.} \tag{20}$$

For simple averaging, we invoke the proof in the appendix that simple averaging is equivalent to the spectral method, with the number of independent parameters equal to P, which gives

$$\mathrm{FPE} = \sigma_{\mathrm{LS}}^2\left(1 + \frac{P}{N}\right) \quad \text{for} \quad \text{simple averaging.} \tag{21}$$

It follows from these results that if the true model for the data contains H harmonics, the spectral method is expected to have less prediction error, on average, than simple averaging when $2H + 1 < P$.

It should be recognized that simple averaging also can be formulated as a least squares estimate. Specifically, if the data is restricted to a single phase of the periodic cycle, and the climatological mean is assumed to be a constant, that is,

$$\mathbf{y}_c = \mu, \tag{22}$$

then the least squares estimate of μ is precisely the sample mean of the data. In this case, the FPE in (19) is found by setting $M = N/P$, which is the number of cycles, and $K = 1$, which is the number of parameters being estimated. Substituting these values in (19) yields (21), as expected.

FPE has the virtue that it is analytic and gives a simple criterion for when the spectral method should have less prediction error than simple averaging. Unfortunately, FPE is not completely satisfactory because it assumes that the true predictors are a subset of the predictors in the model being estimated, and that the residual errors are independent. The latter assumption is most unrealistic. An alternative estimate of the independent prediction error that does not make these assumptions can be obtained by cross-validation methods as follows. For concreteness, assume the period in question is one year. For a given method, one year is withheld and the remaining years are used to estimate the climatological mean at each grid point individually and independently. The mean squared difference between the resulting climatological mean and the withheld year is then computed. Repeating this procedure for each year and then averaging squared differences over the number of withheld years gives the cross-validated mean square error at each grid point. When global averages are desired, the area-weighted mean squared errors are computed over all valid grid points to give a global mean square error. When reporting numerical values, we take the square root of the cross-validated mean square error, which will have the same units as the variable and will be called the cross-validated error (CVE).

The fraction of variance explained by the climatological mean $\mathbf{y}_c$ is defined as

$$\mathrm{EV} = 1 - \frac{(\mathbf{y} - \mathbf{y}_c)^{\mathrm{T}}(\mathbf{y} - \mathbf{y}_c)}{(\mathbf{y}^{\mathrm{T}}\mathbf{y} - \overline{\mathbf{y}}^{\mathrm{T}}\overline{\mathbf{y}}N)}, \tag{23}$$

where $\overline{\mathbf{y}}$ is the mean of the elements of y. Substituting $\mathbf{y}_c = \mathbf{y}_c^{\mathrm{SM}}$ and $\mathbf{y} = \mathbf{y}_c^{\mathrm{SA}}$ gives the variance explained by the climatological mean from the spectral method and simple averaging, respectively.

3. Example

We explore the results above in a simple example using synthetic data. We generate 10 yr of daily data. The data consists of an annual cycle described by four harmonics and random Gaussian noise with mean zero and unit variance. The annual cycle estimated from the first 5 yr of data by the spectral method and by simple averaging is shown in Fig. 1a. The simple averaging method produces an annual cycle that has small but noticeable noise. The annual cycle produced by the spectral method is indistinguishable from the true annual cycle. Two sets of anomalies are formed by removing these two annual cycles from the data. The variance of these anomalies is computed using a 360-day moving average and shown in Fig. 1b. The variance of the simple averaging method anomalies is lower than that of the spectral method anomalies in the first half of the dataset and higher in the second half of the dataset. The discrepancy is well explained by FPE. Specifically, for simple averaging, the number of parameters is $K = 1$, since only the sample mean is estimated, and the 5-yr average is derived from a sample of five numbers, hence $M = 5$. Thus, FPE derived from (17) is $1 + 1/5 = 1.2$, consistent with the values in Fig. 1b in the last half of the time series. The variance of the spectral method anomalies, on the other hand, is close to its true value of 1 and shows no difference between the first and second halves of the data. Consistently, the value of FPE for the spectral method with four harmonics is $1 + 9/(5 \times 365) \approx 1.01$.

4. Data

The dataset used in this study for illustrating the methodologies is the daily, optimally interpolated sea surface temperature for the years 1985–2007, mostly based on the AVHRR infrared satellite but including a blend of in situ data and sea ice concentrations (Reynolds et al. 2007). This dataset is specified on a spatial grid of 0.25° resolution, but to reduce the computational burden we interpolate the data to T85 (approximately 1.4°) resolution. The grid points on land are omitted. Missing data at ocean grid points are handled as described in section 2.

5. Results

The square root of the global mean square cross-validated error (GCVE) for the climatological mean computed from the spectral method, for values up to $H = 10$, is shown in Fig. 2. The GCVE for simple averaging is indicated by the dashed line. For $H \geq 2$, the GCVE for the spectral method is smaller than that for simple averaging; indicating that for such values of H the spectral method gives a more accurate estimate of the climatological mean

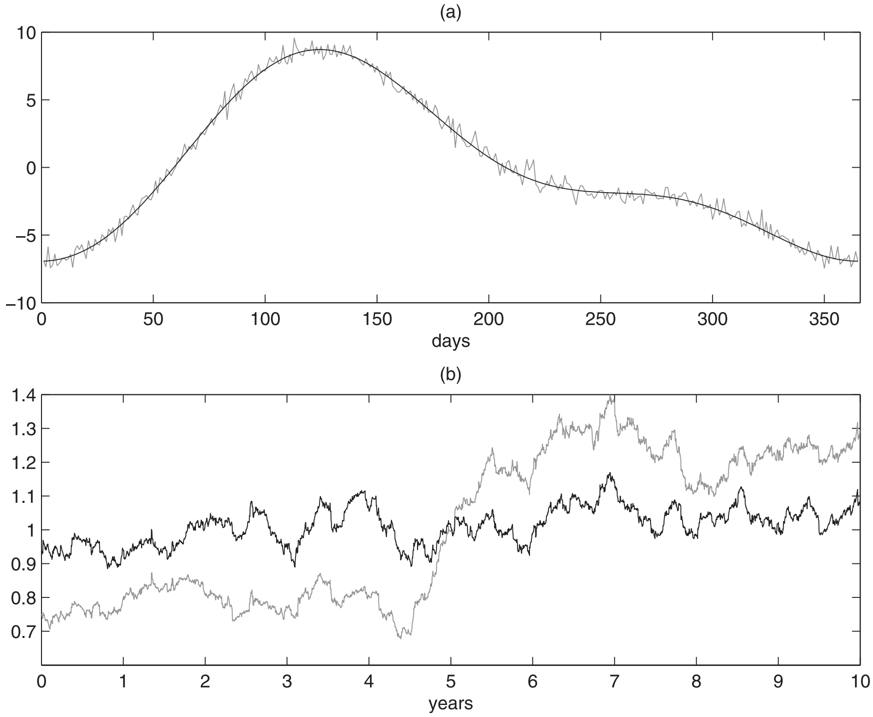

FIG. 1. (a) The annual cycle of the synthetic data. (b) The variance of the synthetic data anomalies computed using a 360-day moving average. The dark line uses the spectral method with four harmonics; the light line uses simple averaging.

in independent data. We emphasize that the climatological mean is described by only five parameters with the spectral method for $H = 2$ but requires 366 parameters using simple averaging. The absolute minimum GCVE for these values of H occurs at $H = 4$, although the GCVEs for values of H between 3 and 10 differ only marginally.

The spatial distribution of the square root of the CVE is shown in Fig. 3 for $H = 2$ and $H = 4$. This constitutes an estimate of the variance of SST anomalies. The differences in CVE for the two cases are virtually indistinguishable. For most points on the globe, the spectral method has smaller CVE than simple averaging, with the main exceptions being in the polar regions, for reasons to be discussed shortly. Comparison of Figs. 2 and 3 also reveals that although the GCVE changes only slightly between $H = 2$ and 4, the local CVE can change significantly in some regions, in the sense of transitioning from being greater than that for simple averaging to less than that for simple averaging. The maps of the CVE for $H > 4$ (not shown) differ only marginally from those for $H = 4$.

The value of H that minimizes the CVE for individual grid points is indicated in Fig. 4. The figure shows that two harmonics are optimal for most of the globe, except in the northern mid to high latitudes and near western boundaries. Six or more harmonics are required to minimize CVE in the northern subtropical highs, the Indian Ocean, and the polar regions. The spatial noisiness of the CVE minimizing H is due to the CVE being a relatively flat function of H for $H > 2$.

The standard deviation of the difference between the two climatological means is shown in Fig. 5. These

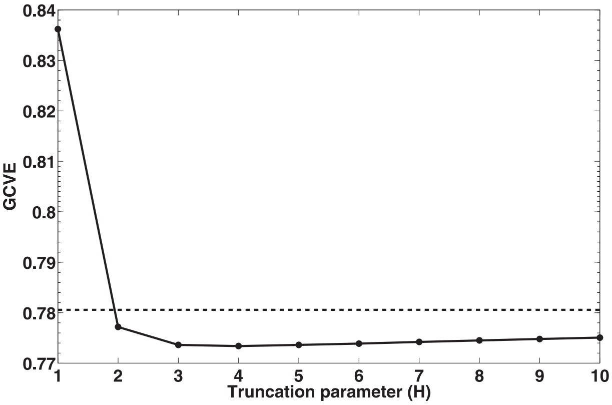

FIG. 2. Square root of the globally averaged cross-validated mean square error (GCVE) obtained with the spectral method (solid) for truncation parameters $H = 1, 2, \ldots, 10$, and the GCVE obtained with the simple averaging method. The minimum GCVE occurs at $H = 4$.

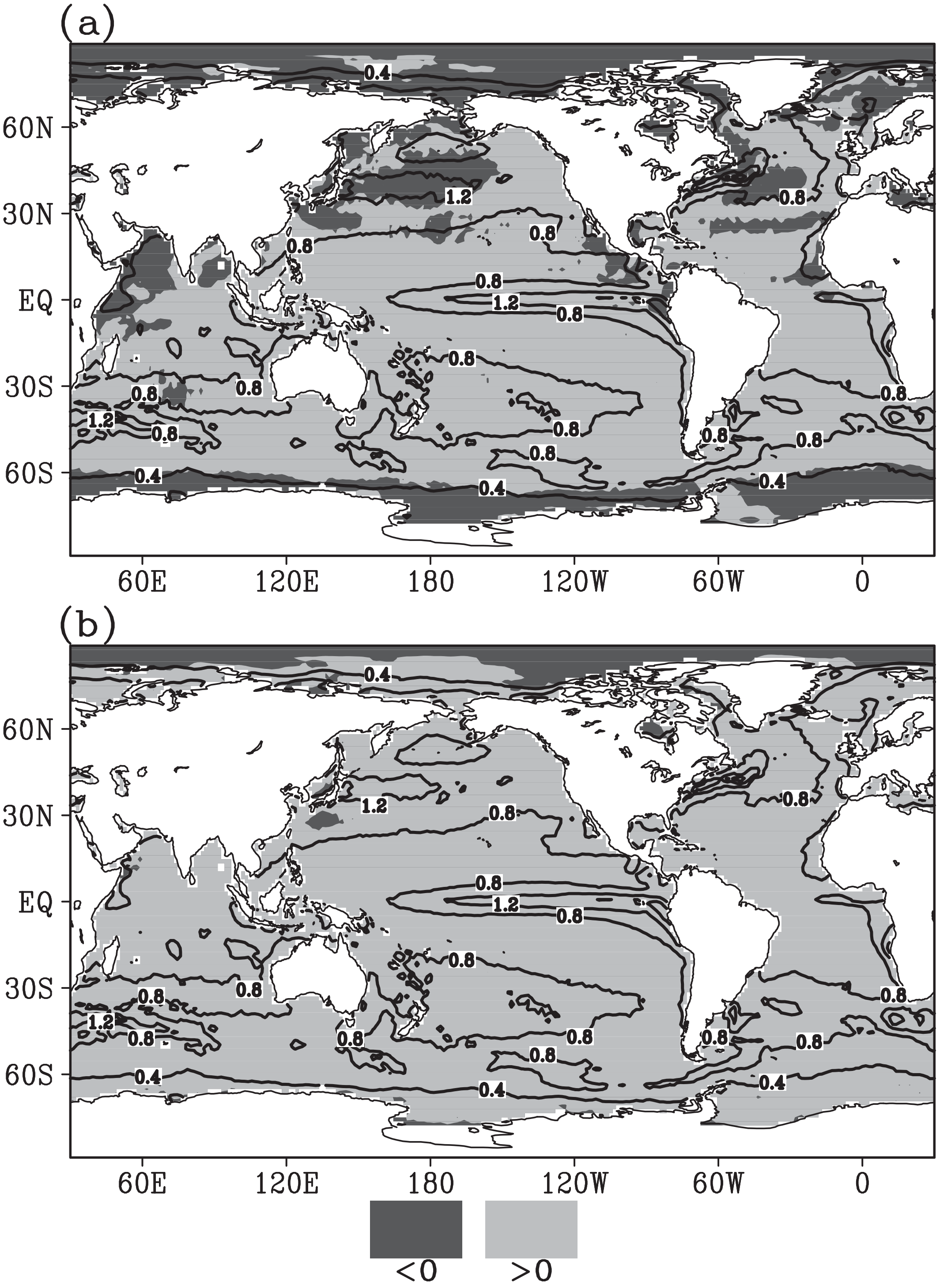

FIG. 3. Square root of the cross-validated mean square error (CVE) (contours) for the climatological mean estimated from the spectral method for H = (a) 2 and (b) 4. The shading indicates the sign of the difference between the CVE for simple averaging minus the spectral method. Positive shading (gray) indicates that simple averaging produces larger CVE than the spectral method.

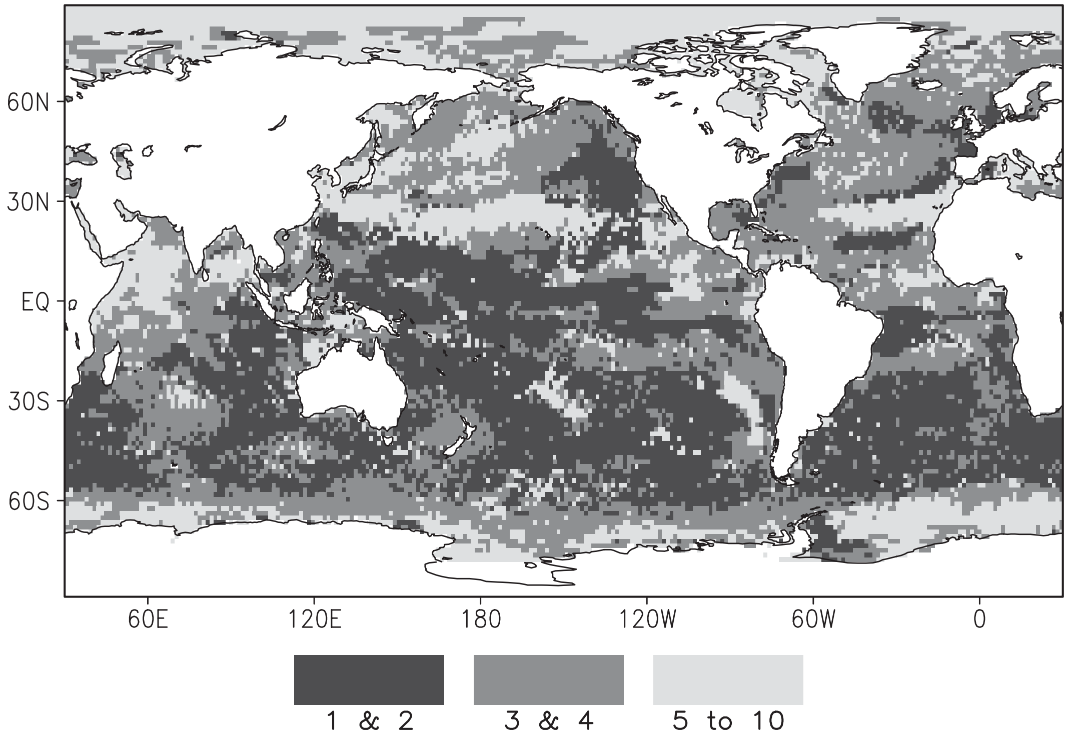

FIG. 4. Number of harmonics H that minimizes the CVE in the spectral method.

differences basically measure the power in the higher harmonics (i.e., the power is spectral coefficients greater than or equal to 5). Interestingly, the standard deviation of the differences is between 0.1 and 0.2 in most regions. Theoretically, we expect this difference to increase as $1/\sqrt{N}$ as the number of years (N) decreases. We see that the greatest differences occur near the western boundary currents and along the tropical eastern Pacific. A comparison of the annual cycle computed from the two methods at selected grid points is presented in Fig. 6. As expected, the simple averaging contains more day-to-day variability than the spectral method. However, the spectral method gives a reasonable smoothed version of the annual cycle computed from simple averaging.

To gain insight into why different regions need different number of harmonics to minimize CVE, we show the raw time series for three selected regions in Fig. 7. The bottom panel shows a time series for a selected region near the poles. A distinctive feature of this latter time series is that the SST never drops below −1.8°C. This feature is related to how the analysis is produced when the sea ice extent exceeds a threshold (Reynolds et al. 2007). Sinusoids are not an efficient basis set for approximating nonsmooth time series. The fact that polar SSTs require more harmonics for minimizing CVE is therefore a consequence of the fact that SSTs in this region have discontinuous first derivatives (i.e., ‘‘corners’’).

The top two panels of Fig. 7 are similar and thus there is no obvious reason why one requires more harmonics to minimize CVE than the other. To explore this last point further, we show in Fig. 8 the CVE as a function of the total number of harmonics for the same regions used in Fig. 7. In general, the CVE drops significantly from one to two harmonics, but the change in CVE from two to higher harmonics is relatively small. This raises questions as to whether the minimum CVE is statistically distinguishable from the CVE at two harmonics. Nevertheless, while the optimal value of H may be uncertain, the corresponding annual cycles and their values of CVE are relatively insensitive to the choice of H. The fraction of variance explained by the annual cycle (Eq. 23) with truncation parameter $H = 4$ is shown in Fig. 9. The annual cycle explains a relatively small fraction of the variance (as low as 10%) in the western Equatorial Pacific and Atlantic but explains over 90% of the variance in the northern midlatitude oceans. In most cases, the annual cycle explains a greater fraction of the variance in the Northern Hemisphere than in the Southern Hemisphere, for the same latitude. The analogous figure showing the fraction of variance explained by the annual cycle determined by simple averaging is very similar to Fig. 9 (not shown).

The fraction of variance explained by the high-frequency component ($H \geq 5$) relative to the low-frequency components ($H \leq 4$) is shown in Fig. 10. The figure shows

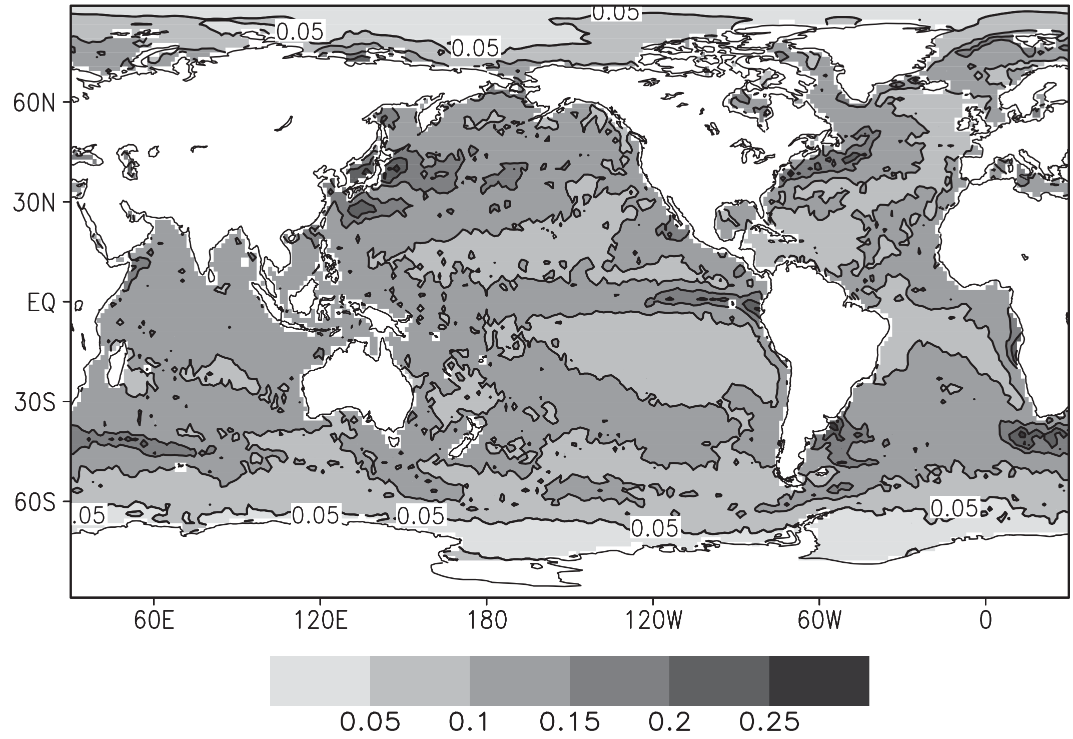

FIG. 5. Standard deviation of the difference between the climatological means computed from simple averaging and the spectral method with $H = 4$.

that higher harmonics of the annual cycle are important primarily along the equator and the North Pole. Quantitatively, the higher harmonics account for about 1% of the total variance in the annual cycle along the equator, indicating that their contribution to the annual cycle is small.

6. Conclusions

This paper compared two methodologies for estimating the climatological mean. In the first method, called simple averaging, the climatological mean is obtained by averaging values with respect to a fixed phase of the annual or diurnal cycle. In the second method, called the spectral method, the climatological mean is estimated by performing a least squares fit of the first H harmonics of the annual or diurnal cycle to the data. We showed in the appendix that simple averaging is a special case of the spectral method. Thus, the question of which method is better reduces to the question of what is the best choice of H. The number of parameters estimated in the spectral method is $2H + 1$, whereas the number of parameters estimated in simple averaging equals the number of samples within an annual or diurnal period. Thus, as the temporal resolution of the data increases, the number of parameters grows for simple averaging but remains constant for the spectral method (for fixed H). As a consequence, the spectral method is expected to be less susceptible than simple averaging to overfitting as the number of samples in the annual or diurnal period increases. The degree of overfitting can be estimated in idealized situations with Akaike's FPE. These expectations were confirmed by cross-validation experiments applied to a multiyear SST dataset. Specifically, the results demonstrate that the spectral method with $H = 4$ harmonics produced less cross-validated error than simple averaging over most points on the globe. Note that in this case the spectral method involves estimation of nine spectral coefficients, whereas simple averaging involves estimation of 366 daily averages—a factor of 40 more parameters. The spectral method also easily accounts for leap years and missing data, in contrast to simple averaging.

The number of harmonics needed to minimize the cross-validated error in the 23-yr SST dataset differs from point to point. Two harmonics suffice in much of the tropics and southern midlatitudes, whereas four or more harmonics are required in certain regions of the northern midlatitudes and polar regions. There appears to be no advantage to using more than four harmonics, however, since the cross-validated error changes only marginally beyond four harmonics.

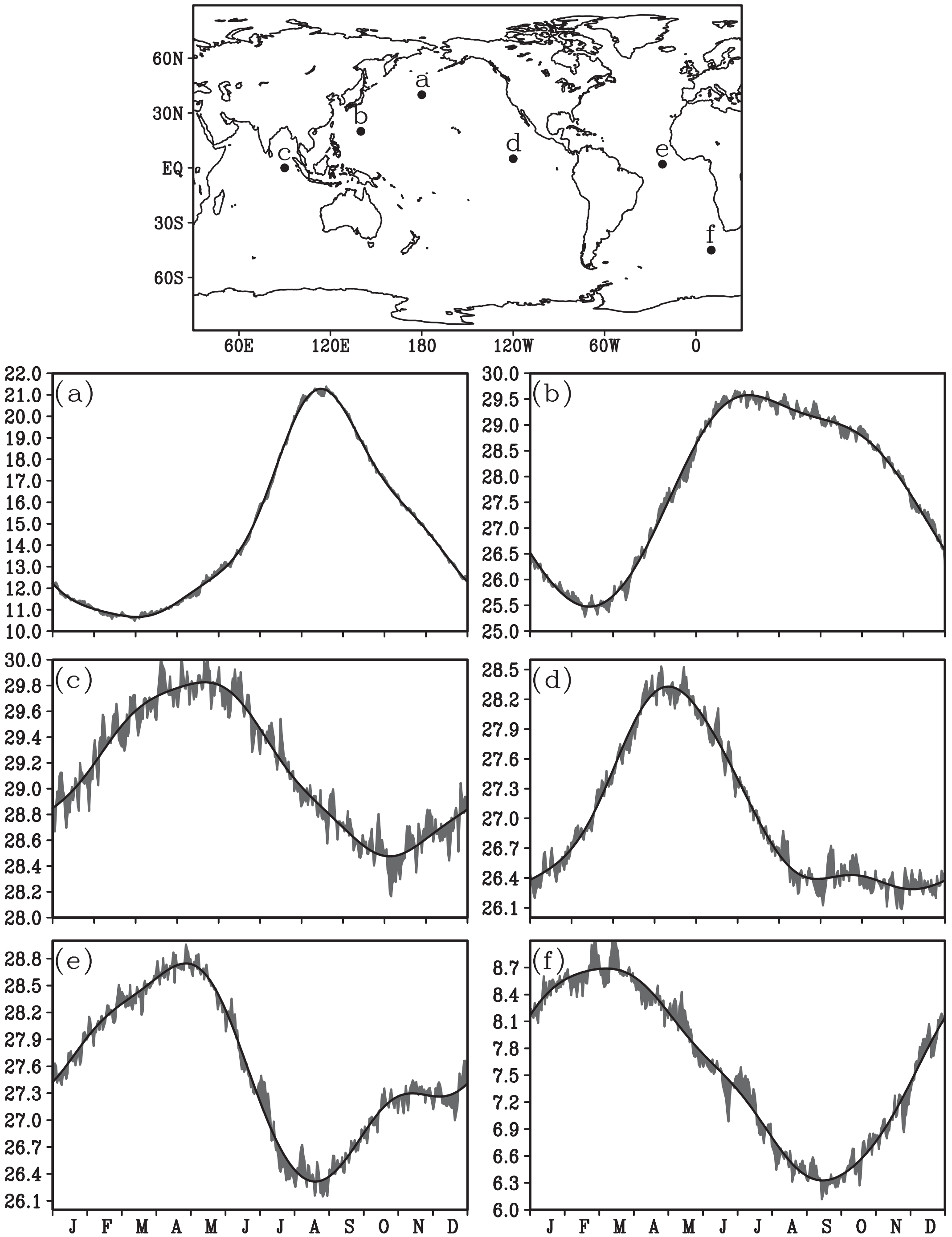

FIG. 6. (top) The location of the thick dotted points identified as a, b, c, d, e, and f, for which the annual cycle is estimated by simple averaging and spectral (of $H = 4$) methods. (bottom) The annual cycle estimations for the simple averaging (gray shaded) and spectral methods (black solid) are shown for points (a) a, (b) b, (c) c, (d) d, (e) e, and (f) f.

The fact that the polar regions require a relatively large number of harmonics to minimize CVE was attributed to the "clipped" nature of the time series, characterized by SST values that never drop below −1.8°C. However, the cross-validated error of the annual cycle diminishes only marginally after four harmonics and differs only slightly from the annual cycle estimated by simple averaging, so there is little advantage to using a large number of harmonics in the polar regions.

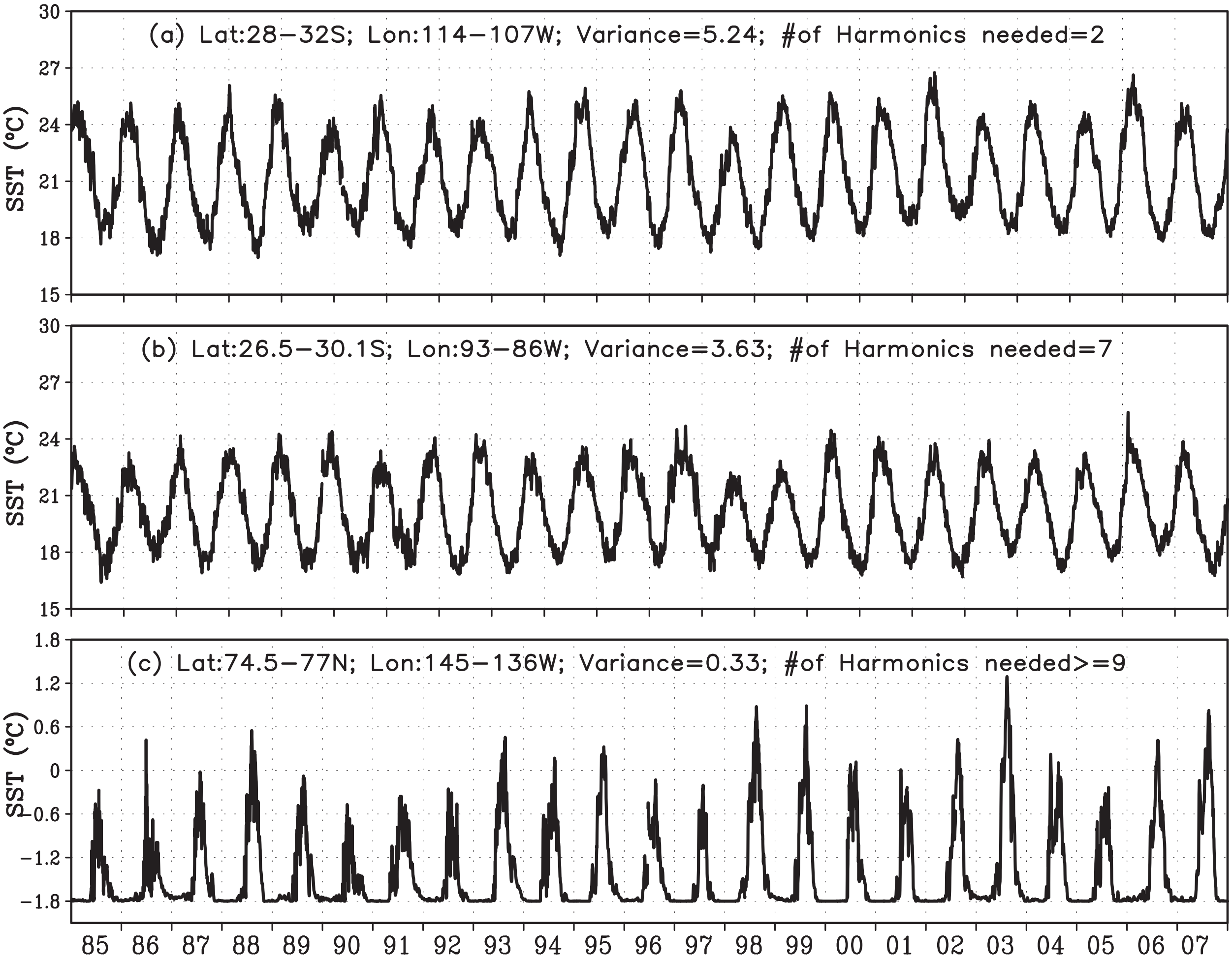

FIG. 7. SST time series averaged over (a) 28°–32°S, 114°–107°W; (b) 26°–30.1°S, 93°–86°W; and (c) 74.5°–77°N, 145°–136°W. The corresponding area-averaged CVE as a function of H is shown in Fig. 8. Note that the SST time series depicted in (a) and (b) are obtained by averaging the SST data approximately on the same latitudes but on different longitudes. The minimum value of SST time series in (c) does not drop below −1.8°C, and the SST value is fixed at −1.8°C throughout the winter.

The difference between 22-yr climatological means of SST, as estimated from the spectral method and simple averaging, has a standard deviation of about 0.1°C. This difference is expected to increase as $1/\sqrt{N}$ as the sample size N decreases. Thus, the spectral method becomes more advantageous for smaller datasets (e.g., datasets from recent, high-resolution observing platforms). In any case, Fig. 6 suggests that these differences do not persist longer than a few weeks. Thus, in terms of modeling the response of the atmosphere to SSTs, the difference in climatological means could produce a significant difference in, say, precipitation, but these differences are expected to last only about a few weeks. Nevertheless, it is not inconceivable that such differences can lead to different conclusions being drawn from the experiments. Another implication of the present results is that an empirical orthogonal analysis (using EOFs) of anomalies with respect to simple averages will have less variance than EOFs of anomalies with respect to climatological means estimated from the spectral method. The reason for this is that simple averages allow more high-frequency variability in the climatological means, which therefore are subtracted from the anomalies. Presumably, differences in the climatological mean can lead to differences in EOFs.

The fraction of variance explained by the annual cycle, shown in Fig. 9, is physically interesting. The regions with the smallest variability due to the annual cycle tend to coincide with regions of maximum precipitation, such as along the equator in the eastern Indian, western Pacific, and Atlantic Oceans. This fact suggests that the relatively weak annual cycle in these locations may be due to the fact that precipitation tends to stabilize the upper ocean owing to freshening. The regions with the

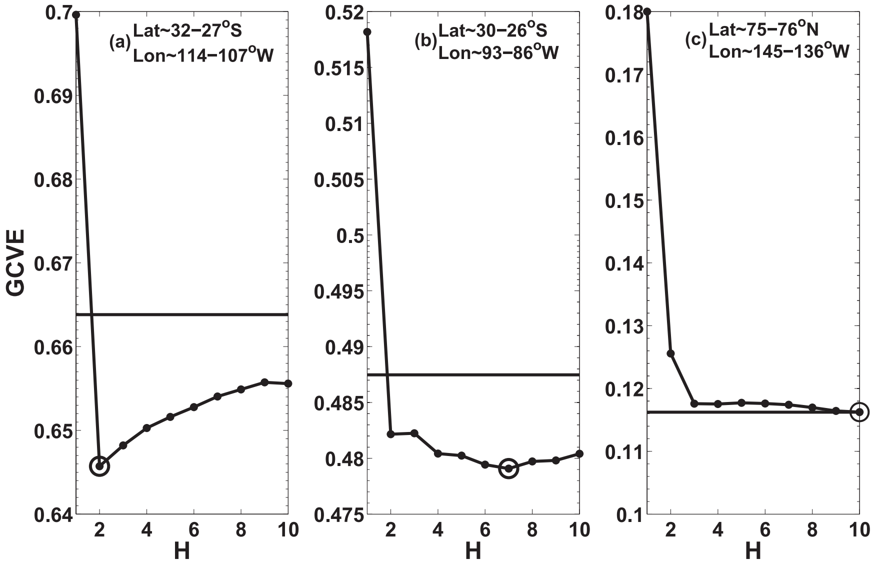

FIG. 8. Square root of the area-averaged CVE over (a) 28°–32°S, 114°–107°W; (b) 26°–30.1°S, 93°–86°W; and (c) 74.5°–77°N, 145°–136°W for each truncation parameter $H = 1, 2, 3, \ldots, 10$. These are the same regions used in Fig. 7. The minimum value is circled in each panel.

largest variability due to the annual cycle tend to be in the northern midlatitudes. These regions are dominated by mixed layer dynamics. The fact that the northern oceans exhibit a stronger ratio than the southern oceans might be due to the role of landmasses—in northern winter, dry, cold continental air blows over the mid-North Pacific and Atlantic during cold surges, which causes large heat loss from the sea surface, whereas this

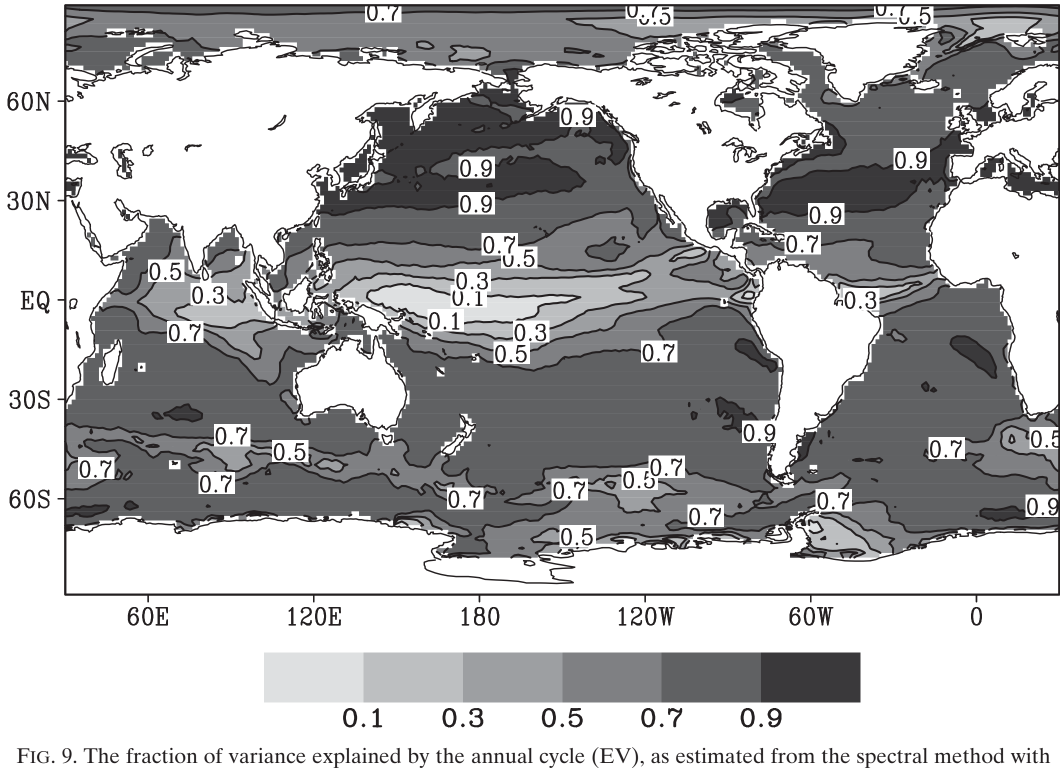

FIG. 9. The fraction of variance explained by the annual cycle (EV), as estimated from the spectral method with truncation parameter $H = 4$.

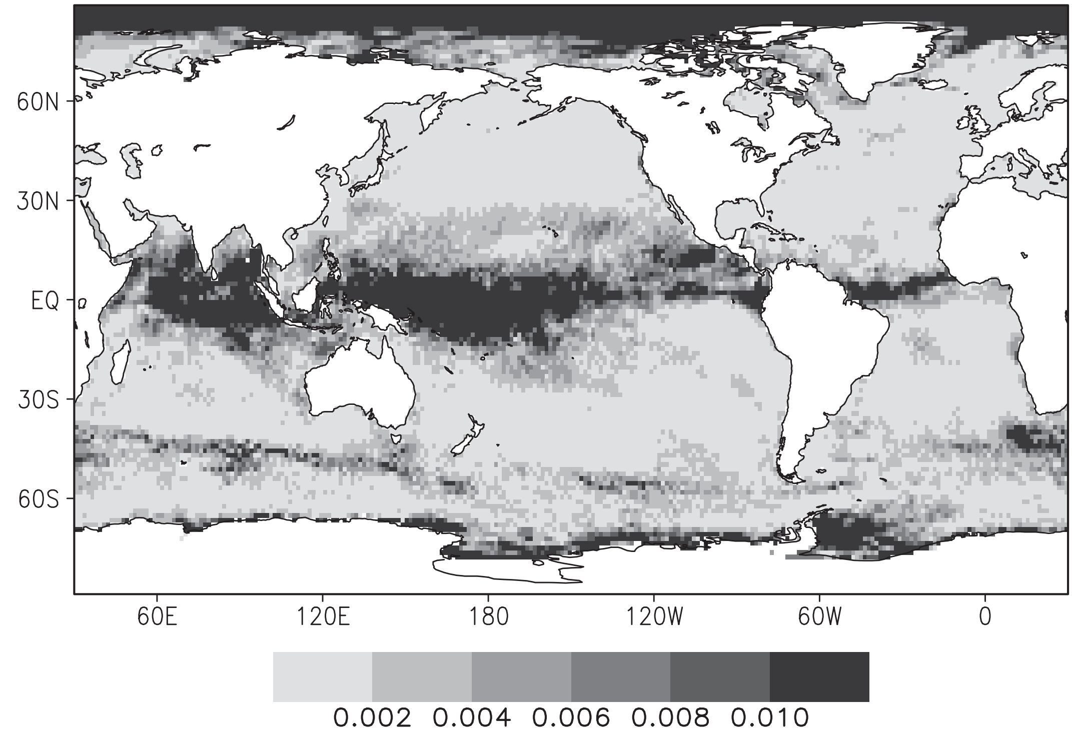

FIG. 10. Ratio of sum square spectral coefficients between 5 and 10 and between 1 and 4. The ratio shows how much variance is explained by the high-frequency components ($H \geq 5$) relative to the low-frequency components ($H \leq 4$).

mechanism is less dramatic in the Southern Hemisphere owing to the lack of landmasses. An intriguing result is that the fraction of variance explained by the annual cycle exhibits a local minimum along a narrow strip in the southern oceans between the southern tips of Africa and Australia. In contrast, the amplitude of the annual cycle is relatively uniform poleward of 45°S (not shown), so that the reduction in fraction of variance explained by the annual cycle is due primarily to enhanced variability that is not described by the annual cycle (consistent with Fig. 3). This minimum may be due to eddy variability induced by sustained wind stresses in this region—recently released QuikScat winds show that wind energy along this strip is about the same from winter to summer, which could sustain eddy variability, whereas the wind energy on the northern and southern boundaries of the strip undergoes a significant change from winter to summer (Liu et al. 2008).

Acknowledgments. We thank Bohua Huang for discussions on the physical implications of the results in this paper. This research was supported by the National Science Foundation (ATM0332910), National Aeronautics and Space Administration (NNG04GG46G), and the National Oceanic and Atmospheric Administration (NA04OAR4310034) and (NA05OAR4311004). The views expressed herein are those of the authors and do not necessarily reflect the views of NOAA or any of its sub-agencies.

APPENDIX

Equivalence between Climatological Means Estimated by the Spectral Method and Simple Averaging

In this section we show that the spectral method gives the same climatological mean as simple averaging, provided the truncation parameter H equals the period of the cycle. Consider a time series $y_n, n = 0, 1, \ldots, N - 1$. Suppose we seek a climatological mean with period P. For simplicity, assume N is an integral multiple of P (i.e., we ignore leap years and datasets with incomplete cycles). Thus, $N/P = L$, where L is an integer.

Simple averaging gives the climatological mean

$$\overline{y}_n = \frac{1}{L}\sum_{l=0}^{L-1} y_{[(n \bmod P)+Pl]}. \tag{A1}$$

Following Press et al. (1986), the discrete Fourier transform pair associated with y_n is

$$y_n = \sum_{k=0}^{N-1} \hat{y}_k e^{2\pi ikn/N}, \tag{A2}$$

$$\hat{y}_k = \frac{1}{N}\sum_{n=0}^{N-1} y_n e^{-2\pi ikn/N}, \tag{A3}$$

where $\hat{y}_k$ denotes the Fourier transform of y_n and $k = 0, 1, \ldots, N - 1$. The transform of $\overline{y}_n$ is

$$\begin{aligned}
\hat{\overline{y}}_k &= \frac{1}{N}\sum_{n=0}^{N-1} \overline{y}_n e^{-2\pi ikn/N} \\
&= \frac{1}{N}\sum_{l=0}^{L-1}\sum_{d=Pl}^{P(l+1)-1} \overline{y}_d e^{2\pi ikd/N} \\
&= \frac{1}{N}\sum_{l=0}^{L-1}\sum_{n=0}^{P-1} \overline{y}_{n+Pl} e^{2\pi ik(n+Pl)/N} \\
&= \frac{1}{N}\sum_{n=0}^{P-1} \overline{y}_n e^{2\pi ikn/N}\left(\sum_{l=0}^{L-1} e^{2\pi ikPl/N}\right),
\end{aligned} \tag{A4}$$

where we have used the fact that $\overline{y}_{n+Pl} = \overline{y}_n$. The term in parenthesis satisfies

$$\sum_{l=0}^{L-1} e^{2\pi ikPl/N} = \sum_{l=0}^{L-1} e^{2\pi ikl/L} = \begin{cases} 0 & \text{if} \quad k \bmod L \neq 0 \\ L & \text{if} \quad k \bmod L = 0 \end{cases}. \tag{A5}$$

Thus, if k is not a multiple of L, then the term in parentheses in (A4) vanishes and we have $\hat{\overline{y}}_k = 0$. On the other hand, if k is a multiple of L, then (A4) becomes

$$\begin{aligned}
\hat{\overline{y}}_k &= \frac{L}{N}\sum_{n=0}^{P-1} \overline{y}_n e^{2\pi ikn/N} \\
&= \frac{L}{N}\sum_{n=0}^{P-1}\left(\frac{1}{L}\sum_{l=0}^{L-1} y_{n\mathrm{mod}P+Pl}\right) e^{2\pi ikn/N} \\
&= \frac{1}{N}\sum_{l=0}^{L-1}\sum_{n=0}^{P-1} y_n e^{2\pi ikn/N} \\
&= \frac{1}{N}\sum_{l=0}^{L-1}\sum_{d=Pl}^{P(l+1)-1} y_d e^{2\pi ik(d-Pl)/N} \\
&= \frac{1}{N}\sum_{l=0}^{L-1}\sum_{d=Pl}^{P(l+1)-1} y_d e^{2\pi ikd/N} e^{-2\pi ikPl/N} \\
&= \frac{1}{N}\sum_{n=0}^{N-1} y_n e^{2\pi ikn/N} \\
&= \hat{y}_k.
\end{aligned} \tag{A6}$$

In summary, we have shown that

$$\hat{\overline{y}}_k = \begin{cases} 0 & \text{if} \quad k \bmod L \neq 0 \\ \hat{y}_k & \text{if} \quad k \bmod L = 0 \end{cases}. \tag{A7}$$

However, if k is a multiple of L, then $k = mL$ for some integer m, and $\overline{y}_n$ can be written as

$$\overline{y}_n = \sum_{m=0}^{P-1} \hat{y}_{mL} e^{2\pi imn/P}. \tag{A8}$$

After some tedious algebra, this expression can be rewritten equivalently as

$$\overline{y}_n = \hat{y}_0 + \sum_{m=1}^{P/2} a_m \cos(2\pi mn/P) + b_m \sin(2\pi mn/P), \tag{A9}$$

where

$$a_m = \begin{cases} \hat{y}_{PL/2} & \text{if} \quad P \text{ is even} \quad \text{and} \quad m = \frac{P}{2} \\ 2\mathrm{Re}\{\hat{y}_{mL}\} & \text{otherwise} \end{cases}, \tag{A10}$$

and

$$b_m = \begin{cases} 0 \quad \text{if} \quad P \text{ is even} \quad \text{and} \quad m = \frac{P}{2} \\ -2\mathrm{Im}\{\hat{y}_{mL}\} \qquad \text{otherwise} \end{cases}, \tag{A11}$$

where $\mathrm{Re}\{y\}$ and $\mathrm{Im}\{y\}$ denote the real and imaginary parts of y, respectively. The derivation of this expression invokes the fact that $\overline{y}_n$ is periodic (i.e., $\overline{y}_n = \overline{y}_{n-P}$) and the fact that $\overline{y}_n$ is real, which in turn implies $\hat{\overline{y}}_k^* = \hat{\overline{y}}_{-k}$, where the asterisk denotes the complex conjugate. Note that (A9) has the same form as the spectral average (3), but with truncation parameter $H = (P/2)$ and with spectral coefficients equal to the spectral coefficients of y_n evaluated at the harmonics $0, L, 2L, \ldots, PL$. Therefore, relation (A9) demonstrates that the simple average estimate $\overline{y}_n$, as represented in (A1), is a special case of the spectral method. A careful count shows that there are P independent parameters in (A9) regardless of whether P is even or odd. Thus, P is the number of parameters in the simple average, consistent with (A1).

REFERENCES

Akaike, H., 1969: Fitting autoregressive models for prediction. *Ann. Inst. Stat. Math.*, **21,** 243–247.

Barnett, T. P., 1981: On the nature and causes of the large-scale thermal variability in the central North Pacific Ocean. *J. Phys. Oceanogr.,* **11,** 887–904.

Box, G. E. P., G. M. Jenkins, and G. C. Reinsel, 1994: *Time Series Analysis: Forecasting and Control*. 3rd ed., Prentice Hall, 598 pp.

Epstein, E., 1988: A spectral climatology. *J. Climate,* **1,** 88–107.

Liu, W. T., W. Tang, and X. Xie, 2008: Wind power distribution over the ocean. *Geophys. Res. Lett.,* **35,** L13808, doi:10.1029/2008GL034172.

Press, W. H., B. P. Fannerty, S. A. Teukolsky, and W. T. Vetterling, 1986: *Numerical Recipes: The Art of Scientific Computing*. 1st ed. Cambridge University Press, 848 pp.

Reynolds, R. W., T. M. Smith, C. Liu, D. B. Chelton, K. S. Casey, and M. G. Schlax, 2007: Daily high resolution blended analysis for sea surface temperature. *J. Climate,* **20,** 5473–5496.

Straus, D. M., 1983: On the role of seasonal cycle. *J. Atmos. Sci.,* **40,** 303–313.