Estimation of Time Averages from Irregularly Spaced Observations: 
With Application to Coastal Zone Color Scanner 
Estimates of Chlorophyll Concentration

DUDLEY B. CHELTON AND MICHAEL G. SCHLAX

College of Oceanography, Oregon State University, Corvallis

A formalism is presented for quantifying the sampling error of an arbitrary linear estimate of a time-averaged quantity constructed from a time series of irregularly spaced observations at a fixed location. The method is applicable to any irregularly sampled time series; it is applied here to satellite observations of chlorophyll from the coastal zone color scanner (CZCS). The two specific linear estimates considered here are the composite average formed from the simple average of all observations within the averaging period and the optimal estimate formed by minimizing the mean squared error of the temporal average based on all of the observations in the time series. The formalism requires a priori knowledge of the variances and correlation functions of the chlorophyll signal and CZCS measurement error. In the usual absence of the necessary detailed information on these parameters, values obtained here from in situ measurements of chlorophyll and fluorescence off the coast of southern California can be used. The resulting estimates are referred to here as "suboptimal estimates," which are optimal only if the assumed values for the parameters are correct. Suboptimal estimates are shown to be much more accurate than composite averages. Moreover, suboptimal estimates are also shown to be nearly as accurate as optimal estimates obtained using the correct signal and measurement error variances and correlation functions for realistic ranges of these parameters. Suboptimal estimation is thus a very useful and practical alternative to the composite average method generally used at present.

1. Introduction

Irregular temporal sampling is a common feature of geophysical and biological time series. The incomplete description of variability resulting from unresolved short time scale fluctuations between the discrete observations is referred to as sampling error. For evenly spaced observations, sampling error is called aliasing. The same effect is present in irregularly spaced time series but is more complex and difficult to diagnose. One method of reducing sampling error is to form time series of temporal averages of the discretely sampled observations in the hope of averaging out random high-frequency variability. The most common form of the temporal average is the composite average, defined here to be the arithmetic mean of all available observations in a specified time window. A time series is then constructed from composite averages over successive nonoverlapping time windows. The effects of sampling variability can still be significant in such time series when the samples are sparsely or irregularly distributed over each averaging period or when the measurement errors are large. A method for quantifying the relative importance of sampling error versus measurement error is presented in this paper. The trade-off between temporal resolution (i.e., the averaging period) and the accuracy of the composite average is addressed. In addition, an improved method of estimating the temporal average based on optimal estimation is introduced, and the results are compared with the composite average method.

A specific example which provided much of the motivation for developing the technique presented here is the estimation of near-surface chlorophyll concentration from satellite measurements of ocean color by the coastal zone color scanner (CZCS). To optimize the availability of visible radiation from the sea surface and undesired solar spectral reflectance from the sea surface, satellites used for measuring color radiances are placed in Sun-synchronous orbits. The minimum time interval between successive observations at a given location is therefore 1 day. In practice, the actual sample interval is generally longer and very irregular owing to the less-than-100% duty cycle of the radiometer and the presence of clouds obscuring the sea surface from the satellite view. Satellite-derived time series of chlorophyll variability are therefore aliased in very complicated ways by unresolved high-frequency variability. A quantitative understanding of the effects of sampling error on the accuracy of such time series is essential for biological applications.

The formalism for determining the statistical error of temporal averages estimated by composite averages and optimal estimation is presented in section 2. An overview of satellite estimates of near-surface chlorophyll concentration is given in section 3. The error formalism requires a priori estimates of the variances and correlation functions of the signal and the measurement errors. For the application to satellite estimates of chlorophyll concentration considered here, a signal correlation function is obtained from spectra of in situ time series of chlorophyll and fluorescence off the coast of southern California as described in section 4. In subsequent sections, the sensitivities of composite averages and optimal estimates to the variances and correlation functions of signal and measurement errors are explored and the formalism is applied to simulated satellite data. Particular emphasis is placed on the importance of the accuracies with which the variances and correlation functions must be specified.
2. Estimation of Temporal Averages

A general formalism can be developed for estimating temporal averages from a time series of irregularly spaced observations at a fixed location. The method presented here is essentially the same as that developed for the estimation of mineral grades, known as “kriging” in the geostatistics literature [e.g., Journel and Huijbregts, 1978]. Let \( C(t) \) be the continuously time-varying process of interest at time \( t \), in this case the chlorophyll concentration in milligrams per cubic meter. It is assumed that the interest is in the temporal average of some function of \( C(t) \) over a time period \( T \) centered at time \( t_0 \). Because chlorophyll concentration is generally approximately lognormally distributed, most analyses of chlorophyll data consider \( \log C(t) \) or \( \log_{10} C(t) \). Since the two are proportional, the linear analysis presented here holds equally well for either quantity. The latter form seems to be used most often in analyses of CZCS data, so this is the function considered here. The quantity of interest is therefore

\[
\bar{z}(t_0) = \frac{1}{T} \int_{t_0-T/2}^{t_0+T/2} c(t) \, dt,
\]

where \( c(t) = \log_{10} C(t) \).

In practice, the formalism presented in this section should be applied only to that part of \( c(t) \) that cannot be easily estimated by other means. Of particular concern in log-transformed chlorophyll data is the seasonal cycle, which can be very energetic at some locations. In other applications, an energetic diurnal cycle may exist. Failure to remove such signals will result in correlation functions used in the estimation described below that are dominated by the cyclical variability. The formalism can then provide little more information about the time series than just the cyclical variabilities, which are often easily estimated by a number of possible methods. A simple method that is often applied to irregularly spaced observations is suggested in the appendix (step 2). It is assumed hereinafter that cyclical variability has been estimated independently and removed (possibly to be added back later; see steps 12 and 13 in the appendix) so that \( c(t) \) for the application here is the nonseasonal variability of log-transformed chlorophyll concentration with zero mean value.

The data available to estimate the temporal average \( \bar{z}(t_0) \) are \( M \) observations at the same geographical location at discrete times \( t_m \). These observations are assumed to be contaminated by measurement errors \( \varepsilon_m \) and can therefore be written as

\[
y_m = c(t_m) + \varepsilon_m, \quad m = 1, \ldots, M.
\]

The measurement errors are assumed to have a mean value of zero and an \( M \times M \) cross-covariance matrix \( N \) with \((m,n)\)th element \( N_{mn} = \langle \varepsilon_m \varepsilon_n \rangle \), where the angle brackets denote the mean value. The form of the estimate \( \hat{z}(t_0) \) for the application here is the nonseasonal variability of log-transformed chlorophyll concentration with zero mean value.

The weights \( \alpha_m \) vary depending upon the type of linear estimate used. Two such estimates are considered below.

One quantitative measure of the error of the estimate \( \hat{z}(t_0) \) is the mean squared error

\[
\phi^2(\alpha, t_0, T) = \left\langle \left[ z(t_0) - \hat{z}(t_0) \right]^2 \right\rangle,
\]

where \( \alpha \) is the vector of the weights in the linear estimate (3) with transpose given by \( \alpha^T = (\alpha_1, \ldots, \alpha_M) \). Using (1), (2) and (3), the mean squared error can be expressed as

\[
\phi^2(\alpha, t_0, T) = \frac{1}{T^2} \int_{t_0-T/2}^{t_0+T/2} R(t, t') dt \int_{t_0-T/2}^{t_0+T/2} R(t, t') dt' - \frac{2}{T} \sum_{m=1}^{M} \alpha_m \int_{t_0-T/2}^{t_0+T/2} R(t, t_m) dt
\]

\[
+ \sum_{m=1}^{M} \sum_{n=1}^{M} \alpha_m N_{mn} \alpha_n,
\]

where the covariance function \( R(t, t') = \langle c(t) c(t') \rangle \) describes the variance and temporal scales of \( c(t) \). It has been assumed in (4) that the signal and measurement errors are uncorrelated, that is, \( \langle \varepsilon_m \varepsilon_n \rangle = 0 \) for all \( t \) and \( m \). The formalism presented here can easily be extended to account for correlation between the signal and measurement errors. The method becomes more difficult in practice, however, since this correlation function must then be prescribed a priori.

It is apparent from (4) that the mean squared error of any estimate of the form (3) can be determined if the signal covariance function \( R(t, t') \) and measurement error covariance matrix \( N \) are known a priori. Estimation of these covariances can be difficult. Sometimes an adequate data base from which \( R(t, t') \) and \( N \) can be estimated already exists. More often, these covariances must be approximated by “best educated guesses.” In practice, it is generally assumed that the process \( c(t) \) is stationary so that the covariance function becomes \( R(t, t') = R(t - t') \) and is symmetric, \( R(t) = R(-t) \). The covariance of \( c(t) \) then depends only on the lag \( \tau = t - t' \) and not on the actual times \( t \) or \( t' \) and can therefore be estimated from a single time series of \( c(t) \).

To accommodate an arbitrary distribution of \( M \) observation times \( t_m \), continuous functional representations of the signal and measurement error covariances are required. The variance of the linear estimate \( \hat{z}(t_0) \) is

\[
\langle \hat{z}^2(t_0) \rangle = \sum_{m=1}^{M} \sum_{n=1}^{M} \alpha_m \alpha_n \langle y_m y_n \rangle + \sum_{m=1}^{M} \alpha_m R(t_m - t_n) \alpha_n
\]

\[
+ \sum_{m=1}^{M} \sum_{n=1}^{M} \alpha_m N_{mn} \alpha_n. \tag{5}
\]

The first quadratic form on the right-hand side of (5) is just the variance of the linear combination \( \sum_{m=1}^{M} \alpha_m c(t_m) \).
of the signal at the observation times \( t_m \). The second quadratic form is similarly the variance of the linear combination \( \sum_{m=1}^{M} \alpha_m \delta(t_m) \) of measurement errors. Since both variances must be nonnegative, each quadratic form on the right-hand side of (5) must be nonnegative. The symmetric \( M \times M \) signal and error covariance matrices in (5) must therefore both be positive definite.

A positive definite signal covariance function, for example, implies a continuous representation that must satisfy

\[
\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} q^*(t) R(t-t') q(t') \, dt \, dt' \geq 0 \tag{6}
\]

for any function \( q(t) \), where \( q^* \) is the complex conjugate of \( q \). The particular function \( q(t) \) of interest in the two terms on the right-hand side of (5) is the discrete sampler

\[
q(t) = \sum_{m=1}^{M} \alpha_m \delta(t_m),
\]

where \( \delta(t_m) \) is the Dirac delta function and \( \alpha_m \) are the arbitrary weights in the linear estimate (3). Then the condition (6) becomes

\[
\sum_{m=1}^{M} \sum_{n=1}^{M} \alpha_m \alpha_n R(t_m - t_n) \geq 0,
\]

i.e., the symmetric \( M \times M \) matrix with \((m, n)\)th element \( R(t_m - t_n) \) is positive definite. The analogous condition for the symmetric \( M \times M \) error covariance matrix \( N \) then assures that the variance given by (5) is nonnegative. The use of covariance functions that are not positive definite can lead to linear estimates (3) with negative mean squared error (4), which is nonphysical.

The requirements that the continuous functional representations of the signal and error covariances be positive definite are equivalent to requiring that the corresponding spectra be nonnegative at all frequencies. This is easily shown by noting that (6) can be written as

\[
\int_{-\infty}^{\infty} q^*(t) \left[ \int_{-\infty}^{\infty} R(t-t') \, dt' \right] \, dt \geq 0.
\]

The term in brackets is the convolution of \( q(t) \) and \( R(t) \). Then, by the convolution theorem,

\[
\int_{-\infty}^{\infty} q^*(t) \left[ \int_{-\infty}^{\infty} F_R(s) e^{i2\pi st} \, ds \right] \, dt \geq 0,
\]

where \( F_q \) and \( F_R \) are the Fourier transforms of \( q(t) \) and \( R(t) \), respectively, and \( s \) is frequency. This last expression can be rearranged to obtain

\[
\int F_q(s) \left[ \int_{-\infty}^{\infty} \left( F_R(s) e^{i2\pi st} \right) \, ds \right] \, ds \geq 0,
\]

where \( F_q(s) \) is the complex conjugate of \( F_q(s) \). Note that \( F_q(s) F_q^*(s) \geq 0 \).

As the function \( q(t) \) is arbitrary, the last expression must be true for any \( F_q(s) \). Consideration of the function \( F_q(s) = a \delta(s) \) for any constant \( a \) and arbitrary frequency \( s_0 \) leads to the conclusion that \( F_R(s_0) \geq 0 \) for all \( s_0 \), i.e., the Fourier transform of the covariance function \( R(t) \) is nonnegative at all frequencies. According to the Wiener-Khinchin theorem, the Fourier transform of \( R(t) \) is the spectrum of \( c(t) \). The constraint that \( R(t) \) be positive definite is therefore equivalent to requiring that the spectrum of \( c(t) \) be nonnegative at all frequencies. Similarly, the requirement that \( N \) be positive definite is equivalent to requiring a nonnegative spectrum of measurement errors.

It is convenient to express the mean squared error in terms of the signal variance \( \eta^2 = \mathbb{E}(0) \) and correlation function \( \rho(t) = R(t) / \eta^2 \) and the measurement error variance \( \sigma^2 \). If we make the definitions

\[
P_{mn} = \rho(t_m - t_n) \tag{7a}
\]

\[
\theta_m = \frac{1}{T} \int_{t_0-T/2}^{t_0+T/2} \rho(t - t_m) \, dt, \tag{7b}
\]

\[
\gamma = \frac{1}{T^2} \int_{t_0-T/2}^{t_0+T/2} \int_{t_0-T/2}^{t_0+T/2} \rho(t - t') \, dt \, dt', \tag{7c}
\]

then the mean squared error (4) expressed as a fraction of the raw signal variance \( \eta^2 \) can be written in compact vector notation as

\[
\psi^2(\alpha, t_0, T) = \frac{\psi^2(\alpha, t_0, T)}{\eta^2} = \alpha^T \left( P + \lambda N' \right) \alpha - 2 \alpha^T \theta + \gamma, \tag{8}
\]

where \( P \) is the \( M \times M \) matrix of cross correlations between the observations with \((m,n)\)th element \( P_{mn} \) given by (7a), \( \theta \) is the vector of integrals (equation (7b)) with transpose given by \( \theta^T = (\theta_1, \ldots, \theta_M) \), \( N' = N / \sigma^2 \) is the \( M \times M \) matrix of cross correlations between the measurement errors, and \( \lambda = \sigma^2 / \eta^2 \) is the measurement error-to-signal variance ratio. If the \( \epsilon_m \) are uncorrelated, then \( N' \) is just the identity matrix \( I \).

Note that a value of \( \psi^2 = 1 \) corresponds to an expected error of 100% of the raw signal variance \( \eta^2 \). In applications, it may be preferable to consider the mean squared error normalized by the variance of the time-averaged signal, \( \psi^2(\alpha, t_0) \equiv \psi^2(\alpha, t_0, T) / \eta^2 \). The normalization (8) is used in this study because it is slightly more convenient mathematically; the mean squared error expressed as a fraction of the variance of the time-averaged signal is just equal to \( \gamma^{-1} \psi^2 \).

It is evident from (8) that the fractional mean squared error of any linear estimate (3) of the time average \( z(t_0) \) can be determined given the correlation functions of the signal \( c(t) \) and the measurement errors \( \epsilon_m \), the measurement error-to-signal variance ratio \( \lambda \) and the weights \( \alpha_m \). Since \( P \) and \( \theta \) depend on the observation times \( t_m \), \( \psi^2 \) also depends implicitly on the temporal distribution of sample observations in the irregularly spaced time series \( y_m \).

Since (8) does not depend on the data values themselves, the expected errors of various hypothetical sampling patterns can be determined before any actual data are acquired. Conversely, the expected errors for a variety of
linear estimates (different choices for the weights \( a_m \)) and varieties of signal and measurement error correlation functions and measurement error-to-signal variance ratio can be determined from a given hypothetical sampling pattern. Both of these approaches are explored in sections 5 and 6.

The weights \( a_m \) in (3) can be chosen in many ways. The simplest example is the composite average of all \( M_T \) data \( y_m \) within the interval \( T \) centered at \( t_0 \). The weights in the composite average are therefore

\[
a_m = \frac{1}{M_T} \Pi \left( \frac{t_m - t_0}{T} \right),
\]

(9)

where \( \Pi \left( \frac{t_m - t_0}{T} \right) \) is the rectangle function, which has a value of 1 if \( |t_m - t_0| \leq T/2 \) and a value of zero otherwise. The accuracy of the composite average depends upon the spectral characteristics (the variances and correlation functions) of the signal \( c(t) \) and measurement errors \( e_m \), as well as on the number and temporal distribution of observations within the averaging period.

An obvious weakness of the composite average is that no information outside of the averaging interval is utilized in the estimate. There is clearly room for improvement in estimating the average since observations outside of the averaging interval at least provide information about the low-frequency components of the variability of \( c(t) \). An important special case is that of no observations in the averaging interval. Then the only possible composite average estimate is \( \bar{c}(t_0) = 0 \), equivalent to setting \( a_m \) equal to the zero vector \( \theta \). In this case, the fractional mean squared error (8) reduces to

\[
\psi^2(\alpha = 0) = \gamma = \frac{\langle z^2(t_0) \rangle}{\eta^2}.
\]

(10)

The value of \( \psi^2 \) for an estimate of zero is thus equal to the variance of the time averaged quantity (1) expressed as a fraction of the raw signal variance \( \eta^2 \). The mean squared error of a composite average must be smaller than (10) to be better than simply estimating \( \bar{c}(t_0) = 0 \) for the time average. Since it is assumed that any energetic cyclical variability has been removed, an estimate of zero is equivalent to just estimating the cyclical variability at the estimation time \( t_0 \). For the application considered in this study, this is just the seasonal cycle of log-transformed chlorophyll concentration at time \( t_0 \).

An improvement over the composite average method is to use a linear estimate (3) that is optimal in some statistical sense. An obvious choice is the estimate \( \bar{z}_{opt}(t) \) that minimizes \( \psi^2 \) with respect to the weights \( a_m \). Differentiating (8) with respect to each of the \( a_m \), \( m = 1, \ldots, M \), and setting each derivative equal to zero yields the \( M \) equations

\[
(P + \lambda N') \alpha_{opt} = \theta
\]

(11)

to be solved for the \( M \) optimal weights \( a_m \). It is evident that the optimal estimate involves the solution of an \( M \times M \) linear system and that this solution requires an a priori knowledge of \( P \), \( N' \), and \( \lambda \). Since the \( M \times M \) matrices \( P \) and \( N' \) are both positive definite and symmetric, so is the matrix \( (P + \lambda N') \). From (11), the fractional mean squared error (8) of the optimal estimate reduces to the simpler form

\[
\psi^2(\alpha_{opt}, t_0, T) = \gamma - \alpha_{opt}^T \theta.
\]

(12)

In practice, the solution of (11) can be cumbersome computationally if the number \( M \) of measurements \( y_m \) is large. In such cases, it may be expedient to utilize only the \( M' \) measurements with observation times \( t_m \) nearest to the estimation times \( t_0 \), thereby reducing the order of the linear system to be solved. In the applications considered here, \( M' \) has been limited to the number of observations within \( \pm 100 \) days of each estimation time \( t_0 \). This typically yields values of \( M' \approx 40 \) for CZCS observations off the west coast of North America.

Because the optimal estimate is expressly designed to minimize the mean squared error, no other linear estimate of the form (3) has lower fractional mean squared error than (12). The optimal estimate is therefore the best possible linear estimate by this measure. Since the optimal estimate utilizes information both from within and outside of the interval \( T \) centered at \( t_0 \), optimal estimation provides an estimate for any time \( t_0 \) regardless of whether any observations actually exist within the interval \( T \) centered on \( t_0 \). Optimal estimates with fractional mean squared error (12) exceeding some prescribed maximum value \( \psi^2_{max} \) can be rejected as being too uncertain to be useful.

3. Satellite Estimates of Chlorophyll Concentration

The potential for estimating near-surface chlorophyll concentration and primary production over the global ocean from satellite data has been clearly demonstrated by numerous analyses of color radiance measurements made by the CZCS. These data have been used to study the spatial structure of phytoplankton distributions from individual images [Abbott and Zion, 1985; Smith et al., 1988] and from composite average images [Esaias et al., 1986]. CZCS data have also been used to investigate the temporal variability of phytoplankton distributions from sequences of individual images [Palafox and McCowan, 1986; Michaelsen et al., 1988] and composite average images [Strub et al., 1990]. The promising results from these and other regional studies have stimulated considerable interest in using satellite ocean color data to investigate the spatial and temporal variability of phytoplankton on ocean basin scales. Since biological studies on such large scales are possible only by satellite remote sensing of ocean color, satellite observations have become an important component of large-scale biological field studies [e.g., Brewer et al., 1986; Rothschild et al., 1989; Johnke, 1990]. A global ocean color data set from the CZCS has recently been made available for such studies [Feldman et al., 1989].

A significant source of error in CZCS data is the accuracy of the instantaneous determination of ocean color after removal of the atmospheric contribution to the radiance measured by the satellite. One aspect of this error is calibration drifts caused by degradation of the radiometer. It appears that most of the calibration drifts in the radiance measurements can be eliminated by empirical corrections [Gordon et al., 1983a]. Of much greater concern are inaccuracies in atmospheric corrections applied to the radiance measurements. The water-leaving radiance in the visible band that is of interest typically accounts for less than 20% of the radiance measured by the satellite. The satellite measurements must therefore be corrected for atmospheric contamination, primarily Mie scattering from aerosol particles and Rayleigh scattering from atmospheric
molecules. The problem is difficult, since the radiative transfer model must account for multiple scattering of the visible radiation [Gordon and Castaño, 1987]. An aspect of the atmospheric correction that has recently raised considerable concern is the accuracy of the radiative transfer model used in early versions of CZCS processing. A linearized approximation to a single-scattering Rayleigh correction yielded erroneously high wintertime estimates of ocean color for solar illumination angles greater than about 50°, resulting in an artificial seasonal cycle in CZCS time series poleward of about 30° latitude that is approximately 180° out of phase with the true seasonal cycle [Fargion, 1989; Strub et al., 1990]. This error has been eliminated in the multiple-scattering Rayleigh correction developed by Gordon et al. [1988] and used in more recent CZCS processing, including the global CZCS data set generated by Feldman et al. [1989].

Another source of error in CZCS estimates of chlorophyll concentration is inaccuracies in the relation between ocean color (the pigment concentrations actually responsible for the water-leaving radiance at the visible wavelengths measured by the radiometer) and chlorophyll concentration. The chlorophyll algorithm issue has received considerable attention in the literature [e.g., Gordon et al., 1980, 1983; Smith and Baker, 1982; Gordon and Morel, 1983; Holligan et al., 1983; Abbott and Zion, 1987]. Most of these studies have concluded that near-surface chlorophyll in so-called "case 1 waters" in which phytoplankton and associated detrital products dominate the optical properties of the water [Morel and Prieur, 1977] can be estimated from the CZCS data to within a factor of two. This estimate does not appear to be rigorous, however, and apparently attributes all of the differences between in situ and CZCS estimates to CZCS measurement error. It is likely that at least part of the factor-of-2 difference results from the inherent difference between a point measurement (in situ) and a spatial average (CZCS) because of spatial patchiness of the phytoplankton distribution. In addition, since the CZCS measures an exponentially weighted vertical average of near-surface pigment concentrations over an "optical depth" which usually includes only 20-30% of the euphotic zone [Smith, 1981], there is no unique relationship between the satellite measurement and any simple characterization of the chlorophyll profile (e.g., the surface value or the vertical integral of chlorophyll concentration). Variations such as a subsurface maximum in the chlorophyll profile, for example, may thus also contribute to errors in the satellite estimates of "near-surface chlorophyll" concentration [Sathyendranath and Platt, 1989; Strauss, 1990].

The combination of color measurement error and chlorophyll algorithm error constitutes the overall CZCS chlorophyll measurement error. Obtaining an a priori estimate of the correlation matrix $N^f$ of these measurement errors is difficult. It is likely that there is some persistence of atmospheric aerosols that give rise to errors in atmospheric corrections applied to the satellite radiance measurements so that errors in successive images may be correlated over a few days or more. On the other hand, the atmospheric path length from the wide-swath CZCS radiometer to a given point on the sea surface changes substantially in successive images because of changes in incidence angle from longitudinal shifts of the satellite ground track. This tends to randomize measurement errors in successive images. In any case, no quantitative estimate of the time scale of measurement errors appears to exist at present. In the absence of such information, the $e_{m}$ will generally be assumed here to be uncorrelated, although the effects of correlated measurement errors are considered in sections 5.4 and 6.3.

A guideline for the appropriate value of the measurement error variance $\sigma^2$ is the conventional wisdom that the CZCS measures chlorophyll concentration to within a factor of 2. We interpret this to mean that "most" of the satellite estimates of a chlorophyll concentration $C$ fall between 0.5$C$ and 2$C$. This is formalized here by assuming that the satellite measurement errors $e_m$ are normally distributed in log$_{10}$ space with standard deviation log$_{10}2 = 0.3$. The variance of log-transformed measurement errors is then $\sigma^2 \approx 0.1$. This estimate is probably somewhat pessimistic, as was discussed above. The true CZCS measurement error probably lies somewhere between a factor of 1.5 and 2 (i.e., $\sigma^2$ between 0.03 and 0.1).

An additional source of error in time series constructed from CZCS data is the sampling error that arises from the fact that not all time scales of variability can be resolved by irregular temporal sampling of the continuously varying phytoplankton distribution. Depending on the degree of unresolved variability, sampling error may be more important than measurement error. Although the effects of sampling error in time series constructed from CZCS data are well recognized qualitatively, estimates of these effects have received little quantitative attention. Aside from maximizing the duty cycle of the instrument, only a limited amount can be done to reduce sampling error. The most important limiting factor is the presence of clouds. A concern raised previously by Michaelson et al. [1988] and Abbott and Zion [1987] is that the sampling is biased in favor of the physical conditions associated with clear atmospheric conditions. Since photoadaptation by phytoplankton populations is correlated with the solar radiation forcing associated with cloud cover, the satellite data are quite likely biased in favor of specific phytoplankton patterns. Unfortunately, the method presented in section 2 can do little to elucidate or correct for such sampling bias.

4. Chlorophyll Correlation Function

In general, the correlation function $\rho(\tau)$ must be geographically dependent. It cannot be estimated with sufficient accuracy from a CZCS time series at the location of interest because the CZCS measurement error is large and the intermittent sampling does not resolve all of the important time scales of variability. Ideally, an accurate estimate of the correlation function would be based upon a chlorophyll time series of at least 3 years’ duration sampled approximately hourly, thus resolving time scales from diurnal to interannual. Documenting the existence of even one such time series, let alone several in varied biogeographical regimes, has proven difficult.

The longest chlorophyll time series that we are aware of is the twice-weekly record from the end of the pier at Scripps Institution of Oceanography (SIO) in La Jolla, California. This ongoing sampling program has been conducted for over 7 years (since February 1983) by J. McGowan and the Marine Life Research Group (MLRG) at SIO. This remarkable data set is a tribute to the enduring efforts
of this organization. It is unfortunate that no other such data set appears to exist anywhere in the world. Even this data set is not completely sufficient for the purposes here, however, since it resolves variability only over periods longer than approximately 1 week (corresponding to the Nyquist frequency for a sample interval of 3.5 days). Chlorophyll variability over shorter periods accounts for much of the satellite sampling error that is the focus of this study. Some estimate of high-frequency variability is therefore required.

We have been unable to locate any calibrated hourly time series of in situ measurements of chlorophyll with sufficient duration to resolve time scales of variability up to 1 week. An alternative source of chlorophyll data from the same biogeographical regime as the SIO pier data is a 23-day time series of in situ fluorescence measurements obtained by C. R. Booth of Biospherical Instruments, Inc., in San Diego, California during the spring of 1984 from a mooring over Scripps Canyon, approximately 1 km offshore from SIO.

The SIO pier data and Scripps Canyon mooring data were combined in a straightforward manner to estimate a correlation function of chlorophyll variability: A composite frequency spectrum was formed from the two individual spectra. The composite spectrum was then inverse Fourier transformed to obtain the covariance function which was normalized by the signal variance to obtain the correlation function. The details of the spectral calculations for both data sets are described in sections 4.1 and 4.2. The composite spectrum and correlation function are described in section 4.3.

4.1. SIO Pier Data

The first 3.5 years of the SIO pier data (February 17, 1983, to June 6, 1986) were kindly provided by J. McGowan. Two bottle casts were taken 30 min apart approximately twice a week off the end of the SIO pier. Prior to November 21, 1983, chlorophyll and phaeophytin concentrations were determined from each cast using the 24-hour cold extraction method [Strickland and Parsons, 1972]; subsequent samples were processed by the 72-hour cold extraction method. The two individual samples of each pigment quantity were averaged. Chlorophyll concentrations were typically larger than phaeophytin concentrations by a factor of 2 or more. For each observation time, the average chlorophyll and phaeophytin were summed to obtain pigment concentration more analogous to the quantity that is actually observed by the satellite.

The time series of raw pigment concentration is shown in Figure 1. In general, the pigment concentrations were highest during late spring and early summer and lowest during the early fall, with a secondary peak during the late fall to early winter. Two notable exceptions to this pattern of seasonal variability are the very energetic spring bloom in 1985 and the lack of any spring bloom in 1984. Since 1986, winter blooms have also been observed in the data record (J. McGowan, personal communication, 1990). The SIO pier record of pigment concentration is thus dominated by interannual variability.

The lognormal character typical of biological quantities is clearly evident in Figure 1. The time series of log-transformed pigment concentration (Figure 2a) is much more "well behaved" in the sense that fluctuations are more symmetrically distributed about the mean value. The seasonal cycle with a primary peak in the spring and a secondary peak in the late fall is still evident. The anomalous conditions during 1984 and 1985 are seen to be of approximately equal but opposite amplitude in the log-transformed data. The record is still dominated by interannual variability. This may not be the case, in general, at other geographical locations. At higher latitudes, for example, primary production is light limited during the winter, often resulting in a strong seasonal cycle of pigment concentrations. After removal of the seasonal cycle, however, the statistical characteristics of pigment variability may be less geographically dependent. The seasonal cycle of log-transformed pigment concentration from the SIO pier was computed by least squares regression of the time series in Figure 2a onto annual plus semiannual harmonics and a constant offset (see smooth line in Figure 2a). The resulting seasonal cycle was then removed to obtain the time series of nonseasonal log-transformed pigment concentration shown in Figure 2b.

Determination of the frequency spectrum of pigment variability from the SIO pier data requires an evenly spaced data set. In the 3.5-year record analyzed here, 90% of the observations were separated by 3 or 4 days.

Fig. 1. The unevenly spaced time series of total pigment concentration (chlorophyll plus phaeophytin) in mg m\(^{-3}\) measured from the end of the SIO pier. The smooth curve represents the seasonal cycle computed by least squares regression onto annual plus semiannual harmonics and a constant offset.
Only five observations were separated by more than 5 days and 14 observations were separated by less than 3 days. The twice-weekly sampling was thus remarkably well maintained over the duration of the data set. The log-transformed SIO pier data were therefore interpolated to evenly spaced 3.5-day intervals by spline interpolation.

The spectrum of the nonseasonal interpolated log-transformed SIO pier time series is shown by the heavy line in Figure 3. In the log-log plot, the spectrum falls off approximately linearly with increasing frequency up to frequencies near the Nyquist frequency of about 0.14 cpd. Using the method described by Schlax and Chelton [1991], the sharp roll-off at frequencies just below the Nyquist frequency was found to be an artifact of the filtering characteristics of the spline interpolation, rather than a real drop in the spectral energy of chlorophyll variability. The spectrum at frequencies higher than about 0.1 cpd therefore cannot be determined from the SIO pier data. The in situ fluorescence data described in the following section are used to estimate the high-frequency portion of the spectrum.

4.2. Scripps Canyon Mooring Data

The in situ fluorescence measurements from the Scripps Canyon mooring are described in detail by Booth et al. [1987]. A fluorometer was moored at a mean depth of 7.7 m for the 23-day period April 17 to May 10, 1984. This fluorescence data set was processed and made available to us as hourly averages by D. Guggenheim of EcoAnalysis, Inc., in Ojai, California. Over the duration of the data set, fewer than 10% of the data values were missing. Short gaps of less than 5 hours had been filled by linear interpolation from neighboring values. Longer periods had been filled by the overall mean fluorescence value.

Since chlorophyll concentration $C(t)$ is proportional to fluorescence $F(t)$, the log-transformed chlorophyll concentration is

$$c(t) = \log_{10} \{\kappa(t) + f(t)\},$$

where $f(t) = \log_{10} F(t)$ and $\kappa(t)$ is the proportionality constant for the in situ fluorometer. This calibration constant is known to vary as a result of physiological changes in the phytoplankton and changes in nutrient status, light history, species composition, and other unexplained factors [Kiefer, 1973; Abbott et al., 1982] and may therefore vary with time over the 23-day duration of the mooring time series.

The time series of log-transformed fluorescence (Figure 4a) shows strong diurnal variability. Some of this variability may actually represent diel variations in near-surface chlorophyll abundance. However, since the mean mooring depth of approximately 7.7 m was fixed relative to the seafloor, much of the variability is probably due to vertical motion of the subsurface chlorophyll maximum associated with tides [Cullen et al., 1983]. Part of the diurnal signal is also associated with temporal variations in the calibration constant $\kappa$ due to diel physiological changes in the phytoplankton [Kiefer, 1973; Harris, 1980]. The two sources of diurnal variability associated with the measurement technique cannot be distinguished from actual diurnal variability of chlorophyll abundance. Chlorophyll variability on these time scales thus cannot be investigated from the mooring data. The raw hourly time series in Figure 4a was therefore regressed onto diurnal and semidiurnal harmonics and a constant offset. The resulting
harmonic variability was removed to obtain the anomaly log-transformed time series shown in Figure 4b.

For a relatively short in situ fluorescence time series at a fixed location, the primary source of temporal variability of the fluorometer calibration constant $\kappa$ is probably the diurnal signal. After removal of the diurnal cycle, the calibration constant may therefore be nearly constant for the 23-day moored fluorometer residual time series considered here. In this case, the Fourier transform of $c(t)$ given by (13) is

$$\mathcal{F}_c(s) = \mathcal{F}_f(s) + 6(s) \log_{10} \kappa.$$ \hspace{1cm} (13)

Since the spectrum of log-transformed chlorophyll variability at frequency $s$ is $S_c(s) = \mathcal{F}_c(s) \mathcal{F}_c^*(s)$, it is apparent that calibration of the in situ fluorometer is not a problem for spectral analysis of log-transformed data as long as the temporal variability of $\kappa$ is small after removing the diurnal cycle; the frequency spectra of $c(t)$ and $f(t)$ are identical in this case except at the zero frequency, which corresponds to the mean value and is therefore of no interest here.

The spectrum of the anomaly log-transformed moored fluorometer time series is shown by the thin line in Figure 3. It is apparent that the mooring data provide the information needed about variability at frequencies higher than can be resolved by the SIO pier data. In the log-log plot, the mooring spectrum falls off approximately linearly with increasing frequency with very nearly the same slope as the SIO pier spectrum.

4.3. Composite Spectrum and Correlation Function

The approximate factor-of-4 difference between the spectral density levels of the SIO pier and Scripps Canyon mooring time series (the vertical offset between the two spectra in Figure 3) could be due to several factors. Part of the difference may be due to the Hanning data windows (see section 11.5.2 of Bendat and Piersol [1986]) applied to the two different-length time series to reduce spectral leakage, although an attempt was made to correct for the reduction of sample variance in the windowed time series. Most of the difference is probably due to statistical sampling variability.

The duration of the mooring time series is only 23 days, coincidentally occurring during the period of lowest chlorophyll abundance in the 3.5-year SIO pier time series. This was also the period of lowest zooplankton biomass on record in the California Current. The low zooplankton and phytoplankton biomasses during 1984 have been associated with the major 1983 E1 Niño episode [McGowen, 1985; Strub et al., 1990], but there is no obvious explanation for why this timing would result in a higher variance at the high frequencies resolved by the mooring data. A more likely explanation is just normal sampling variability in the statistics. While the spectral shape (i.e., the linear roll-off in the log-log plot) is quite likely fairly stable, the absolute energy level is expected to vary somewhat with time. The sample variance in any particular 23-day period
will therefore vary, thus shifting the spectrum of mooring data up or down in Figure 3.

Statistical sampling variability similarly affects the spectrum of pier data. Indeed, the spectra of three non-overlapping subrecords of 1-year duration constructed from the full 3.5-year pier time series all had very nearly the same linear slope, but the sample variances for the three subrecords were found to differ by a factor of 2–3, thus shifting the spectra up or down by the same amount.

Given the uncertainty in the sample variances from the finite record lengths, it is easy to justify vertical shifts of the two individual spectra in Figure 3 so that they match at the overlapping frequencies. This is equivalent to hypothesizing a smooth, continuous spectral roll-off from low to high frequencies, an assumption that does not seem unreasonable. We have patched the two spectra together by shifting the pier spectrum up by about a factor of 2 and shifting the mooring spectrum down by about the same amount. Since none of the individual peaks in the two spectra are statistically significant, the composite spectrum is closely approximated by the dashed straight line in Figure 3. This line corresponds to a spectrum that varies with frequency as $s^{-1.35}$.

The procedure followed to obtain an estimate of the covariance function was to sample the two-sided continuous composite spectrum defined by

$$S_c(s) = \begin{cases} 0 & s < 5 \times 10^{-4} \text{ cpd} \\ s^{-1.35} & 5 \times 10^{-4} \text{ cpd} \leq s \leq 1 \text{ cpd} \\ 0 & s > 1 \text{ cpd} \end{cases} \quad (14)$$

(i.e., an $s^{-1.35}$ spectrum for periods ranging from 1 day to 2000 days) at 212 evenly spaced frequencies ranging from 0 to 10 cpd. The discrete spectrum was then inverse Fourier transformed to obtain the covariance function $R(\tau)$ at 212 lags $\tau$ separated by lag intervals of 0.1 day. The inverse Fourier transform must be handled carefully, as is described in section 11.4.2 of Bendat and Piersol [1986]. Increasing the upper limit of the frequency band over which the spectrum was assumed to have the form $s^{-1.35}$ has little effect, since the energy level is so low at high frequencies. However, the calculated covariance function is rather sensitive to the lower limit used for this frequency band. The value used here gives a good fit to the covariance function computed from the SIO pier data (see below).

Since the spectrum as defined by (14) is nonnegative at all frequencies, the covariance function derived in this manner is positive definite, as discussed in section 2. The computed covariance $R(\tau)$ at each lag $\tau$ was normalized by the variance $R(0)$ to obtain the correlation function $\rho(\tau)$. The signal variance $R(0) = \eta^2$ necessary to relate the correlation and covariance functions by $R(\tau) = \eta^2 \rho(\tau)$ must be estimated by other means (see step 4 in the appendix). The variance of the log-transformed SIO pier data was $\eta^2 = 0.065$.

The procedure followed to obtain an estimate of the correlation function $\rho(\tau)$ was to sample the two-sided continuous composite spectrum defined by
that are needed to solve (8) for the mean squared error of any linear estimate and to solve (11) for the weights in the optimal estimate are the $M^2$ correlations $P_{mn}$ given by (7a), the $M$ integrals $\delta_m$ given by (7b), and the double integral $\gamma$ given by (7c). The $P_{mn}$ require the signal correlation only at the discrete lags separating the observation times. For the CZCS measurements, which are separated by multiples of 1 day, these can easily be obtained from the derived correlation function with 0.1-day lag interval. The functions $\delta_m$ and $\gamma$ require first and second integrals of the correlation function over continuous lags. These integrals can probably be evaluated with acceptable accuracy by integrating the 0.1-day sampled correlation function using the simple trapezoidal rule. In practice, however, we fit the discretely sampled correlation function with a cubic spline, which immediately provides $\rho(\tau)$ at any arbitrary lag $\tau$ and its first integral (equation (7b)) between any two limits of integration. The second integral (equation (7c)) was then evaluated by Romberg integration of the spline-derived integral (equation (7b)) [e.g., Press et al., 1986].

An obvious concern in application is how representative the composite spectrum in Figure 3 (and hence the correlation function shown by the heavy line in Figure 5) is of chlorophyll variability at locations other than off southern California. Surely it is too much to hope for a universal frequency spectrum of chlorophyll variability. On the other hand, it may not be unreasonable to expect generally similar spectral characteristics (roll-off rates) at other locations after removing seasonal and diurnal variability. In any case, there seems to be an acute shortage of in situ chlorophyll time series from which to examine this question at present. The dependence of the estimated mean squared error on the form of the frequency spectrum is investigated in section 5.3, and the error introduced by assuming the $s^{-1.35}$ chlorophyll frequency spectrum deduced from Figure 3 is investigated in section 6.2. It is concluded that use of the correlation function shown by the heavy line in Figure 5 is adequate within the likely bounds of the spectral characteristics of chlorophyll variability.

For general application, discretely sampled values of this correlation function are listed in Table 1 for lags $\tau$ ranging from 0 to 100 days. A simple cubic spline fit to the 25 values in the table yields a continuous correlation function $\rho(\tau)$ that differs from the correlation function deduced from Figure 3 by a root mean square of 0.001 with a maximum of 0.015 over 0.1-day lag intervals between 0 and 100 days. The reader can therefore obtain an acceptably accurate continuous representation of the signal correlation function from a cubic spline fit to the 25 discrete values in Table 1.

In practice, if the spectrum of log-transformed chlorophyll variability is found to deviate significantly from that shown in Figure 3 at the particular geographical location of interest, the procedures outlined in this section can be followed to derive the appropriate correlation function for estimating the effects of sampling error.

5. Composite Averaging and Optimal Estimation

To illustrate the sampling error in time-averaged log-transformed chlorophyll variability, two 4.5-year time series of CZCS estimates of chlorophyll concentration were obtained for the geographical regions shown in Figure 6 from A. Thomas at the Atlantic Center for Remote Sensing of the Ocean in Canada. These data had been spatially averaged over approximately 0.07ø of latitude by 1.4ø of longitude. Recall from section 2 that the mean squared error of an estimated temporal average can be determined without actually knowing the data values. For the purposes of this study, only the observation times at the two locations are therefore considered.

The two geographical regions considered characterize two quite different CZCS sampling regimes. Clear skies

<table>
<thead>
<tr>
<th>$\tau$, days</th>
<th>$\rho(\tau)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.0000</td>
</tr>
<tr>
<td>1</td>
<td>0.8922</td>
</tr>
<tr>
<td>2</td>
<td>0.8400</td>
</tr>
<tr>
<td>3</td>
<td>0.8028</td>
</tr>
<tr>
<td>4</td>
<td>0.7729</td>
</tr>
<tr>
<td>5</td>
<td>0.7475</td>
</tr>
<tr>
<td>10</td>
<td>0.6548</td>
</tr>
<tr>
<td>15</td>
<td>0.5893</td>
</tr>
<tr>
<td>20</td>
<td>0.5369</td>
</tr>
<tr>
<td>25</td>
<td>0.4926</td>
</tr>
<tr>
<td>30</td>
<td>0.4539</td>
</tr>
<tr>
<td>35</td>
<td>0.4193</td>
</tr>
<tr>
<td>40</td>
<td>0.3878</td>
</tr>
<tr>
<td>45</td>
<td>0.3590</td>
</tr>
<tr>
<td>50</td>
<td>0.3323</td>
</tr>
<tr>
<td>55</td>
<td>0.3074</td>
</tr>
<tr>
<td>60</td>
<td>0.2840</td>
</tr>
<tr>
<td>65</td>
<td>0.2620</td>
</tr>
<tr>
<td>70</td>
<td>0.2412</td>
</tr>
<tr>
<td>75</td>
<td>0.2215</td>
</tr>
<tr>
<td>80</td>
<td>0.2027</td>
</tr>
<tr>
<td>85</td>
<td>0.1848</td>
</tr>
<tr>
<td>90</td>
<td>0.1677</td>
</tr>
<tr>
<td>95</td>
<td>0.1513</td>
</tr>
<tr>
<td>100</td>
<td>0.1355</td>
</tr>
</tbody>
</table>
are more common in region 1 off southern California than in region 2 off central Oregon, especially during the wintertime. Consequently, long time intervals between successive clear CZCS images are generally more frequent for region 2. This is quantified in Figure 7 by histograms of sample interval for the 4.5-year period beginning July 11, 1979. The sampling characteristics for the two regions can be briefly summarized as follows. Overall, there were about 15% more observations in region 1 than in region 2, although this was highly variable from year to year. During 1980, for example, there were nearly twice as many observations in region 1 as in region 2. During 1981 and 1982, there were about the same number of observations in each region. Clear images tended to occur in bursts rather than being evenly distributed over the year; the most frequent sample interval was 1 day, accounting for more than one fourth of the observations in each region. There were secondary peak sample intervals at about 5 and 11 days. For both regions, the overall average sample interval was about a week (see Figure 7) and more than two thirds of the observations were separated by a week or less.

The mean squared errors of composite averages and optimal estimates are examined in section 5.1 as functions of the averaging period $T$ and the measurement error-to-signal variance ratio $\lambda = \sigma^2/\eta^2$. Attention is focused on 30-day and 10-day averages in the remainder of section 5. The mean squared errors are examined as a function of time in section 5.2. The dependencies of the mean squared errors on the signal and measurement error correlation time scales are examined in sections 5.3 and 5.4, respectively. Throughout this section, it is assumed that the seasonal cycle of log-transformed chlorophyll variability has been estimated independently and removed from the observations (see steps 2 and 3 in the appendix). An estimate of $\tilde{z}(t_0) = 0$ is thus equivalent to estimating just the seasonal mean value at time $t_0$.

5.1. Averaging Period and Measurement Error-to-Signal Variance Ratio

Attention is restricted in this section to the sampling pattern for calendar year 1980. This period is characteristic of the first 3 years of the CZCS data record. Sampling was generally much less frequent after 1981 because of a combination of CZCS sensor degradation and reduced power availability onboard the Nimbus 7 satellite (see Fu et al. [1990] for a detailed discussion). Averaging periods $T$ ranging from 5 to 50 days and measurement error-to-signal variance ratios of $\lambda = \sigma^2/\eta^2 = 3.5, 1.5, 0.5, and 0$ are considered. All error estimates in this section were computed using the signal correlation function $\rho(\tau)$ shown by the heavy line in Figure 5, and the measurement errors were assumed to be uncorrelated. For the SIO pier signal variance of $\eta^2 = 0.065$, the four values of $\lambda$ correspond to log-transformed measurement error variances of $\sigma^2 = 0.23,$
These values of $\sigma^2$ represent factors of 0.1, 0.03, and 0. These values of $\sigma^2$ represent factors of 3, 2, 1.5, and 1 chlorophyll measurement errors, with the latter equivalent to zero measurement error. As discussed previously, the actual CZCS measurement error probably lies between a factor of 1.5 and 2, corresponding to $\sigma^2$ between 0.03 and 0.1. For the SIO pier variance of $\eta^2 = 0.065$, this corresponds to a value of $\lambda$ between 0.5 and 1.5.

For each combination of $T$ and $\lambda$, the fractional mean squared errors $\bar{\psi}(\alpha, t_0, T)$ of composite averages and optimal estimates were computed from (8) at times $t_i$, $i = 1, ..., N_T$ at the centers of the $N_T$ non-overlapping periods $T$ during 1980. Since composite averages cannot be constructed when there are no observations in the averaging period $T$, only those $N_T$ periods with at least one CZCS observation are considered here. Optimal estimates can be constructed for periods $T$ with no observations, but the mean squared errors for these periods are excluded from the analysis presented here to make the comparison between composite averages and optimal estimates compatible. The average fractional mean squared error over the period considered is

$$\bar{\psi}^2(\alpha, T) = \frac{1}{N_T} \sum_{i=1}^{N_T} \psi^2(\alpha, t_i, T).$$  (15)

The square root of the average fractional mean squared error (equation (15)), which will be denoted shorthand as $\bar{\psi}$ with the dependence on $\alpha$ and $T$ implicit, is plotted in Figure 8 as a function of $T$ for each value of $\lambda$ considered.

The errors of the optimal estimates decrease smoothly and monotonically with increasing averaging period. The errors are somewhat larger for the infrequently sampled region 2 than for the comparatively densely sampled region 1. In both regions, the dependence of $\bar{\psi}$ on averaging period is weak; the errors of optimal estimates of 10-day averages, for example, are less than 20% larger than those of 30-day averages (see, however, the discussion at the end of section 7). The errors of composite averages can be seen from Figure 8 to be much more dependent on $T$ than those of optimal estimates. For each value of $\lambda$, the $\bar{\psi}$ for the composite averages at each location are larger for all averaging periods than those for optimal estimates with the corresponding value of $\lambda$. The expected errors of composite averages approach those of optimal estimates when $T$ is large but generally increase much more rapidly with decreasing $T$.

It is also apparent from Figure 8 that the errors of composite averages are much more dependent on the distribution of CZCS observations than those of optimal estimates. For region 1, the errors decrease with increasing averaging period as anticipated. The same is generally true for region 2. The increase in error for averaging periods longer than 40 days for region 2 is because these averaging periods are long enough to include wintertime averages with a very small number of observations, as will be seen from the distribution of CZCS observations in Figure 10. For very small $\lambda$, however, the composite average errors for region 2 exhibit a rather surprising behavior; the $\bar{\psi}$ for small $T$ are actually somewhat smaller than those for large $T$. This is because the CZCS observations tend to occur in bursts of several samples over a short period, as was noted previously from Figure 7. Composite averages over these well-sampled short periods are therefore reasonably accurate as long as the measurement error is small (i.e., $\lambda$ is small). As the averaging period increases beyond the span over which bursts of samples occur, the composite averages become less reliable. The $\bar{\psi}$ for region 1 are less sensitive to this effect than those for region 2 because there were fewer long gaps between sampling bursts (i.e., more short intervals with a small number of samples).

The dotted lines in Figure 8 represent the standard deviation of time-averaged log-transformed chlorophyll concentration $z(t_0)$ normalized by the square root of the signal variance $\eta^2$. As shown in section 2, this quantity is given by the square root of (10), which is the fractional expected error of an estimate of $z(t_0) = 0$ and will therefore be denoted shorthand as $\psi_0$. The expected error $\bar{\psi}$ must be

![Fig. 8. The square root of the average fractional mean squared error (equation (15)) of composite averages (thin lines) and optimal estimates (heavy lines) for calendar year 1980 as a function of averaging period $T$ for the two regions shown in Figure 6. The four individual curves of each type correspond to measurement error-to-signal variance ratios $\lambda$ of (top to bottom) 3.5, 1.5, 0.5, and 0. The dotted lines represent the fractional expected error of an estimate of zero, equal to $\psi_0$ given by the square root of (10), as a function of averaging period $T$.](image-url)
smaller than \( \psi_0 \) for a linear estimate of the form (3) to be better than simply estimating \( \hat{z}(t_0) = 0 \). At both locations, the \( \psi \) for optimal estimates are considerably smaller than \( \psi_0 \) for all \( T \) and \( \lambda \). The most important conclusion from Figure 8 is that for a value of \( \lambda = 1.5 \) (corresponding to a factor-of-2 measurement error for the SIO pier variance of \( \eta^2 = 0.065 \)), the \( \psi \) for composite averages exceed \( \psi_0 \) for averaging periods of less than 2–3 weeks for region 1 and for all averaging periods for region 2.

The effects of changing \( \lambda \) are shown by the plots of \( \psi \) as a function of \( \lambda \) in Figure 9 for averaging periods of \( T = 10 \) and 30 days. For composite averages and optimal estimates, \( \psi \) initially increases rapidly with increasing \( \lambda \). The rate of increase drops by about a factor of 2 as \( \lambda \) increases from 0 to 1 and remains nearly constant thereafter (i.e., \( \psi \) increases approximately linearly with increasing \( \lambda \)). Over the full range of \( \lambda \), \( \psi \) for composite averages is more than twice as sensitive to \( \lambda \) as that for optimal estimates.

The greater sensitivity of composite averages to \( \lambda \) is a result of the relatively small number of CZCS samples within the averaging interval. The estimates are therefore more susceptible to errors in the observations. A quantitative understanding of the accuracy of composite averages thus depends critically on knowledge of the measurement error-to-signal variance ratio \( \lambda \). This requires detailed knowledge of the signal variance \( \eta^2 \) and the measurement error variance \( \sigma^2 \). As discussed previously, the latter is presently not well understood for the CZCS. Optimal estimates are computed from observations both within and outside of the averaging interval and are therefore based on a much larger number of observations than composite averages. This helps to reduce the effects of random measurement errors in the optimal estimates. Consequently, when compared with composite averages, the accuracies of optimal estimates are less dependent on \( \lambda \).

It can be concluded that the accuracies of optimal estimates are only weakly dependent on \( T \), \( \lambda \), and the temporal distribution of sample observations, while the accuracies of composite averages are very sensitive to all three factors. Indeed, without accurate knowledge of \( \lambda \), it is not even possible to generalize whether a composite average for a given \( T \) will be more accurate than just estimating \( \hat{z}(t_0) = 0 \). Similarly, the accuracy of a composite average for a given \( \lambda \) depends critically on \( T \) and the temporal distribution of sample observations.

5.2. Temporal Dependence

In this section, the expected errors of composite averages and optimal estimates are examined as a function of time for the 1980 and 1981 calendar years. The square root of the fractional mean squared error (8) at time \( t_0 \) will be denoted shorthand as \( \psi_0 \), with the dependence on \( \alpha \) and \( T \) implicit. Optimal estimates were computed for all intervals and a composite average of zero (with corresponding error \( \psi = \psi_0 \)) was used for intervals with no observations. The signal correlation function \( \rho(\tau) \) shown by the heavy line in Figure 5, the SIO pier signal variance of \( \eta^2 = 0.065 \), and a factor-of-2 measurement error (\( \sigma^2 = 0.1 \)) were used and the measurement errors were assumed to be uncorrelated. These values of \( \eta^2 \) and \( \sigma^2 \) correspond to a measurement error-to-signal variance ratio of \( \lambda = 1.5 \).

The errors of 30-day averages are shown as a function of time in Figure 10a. For both regions, the \( \psi \) exhibit a distinct seasonal cycle related to the temporal distributions of observations. The range of the seasonal variability of \( \psi \) is much larger for composite averages than for optimal estimates. For region 1, the errors are large for the approximate 2-month winter rainy season off southern California and again for 2–3 months during the summer season of coastal fog and stratus cloud cover. For region 2, the errors are large for an extended period of 4–6 months during the northwest winter rainy season and small during the summer season of prevailing clear skies. Except for the foggy periods in region 1 and clear periods in region 2 during late summer of both years, the \( \psi \) for each estimate are everywhere smaller for region 1 than for
the less frequently sampled region 2. The $\psi_i$ for optimal estimates of 30-day averages are smaller than those for composite averages by 20–30% during periods of frequent CZCS samples and by more than 100% during periods of infrequent samples. For the value of $\lambda = 1.5$ used here, the $\psi_i$ for composite averages often exceed $\psi_0$ (shown by the dotted lines), especially for region 2. By comparison, the $\psi_i$ for optimal estimates are always smaller than $\psi_0$.

The errors of 10-day averages are shown as a function of time in Figure 10b. The $\psi_i$ for 10-day composite averages are generally smaller than $\psi_0$, but only marginally so. Since the actual CZCS measurement error probably falls somewhere between $\sigma^2 = 0.03$ and 0.1 (corresponding to $\lambda$ between 0.5 and 1.5 for the SIO pier variance of $\eta^2 = 0.065$), it can be concluded that 10-day composite averages are not sufficiently more accurate for most purposes than simply estimating $\hat{z}(t_0) = 0$.

5.3. Signal Correlation Time Scale

The error estimates presented in sections 5.1 and 5.2 were based on the correlation function $\rho(\tau)$ shown by the heavy line in Figure 5. As was described in section 4.3, this correlation function was computed from a frequency

---

**Fig. 10a.** The square root of the fractional mean squared error (equation (8)) of non-overlapping 30-day composite averages (thin lines) and optimal estimates (heavy lines) as a function of time during 1980 and 1981 for a measurement error-to-signal variance ratio of $\lambda = 1.5$. The times of CZCS observations are shown by the dots along the bottom of each plot and the short vertical lines separate the non-overlapping 30-day averaging periods. The dotted lines correspond to $\psi_0$ given by the square root of (10) for 30-day averages.
Fig. 10b. The same as Figure 10a, except for non-overlapping 10-day averaging periods.

spectrum with $s^{-1.35}$ roll-off. The formalism in section 2 can be applied to investigate the dependence of mean squared error on the signal spectral roll-off rate (or, equivalently, the signal correlation time scale) by changing $\rho(\tau)$ appropriately. The method summarized in section 4.3 was used to obtain correlation functions for spectra of the form $S_c(s) = s^{-r}$ for spectral roll-off rates $r$ ranging from 0 to 3. These correlation functions were used to compute $\psi$ for composite averages and optimal estimates during 1980 as a function of $r$ using the same signal and measurement error characteristics as in section 5.2.

The values of $\psi$ for 10-day and 30-day averages are shown as a function of $r$ in Figure 11. As anticipated, $\psi$ for composite averages are always much larger than those for optimal estimates. For all values of $r$, the errors are smaller for region 1 than for region 2, reflecting the different sampling densities at the two locations. The values of $\psi$ for the composite average decrease with increasing $r$. For $r \geq 1.4$, the $\psi$ for optimal estimates also decrease with increasing $r$. Thus the accuracies of both composite averages and optimal estimates improve as the chlorophyll signal becomes progressively more dominated by low-frequency variability.

A perhaps surprising result of Figure 11 is that the $\psi$ for optimal estimates decrease for $r \leq 1.4$. This can be explained as follows. In the extreme case of $r = 0$, the chlorophyll variability is a random, "white noise" process over all time scales down to the 1-day period included in the spectrum (equation (14)) from which the correlation function was derived. The true average is then very nearly zero. The optimal estimate takes this random character into account in the evaluation of the weights...
random nature of the measurement error. Measurement errors serve to make the measured chlorophyll time series which the b for composite averages are smaller than the crossover point above which 30-day composite averages both shift to smaller values of r because of the assumed optimal estimates. The critical roll-off rates for optimal increasing r is greater both for composite averages and for values of b decrease and the rate of decrease of b with approximately r -- 1.2 for region I and r - 1.6 for region 2. become more accurate than an estimate of zero occurs at exceed it for small r. For a value of A - 1.5 used here, with increasing r. For 30-day composite averages, the estimate of zero, with the degree of improvement of the measurement errors are examined in section 5.4.
The ψ for 10-day averages can be seen from Figure 11 to depend on r similarly to the ψ for 30-day averages. For a given value of r, the values of ψ for optimal estimates increase somewhat, but the ψ for composite averages increase substantially for the shorter averaging period. The most significant change is that the crossover point above which the ψ for composite averages are smaller than ψ0 shifts to a much higher value of r. For a value of λ = 1.5, the composite averages are less accurate than an estimate of zero even for a spectral roll-off rate as large as r = 3.
It can be concluded that because composite averages are based on the relatively small number of observations within each averaging period, they can only benefit in a very limited way from low-frequency variability of chlorophyll. This is especially true for large λ. Since optimal estimates are based on a much larger number of observations spanning a much wider period of time, they more effectively utilize information on the spectral characteristics of chlorophyll variability to improve the accuracy of the estimate ẑ(t0).

5.4 Measurement Error Correlation Time Scale
In the preceding sections, the measurement errors have been assumed to be temporally uncorrelated. The actual correlation time scale of CZCS measurement errors is very poorly understood at present. The effects of temporally correlated measurement errors are investigated in this section by assuming a Gaussian measurement error correlation function with a value of 0.5 at lag τ = τ1/2 for a range of values of τ1/2: a value of τ1/2 = 0 corresponds to uncorrelated measurement errors. The Gaussian measurement error correlation functions were used to compute the error correlation matrix N' for each value of τ1/2 considered.

The ψ were then computed from (8) for composite averages effectively equivalent to an error-free time series with lower spectral roll-off rate. Decreasing λ thus has the same effect as increasing r. The effects of serial correlation of the measurement errors are examined in section 5.4.

Since ψ for optimal estimates becomes increasingly smaller than ψ0 as the spectral roll-off rate r increases, optimal estimates are always more accurate than an estimate of zero, with the degree of improvement with increasing r. For 30-day composite averages, the values of ψ are smaller than ψ0 for large values of r but exceed it for small r. For the value of λ = 1.5 used here, the crossover point above which 30-day composite averages become more accurate than an estimate of zero occurs at approximately r = 1.2 for region 1 and r = 1.6 for region 2.
The ψ for 30-day averages and smaller values of λ (not shown) are qualitatively similar to Figure 11. The values of ψ decrease and the rate of decrease of ψ with increasing r is greater both for composite averages and for optimal estimates. The critical roll-off rates for optimal estimates and the value of the roll-off crossover point above which the ψ for composite averages are smaller than ψ0 both shift to smaller values of r because of the assumed random nature of the measurement error. Measurement errors serve to make the measured chlorophyll time series

\[ \alpha_m \] by (11) and yields an estimate of approximately zero with expected error thus equal to ψ0. Recall from (10) that the fractional error ψ0 of an estimate of ẑ(t0) = 0 is also the fractional standard deviation of the time-averaged signal z(t). The value of ψ0 (shown by the dotted lines in Figure 11) therefore increases as the signal spectral roll-off rate r increases (i.e., the spectrum becomes less "white"). The optimal estimate takes into account the nonrandom nature of the signal variability when r > 0 and yields an estimate that is better than ẑ(t0) = 0. The value of ψ is consequently less than ψ0 and therefore increases with increasing r at a lower rate than ψ0. The critical roll-off rate r ≈ 1.4 apparently represents the point at which the spectrum of chlorophyll variability becomes sufficiently "red" for the optimal estimate to separate the low-frequency signal from the random measurement error. The values of ψ then begin to decrease with increasing r above the critical roll-off rate.

Since ψ for optimal estimates becomes increasingly smaller than ψ0 as the spectral roll-off rate r increases, optimal estimates are always more accurate than an estimate of zero, with the degree of improvement with increasing r. For 30-day composite averages, the values of ψ are smaller than ψ0 for large values of r but exceed it for small r. For the value of λ = 1.5 used here, the crossover point above which 30-day composite averages become more accurate than an estimate of zero occurs at approximately r = 1.2 for region 1 and r = 1.6 for region 2.

The ψ for 30-day averages and smaller values of λ (not shown) are qualitatively similar to Figure 11. The values of ψ decrease and the rate of decrease of ψ with increasing r is greater both for composite averages and for optimal estimates. The critical roll-off rates for optimal estimates and the value of the roll-off crossover point above which the ψ for composite averages are smaller than ψ0 both shift to smaller values of r because of the assumed random nature of the measurement error. Measurement errors serve to make the measured chlorophyll time series effectively equivalent to an error-free time series with lower spectral roll-off rate. Decreasing λ thus has the same effect as increasing r. The effects of serial correlation of the measurement errors are examined in section 5.4.
The ψ for 10-day averages can be seen from Figure 11 to depend on r similarly to the ψ for 30-day averages. For a given value of r, the values of ψ for optimal estimates increase somewhat, but the ψ for composite averages increase substantially for the shorter averaging period. The most significant change is that the crossover point above which the ψ for composite averages are smaller than ψ0 shifts to a much higher value of r. For a value of λ = 1.5, the composite averages are less accurate than an estimate of zero even for a spectral roll-off rate as large as r = 3.

It can be concluded that because composite averages are based on the relatively small number of observations within each averaging period, they can only benefit in a very limited way from low-frequency variability of chlorophyll. This is especially true for large λ. Since optimal estimates are based on a much larger number of observations spanning a much wider period of time, they more effectively utilize information on the spectral characteristics of chlorophyll variability to improve the accuracy of the estimate ẑ(t0).
and optimal estimates during 1980 as a function of $\tau_{1/2}$ using the same values of $\rho(\tau)$ and $\lambda$ as in section 5.2.

The values of $\psi$ for averaging periods of $T = 10$ days and 30 days are shown as a function of error correlation time scale $\tau_{1/2}$ in Figure 12 for $\tau_{1/2}$ ranging from 0 to 150 days. Compared with optimal estimates, the accuracies of composite averages can be seen to be much more sensitive to correlated measurement errors. For small values of $\tau_{1/2}$, $\psi$ increases with increasing $\tau_{1/2}$ at a rate more than 3 times greater for composite averages than for optimal estimates. The $\psi$ for composite averages are nearly twice as large as those for optimal estimates for all $\tau_{1/2}$.

For $\tau_{1/2}$ longer than approximately the averaging period $T$, the expected errors of composite averages become relatively constant because only CZCS observations within the averaging interval are included in the composite average, and $\tau_{1/2} > T$ therefore cannot increase $\psi$ substantially. A similar flattening of the error curve for optimal estimates occurs because observations nearest the averaging period $T$ are weighted most heavily. An interesting result from Figure 12 is that the $\psi$ for optimal estimates continue to increase until $\tau_{1/2} \approx 50$ days, regardless of the averaging period. For larger $\tau_{1/2}$, the $\psi$ for optimal estimates decrease somewhat, evidently because such very long time scale measurement errors are better resolved with the large number of observations included in optimal estimation and can therefore be better accounted for in the estimate.

A curious feature of Figure 12 is the crossover of the error curves for 10-day and 30-day composite averages for region 2. For small $\tau_{1/2}$, the $\psi$ for 10-day averages are smaller than those for 30-day averages as expected. For $\tau_{1/2} \approx 10$ days, however, 10-day composite averages are more accurate than 30-day composite averages. This is related to the burst sampling discussed previously which was more prevalent in region 2 than in region 1 during 1980. The 10-day periods included in the estimate $\psi$ computed by (15) are generally better sampled (i.e., the observations are more evenly distributed over the averaging interval) than for 30-day periods. The measurement error is therefore better resolved in 10-day composite averages for sufficiently long error decorrelation time scale $\tau_{1/2}$.

### 6. Suboptimal Estimation

The error analyses in section 5 clearly demonstrate the superiority of optimal estimation over composite average estimates of time-averaged chlorophyll variability. The primary difficulty in application is the need for a priori estimates of the chlorophyll signal and measurement error variances and correlation functions. An important question is how sensitive the optimal estimates are to assumed values for these parameters. The sensitivities to assumed values for the measurement error-to-signal variance ratio $\lambda$, the spectral roll-off rate $r$ of the signal (or equivalently, the signal correlation time scale) and the measurement error correlation time scale $\tau_{1/2}$ are examined in this section.

The optimal estimation formalism in section 2 is used to compute the weights $a_m, r, m = 1, \ldots, M$ by solving the linear system (11) using assumed values for $\lambda$, $r$, and $\tau_{1/2}$. The resulting estimates $\hat{z}(t_0)$ will be referred to here as "suboptimal estimates" and are equal to the optimal estimates only if the three assumed parameters are correct. The expression (8) for the fractional mean squared error is valid for any prescribed weights $a_m$ and is thus applicable whether or not the correct values of $\lambda$, $r$, and $\tau_{1/2}$ are used in (11) to compute the $a_m$. The only requirement is that the correct values for these parameters must be used in all terms in (8) except the $a_m$.

A point that may need clarification is the distinction between the true error of suboptimal estimates and the error estimated using assumed values for $\lambda$, $r$, and $\tau_{1/2}$ in (8) as well as in (11). If the three parameters are unknown, one might consider computing the fractional mean squared error...
from (8) using the assumed values. However, if the true values differ from the assumed values, this error estimate may be overly optimistic or pessimistic. Unfortunately, the relationship between the true error of the suboptimal estimate and the error computed using the assumed values for \( \lambda, r, \) and \( \tau_{1/2} \) cannot be determined without accurate knowledge of all three parameters. However, if the true error of the suboptimal estimate can be shown to differ little from the error of the optimal estimate, one can at least take solace in the fact that the accuracy of the estimate \( \hat{z}(t_0) \) would not improve much from more accurate knowledge of \( \lambda, r, \) and \( \tau_{1/2}. \) The goal of this section is to investigate this question.

6.1. Measurement Error-to-Signal Variance Ratio

Primarily because of ignorance of the measurement error variance \( \sigma_e^2, \) the measurement error-to-signal variance ratio \( \lambda \) is generally not well known. A value of \( \lambda = 1.5 \) has been used throughout most of this study in order to isolate the effects of other factors on the estimation error. This particular value of \( \lambda \) was based on the SIO pier signal variance of \( \sigma_e^2 = 0.065 \) and the worst-case scenario of a factor-of-2 measurement error, corresponding to \( \sigma_e^2 = 0.1. \)

The effects of \( \lambda \) on \( \psi \) for optimal estimates were examined in section 5.1. The \( \psi \) for suboptimal estimates computed using \( \lambda = 1.5 \) to solve (11) for the weights \( \alpha_m \) are examined in this section as a function of the true value of \( \lambda \) assuming \( r = 1.35 \) and \( \tau_{1/2} = 0. \)

The results for calendar year 1980 are shown by the heavy dashed lines in Figure 11 for suboptimal estimates of 10-day and 30-day averages. As it must be, \( \psi \) for the suboptimal estimate is exactly the same as that for the optimal estimate when the assumed value of \( \lambda = 1.5 \) is correct. It is evident from Figure 9 that the values of \( \psi \) for the suboptimal estimates increase approximately linearly with increasing \( \lambda, \) but the dependence on \( \lambda \) is remarkably weak; the \( \psi \) for suboptimal estimates increase by less than 30% as \( \lambda \) increases from 0 to 3. Moreover, the \( \psi \) for suboptimal estimates differ little from the values for optimal estimates when \( \lambda \gtrsim 0.6. \) For large enough \( \lambda, \) then, the suboptimal estimates obtained by assuming that \( \lambda = 1.5 \) are nearly as accurate as the optimal estimates that could be obtained if the true value of \( \lambda \) were known. For \( \lambda \lesssim 0.6, \) the suboptimal estimates become increasingly less accurate than the optimal estimates but are still more accurate than composite averages (except for very small \( \lambda \) for region 1).

Because of the generally small log-transformed chlorophyll signal variance after the seasonal cycle is removed (only 0.065 for the SIO pier data) and the inherent noisiness of CZCS estimates of chlorophyll concentration (measurement errors somewhere between a factor of 1.5 and 2), a measurement error-to-signal variance ratio of \( \lambda \approx 0.6 \) probably represents a lower bound on the true value in most applications. Thus unless the signal variance \( \sigma_e^2 \) is very large (resulting in a small value of \( \lambda \)), use of the value \( \lambda = 1.5 \) to estimate time averages by suboptimal estimation should produce estimates that are acceptably close to the actual optimal estimates for most applications.

6.2. Signal Correlation Time Scale

In application, little may be known about the nature of chlorophyll variability at the location of interest. This necessitates using an assumed form for the signal correlation function \( \rho(r). \) The correlation function shown by the heavy line in Figure 5 has been used throughout most of this study in order to isolate the effects of other factors on the estimation error. This particular form for \( \rho(r) \) was derived from the \( s^{-1.35} \) frequency spectrum obtained in section 4 from in situ measurements of chlorophyll and fluorescence off southern California. The effects of different spectral roll-off rate \( r \) on \( \psi \) for optimal estimates were examined in section 5.3. The \( \psi \) for suboptimal estimates computed using \( r = 1.35 \) to solve (11) for the weights \( \alpha_m \) are examined in this section as a function of the true spectral roll-off rate \( r \) assuming \( \lambda = 1.5 \) and \( \tau_{1/2} = 0. \)

The results for calendar year 1980 are shown by the heavy dashed lines in Figure 11 for suboptimal estimates of 10-day and 30-day averages. As it must be, \( \psi \) for the suboptimal estimate is exactly the same as that for the optimal estimate when \( r = 1.35; \) this is the spectral roll-off for which the assumed correlation function was generated. It is evident from the figure that the \( \psi \) for suboptimal estimates are nearly constant and differ little from the \( \psi \) for optimal estimates when \( r \gtrsim 1. \) For large enough \( r, \) then, the suboptimal estimates obtained by assuming that \( r = 1.35 \) are nearly as accurate as the optimal estimates that could be obtained if the true value of \( r \) were known. The agreement is especially close for values of \( r \) ranging from about 1 to 2. This was found to be true regardless of the measurement error-to-signal variance ratio \( \lambda. \)

For smaller values of \( r, \) the suboptimal estimates become increasingly less accurate than the optimal estimates but are still much more accurate than the composite averages.

The correlation functions \( \rho(r) \) for \( s^{-1} \) and \( s^{-2} \) spectral roll-offs are shown in Figure 5 by the dashed and dotted lines, respectively. It is evident from Figure 5 that the \( \rho(r) \) have a value of 0.5 are about 4 and 76 days for \( s^{-1} \) and \( s^{-2} \) spectra, respectively, compared with about 25 days for the \( s^{-1.35} \) spectrum. These are probably reasonable bounds on the true correlation time scale of log-transformed chlorophyll variability at most locations. Since the suboptimal estimates are nearly as accurate as optimal estimates for \( r \) between 1 and 2, use of the correlation function for \( s^{-1.35} \) spectral roll-off (heavy line in Figure 5) to estimate time averages by suboptimal estimation should produce estimates that are acceptably close to the actual optimal estimates for most applications. For practical application, the reader can obtain a usefully accurate representation of this correlation function from a cubic spline fit to the 25 discrete samples listed in Table 1.

6.3. Measurement Error Correlation Time Scale

As was discussed in section 3, the correlation time scale \( \tau_{1/2} \) of CZCS measurement errors is presently very poorly understood but is generally speculated to be small. A value of \( \tau_{1/2} = 0 \) has therefore been used throughout most of this study in order to isolate the effects of other factors on the estimation error. The \( \psi \) for suboptimal estimates computed using a value of \( \tau_{1/2} = 0 \) to solve (11) for the weights \( \alpha_m \) are examined in this section as a function of the true value of \( \tau_{1/2} \) assuming \( r = 1.35 \) and \( \lambda = 1.5 \).

The results for calendar year 1980 are shown by the heavy dashed lines in Figure 12 for suboptimal estimates of 10-day and 30-day averages. As it must be, \( \psi \) for the suboptimal estimate is exactly the same as that for the
optimal estimate when the assumed value of $\tau_{1/2} = 0$ is correct. In contrast to the relatively small sensitivity to the accuracies of assumed forms for $\lambda$ and $r$ demonstrated in sections 6.1 and 6.2, the estimated time average can be seen to be very sensitive to $\tau_{1/2}$. The $\bar{\psi}$ for suboptimal estimates increase with increasing $\tau_{1/2}$ at approximately twice the rate as the $\bar{\psi}$ for optimal estimates but are still much more accurate than composite averages, even for very large $\tau_{1/2}$. Note, however, that suboptimal estimates become worse than an estimate of zero (i.e., $\bar{\psi}$ becomes greater than $\psi_0$) for $\tau_{1/2} \gtrsim 30$ days.

It can be concluded that the time scale of measurement errors is an important factor in the accuracy of suboptimal estimates. As long as $\tau_{1/2}$ is less than a week or so, the suboptimal estimates obtained by assuming that $\tau_{1/2} = 0$ should produce estimates that are acceptably close for most applications to the actual optimal estimates that could be obtained if the true value of $\tau_{1/2}$ were known. For longer measurement error correlation time scales, however, the expected errors of suboptimal and optimal estimates diverge, and the difference becomes significant.

7. **APPLICATION TO SIMULATED CZCS DATA**

The formalism presented in this paper can be demonstrated by applying the procedure outlined in step-by-step detail in the appendix to the 3.5-year SIO pier data. For this application, the time-averaged chlorophyll concentration can be determined exactly and compared with composite averages and optimal estimates constructed from irregularly spaced samples of the pier data. The in situ SIO pier observations described in section 4.1 were log transformed and interpolated to obtain a nearly continuous time series (evenly spaced data values with a 0.25-day sample interval). This time series was then seasonally corrected and sampled at the CZCS observation times for each of the two regions shown in Figure 6. It will be assumed for the purposes of demonstration that the interpolated pier data are error free. CZCS observations were simulated by adding normally distributed random noise with variance $\sigma^2 = 0.1$ (corresponding to a factor-of-2 measurement error) to the interpolated log-transformed pier time series at the times of CZCS observations for regions 1 and 2. Since the signal variance of the pier data is $\eta^2 = 0.065$, this value of $\sigma^2$ represents a measurement error-to-signal variance ratio of $\lambda = 1.5$. The portion of the CZCS record considered spans the 3.5-year period beginning January 1, 1980.

For each region, 30-day averages of log-transformed chlorophyll concentration were estimated at non-overlapping 30-day intervals by composite averages and optimal estimation. For 30-day periods in which there were no CZCS observations, a composite average of zero was used, equivalent to estimating the seasonal mean value for that period. Optimal estimates were formed using the correlation function $\rho(t)$ shown by the heavy line in Figure 5. The estimated time averages were then compared with the "true" time averages computed by integration of the 0.25-day pier data. In order to focus only on the estimation procedure, results are presented here for the nonseasonal log-transformed chlorophyll concentration from step 10 of the procedure outlined in the appendix.

Time series of simulated CZCS 30-day composite averages and optimal estimates for the region 1 and region 2 sampling patterns are shown in Figure 13a by the thin and heavy lines, respectively, along with the true 30-day averages (dotted lines). The composite averages exhibit approximately twice as much scatter about the true values as do the optimal estimates. For this particular 3.5-year sample record, the root-mean-square errors of the composite averages and optimal estimates for region 1 are 0.26 and 0.16, respectively; the corresponding errors for region 2 are 0.23 and 0.12. The occasional large positive and negative spikes in the composite average time series, which have no basis in reality, are artifacts of sampling error from infrequent sampling of chlorophyll variability.

Time series of simulated CZCS 10-day averages are shown in Figure 13b. The positive and negative fluctuations of the composite averages are much more erratic than those for 30-day averages. The risk of interpreting these features as real variations in chlorophyll concentration is clearly a concern.

An important observation from Figures 13a and 13b is that for the value of $\lambda = 1.5$ used here, the 10-day optimal estimates do not differ substantially from the 30-day optimal estimates. This results in part from the fact that the use of the correlation function shown by the heavy line in Figure 5 effectively defines the filtering characteristics of the optimal estimation. In addition, the white noise spectral energy of measurement errors becomes larger than the spectral energy of the signal above some threshold frequency so that high-frequency signal and measurement error cannot be distinguished. Because of these two effects, the optimal estimate gives overly smooth time series of short time averages. The conclusion is that chlorophyll variability on 10-day time scales cannot be resolved either by composite averages or by optimal estimation for the signal spectrum and measurement error-to-signal variance ratio of $\lambda = 1.5$ considered here. For composite averages, the symptom of this inability to resolve the signal is erratic fluctuations in the chlorophyll estimates. For optimal estimation, the symptom is an overly smooth time series. As $\lambda$ decreases, the ability to resolve chlorophyll variability on short time scales improves. The improvement is much greater for optimal estimation than for composite averages. This once again emphasizes the need for accurate knowledge of $\lambda$.

8. **CONCLUSIONS**

A technique has been presented for quantifying the sampling error in estimates of a time-averaged quantity constructed from a time series of irregularly spaced observations at a fixed location. Although the emphasis here is on chlorophyll concentration estimated from satellite measurements of ocean color, the methodology is equally applicable to any irregularly sampled data set. In applications of the technique to chlorophyll data, it is advantageous to work with log-transformed data because the approximately normal distribution of the resulting data values is more desirable, and it is easier to simulate the measurement errors believed to be characteristic of log-transformed satellite estimates of chlorophyll concentration.
Fig. 13a. Simulated CZCS time series of 30-day composite averages (thin lines) and optimal estimates (heavy lines) of nonseasonal log-transformed chlorophyll concentration constructed as described in the text for the two regions shown in Figure 6. The true 30-day averages are shown by the dotted lines. The times of CZCS observations for the period January 1980 through May 1983 are shown by the dots along the bottom of each plot, and the short vertical lines separate the non-overlapping 30-day averaging periods. The units on the left axes are nonseasonal log-transformed chlorophyll concentration. The corresponding nonseasonal chlorophyll concentrations in mg m$^{-3}$ are shown on the right axes.

Fig. 13b. The same as Figure 13a, except for non-overlapping 10-day averaging periods.
It has been assumed that the interest is in time averages, rather than instantaneous values, of chlorophyll concentration at a particular location. A method that is frequently used in analyses of CZCS data is to composite average the irregularly spaced observations within the specified averaging period. This method was considered in detail in section 5. The accuracy of composite averages was shown to be very sensitive to the temporal distribution of CZCS observations. The expected error of composite averages increases rapidly with increasing measurement error-to-signal variance ratio \( \lambda \), decreasing averaging period \( T \), and increasing measurement error correlation time scale \( \tau_{1/2} \). For realistic values of these parameters, composite averages over periods less than about a month are often less accurate than an estimate of zero. This conclusion depends most critically on the value of \( \lambda \), which requires more accurate knowledge of the measurement error variance than is presently available. Since the method outlined here is applied to nonseasonal data (see the appendix), an estimate of zero is equivalent to simply estimating the seasonal mean value.

A much better method of estimating time averages was introduced in section 2 and examined in detail in section 5. The method is based on optimal estimation, which estimates the time average as a weighted sum of the irregularly spaced observations with weights \( \alpha_m \) computed from the signal and measurement error variances and correlation functions. Optimal estimates were shown to be much more accurate than composite averages. Moreover, the accuracy of optimal estimates is much less sensitive to \( \lambda \), \( \tau_{1/2} \), and the temporal distribution of CZCS observations. Furthermore, optimal estimation is much more effective at utilizing information on the signal correlation time scale to reduce the expected error of the estimate.

The primary difficulty in application of optimal estimation, aside from a greater computational burden than simple composite averaging, is the need for a priori specification of the signal and measurement error variances and correlation functions. In the absence of detailed descriptions of these parameters, it was suggested that the optimal estimation formalism be applied using assumed values of \( \lambda = 1.5 \), \( \tau_{1/2} = 0 \), and the signal correlation function derived from in situ measurements of chlorophyll and fluorescence off southern California. The resulting estimates (termed "suboptimal estimates" here) are optimal only if the assumed values of the parameters are correct. Discrete samples of the signal correlation function are listed in Table 1. A simple spline fit to the 25 values in the table yields a continuous representation of \( \rho(r) \) with acceptable accuracy for general application. The 3.5-year time series of optimal estimates of non-overlapping 30-day averages in section 7 were computed based on the \( M' \) observations within \( \pm 100 \) days of each estimation time in about 15 seconds of CPU time on a MicroVAX II computer.

The agreement between suboptimal and optimal estimates was investigated in section 6. It was concluded that the accuracy of suboptimal estimates is only weakly dependent on \( \lambda \) and the signal correlation function but strongly dependent on the measurement error correlation time scale \( \tau_{1/2} \). The error correlation time scale for CZCS data is not well understood at present but is believed to be relatively short. It was shown in section 6.3 that suboptimal estimates become significantly less accurate than optimal estimates if measurement errors are correlated over time scales longer than about a week. However, the suboptimal estimates were shown to be much more accurate than composite averages, even for long measurement error correlation time scales. Suboptimal estimation using the recommended values for the parameters is thus a significant improvement over the composite averaging method presently used.

The conclusions of sections 6.1 and 6.2 have important implications for practical applications of the formalism presented here. While it is easy to criticize the in situ chlorophyll measurements off southern California as not being representative of chlorophyll variability at other locations, it is much more difficult to support this concern with hard evidence; the unfortunate fact is that chlorophyll time series of sufficiently long duration and short sampling interval necessary to test this objection appear to be lacking at locations other than off southern California. The weak sensitivities of the estimate to the measurement error-to-signal variance ratio \( \lambda \) and the signal correlation time scale (or, equivalently, the signal spectral roll-off rate \( r \)) imply that there is wide latitude in the specification of these parameters; use of the values deduced from the in situ data off southern California should therefore be acceptable for applications at most other locations. If evidence is found to the contrary, it is straightforward to apply the formalism of section 2 with different values for these parameters.

The procedure for applying the formalism presented here is outlined in detail along with some cautionary notes in the appendix. Application of the technique in section 7 to two simulated CZCS data sets for which the true time averages are known clearly illustrates the superior performance of optimal estimates compared with composite averages. It also demonstrates that a large measurement error-to-signal variance ratio \( \lambda \) and specification of a long signal correlation time scale fundamentally limit the temporal resolution of estimated time averages. For the chlorophyll signal spectrum considered here and a value of \( \lambda = 1.5 \), 30-day optimal estimates are adequate, but 10-day optimal estimates are overly smooth. The ability to resolve short time averages improves with decreasing \( \lambda \) and specification of a shorter signal correlation time scale.

In principal, the temporal interpolation technique presented here, which uses only observations at the geographical location of interest, could be extended to include spatial interpolation using CZCS observations at nearby locations. This would require a priori knowledge of the spatial correlation function of chlorophyll variability over scales ranging from about 1 km to perhaps 500 km. Such information probably does not exist from presently available in situ data. Moreover, since the spatial scales of cloudy skies are generally large, it is unlikely that spatial interpolation would improve the estimated time averages by much; statistically, when a satellite observation is missing at the location of interest, nearby observations are generally also missing. Furthermore, the spatial scales of errors in satellite estimates of chlorophyll concentration are also large because of the large-scale nature of atmospheric aerosols responsible for most of the measurement error. Spatial interpolation using CZCS observations from nearby locations would therefore not substantially reduce the effects of measurement errors. We conclude that
temporal interpolation alone is probably adequate for most applications.

**Appendix: Summary of the Application Procedure**

The procedure for applying the formalism described in section 2 to obtain optimal estimates of time-averaged chlorophyll concentration is summarized in detail in this appendix. Consider a set of \( M \) irregularly spaced observations \( Y_m \) of the chlorophyll concentration \( C(t) \) in milligrams per cubic meter with measurement errors \( E_m \),

\[
Y_m = C(t_m) + E_m, \quad m = 1, \ldots, M. \tag{16}
\]

Define the chlorophyll concentration \( Z \) in milligrams per cubic meter averaged over a prescribed period \( T \) centered at time \( t_0 \) to be

\[
Z(t_0) = \frac{1}{T} \int_{t_0-T/2}^{t_0+T/2} C(t) \, dt. \tag{17}
\]

The procedure for deriving a time series of optimal estimates \( \hat{Z}(t) \) of the true time average (17) at a collection of estimation times \( t_0, i = 1, \ldots, N_f \) is as follows:

1. Log transform the observations \( Y_m \) to obtain

\[
\hat{y}_m = \tilde{c}(t_m) + \tilde{\epsilon}_m \tag{18}
\]

where \( \hat{y}_m \), \( \tilde{c}(t_m) \) and \( \tilde{\epsilon}_m \) are the log10 transforms of \( Y_m \), \( C(t_m) \) and \( E_m \).

2. Estimate the seasonal cycle of the log-transformed data \( \hat{y}_m \). One method of achieving this is to regress the \( \hat{y}_m \) onto annual plus semiannual harmonics plus a constant offset,

\[
\hat{y}_{\text{seas}}(t) = \hat{c}_{\text{seas}}(t) + \epsilon_{\text{seas}}(t) = a_0 + a_1 \sin(2\pi f_1 t + \zeta_1) + a_2 \sin(2\pi f_2 t + \zeta_2), \tag{19}
\]

where \( f_1 (= 1 \text{ cycle yr}^{-1}) \) and \( f_2 (= 2f_1) \) are the annual and semiannual frequencies and \( a_0, a_1, a_2, \zeta_1, \text{ and } \zeta_2 \) are the amplitudes and phases estimated by regression analysis. Higher-order harmonics can be included in the regression if there is reason to believe that this is warranted, but it has been the experience of the authors that annual plus semiannual harmonics are almost always sufficient to define the seasonal cycle. If the measurement errors are random, the seasonal cycle of \( \hat{y}_m \) is just the seasonal cycle of \( \hat{c}(t) \). Any seasonal errors \( \epsilon_{\text{seas}}(t) \) in the satellite measurements must be estimated independently and removed from \( \hat{y}_{\text{seas}}(t) \) to isolate the seasonal cycle of chlorophyll concentration, \( \hat{c}_{\text{seas}}(t) \). Of particular concern is the seasonal error known to exist due to incorrect modeling of Rayleigh scattering in early processing of CZCS data as was discussed in section 3. Strub et al. [1990] have proposed a method for estimating the resulting seasonal cycle \( \epsilon_{\text{seas}}(t) \) off the west coast of the United States based on analysis of data far offshore where they hypothesize that \( \epsilon_{\text{seas}}(t) \) is relatively small.

3. Remove the seasonal cycle (equation (19)) to obtain an irregularly spaced time series of nonseasonal log-transformed chlorophyll measurements

\[
y_m \equiv \hat{y}_m - \hat{y}_{\text{seas}}(t_m) = c(t_m) + \epsilon_m, \tag{20}
\]

where \( c(t_m) \) is the nonseasonal log-transformed chlorophyll and \( \epsilon_m \) is the nonseasonal log-transformed measurement error. Note that the formalism of section 2 can be applied without removing the seasonal cycle (or other energetic cyclical variability) or after removing a poor estimate of the seasonal cycle, but the correlation functions used in the optimal estimation may then be dominated by the seasonal variability. The resulting optimal estimates would yield little more information than just the seasonal cycle.

4. Estimate the measurement error-to-signal variance ratio \( \lambda \) for the log-transformed chlorophyll data. This step requires careful consideration, since it is unlikely that either the signal variance nor the measurement error variance are known accurately a priori. It is easy to compute an estimate of the total variance (signal plus measurement error) from the sample variance of the observations \( y_m \). The accuracy of this estimate depends on the sample size \( M \). Separating the signal variance \( \eta^2 \) and the measurement error variance \( \sigma^2 \) contributions to the total variance can be a problem. If some information is known about the measurement error variance (e.g., that the measurements are accurate to within a factor of 2), then \( \sigma^2 \) can be prescribed a priori; a factor-of-2 measurement error corresponds to \( \sigma^2 = (\log_{10} 2)^2 \approx 0.1. \) As was discussed in section 3, this value of \( \sigma^2 \) is probably pessimistic. The true value probably lies between 0.03 and 0.1, corresponding to measurement errors between a factor of 1.5 and 2. Even if \( \sigma^2 \) is known a priori, estimating \( \eta^2 \) is difficult. If the signal and measurement error are known to be uncorrelated, then \( \eta^2 \) can be approximated as the difference between the sample variance and the assumed measurement error variance. This method can unfortunately give a negative signal variance, since the sample size \( M \) is typically small, resulting in large uncertainty in the sample variance, and \( \sigma^2 \) for CZCS data is generally relatively large compared with the signal variance \( \eta^2 \) after removal of the seasonal cycle. It was shown in section 6.1 that a practical solution is to use a value of \( \lambda = 1.5. \) As long as the true value of \( \lambda \) is larger than about 0.6, the suboptimal estimate obtained assuming \( \lambda = 1.5 \) is very nearly the same as the optimal estimate obtained using the correct value of \( \lambda. \)

5. Choose an appropriate form for the signal correlation function \( \rho(\tau) \). An adequate estimate of \( \rho(\tau) \) cannot be obtained from CZCS observations themselves because of their inherently large measurement errors and poor temporal resolution from the irregular sample spacing. As was described in section 4, an accurate estimate of \( \rho(\tau) \) requires a long, densely sampled time series of chlorophyll at the location of interest. In the absence of such information, the correlation function shown by the heavy line in Figure 5 derived from the composite spectrum constructed from in situ chlorophyll measurements and fluorometer data off southern California could be used. It was shown in section 6.2 that the mean squared error of the resulting suboptimal estimate is only weakly dependent on the form of the correlation function; for log-transformed chlorophyll variability with spectral roll-off ranging from \( s^{-1} \) to \( s^{-2} \) (corresponding to a signal correlation time scale between 4 and 76 days), the suboptimal estimate obtained assuming an \( s^{-1.35} \) spectrum is very nearly the same as the optimal estimate obtained using the correct correlation function. We therefore conclude that the correlation function in Figure 5 should be suitable for most.
applications. The reader can obtain a usefully accurate representation of this correlation function from a simple cubic spline fit to the 25 discrete samples listed in Table 1.

6. Calculate the $M \times M$ data-data cross correlation matrix $P$ with $(m,n)$th element given by (7a). If only the $M'$ observations $y_m$ nearest each estimation time $t_i$, rather than all $M$ observations, are used in order to reduce the computational burden, the $M' \times M'$ matrix $P$ will in general be different for each $t_i$, $i = 1, \ldots, NT$.

7. For each time $t_i$ at which the time-averaged chlorophyll concentration is to be estimated, compute the integrals $\theta_m$, $m = 1, \ldots, M$ and $\gamma$ given by (7b) and (7c).

8. Choose an appropriate form for the measurement error correlation function $N'$. This is probably the most difficult step in the procedure, since the quality of the estimate depends strongly on the correlation time scale of the measurement errors. Such information does not exist at the present time, but the CZCS measurement error time scale is generally believed to be relatively short, as was discussed in section 2. It was shown in section 6.3 that the suboptimal estimate obtained by assuming that the measurement errors are uncorrelated should be adequate as long as the actual measurement error correlation time scale is less than about a week.

9. For each estimation time $t_i$, solve the linear system (11) to obtain the weights $\alpha_m(t_i)$, $m = 1, \ldots, M$.

10. For each estimation time $t_i$, compute the estimate $\tilde{z}(t_i)$ of the time average (equation (1)) using the linear form (3) with weights $\alpha_m(t_i)$ determined in the preceding step. The fractional mean squared error $\psi^2$ of this estimate could then be computed by (12) using the weights $\alpha_m(t_i)$ and the integrals $\theta_m$ and $\gamma$ determined in step 7. However, the mean squared error computed in this manner for the suboptimal estimate based on assumed values for the measurement error-to-signal variance ratio, the signal correlation function, and the measurement error correlation time scale may be overly optimistic or pessimistic, depending on the accuracies of the assumed values. The true mean squared error cannot be determined without accurate knowledge of all three parameters (see further discussion at the beginning of section 6).

11. Choose some threshold fractional mean squared error $\psi^2_{\text{max}}$ above which it is felt that the uncertainty in the estimated time average is too large. Estimates with $\psi^2_{\text{max}}$ larger than $\psi^2_{\text{max}}$ should be rejected.

12. Compute the time average of the seasonal cycle of log-transformed chlorophyll concentration at each estimation time $t_i$,

$$z_{\text{season}}(t_i) = \frac{1}{T} \int_{t_i-T/2}^{t_i+T/2} \tilde{z}_{\text{season}}(t) \, dt.$$  

(21)

As was noted in step 2 above, care must be taken to eliminate any seasonally dependent errors $\tilde{z}_{\text{season}}(t)$ so that $z_{\text{season}}(t)$ can be obtained from the seasonal cycle (equation (19)) computed from the log-transformed CZCS observations $y_m$.

13. Add the time averaged seasonal cycle (equation (21)) to the estimates $\tilde{z}(t_i)$ at each estimation time $t_i$ to obtain

$$\tilde{z}(t_i) = \tilde{z}(t_i) + z_{\text{season}}(t_i).$$  

(22)

14. Inverse log transform the $\tilde{z}(t_i)$ to obtain the estimated time-averaged chlorophyll concentration in milligrams per cubic meter at each estimation time $t_i$, $i = 1, \ldots, NT$.

$$\tilde{z}(t_i) = 10^{\tilde{z}(t_i)}.$$  

(23)

Acknowledgments. We thank John McGowan for the SIO pier data and Rocky Booth for the moored fluorescence data used in section 4. David Guggenheim is gratefully acknowledged for his processing of the mooring data and for providing the cleaned-up data set used in section 4.2. Financial support for the exceptionally long (presently more than 7 years) SIO pier time series of chlorophyll and phaeopigment measurements has been provided by the Marine Life Research Group at SIO. Discussions of the fluorescence data with Ray Smith and Mark Abbott were very helpful. We also thank Andrew Thomas for providing the CZCS observation times used in this study. Comments from Andrew Thomas, Mark Abbott, John McGowan, Donna Witter, Kathy Kelly, Colin Roessler, and three anonymous reviewers are greatly appreciated. The research summarized in this paper was supported by NASA grants NAGW-869 and NAGW-1883.

REFERENCES


D. B. Chelton and M. G. Schlax, College of Oceanography, Oceanography Administration Building 104, Oregon State University, Corvallis, OR 97331-5503.

(Received July 9, 1990; accepted January 30, 1991.)