REDUCED SPACE APPROACH TO THE OPTIMAL ANALYSIS OF HISTORICAL MARINE OBSERVATIONS: ACCOMPLISHMENTS, DIFFICULTIES, AND PROSPECTS

A. Kaplan, M.A. Cane, and Y. Kushnir

Lamont–Doherty Earth Observatory of Columbia University Palisades, New York, USA

WMO Guide to the Applications of Marine Climatology,

World Meteorological Organization, Geneva, Switzerland, 2002,

in press

January 28, 2002

Abstract

Observed historical climate fields are characterized by comparatively precise data and good coverage in the last few decades, and poor observational coverage prior. The technique of the reduced space optimal analysis of such fields (i.e. estimating them in projections onto a low-dimensional space spanned by the leading patterns of the signal variability) is presented in the context of more traditional approaches to data analysis. Advantages of the method are illustrated on examples of reconstructions of near-global monthly fields of sea surface temperature and sea level pressure from 1850s to present along with verified error bars. The limitations of the technique concerning quality and robustness of estimating a priori parameters, representation of long-term and small-scale types of variability, assumption of stationarity of means and covariances, and incompleteness of coverage are discussed, and possible ways to overcome these problems are suggested.

1. Introduction

Less than two centuries of observational records which have made their way from the hand-written ship logs to the modern data banks constitute the main source of our knowledge of the variability associated with the ocean-atmosphere interaction. For the use in climate research the ship measurements are customarily being compiled into monthly binned averages on regular longitude–latitude grids with quality control and other statistics (e.g. Comprehensive Ocean–Atmosphere Data Set (COADS) – Woodruff et al. 1987; Global Ocean Surface Temperature Atlas (GOSTA) – Bottomley et al. 1990). Resulting products still reflect the historical variations in the intensity of marine traffic, being incomplete at present, quite gappy before 1950s, and extremely sparse for the most of the 19th century. Satellite observations can complete the modern part of this record (almost two decades for sea surface temperature (SST), much less for other climate variables), but cannot provide a record lengthy enough for the studies of decadal and longer time scale climate variability. As a result, in climate studies one faces the necessity to use very incomplete and affected by observational and sampling error data fields. In contrast, the two main approaches to modern climate research, namely statistical techniques (like principal components analysis, singular vector decomposition, singular spectrum analysis, etc) and model experiments (use of observed fields for boundary conditions), both expect gapless and error-free input data. Because of that a great deal of attention was paid in the last few decades to various methods of data analysis, to those which are supposed to interpolate gaps and suppress data error.

Majority of existing approaches to interpolating historical data are drawn from the idea of minimization of least squares. It is well-known (by Gauss–Markov theorem; e.g. Mardia et al. 1979, Rao 1973) that the systematic use of it allows to produce an optimal estimate (unbiased one with the smallest error among all linear estimates). This involves, however, a few assumptions, including the knowledge of error covariances. Lacking this knowledge, some additional assumptions are usually being made. Statistical techniques like kriging (e.g. Cressie, 1991) or successive corrections (Daley 1993) normally assume some "localized" covariance structure and produce useful results if the gap size does not exceed the data decorrelation scale (e.g. Da Silva et al. 1994, Levitus and Boyer 1994).

A seemingly different approach to the historical data analysis (often also called data reconstruction, to emphasize the scarcity of the input data), which is based on the use of empirical orthogonal functions (EOFs), has become quite popular in recent years (Shriver and O'Brien (1995), Smith et al. (1996), Rayner et al. (1996), Mann et al. (1998)). In fact this technique can be derived as a straightforward application of a classic least squares estimate with a special EOF-based reduced rank approximation of a signal (or model error) covariance matrix. In this venue Kaplan et al. (1997) formulated reduced space analogues of the traditional technique of optimal analysis (optimal interpolation, Kalman filter, optimal smoother). Applications of this technique to the historical data sets of SST and marine sea level pressure (SLP) resulted in near-global monthly analyses of these variables going back to more than 140 years accompanied by the error bars (Kaplan et al. (1998,2000)) which are publicly available. The assumptions underlying the method, namely the stationarity of the mean field and covariance of the signal, have been recently criticized (Hurrell and Trenberth (1999)). Additionally, the current settings of the analysis result in globally incomplete fields of comparatively sparse resolution ($4\div5^o$ grid size) which limits considerably the utility of such analyses in climate model experiments. In section 2 we bring the reduced space optimal analysis into the context of more traditional objective data analyses and summarize its advantages and existing applications. Section 3 discusses the current difficulties in applications of the method and suggests the ways of resolving them. Section 4 concludes the paper by emphasizing the prospects of the method and directions for the further applications.

2. Accomplishments

2.1 Theoretical background

The generic problem of optimal analysis of time-evolving fields $\mathcal{T}_n$ (n is the time index) requires reconciliation of information coming from two sources: an imperfect model of time transitions A_n and incomplete and erratic observations $\mathcal{T}_n^o$ connected to the estimated field via a linear (or linearized) operator H_n (Figure 1). Note that error of linearization (or interpolation) of the operator H_n is included into effective observational error e^{obs}.

This problem is central for two areas of climate research which traditionally are considered separately: assimilation of data into numerical models and objective analyses (reconstructions) of data sets of historical observations. In fact the main difference between these two types of problems is in relative amount of information brought by the model versus observations: it is high in the former problem and low in the latter. If model and observational error in the equations shown in Figure 1 are white in time, uncorrelated with each other and the estimated fields $\mathcal{T}_n$, have zero mean and known spatial covariances Q_n and R_n respectively, we have a classic Gauss–Markov estimation problem for $\mathcal{T}_n$, whose solution can be found as a minimizer of the quadratic cost function

$$S[\mathcal{T}_1, \mathcal{T}_2, \cdots, \mathcal{T}_N] =$$
$$\textstyle\sum_{n=1}^{N}(H_n\mathcal{T}_n - \mathcal{T}_n^{\text{o}})^T R_n^{-1}(H_n\mathcal{T}_n - \mathcal{T}_n^{\text{o}}) + \sum_{n=1}^{N-1}(\mathcal{T}_{n+1} - A_n\mathcal{T}_n)^T Q_n^{-1}(\mathcal{T}_{n+1} - A_n\mathcal{T}_n) \qquad (1)$$

(For a detailed explanation of notation, terminology and basic facts of optimal estimation a reader is referred to Kaplan et al. (1997)). According to the Gauss–Markov theorem (e.g., Mardia et al. (1979), Rao (1973)), this solution has minimum error variance among all linear estimates of $\mathcal{T}$, and because of that it is usually referred to as the "optimal" solution. In fact, if additional assumptions on the Gaussian distribution of errors or the signal are made, the same solution receives an interpretation as the maximum likelihood estimate, or becomes the best solution among all, even nonlinear estimates, for a wide class of optimality criteria. Because the solution minimizes a quadratic cost function, it is often referred to as a "least-squares solution".

There exist well-known algorithms to find a minimizer of (1) in its complete form (fixed-interval optimal smoother (OS)), or somewhat truncated forms (fixed-lag optimal smoother, Kalman filter (KF)), or its simplification for a single-time estimation (optimal interpolation (OI)). They are supposed to give optimal solutions if assumptions on errors are satisfied, including the requirement that the covariance matrices of errors (Q and R) are known. However, in actual applications to the problems of climate research the realistic dimensions of data are usually large enough to warrant two outcomes:

1. error covariance matrices are not known in all their details since there are not enough data to resolve them completely, so some crude parameterizations are used instead,

2. if no simplifications are done, optimal data analysis procedures are very expensive (OI), extremely expensive (KF), or prohibitively expensive (fixed-interval OS).

Both these difficulties, however, can be dealt with at once if certain features of optimal solutions of realistic climate fields are taken into account.

Consider as an example a standard OI problem whose solution is a minimizer $\mathcal{T}$ of the cost function

$$\mathbf{S}[\mathcal{T}] = (H\mathcal{T} - \mathcal{T}^o)^T R^{-1}(H\mathcal{T} - \mathcal{T}^o) + (\mathcal{T} - \mathcal{T}^b)^T C^{-1}(\mathcal{T} - \mathcal{T}^b). \tag{2}$$

Here $\mathcal{T}^o$ is a (column-) vector of observations, $\mathcal{T}^b$ is a first guess (background) solution, H is a transfer matrix from a complete field to the set of observed points, R and C are covariances of observational and first guess errors respectively. The two terms of the cost function $\mathbf{S}$ "punish" the solution $\mathcal{T}$ for deviation from observations and from the background values.

Solution to this OI problem is

$$\mathcal{T} = P\left(H^T R^{-1}\mathcal{T}^o + C^{-1}\mathcal{T}^b\right),$$

where

$$P = \left(H^T R^{-1} H + C^{-1}\right)^{-1}$$

is estimated covariance of its error.

Let us subtract the first guess solution from the estimated field, so that the new $\mathcal{T}$ is $\mathcal{T} - \mathcal{T}^b$ and new $\mathcal{T}^o$ is $\mathcal{T}^o - H\mathcal{T}^b$. If the first guess solution is a climatological field, then we have redefined the signal to be a field of anomalies. After such a change in definitions the first guess solution equals zero, so that the first guess error equals the entire value of the signal $\mathcal{T}$, and the matrix C becomes the covariance of the signal $\langle \mathcal{T}\mathcal{T}^T \rangle$. It can be expanded into its canonical representation

$$C = E\Lambda E^T, \tag{3}$$

E being a matrix of eigenvectors (EOFs if C is effectively a sample covariance estimate), and Λ is a diagonal matrix of eigenvalues. We can use eigenvector patterns to rotate an estimated field

$$\mathcal{T} = E\alpha \tag{4}$$

so that $\alpha = E^T\mathcal{T}$ becomes a new unknown: a vector of projections of a target field on eigenvectors.

For simplicity let us consider the case of a completely observed system ($H = I$) with white uniform error ($R = rI$). OI solution for such a system has a closed form for each component of α:

$$\alpha_i = \frac{\lambda_i}{\lambda_i + r}\alpha_i^o \tag{5}$$

($i = 1 \cdots N$ is an index of components, eigenvalues and eigenvectors, $\alpha^o = E^T\mathcal{T}^o$ is a vector of projections of the observed field $\mathcal{T}^o$ on eigenvectors). We assume that eigenvalues are arranged in a descending order. The usual case then is that $\lambda_N \ll r \ll \lambda_1$. This means that $\alpha_1 \approx \alpha_1^o$, and $\alpha_N \approx 0$. In other words, the standard least squares procedure of OI in its search for the optimal solution will damp the observed values of all eigenvector amplitudes whose energy in the signal does not dominate over the observational error. Eigenvector modes which are expected to have energy much below the level of observational error will not be represented in the OI solution. In the case of global SST anomaly fields realistic observational error level of 0.5^oC (see Kaplan et al. 1998 for details on the data set and its error model) results in the reduction by the factor of 2 or more the variance in the modes beyond top 100 (Figure 2).

Consequently, computing the OI solution in all its details (projection to all EOFs) is superfluous: results as good can be achieved by computing only projections on some set of leading eigenvectors. It should be noted that for many physical variables the most energetic modes are those of the largest spatial scale. Details of the solution on small scales (projection to high number eigenvectors) is controlled by the fine details of the covariance matrix C which usually cannot be reliably estimated from the data. Large scale patterns of C (leading eigenvectors), however, can be estimated in a more reliable way. Approximation of C in (3) by only a few leading terms (truncation) results in infinite coefficients in the second term of the cost function (2) which totally disallow projection of the solution on truncated modes (in terms of the solution (5), if λ_i is assumed to be zero, then α_i will be zero as well). The same result, of course, can be achieved by truncating the eigenvector representation of the solution (4) to begin with. We call such a truncation a reduced space representation of the solution; inserting the truncated form of (4) into cost functions followed by their minimization with regards to the low-dimensional vector α allowed to develop the reduced space analogues of the OI, KF, and OS algorithms (Cane et al. (1996), Kaplan et al. (1997)). If certain assumptions are held, these solutions are in fact projections of the "complete" full-grid optimal solutions onto the low-dimensional reduced space.

Figure 3 emphasizes the contrast between the reduced space solutions and those obtained via more traditional kriging (or successive corrections) approach. Both types of a solution are based on the least-squares, the difference being in the approximation used for the baseline error covariance. The reduced space approach uses the most effective way of low-rank covariance approximation: via its leading eigenvectors in its canonical expansion (Golub and Van Loan (1996)). For most climatic fields this approximation will retain the part of the covariance with the longest spatial (and often temporal) scales, i.e. it corresponds to that part of the signal which we usually presume "climatic". The residual of this representation will have predominantly short decorrelation scales and in fact won't be an effective representation of the true climatic covariance in any matrix norm. Yet it is being used in the standard applications of kriging and successive correction techniques for the sole reason that such "localized" covariance structures is easy to model statistically.

While the reduced space solutions are formally suboptimal among full grid solutions, they are optimal among all reduced space solutions, being also far cheaper and much easier to feed by a priori error covariance information. For the settings which allow direct comparison, the solutions in the reduced space prove to be not inferior to the actually existing full grid solutions (Cane et al. (1996)). The reason for that is the poor representation (or inadequate parameterization) of small scales in full grid error covariance estimates. As a result, the full grid data analysis of small scales sometimes does more harm than good. Moreover, the analysis for those scales represents the major computational expense of the entire procedure. Hence the savings of reduced space analysis occur at the scales which are not really constrained by the data. Estimation on such scales is often meaningless, but the traditional schemes cannot selectively cut off computation there. The tunable nature of the dimension of a reduced space allows to put into the solution all scales down to the smallest resolved by available data, and the choice of leading EOFs for a basis guarantees to some extent the minimal dimension of the analysis space.

When the covariance of a climatic variable is dominated by a few large-scale modes, the generic objective analysis with correctly estimated covariance structures will predominantly reconstruct the patterns manifested in the large-scale climate dynamics. This is true for both full-grid and reduced space analyses, the latter being particularly effective in such settings. When this is not the case, the results of covariance estimation and of the full-grid objective analysis applied to the sparsely observed data are likely to be less robust and more error-prone, space reduction not being effective either.

2.2 Applications

We applied the reduced space OS to produce the near-global analysis of $5^o \times 5^o$ SST monthly anomaly grids for the period 1856–1991 (Kaplan et al. (1998)). For a model of time transitions we used an empirically fitted first order autoregressive model which was assumed to be diagonal in the reduced space coordinates. Observational data used in this work is known as MOHSST5 compilation of ship observations produced by the U.K. Meteorological Office (Bottomley et al. (1990), Parker et al. (1994)). Covariance of the SST field was derived from the period 1951-1991, then its leading 80 EOFs were used for the optimal estimation in the entire time range from 1856 to 1991.

Extensive tests proved the analysis to be robust and self-consistent. As Figure 4 illustrates, even under the sparse spatial coverage of December 1877 (known to be a strong warm ENSO event), the analysis produces a believable structure for a very strong El Niño known to have occured that year (panels a, b). We verified the credibility of that reconstruction by taking data for December 1986 (a well sampled month, panels c, d), sampled them per the 1877 sampling pattern, and corrupted them by noise (to reflect the increase in the error at each grid box due to less frequent sampling). The OS analysis produced the 1986 El Niño pattern with only slightly weaker amplitude than that obtained with the full data (panels e, f). As expected, the magnitude of the large-scale estimated error is much larger for the reconstruction from the December 1986 reduced quality simulation, than for the reconstruction from the complete data (panels g, h). Further tests show that our reconstructions are very similar to the Reynolds and Smith (1994) NCEP OI estimates of December 1986 SST anomaly (the NCEP OI combines in situ and AVHRR satellite data), though the latter is richer in small-scale details, particularly when presented in its full $1^o \times 1^o$ resolution (panels i, j).

To test the analysis for a period not used in estimating the covariance structures, we carried out additional experiments as follows: the Reynolds and Smith (1994) NCEP OI SST anomaly fields for 1992-1996 were chosen as the "true" solution. These "true" data were resampled and corrupted by noise according to the data availability and our estimates of observational error for the 1916-1920 period (Figure 5). The average rms error for available observations is 0.74^oC, and there are many locations where the SST is not observed at all (panel a). The analysis of the simulated data differs from the NCEP OI fields by 0.48^oC on average (panel b). However, the major part of this difference is in the error of truncation: the variance of NCEP OI fields which cannot be represented by the 80 EOFs used in our reconstruction (cf. Fig. 6f from Kaplan et al. (1998)). Projecting the NCEP OI fields on the linear subspace defined by the 80 EOFs from our analysis provides the "reduced space version" of the true SST field (and incidentally allows for a statistically homogeneous extension of reduced space historical analyses by a higher quality modern period data sets: extension of our OS by the reduced space projection of the NCEP OI is now publicly accessible, see Acknowledgments). Our simulated analysis differs on average by 0.31^oC from this reduced space version of truth (panel c), which is in a good agreement with the average theoretical error estimate, 0.28^oC (panel d). The years 1992-1996 are outside the period used in constructing the covariance estimate and are marked by strikingly different behavior. Thus these experiments demonstrate that even with limited data the reduced space OS is able to reconstruct the global SST in a period when the covariance structure is somewhat different from the one used by the analysis procedure.

We also applied the reduced space OI analysis to the SLP data of Comprehensive Ocean–Atmosphere Data Set (COADS, release 1 extended by standard release 1a; Woodruff et al. (1987, 1993)) to produce $4^o \times 4^o$ fields of SLP monthly anomaly for the period 1854–1992 (Kaplan et al. 2000). Note that both our SST and SLP analyses utilize only ship observations presented in the form of monthly "superobservations" (Smith et al. (1996)) – mean values for $5^o \times 5^o$ (MOHSST5) or $2^o \times 2^o$ (COADS) bins. The U.K. Met Office applies so called "winsorization" (Bottomley et al. (1990)) to the content of their bins which makes the bin average more similar to a median. The COADS maintains variety of statistical characteristics of the bin contents in their "monthly summaries": in addition to the mean they provide number of observations, their standard deviation, median, sextiles, etc. Pre-war SST data of the U.K. Met Office has Folland and Parker (1995) "bucket corrections" applied to it.

Figure 6 shows the monthly values of the analyzed NINO3 (mean SST for the eastern equatorial Pacific 5^oS–5^oN, 150^o–90^oW), a familiar El Niño – Southern Oscillation index, with 3σ error bars supplied by the analysis. Obviously, the analysis eliminates a great deal of noise present in direct NINO3 estimates from the observed data, and agrees well with the Quinn (1992) list of El Niño events, which is based on a variety of land-based, historical factors known to be associated with El Niño. The summary comparison of annual mean NINO3 with Quinn's data is shown in Figure 7. The relation is strong but not perfect: six El Niño events, rated as "moderate" or weaker by Quinn have in fact negative (as large as -1^{o}C for 1874) annual NINO3 from our analysis. The latest of them has happened in 1943, others in the 19th century. However, the analysis of the Southern Oscillation (SO) and associated coastal phenomena for the period 1926-1986 by Deser and Wallace (1987) suggests that the coastal SST index might show stronger connection to Quinn's index of El Niño events. For this purpose we created a coastal SST index by averaging the results of the OS analysis over the NE triangular half of [15^oS–0^oN, 90^o–75^oW] square (the diagonal included). Indeed, the comparison presented in Figure 7 qualitatively supports this suggestions: only 2 of events (1871 and 1907) have a negative coastal SST value. The ability of the analysis to distinguish between geographically close but functionally different area averages (like NINO3 and the coastal average) is encouraging, because the small-scale differences between close areas could be lost in the sampling and observational noise which our technique is filtering out in the attempt to reconstruct the large-scale dominant structures.

Kaplan et al. (2000) present the comparison of SLP time series measured at a few island or coastal stations (Darwin, Tahiti, Reykjavik, and Gibraltar) affected by large-scale atmospheric phenomena (SO and North Atlantic Oscillations) with their "marine-based proxies" – averages of analyzed COADS SLP over a few analysis grids surrounding a station. The proxies compared favorably with the land-based measurements, despite being produced from the much inferior quality ship report data. Significant portion of this success should be attributed to the fact that the major part of the SLP signal on the stations we considered comes from the large-scale atmospheric oscillations which are being predominantly reconstructed by the global analysis of marine data.

Figure 8 compares correlation coefficients between Darwin and Tahiti station data (Können et al (1998), Ropelewski and Jones (1987)) with the same for these station's marine-based proxies (Kaplan et al. (2000)). Both coefficients are computed in different width time windows and presented as functions of time. The correlation coefficients are close for the modern period, but the land station values are lower during earlier times. We suggest that the correlation between the land-based data weakens for the early part of the record because of degraded data quality. Factors like instrument defects and replacements, changes in observational times and location can create systematic problems in early fragments of station records; some of such problems for Darwin and Tahiti records are documented, and corrections for them are customarily applied (Ropelewski and Jones (1987), Allan et al. (1991)). Most likely, however, there are uncorrected biases still left in these records, particularly in the one for Tahiti (Kaplan et al. 2000).

On the other hand, the sparser and more erratic marine data forces the analysis to reproduce less of smaller scale (and thus more error-prone) phenomena, and to leave mostly the large-scale SO-associated pressure changes in the reconstruction. That strengthens the correlation between the analysis proxies for Darwin and Tahiti which are located near antinodes of the SO. Note that this correlation increase happens as the response of the analysis procedure on systematic decrease in the quality of marine data, despite the underlying assumption of constant covariance for an estimated field.

In fact, correlation between Darwin and Tahiti SLP records has traditionally been interpreted (Trenberth (1984)) as an indicator of the signal-to-noise ratio when these station records are used as the indices of the SO (in this case the "signal" is the SO, everything else is the "noise"). Note that the most of the weakening episodes in 6-year window correlations exhibited by the land stations in the early part of the record are mimicked by the marine proxy correlations. Those episodes are most likely the realistic changes in the strength of SO relative to the background atmospheric noise. Those which are present only in the land records might be either spurious or missed in the marine records because of the sparsity of COADS coverage at those particular times. The level of certainty of the latter possibility may change significantly when the SLP from the "Dutch" deck, a major COADS component prior to World War II, is included in the monthly summaries in further COADS releases (Woodruff et al. (1998)). Even at the present level of coverage the indices based on ship observations may provide cleaner indication of the large-scale phenomena than the local land-based records.

3. Difficulties and ways to resolve them

3.1 Specification of observational uncertainty

Advantages of the reduced space optimal analysis do not come for free: they are based on our knowledge of a priori estimates, namely covariances of observational error R and of the first two statistical moments of the solution: its mean field $\mathcal{T}_m$ and its covariance C. All these necessary values only approximately can be computed from the observations.

In computing R (which allows the analysis to distinguish between poor and high quality superobservations) we use intrabox variability and number of observations for the superobservational bins. When we analyze the U.K. Met Office SST data, we have to estimate their intrabox standard deviations from the COADS monthly summaries because the U.K. Met Office does not maintain any intrabox statistics but winsorized means in its official data format. Our estimates of observational error are far from being perfect. Figure 9 shows the map of our estimated single ship observational error (values used in the analysis by Kaplan et al. (1998)). These errors are standard deviations of individual measurements taken during one month within a given $5^o \times 5^o$ box. Such deviations account for both instrumental and sampling error (for a single measurement the latter is equal to the natural variability of SST in the given space-time box). Note that these deviations from mean values are computed for monthly bins, so they do not reflect month-to-month or longer climate variability. These values can be easily computed for larger bins, if mean and standard deviation statistics are available for their parts (Kaplan et al. 2000, p. 2989).

Comparison of Figure 9 with the map of random error estimates by Kent et al. (1999, their Figure 3d) brings uneven conclusions. The latter map does not include any kind of sampling error. This explains much larger values of Figure 9 in the regions of Gulf Stream and Kuroshio Current. However, outside of these areas the map of Figure 9 should be giving larger values as well. Not everywhere this seems to be the case: insufficient density of observations does not allow for an adequate sampling of the SST natural variability in many areas of the world ocean. For the SLP the contribution of sampling variability into our estimates of a single ship error (not shown) is so large, that our COADS-based estimates (used by Kaplan et al. (2000)) exceed those of Kent et al. (1999, their Figure 3b) by the factor of 3 in the midlatitudes and marginally in the tropics.

Clearly a lot more work should be done in this direction till really reliable observational error estimates will enter gridded analyses of climate variables. Important step on this way would be to bring to the attention of all data centers the necessity to include the statistics of intrabox distributions into their standard data formats, rather than just provide box mean values. This seems to be particularly crucial in the planned blending project of the COADS and UKMO data banks (Woodruff et al. (1998)). Comparison of ship based estimates like that of Figure 9 with those obtained from satellite data suggests that the ship-based estimates are affected by the sampling error even for the periods of the best coverage. Hence the satellite data has to be used to supplement the ship-based estimates of the small-scale variability.

3.2 Characterization of the signal

The problems with the reliable estimation of $\mathcal{T}_m$ and C are even more fundamental. Ideally, these statistical characteristics of the signal are supposed to be applicable to the entire period of the analysis. In fact, poor data quality and sparse coverage in the early part of the record forces us to use only the modern data period for the derivation of $\mathcal{T}_m$ and C. In the applications described above we used climatological means for the period 1951-1980 and estimated the covariance for the period from 1950 to the beginning of 1990s. Analysis is then done using these values for as far back as the middle of the 19th century.

Problems with the mean. It is been observed by Hurrell and Trenberth (1999) that a linear trend for the 20th century computed from our SST analysis shows somewhat less of the warming than other estimates. They suggested that it happens due to the "stationarity" assumption: the hypothesis that the modern-period mean and covariance are applicable for the entire record. If in fact the long-term variability of SST (e.g. trend) resulted in a much different mean SST state for the first half of the century, and the pattern of this change is not well-represented by the modern-period covariance, the analysis might underestimate this change.

At present we are addressing this issue via the analysis of data residuals, the difference between the observed data and our analysis. These residuals presumably consist of two major components: observational and sampling error and part of long-term variability unresolved by the analysis. Because of very different characteristics of these components, it should be relatively easy to isolate the latter. Prospective methods of isolation include application of the reduced space OI and OS technique to the residuals and covariance reestimation (Kaplan et al. 1997, 2000) and polynomial spline smoothing of the residuals (Wahba 1990). Note that bestfitting a straight line or other slowly changing function of time to the residuals can be brought into the prospect of optimal estimation and provide error bars for the trend estimates, because all other variability in the residuals is expected to be temporally uncorrelated errors. The same approach does not work for fitting a slowly changing functions of time to the actual temperature changes, as the latter contains complete spectrum of temporally correlated variability, from secular to intermonthly. If those are not removed, one should not assume the "whiteness" (mutual statistical independence) of errors, for such an assumption will result in unrealistically low theoretical estimates for the uncertainty of the fit. Figure 10 presents a cartoon emphasizing the advantage of removal of the interannual climate variability from the observed data prior to estimating long-term changes.

Our preliminary analysis indeed shows some long-term variability left in the residuals. Once we are done with its complete estimation, we will be able to estimate the total long-term variability in the SST record and measure its uncertainty. Then we will either separate it from the raw observations before applying the analysis procedure, or make sure that it is properly represented in the covariance structure.

Problems with the covariance: stationarity. Assumption of covariance stationarity and possibility of its negative consequences comes up quite often in discussions, but has not been, to the best of our knowledge, systematically researched. As the first attempt at it, we compared covariance matrices estimated in different 40 year time windows for our SST analysis. Since the current SST analysis was performed under a conservative assumption of stationary covariance, this comparison probably underestimates the actual covariance variability. As a measure of distance between two covariance matrices we use the Frobenius norm: the square root of the sum of squares of all elements in the matrix difference (Golub and Van Loan (1996)). Even the norm of covariance matrix itself seem to change over time dramatically, with the minimum in 1930-1970, and the maximum in 1890-1920 (Figure 11, top). The diagram of normalized distances (norm of a difference divided by the norm of the covariance matrix computed for the recent 40 years) suggests that every period in the last 1.5 centuries was in some sense unique: the farther from each other middles of a sample periods get, the larger is the normalized distance between matrices (Figure 11, bottom). In some periods (like the one centered on 1930) this change happens very fast, in others (1910 and 1950) more slowly.

Successful validation of many aspects of our SST analysis so far shows that the exhibited instability of the covariance matrix does not render the analysis completely wrong or useless: the inherent robustness of the least squares estimates can absorb some level of inadequacy of a priori estimates. Moreover, all the different covariance matrices compared in the Figure 11 were produced by the analysis of Kaplan et al. (1998) under the assumption that the covariance of the SST field is constant and equal to the sample covariance of 1951–1991. It seems reasonable, however, to involve data of all time periods in the computation of the covariance and either use the estimate which would be applicable to the entire analysis period, or account for slow changes in time of the covariance structure in our analysis methodology.

Problems with the covariance: resolution and coverage. Significant volatility of the covariance structure discourages uses of only the modern period of particularly good (helped by satellite coverage) data for covariance estimation. If we are determined to estimate the large-scale covariance structures from the period no shorter than a few decades, that imposes certain restrictions on the spatial resolution with which covariance can be estimated. Before analyzing COADS SLP data we tried to estimate covariance for $2^o \times 2^o$ spatial bins, and found that the analysis domain had large holes (shown in black color in Figure 12) in the tropical

Pacific. It took averaging to $4^o \times 4^o$ grid to "close" these holes. As a result the analysis domain we obtain has quite coarse resolution and still is globally incomplete. This severely limits the usage of such analyses in the climate model studies. It seems important to be able to generalize the technique of the reduced space optimal estimation to the stage at which it can produce high resolution and globally complete analyses.

In fact, the reduced space reconstruction technique can be empowered by the multivariate approach. Principal modification of the reduced space optimal analysis that can produce high- resolution globally complete fields is to separate an estimated field into a few terms which correspond to different scales of resolution (and thus variability). Different terms can be observed through different sources. For example, the most of the ocean $5^o \times 5^o$ resolution term is well observed by ships during last 50 years, and $1^o \times 1^o$ variability within $5^o \times 5^o$ boxes plus all variability in the Southern ocean can be estimated from the NCEP OI (Reynolds and Smith 1994) for the last 15 years, etc. The set of all terms can be subjected to multivariate EOF analysis, each piece being a separate variable in this analysis. These multivariate EOFs then used for the reconstruction of all pieces together, and thus for the entire high-resolution globally-complete field. This approach has certain "modular" nature because of which it allows to push further in both directions: very large scale variability can be estimated for very long periods from the paleodata, extending the analysis to very long periods, and certain areas of high gradients and/or good observational networks can be "refined" by adding special high resolution "patches."

Problems with the covariance: representing small scales. Seemingly fruitful direction for producing high-resolution objective analyses is to literally combine analyses represented by left and right parts of Figure 3. Note that the exact solution for the full grid OI can be separated into two parts:

$$\mathcal{T} = (H^T R^{-1} H + C^{-1})^{-1} H^T R^{-1} \mathcal{T}^o = C H^T (R + H C H^T)^{-1} \mathcal{T}^o =$$

$$= E \Lambda E^T H^T (H E \Lambda E^T H^T + H E' \Lambda' E'^T H^T + R)^{-1} \mathcal{T}^o + E' \Lambda' E'^T H^T (H E \Lambda E^T H^T + H E' \Lambda' E'^T H^T + R)^{-1} \mathcal{T}^o =$$

$$= E\alpha + C' H^T (H C' H^T + R)^{-1} \Delta\mathcal{T}^o = E\alpha + \Delta\mathcal{T}$$

The first term $E\alpha$ here is our standard reduced space OI solution. The second part represents a correction to it towards the complete (exact) solution. This correction is defined by the covariance piece C' and contributes predominantly to the small-scale variability. It is easy to check that $\Delta\mathcal{T}$ is a formal OI solution to the estimation problem

$$\Delta\mathcal{T}^o = H\Delta\mathcal{T} + \varepsilon^o, \qquad \langle \Delta\mathcal{T}\Delta\mathcal{T}^T \rangle = C', \qquad \langle \varepsilon^o \varepsilon^{o\,T} \rangle = R$$

where $\Delta\mathcal{T}^o = \mathcal{T}^o - HE\alpha$ is an observational residual to the reduced space OI solution.

We do not expect to be able to estimate C' from the data without any special assumptions. However, this part of covariance can be modeled statistically under certain assumptions of spatial stationarity, e.g. as a function of spatial lag, in the style of traditional kriging or successive correction approach. Thus these traditional techniques can be successfully used for complementing the reduced space solution with small-scale corrections.

Problems with the covariance: consistent estimation and uncertainty. When an a priori estimate of the signal covariance is correct, the statistics of the solution should be consistent with it, i.e. certain balance equations should be satisfied. If it is found not to be the case, a priori values can be reestimated to satisfy the balance, and then the analysis solution can be recalculated. These steps can be repeated iteratively until the solution satisfies the balance. However, use of different balance formulations might result in somewhat different solutions.

Kaplan et al. (1997) introduced the balance in the form of the system of equations

$$A_p \stackrel{\text{def}}{=} \langle \alpha^p \alpha^{p\,T} \rangle = \Lambda + \mathcal{P}^p,$$

$$A_{\text{OI}} \stackrel{\text{def}}{=} \langle \alpha^{\text{OI}} \alpha^{\text{OI}\,T} \rangle = \Lambda(\Lambda + \mathcal{P}^p)^{-1}\Lambda.$$

which ties together covariances of the projection and reduced space OI solutions (α^p and α^{OI} respectively), error covariance for the projection solution $\mathcal{P}^p$, and the reduced space representation of the covariance. The projection solution α^p consists of the best fit coefficients of EOF patterns to the observed data. $\mathcal{P}^p$ is the theoretical covariance of the error in these coefficients. Originally they were using one-parametric heuristic formula for "redistributing" the spectrum of Λ. This seemed to give satisfactory results for SST analyses, but failed in application to the SLP analysis by Kaplan et al. (2000). Because of that, the latter work reduced the system to a single nonlinear matrix equation for Λ

$$A_p = \Lambda A_{\mathrm{OI}}^{-1} \Lambda$$

and presented an exact solution to it. The results of the analysis satisfied the balance after the first iteration.

An alternative way to state the analysis balance can be based on the Expectation Maximization (EM) procedure (Schneider 2000 and references therein). In the reduced space version and taking into account the observational error, the EM balance for the OI solution can be written as

$$\overline{\alpha} = \langle \alpha^{\mathrm{OI}} \rangle, \qquad \Lambda = \langle (\alpha^{\mathrm{OI}} - \overline{\alpha})(\alpha^{\mathrm{OI}} - \overline{\alpha})^T \rangle + \mathcal{P}^{\mathrm{OI}}.$$

Our initial trials of this procedure for the SST analysis have shown convergence after approximately 10 iterations.

It should be noted that because of their reduced space nature the procedures described above cannot bring the estimates of the leading EOFs outside of the initially defined reduced space. However, if the small-scale correction is added after every iteration, and the full-grid covariance is reestimated, that might result in substantially better estimates of the signal covariance and perhaps overcome the limitation of gappy and erratic data from which it is derived.

On the other hand, however complicated technique we use, the covariance is always estimated with some uncertainty. Explicit modeling of this uncertainty, transferring it into the uncertainty of the analyzed fields, perhaps in the Bayesian framework, is an important task for the future.

4. Conclusions and prospects

We have shown that the reduced space optimal estimation is a computationally effective restructuring of the process of obtaining the full-grid optimal solution, and that it delivered verifiable analyses of climatic fields in both systematic applications to date (for SST and SLP).

The problems of the method are the same as those of any objective analysis technique:

difficulty to derive reliable a priori estimates from the sparse and erratic data. Partially these problems might get relieved if new significant volumes of the data for the early periods become available (Woodruff et al. (1999)). It is very important that all data centers involved provide extensive statistics of intrabin distributions (as opposed to providing means only), like e.g. current COADS model of monthly summaries. Use of satellite data is another prospective way of improving a priori estimates of in situ error statistics.

Land station data is another powerful information resource which can be combined in the analyses with marine observations to the advantage of the product (cf. recent SLP analysis of the U.K. Met Office by Basnett and Parker (1997))

Further improvement of the analysis technique should include systematic a priori estimation of mean, covariance, or long-term variability and changing covariance structure from the entire period of available data. Separation of the estimated fields into large and small scale varying components allows for the generalization of the technique which can produce high resolution globally complete products.

The technique of reduced space optimal estimation should be used more systematically in application to all climate variables for which historical (COADS) data sets are available, e.g. meridional and zonal winds, marine air temperature, humidity, or (non-COADS) precipitation, sea ice concentration, and possibly sea surface height. It also opens interesting prospects for historical analyses of ocean-atmosphere fluxes with possible modification of applying the analysis to the system of a few physical variables (e.g. surface wind components and SLP) and using a linearized physical model (e.g. geostrophic or frictional balance) as an additional analysis constraint.

5. Acknowledgments

This work was supported by NOAA grant UCSIO-10775411D/NA47GPO-188, NOAA/NASA Enhanced Data Set Project grant NA06GP0567, and NSF/NOAA Earth System History grant NA86GP0437. AK is thankful to the NOAA OGP for the CLIMAR99 travel grant (administered through the UCAR) that served as an encouragement and facilitated this work. Discussions with the CLIMAR99 participants Scott Woodruff, Dick Reynolds, David Parker, Liz Kent, and Henry Diaz are gratefully acknowledged. The hospitality of the

NCAR Geostatistics Project during AK's July 2000 visit (supported by the NSF grant DMS-9815344), their workshop on "Statistics for Large Data Sets" and stimulating discussions with Doug Nychka, Dave Higdon, Craig Johns, and Chris Wikle helped to bring this work into its final shape. Discussions with Misha Chechelnitsky and his technical help with Figure 1 were invaluable. Remarks of both anonymous reviewers helped to improve the clarity of the manuscript. Benno Blumenthal's Ingrid software is responsible for all ocean domain plots in this work, and his Data Library system uses Senya Basin's CUF format for providing easy public access to the analyzed data presented in this work at http://ingrid.ldgo.columbia.edu/SOURCES/.KAPLAN/.RSA_MOHSST5.html for the SST analysis, http://ingrid.ldgo.columbia.edu/SOURCES/.KAPLAN/.RSA_COADS_SLP1.html for SLP analysis, and http://ingrid.ldeo.columbia.edu/SOURCES/.KAPLAN/.EXTENDED/ for the OS SST analysis monthly extensions to the present (this is achieved by concatenating the NCEP OI projections onto the analysis' reduced space). This work is a Lamont–Doherty Earth Observatory contribution number 6149.

References

Allan, R.J., Nicholls, N., Jones, P.D. and Butterworth, I.J., 1991: A further extension of the Tahiti-Darwin SOI, early SOI results and Darwin pressure. J. Climate 4, 743-749.

Basnett, T.A. and D.E. Parker, 1997: Development of the global mean sea level pressure data set GMSLP2. Hadley Centre of the UK Met Office. Clim. Res. Tech. Note 79. Unpublished document available from the Hadley Centre for Climate Prediction and Research, Meteorological Office, London Road, Bracknell, RS12 2SY, U.K.

Bottomley, M., C.K. Folland, J. Hsiung, R.E. Newell, D.E. Parker, 1990: *Global Ocean Surface Temperature Atlas*. HMSO, London

Cane, M.A., A. Kaplan, R.N. Miller, B. Tang, E.C. Hackert, and A.J. Busalacchi, 1996: Mapping tropical Pacific sea level: data assimilation via a reduced state space Kalman filter. *J. Geophys. Res.*, **101**, 22599–22617.

Cressie, N. A. C., 1991: *Statistics for Spatial Data*, Wiley-Interscience.

Deser, C. and J.M. Wallace, 1987: El Niño events and their relation to Southern Oscillation: 1925-1986, *J. Geoph. Res.*, **92**, 14,189–14,196.

Daley, R., 1993: *Atmospheric Data Analysis*, Cambridge University Press, 457pp.

Da Silva, A.M., C.C. Young, and S. Levitus, 1994: *Atlas of Surface Marine Data.* Vol. 1, *Algorithms and Procedures*, NOAA, 83 pp. [Available from U.S. Department of Commerce, National Oceanographic Data Center, User Services Branch, NOAA/NESDIS E/OC21, Washington, DC 20233.]

Folland, C.K. and D.E. Parker, 1995: Correction of instrumental biases in historical sea surface temperature data, *Q. J. R. Meteorol. Soc.*, *121*, 319-367.

Golub, G.H. and C.F. Van Loan, 1996: *Matrix Computations.* Third edition, The John Hopkins University Press, Baltimore. 694pp.

Hurrell, J.W., and K.E. Trenberth, 1999: Global sea surface temperature analyses: multiple problems and their implications for climate analysis, modeling, and reanalysis. *Bull. Amer. Meteor. Soc.*, **80**, 2661-2678.

Kaplan, A., Y. Kushnir, M.A. Cane, and M.B. Blumenthal, 1997: Reduced Space Optimal Analysis for Historical Datasets: 136 years of Atlantic Sea Surface Temperatures. *J. Geoph. Res.*, **102**, 27,835–27,860.

Kaplan, A., M. Cane, Y. Kushnir, A. Clement, M. Blumenthal, and B. Rajagopalan, 1998: Analyses of global sea surface temperature 1856-1991, *J. Geoph. Res.*, **103**, 18,567-18,589.

Kaplan, A., Y. Kushnir, and M.A. Cane, 2000: Reduced space optimal interpolation of historical marine sea level pressure: 1854–1992. *J. Climate*, **13**, 2987-3002.

Kent, E.C., P. Challenor, and P. Taylor, 1999: A statistical determination of the random observational errors present in voluntary observing ships meteorological reports. *Journal of Atmospheric and Oceanic Technology*, **16**, 905-914.

Levitus, S., and T.P. Boyer, 1994: *World Ocean Atlas 1994.* Vol. 4, *Temperature.* NOAA Atlas NESDIS 4, U.S. Department of Commerce, Washington, DC, 117pp. [Available from U.S. Department of Commerce, National Oceanographic Data Center, User Services Branch, NOAA/NESDIS E/OC21, Washington, DC 20233.]

Können, G.P., Jones, P.D., Kaltofen, M.H. and Allan, R.J., 1998: Pre-1866 extensions of the Southern Oscillation Index using early Indonesian and Tahitian meteorological readings. *J. Climate*, **11**, 2325-2339.

Mann, M.E., R.S. Bradley, M.K. Hughes, 1998: Global-scale temperature patterns and climate forcing over the past six centuries. *Nature*, **392**, 779-787.

Mardia, K.V., J.T. Kent, and J.M. Bibby, 1979: *Multivariate Analysis*, Academic, San Diego, Calif, 521 pp.

Parker, D.E., P.D. Jones, C.K. Folland, and A. Bevan, 1994: Interdecadal changes of surface temperature since the late nineteenth century. *J. Geophys. Res.*, **99**, 14,373–14,399.

Parker, D.E., C.K. Folland, and M. Jackson, 1995: Marine surface temperature: observed variations and data requirements, *Clim. Change*, *31*, 559-600.

Quinn, W.H., 1992: A study of Southern Oscillation–related climatic activity for A.D. 622–1900 incorporating Nile River flood data, in *El Niño Historical and Paleoclimatic Aspects of the Southern Oscillation*, edited by H.F. Diaz and V. Markgraf, pp. 119-149, Cambridge Univ. Press, New York.

Rao, C.R., 1973: *Linear Statistical Inference and Its Applications*, John Wiley, New York, 625 pp.

Rayner, N.A., E.B. Horton, D.E. Parker, C.K. Folland, and R.B. Hackett, 1996: Version 2.2 of the global sea-ice and sea surface temperature data set, 1903-1994. Clim. Res. Tech. Note 74. Unpublished document available from the Hadley Centre for Climate Prediction and Research, Meteorological Office, London Road, Bracknell, RS12 2SY, U.K.

Reynolds, R.W., and T.M. Smith, 1994: Improved global sea surface temperature analysis using optimum interpolation. *J. Climate*, **7**, 929-948.

Ropelewski, C.F. and Jones, P.D., 1987: An extension of the Tahiti-Darwin Southern Oscillation Index. *Mon. Weather Rev.*, **115**, 2161-2165.

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

Smith, T. M., R. W. Reynolds, R. E. Livezey, and D.C. Stokes, 1996: Reconstruction of historical sea surface temperatures using empirical orthogonal functions. *J. Climate*, **9**, 1403-1420.

Schneider, T., 2001: Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *J. Climate*, in press

Trenberth, K. E., 1984: Signal versus noise in the Southern Oscillation. *Mon. Wea. Rev.*, **112**, 326–332.

Wahba, G., 1990: *Spline Models for Observational Data.* CBMS-NSF, Regional Conference Series in Applied Mathematics, Vol. 59, Society of Industrial and Applied Mathematics, 169 pp.

Woodruff, S.D., R.J. Slutz, R.L. Jenne, and P.M. Steurer, 1987: A comprehensive ocean–atmosphere data set. *Bull. Amer. Meteor. Soc.*, **68**, 521-527.

Woodruff, S.D., S.J. Lubker, K. Wolter, S.J. Worley, and J.D. Elms, 1993: Comprehensive Ocean-Atmosphere Data Set (COADS) Release 1a : 1980-92. *Earth System Monitor*, **4**, No. 1, 1-8.

Woodruff, S.D., H.F. Diaz, J.D. Elms, and S.J. Worley, 1998: COADS Release 2 data and metadata enhancements for improvements of marine surface flux fields. *Phys. Chem. Earth*, **23**, 517-526.

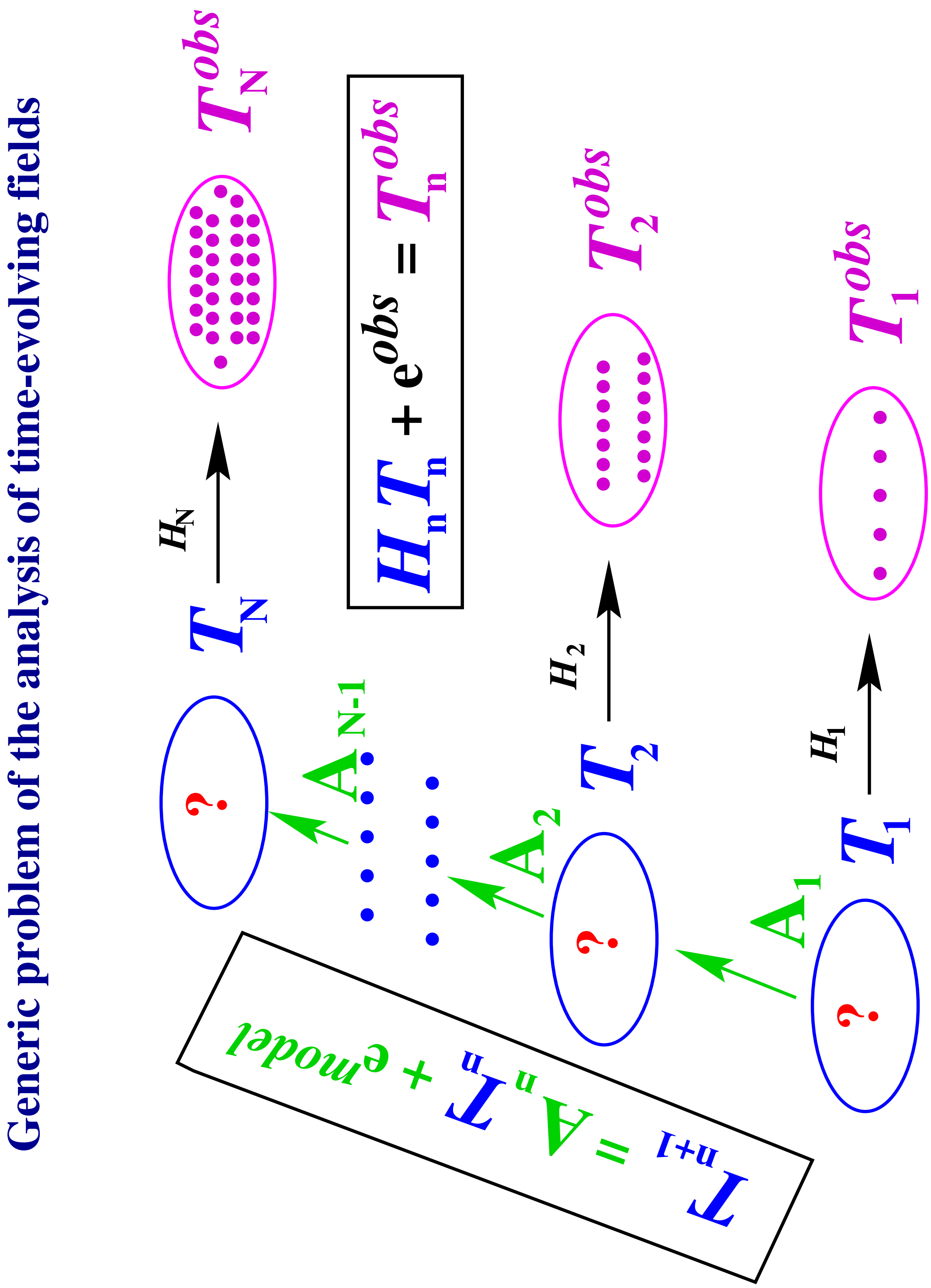

Figure 1: Generic scheme of the informational content for the analysis of time-evolving fields.

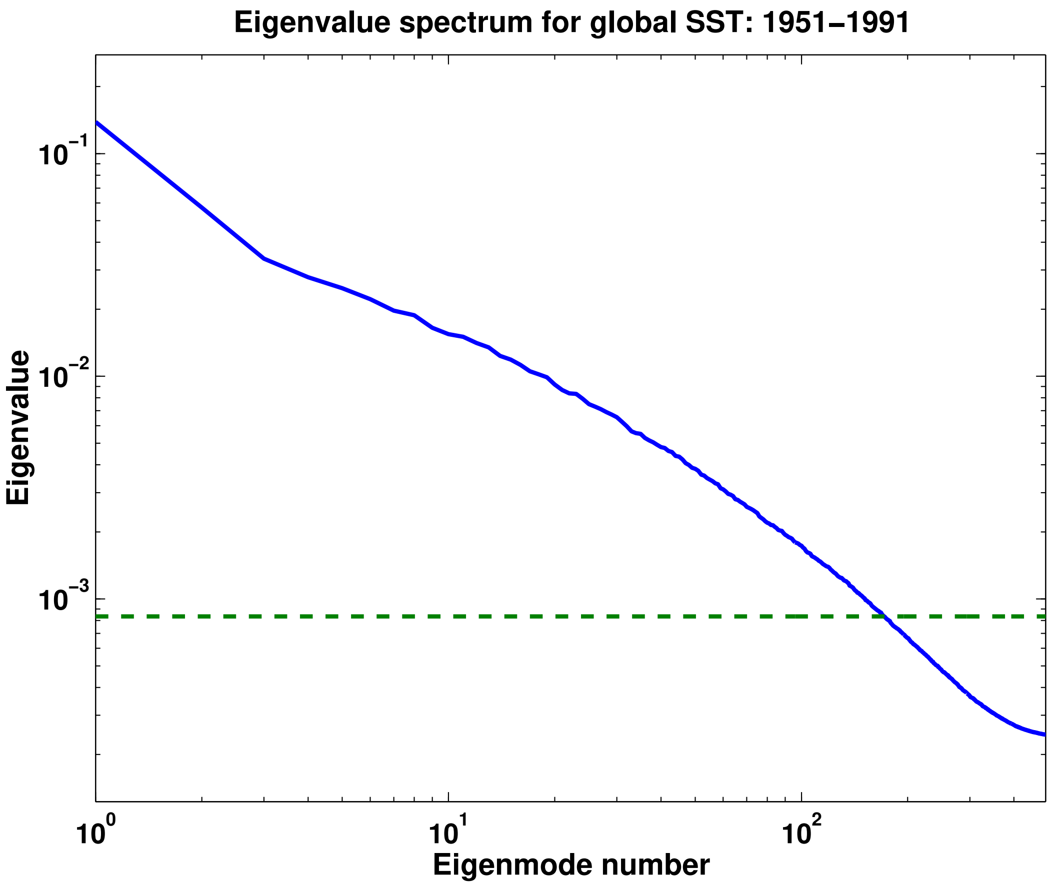

Figure 2: Normalized eigenvalue spectrum of global SST fields. Dashed line corresponds to the magnitude of the characteristic observational error in the historical data (standard deviation of 0.5^{o}C): variance in the eigenmodes with the eigenvalues below this level are reduced by the factor of 2 or more in the least-squares optimal estimates.

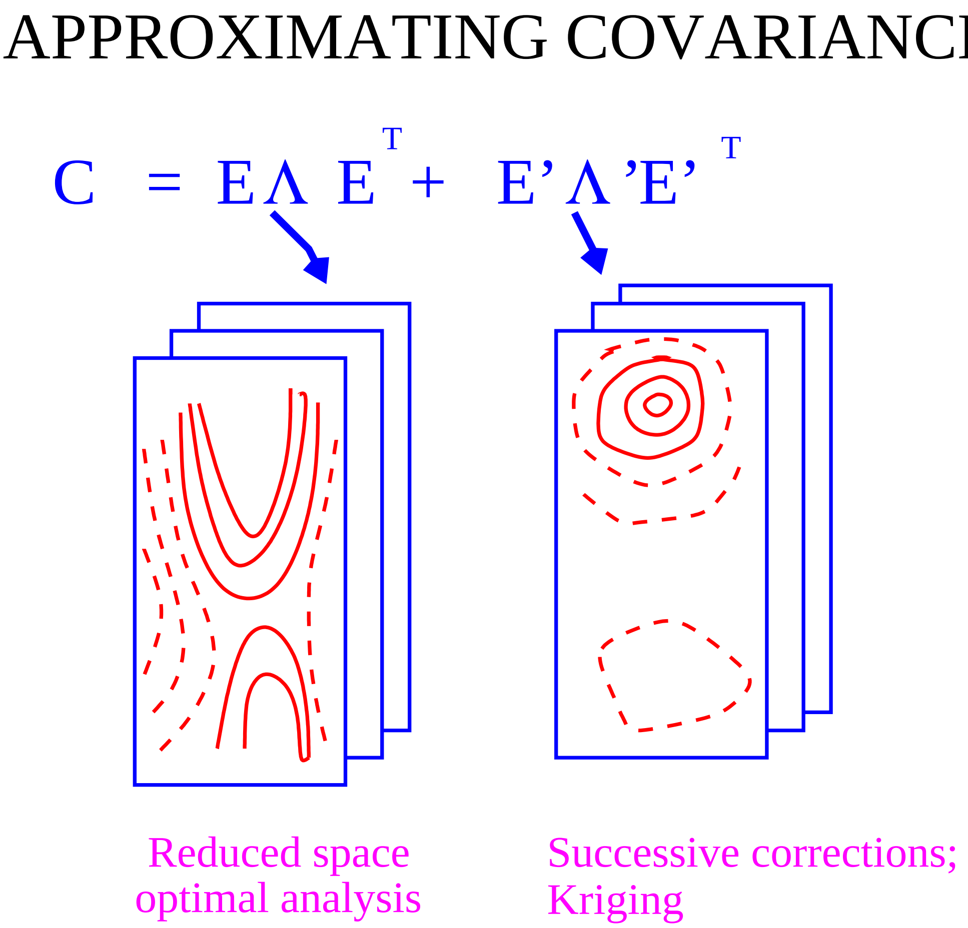

Figure 3: Separation of covariance into large- and small-scale portions in various optimal analysis techniques.

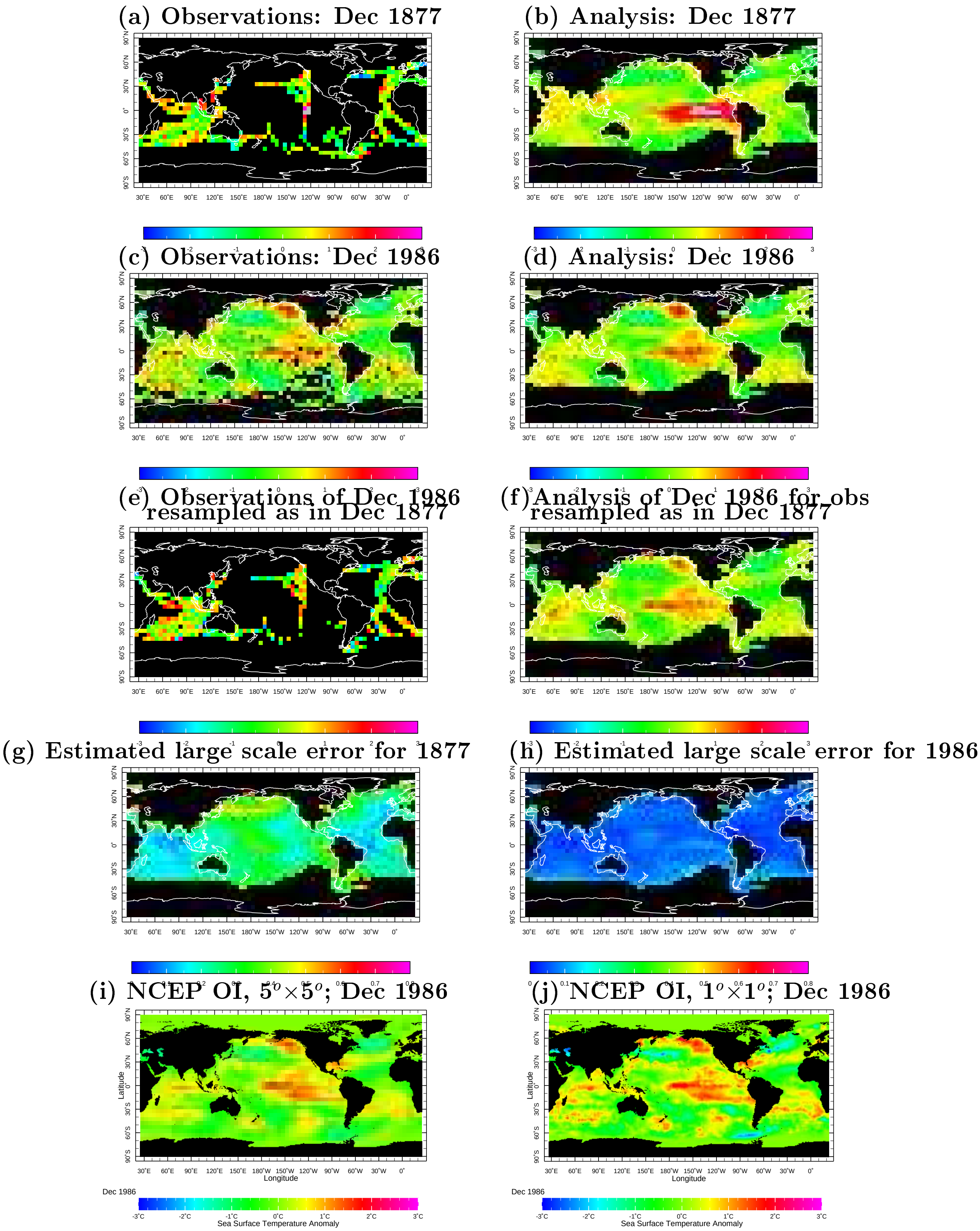

Figure 4: Available SST observations and their reduced space OS analysis for December 1877 (panels (a) and (b)) with verification through the experiment with 1986 data: simulated OS analysis for December 1986 using the data distribution of 1877 (panels (e) and (f)) versus the standard OS analysis for December 1986 with all available data (panels (c) and (d)). Also shown are large scale errors in the two reconstructions (panels (e) and (f)) and the NCEP OI December 1986 field presented in (i) $5^o \times 5^o$ and (j) $1^o \times 1^o$ resolution. Units are ^oC.

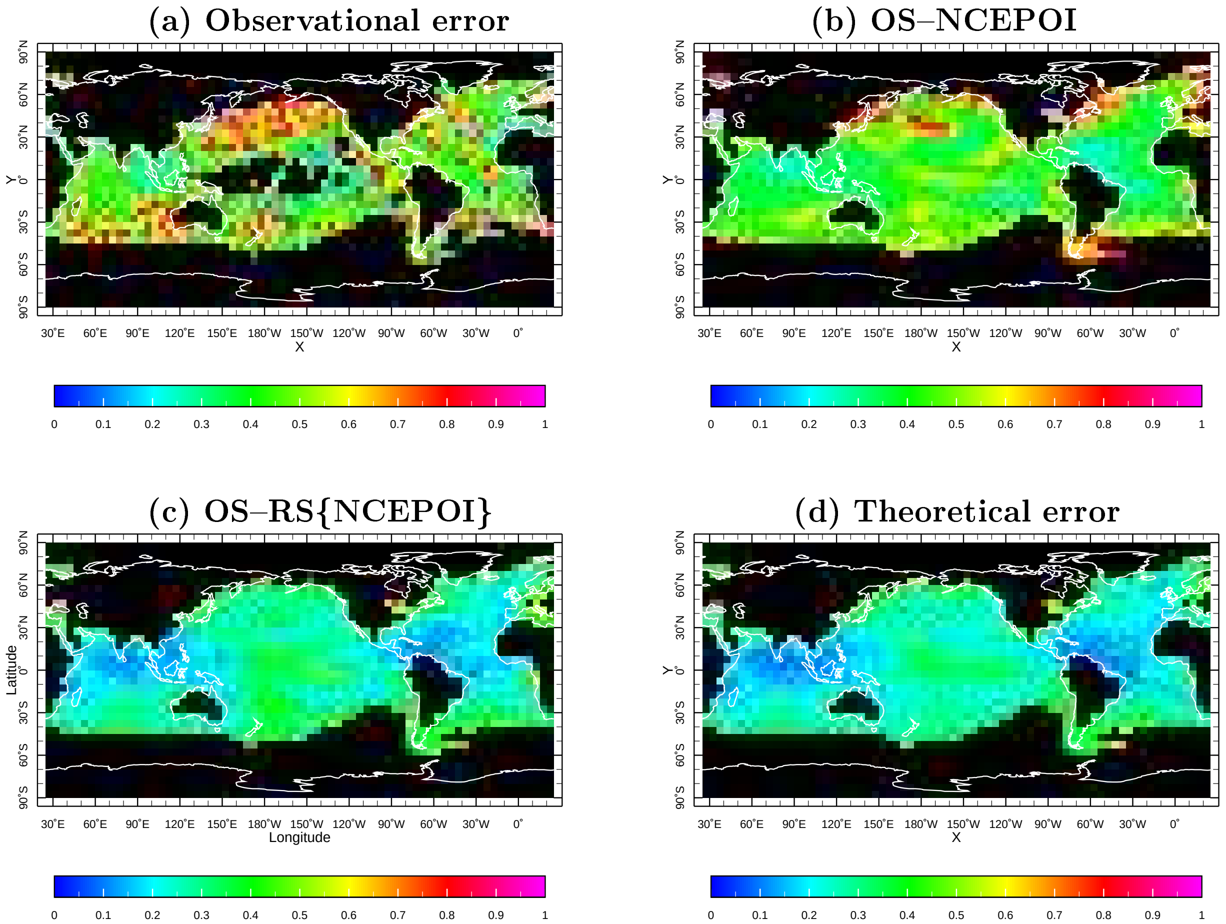

Figure 5: Statistics of the experiment with the NCEP OI data for 1992-1996 period resampled according to the 1916-1920 observational coverage. See text for the explanation.

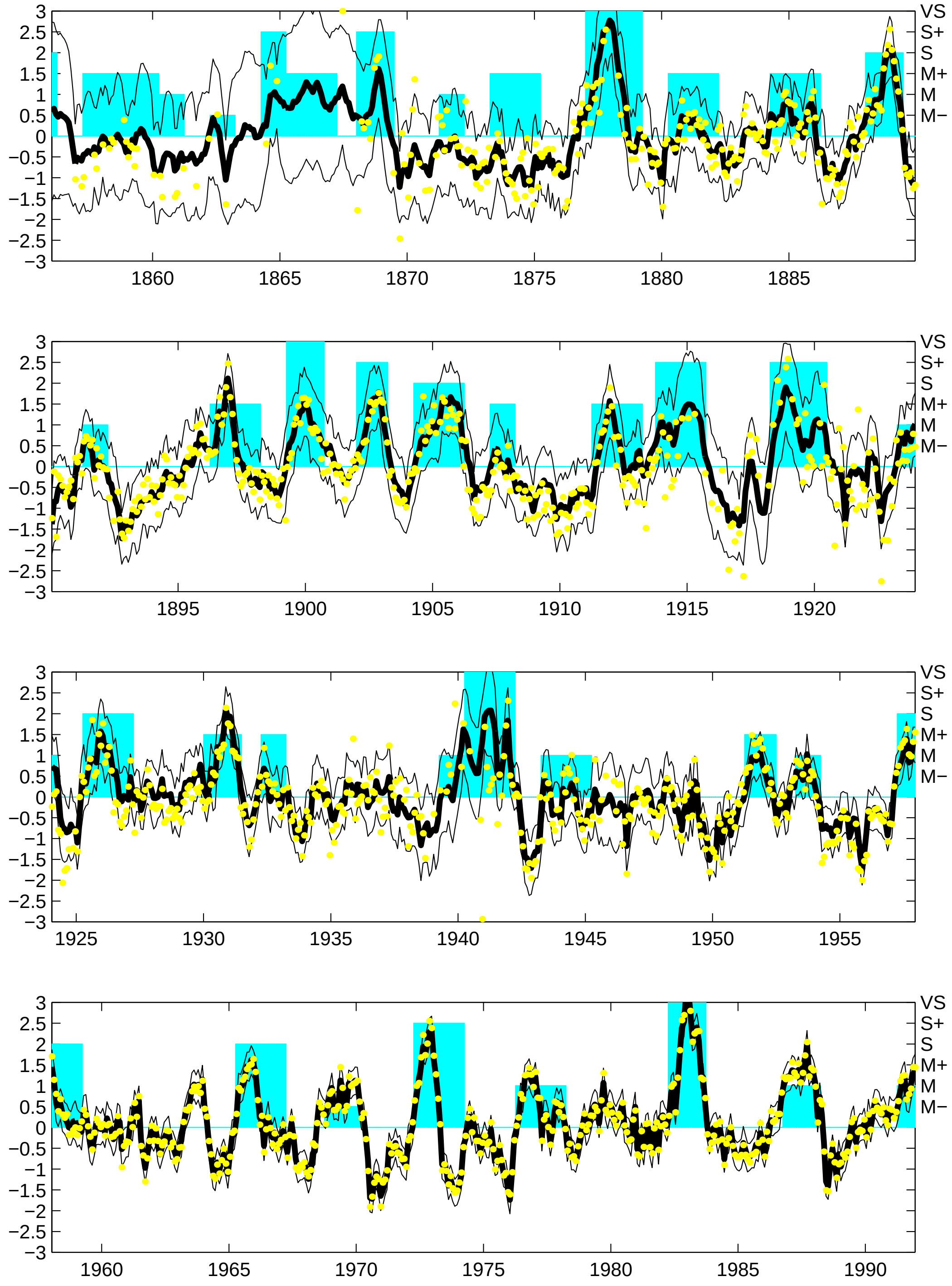

Figure 6: The reduced space optimal smoother reconstruction of the NINO3 index based on ship observations of SST (thick solid line) (Kaplan et al. (1998)). Also shown are 3σ error bars on the analysis values (thin lines), the straight estimates of NINO3 from raw data (dots), and the ENSO event ratings of Quinn et al. (1992) (histogram bars). Histogram bars are scaled to the Quinn et al. (1992) ratings: M-, M (moderate), M+, S (strong), S+, and VS (very strong).

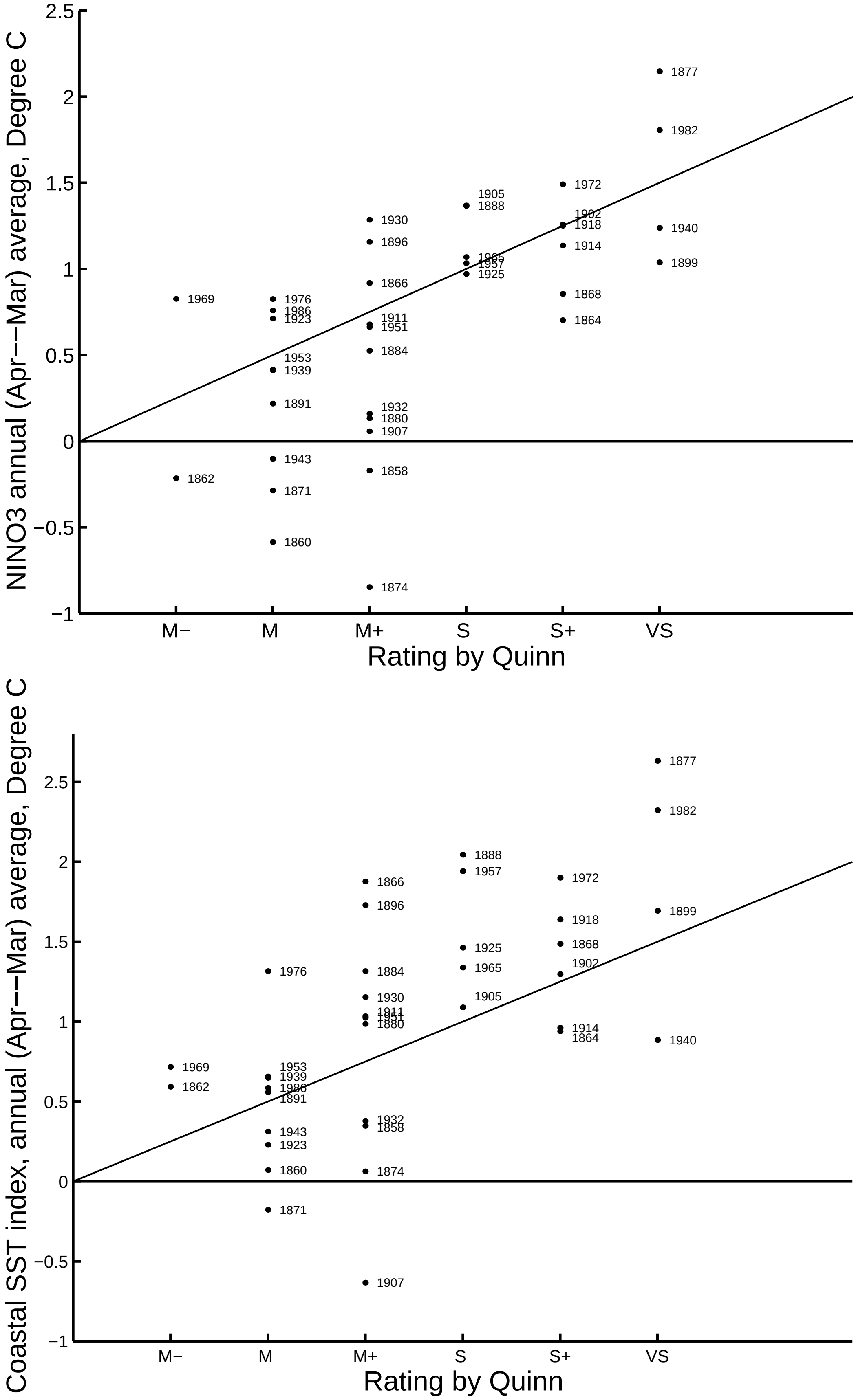

Figure 7: (top) Summary of shown in Figure 6 comparison of reconstructed NINO3 with the Quinn et al. (1992) ratings of El Niño events: events are represented by points on a (Quinn's rating, annual NINO3) plane; (bottom) same for the coastal SST index.

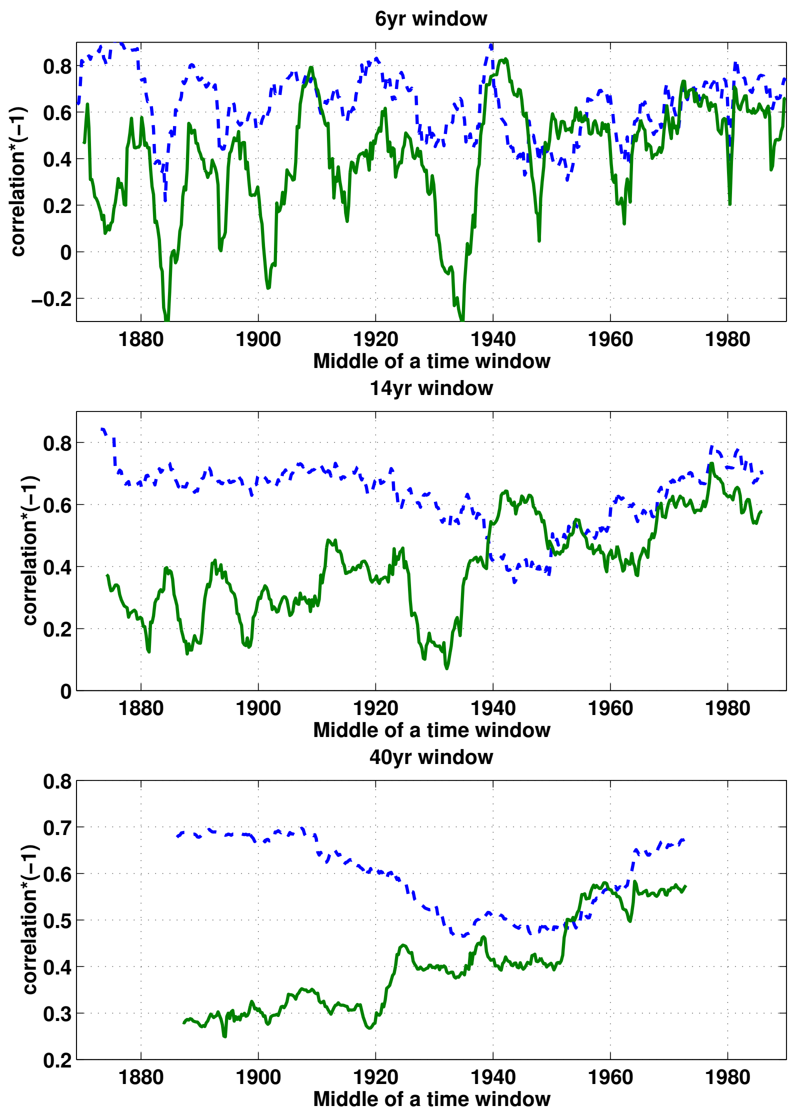

Figure 8: Correlation coefficient (multiplied by -1) between Darwin and Tahiti land station records (solid lines) and their marine-based proxies (dashed lines) shown as a function of time for time windows of 6, 14, and 40 years.

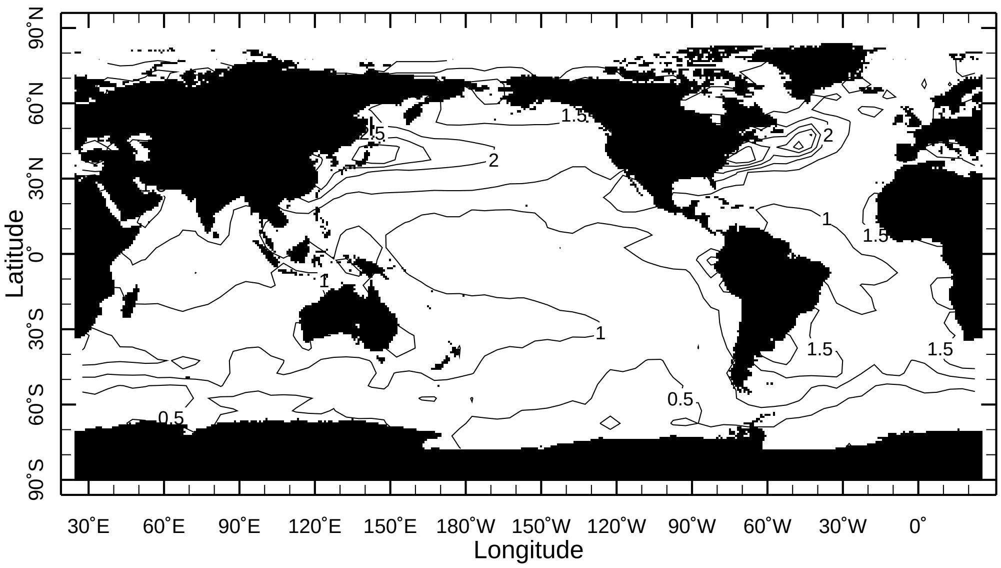

Figure 9: Intrabox SST variability, ^{o}C, inside $5^{o}\times 5^{o}$ bins, estimated from COADS, and used as a single ship error by Kaplan et al. (1998).

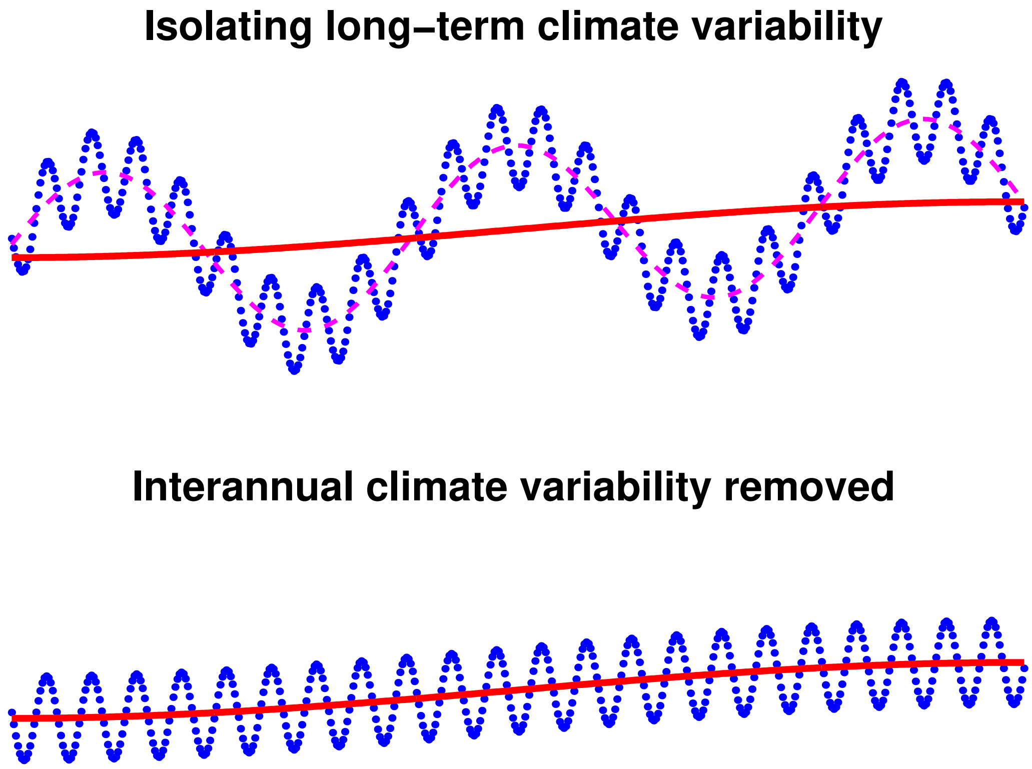

Figure 10: Removal of interannual climate variability (dashed line) from the observed data leaves the mixture of long-term variability (solid line) and error (dots).

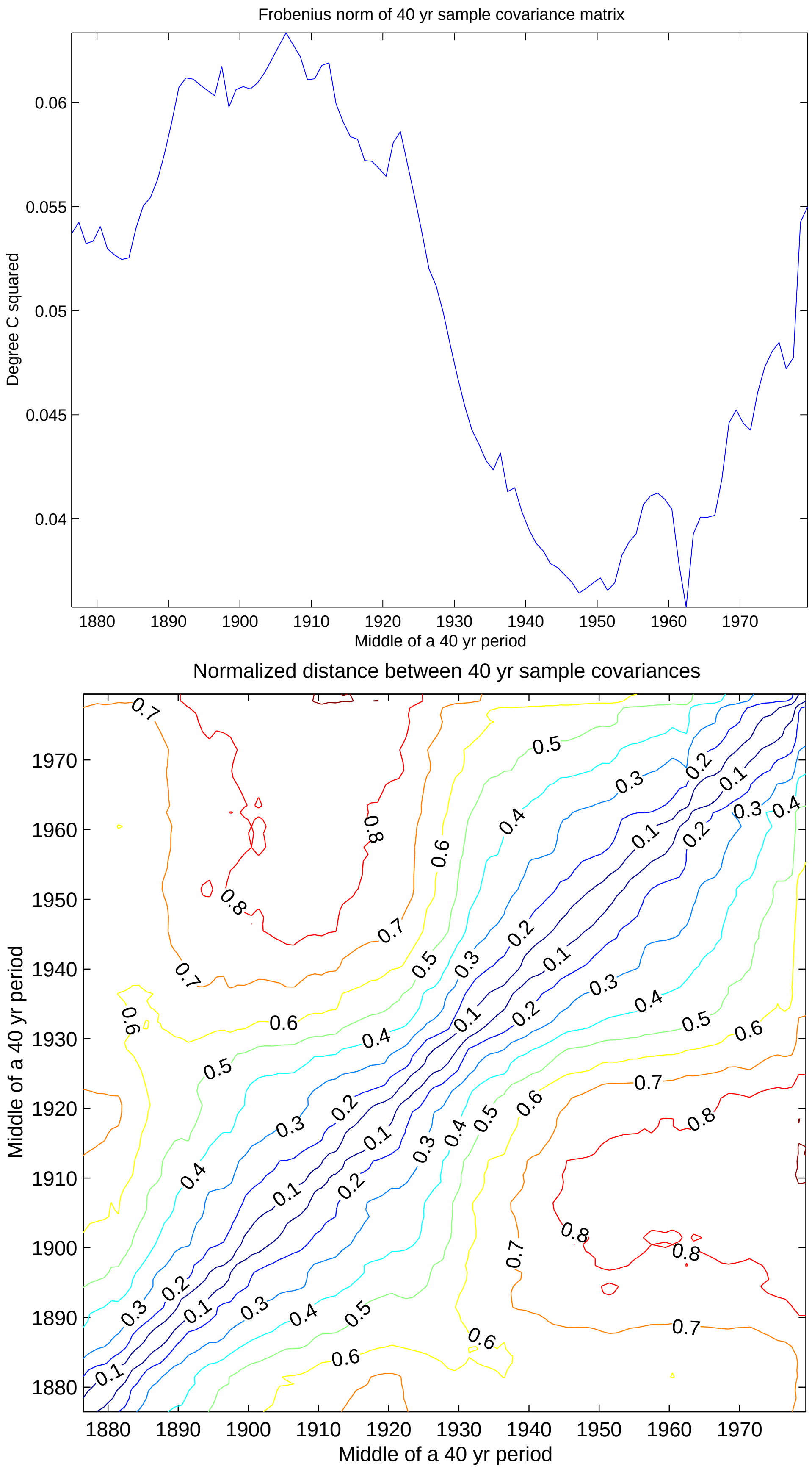

Figure 11: (top) The Frobenius norm of SST covariance estimated in 40 year time windows, (bottom) normalized distance between sample covariance matrices estimated for different 40 year windows.

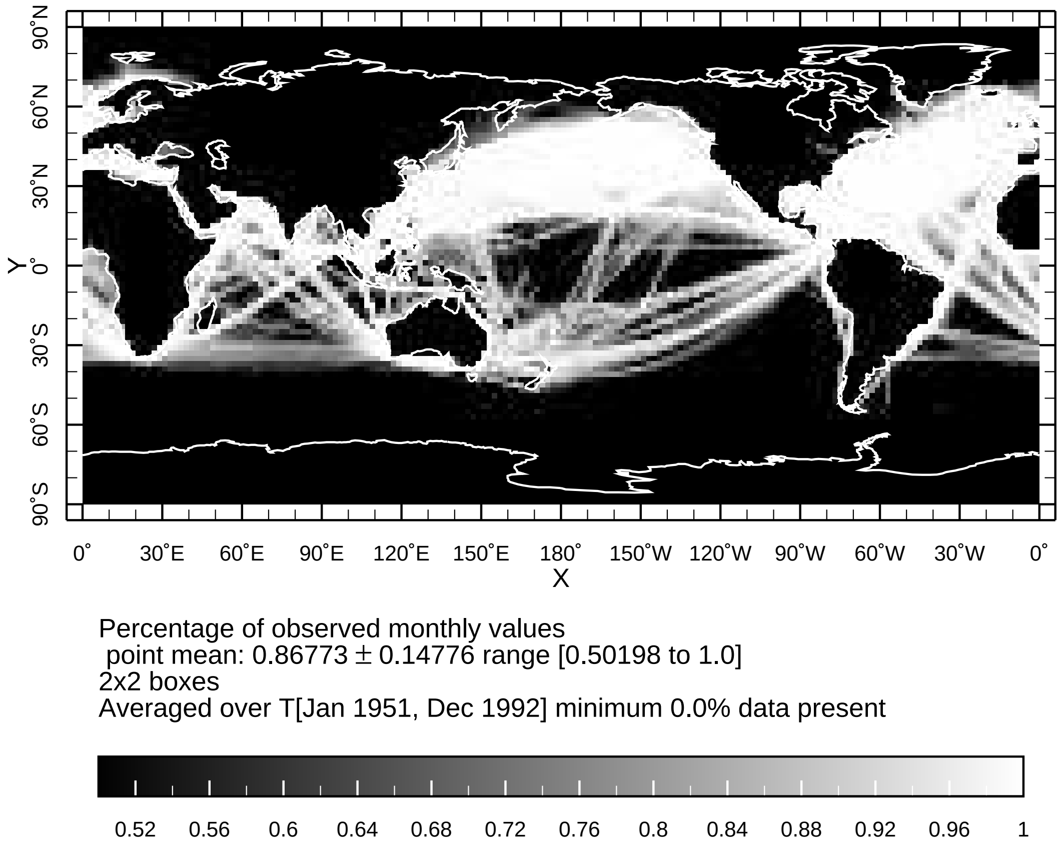

Figure 12: Percentage of months with observed SLP in COADS $2^o \times 2^o$ monthly summaries for the period 1951-1992.