Climate signals from station arrays with missing data, and an application to winds
Steven C. Sherwood
Universities Space Research Association, Seabrook, Maryland

Abstract. Typical approaches to climate signal estimation from data are susceptible to biases if the instrument records are incomplete, cover differing periods, if instruments change over time, or if coverage is poor. Here a method (Iterative Universal Kriging, or IUK) is presented for obtaining unbiased, maximum-likelihood (ML) estimates of the climatology, trends, and/or other desired climatic quantities given the available data from an array of fixed observing stations that report sporadically. The conceptually straightforward method follows a mixed-model approach, making use of well-known data analysis concepts, and avoids gridding the data. It is resistant to missing data problems, including “selection bias,” and should also be helpful in dealing with common data heterogeneity issues and gross errors. Perhaps most importantly, the method facilitates quantitative error analysis of the signal being sought, assessing variability directly from the data without the need for any auxiliary model. The method is applied to rawinsonde data to examine weak meridional winds in the equatorial lower stratosphere, providing some improvements on existing climatologies.

1. Introduction

Significant attention has recently been paid to the problems of natural and anthropogenic climate signal detection. The former often exploits empirical orthogonal function (EOF) and/or linear association techniques to find coherent fluctuations, while the latter has been approached by methods such as optimal detection [e.g., Bell, 1982] and fingerprinting [Hasselmann, 1979]. Such methods frequently rely on regular sampling of some variable of interest from an array of measuring sites, prior knowledge of the mean (climatology), or both.

Unfortunately, data available for climate research typically suffer from uneven sampling due to the addition, relocation, and/or loss of observing stations or platforms; sporadic instrument failure; or other interruptions during the period of interest. Even carefully planned field programs experience similar problems. Missing data can, for many purposes, be ignored or filled in by ad hoc methods with little impact on the final signals extracted from the data. However, biases can be introduced into the climatology or signal estimates in at least three ways: (1) Missing data aliasing: The data or platform availability pattern can interact with natural variations of the field in time and space, for example, instruments can fail during foul weather (“censoring”). (2) Interpolation bias: Field interpolations (explicit or implicit) at unobserved locations in the domain can be biased with respect to a desired signal if the interpolation or fitting procedure is designed to optimize a generic figure of merit not tailored to that signal [Sherwood, 2000b]. This problem can occur whenever some areas are persistently underobserved or unobserved. (3) Instrument inhomogeneity: Instruments can change sensors or characteristics or be moved to new sites during the period of interest; also, some instruments may be biased or simply unreliable.

These problems may be accidental (e.g., a heavily instrumented time period or region may happen to be abnormal) or systematic. Missing data aliasing can be caused by fair-weather bias in ship observations or foul-weather bias in operational radiosonde launches [see Elliot and Ross, 2000], while interpolation biasing could result from (for example) the preferential location of tropical rawinsonde stations in regions of cyclogenesis. These first two problem types may be thought of as data coverage problems and the third as a data quality problem, though we shall exploit the possibility in many cases of converting the latter into the former. In addition to these problems is the more well known occurrence in most data sets of occasional gross instrument errors,
which ideally should be recognized and discarded.

For atmospheric (and soon, ocean) work, one way to avoid these problems is to rely on products from data assimilation centers. These are appealing but unsuitable for some investigations: those whose scope exceeds that of available analyses (such as geophysical or long-term climate processes), where the analyses are believed to be poor, or where it is desired to keep the result independent of the numerical forecast models used in assimilation (e.g., for validation purposes or to avoid building assumed model physics into the result). In these cases some way must be found to analyze the data on their own.

Investigators doing this have noticed the appearance of analysis-dependent biases in mapping many climate signals, including interdecadal variations in temperature [Karl et al., 1994] and long-term means in rainfall [Hulme and New, 1997] and surface pressure [Jones et al., 1999]. Global climatologies of all types suffer from uneven coverage, particularly in the Southern Hemisphere. Instrument inhomogeneity problems can also be a great headache for climate analysts. They have been the object of much effort using a wide variety of methods [Peterson, 1998] and are a greater impediment to accurate calculations of surface temperature increase, for example, than are biases due to coverage [Intergovernmental Panel on Climate Change, 1995, chapter 3]. Finally, an underemphasized shortcoming of typical approaches is the difficulty of obtaining error bars on the end result, for which investigators often must turn to numerical climate simulations.

The common analysis approaches are well illustrated by the example problem of estimating surface temperature variations and warming over the last century. Jones et al. [1986] and Jones [1994] collected surface data in large boxes and intercalibrated them in missing data periods using a “reference period” in which all data are present. One shortcoming of this approach is the inability to treat heterogeneity within boxes, which Hansen and Lebedeff [1987] addressed by means of an ad hoc scheme in which the bias at each station was estimated approximately by successive pairwise comparisons. Vinnikov et al. [1990] employed statistically optimal techniques to map the data synoptically onto a regular grid prior to estimation of climate signals. Though the warming signal itself is sufficiently robust to be well captured by each of these methods, interdecadal temperature variations apparently are not [Karl et al., 1994].

Recently, Sherwood [2000b] pointed out that approaches that avoid synoptically mapping or gridding the data could both improve accuracy and facilitate rigorous error estimates. Unfortunately, with an incomplete or heterogeneous space-time array of observations, such approaches are straightforward and accurate only when the signal is well sampled at all sites and other variability can be filtered out. This excludes most cases of interest. Here a method is presented which relaxes this constraint considerably. The method is first described in section 2, section 3 uses simulations to show how this method avoids problems that afflict other strategies for handling the missing observations. Finally, section 5 briefly compares lower-stratospheric wind estimates with operational products.

2. IUK method

This method (Iterative Universal Kriging (IUK)) actually works in a manner somewhat similar to traditional data assimilation techniques except that (1) the “background guess” field, anomalies from which are treated stochastically, can be obtained from the data themselves rather than a physical forecast model; and (2) the model steps are iterated, automatically converging to the maximum-likelihood (ML) estimates of the desired parameters given the available data and the model. The method exploits both the sampling characteristics of the data network and the spatiotemporal coherence of the field.

The general approach is that of “signal mapping” [Sherwood, 2000b]. This means that the user first decides which statistics or parameters \( a = \{a_1, \ldots, a_n\} \) are desired as end products of the work, then estimates \( a \) at observation sites directly from the data, and finally (if desired) grids or interpolates \( a \) to unobserved locations. The parameters \( a \) can be means, trends, cycle amplitudes, etc. At present, however, they must be additive (linear) components of the estimand \( Z \).

The IUK method is based on “universal kriging” [e.g., Cressie, 1993], or the representation of the data as a parametric model plus a random process,

\[
Z = \mu + \epsilon,
\]  

where each of these quantities is a function of space and time. The parametric model is a linear superposition of basis
functions:
\[ \mu(s, t) = \sum_{i=1}^{m} a_i f_i(s, t) + \sum_{i=1}^{m} b_i g_i(s, t). \]

The variables \( s \) and \( t \) are the discrete location and time coordinates, respectively, at which measurements of \( Z \) are nominally available. The observable \( Z \) and each of its basis functions \( f_i \) and \( g_i \) can be scalar or vector quantities. The functions \( f \) represent the desired signal patterns and should be chosen accordingly by the analyst; for example, if the trend is desired then one of the \( f_i \) should be a linear function of time. The signal amplitudes \( a \) are estimated by the method. The additional functions \( g \), which may be determined empirically or a priori, represent extraneous variability that is coherent over scales spanning a number of locations \( s \) and times \( t \). Less coherent variability is represented by the field \( \epsilon \), which is modeled as a Gaussian random field (GRF) having a heterogeneous and stationary autocovariance function \( \sigma_e(dx, dt) \) [e.g., Daley, 1991].

Note that what is usually called the “noise” is here partitioned into \( g \) and \( \epsilon \) according to its coherence in space and time. The treatment (1)-(2) is a generalization of curve-or spline-fitting techniques (which would simply treat \( \epsilon \) as a white noise process) and standard statistical interpolation (SI) techniques (which would treat \( \mu \) as a constant). North et al. [1995] provided an elegant solution for the case where all \( \{f, g\} \) are known a priori (e.g., from a numerical model) but provided little discussion of incomplete data and how this would affect the analysis. We too treat \( g \) as given, until section 4. There the benefits of, strategies for, and cautions concerning noise partitioning (i.e. determining \( g \) empirically) are discussed.

I assume throughout that the complete set of observables \( Z \) at all \( s \) and \( t \) is made up of an available subset \( Z_A \) and a missing or unknown subset \( Z_M \) (similar notation is used for \( \epsilon \)). I also assume that a straightforward method is available for obtaining the ML estimate of \( a \) from a complete data set \( Z \) (for the linear model here, this would just be the least squares projection of \( Z \) onto \( f \)). Note that most such methods cannot be applied directly to incomplete data sets. The goal here is to find the best estimates of \( a \) given only \( Z_A \).

The method proceeds in a series of steps which are iterated to convergence. The basic idea of iterating the universal kriging technique was previously suggested by Ripley [1991], and probably others.

The IUK steps are illustrated in Figure 1 and listed in Table 1. This series of steps is a direct implementation of the EM algorithm due to Dempster et al. [1977]. They demonstrated that an iterative algorithm could converge to the ML estimates of the parameters of a wide class of models, given an incomplete set of data. Their algorithm iterates between an E-step (in which missing data are filled in by their conditional expectations given the observations and the currently fitted model parameters, and expectation values of sufficient statistics to fit the model are computed), and an M-step (in which model parameters are fitted to the complete data and sufficient statistics as though they were all observed). This algorithm is absolutely convergent, with initially rapid convergence that slows considerably as the solution is approached. EM has been used previously to fill missing data in geophysical problems before fitting a model [Carroll et al., 1997; Shumway and Stoffer, 1982], but here it is used to embrace the entire modeling process. The resulting estimates of parameters \( a \) are maximum likelihood, given \( \epsilon \) and \( \{f, g\} \).

Steps 2 and 3 apply the technique of statistical interpolation (SI) to predict unknown values of the residual field \( \epsilon \) which are optimal given a stochastic model of \( \epsilon \). Variations of this technique are also known as kriging in the geological sciences, the Gauss-Markov method in statistics and oceanography, and optimal interpolation in atmospheric data.

<table>
<thead>
<tr>
<th>Table 1. Steps in the Iterative Universal Kriging procedure.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0. Initialize ( a ) and ( Z_M ) by any approximate method (e.g., statistical interpolation).</td>
</tr>
<tr>
<td>1. Fit ( \mu ) to this estimate of ( Z ); obtain residuals ( \epsilon ).</td>
</tr>
<tr>
<td>2. Estimate ( \sigma_e ) from ( \epsilon_A ) by a standard empirical method.</td>
</tr>
<tr>
<td>3. Use statistical interpolation to fill in ( \epsilon_M ) values based on ( \epsilon_A ).</td>
</tr>
<tr>
<td>4. Evaluate ( \mu ) at the missing observation sites and add to ( \epsilon_M ) to reconstruct ( Z_A ).</td>
</tr>
<tr>
<td>5. Go to 1 unless ( a ) has become sufficiently stable.</td>
</tr>
</tbody>
</table>

Figure 1. Iterative Universal Kriging method proceeds by iteration of three steps until desired accuracy is reached.
assimilation. Some of these are available in commercial software packages. Daley [1991] gives a clear presentation of SI and its atmospheric applications. The implementation in this paper is identical to that described by Sherwood [2000b] except as noted. One unusual feature here is that observations at all times (“asynoptic” data) are used to predict the value at a target \((s, t)\), rather than merely those from the same time \(t\). Though uncommon, there is nothing preventing the use of SI as an interpolator in both space and time, as long as \(\sigma_e\) is valid for this many dimensions.

2.1. Details of the Steps

2.1.1. Step 1 (\(\mu\)). A straightforward method for fitting \(\mu\) to a complete data set is assumed to be available. This will usually involve computing sample means, linear regression, or Fourier transforms/linear projections, depending on \(\{f, g\}\). If \(g\) values are to be determined empirically, then the data are first projected onto \(f\) to obtain \(a\). The residual “noise” from this fit is analyzed to obtain \(g\) and projected thereon to yield \(b\). Calculation of \(\mu\) and subtraction from \(Z_A\) to give \(\epsilon_A\) completes the step.

2.1.2. Step 2 (\(\sigma_e\)). The autocovariance function \(\sigma_e\) is parameterized here using the “spherical” form, one of several common, simple choices defined by a few parameters \(\Theta\). These parameters include the “ranges” or distances in each direction beyond which \(\epsilon\) values become decorrelated, a “nugget” representing white noise in \(\epsilon\), and the overall variance of \(\epsilon\). The \(\Theta\) must be estimated empirically from the \(\epsilon_A\), which is done here using standard methods [see Gunst, 1995; Sherwood, 2000b]. Here \(\sigma_e\) is a function of normalized distance (the quadrature sum of the distances in each dimension normalized by their respective ranges). The spherical form for \(\epsilon\) is valid in up to three dimensions [Cressie, 1993], i.e., two horizontal and time.

It is important to perform this step using only the \(\epsilon_A\), to avoid the large biases in second-order statistics that would ensue if the estimated residuals \(\epsilon_M\) were used. This “naive estimate” of \(\Theta\) is not ML, but as long as the data set is reasonably large, it should be very close, as found, for example, by Carroll et al. [1997]. The estimate is also not strictly unbiased, since \(\mu\) has been determined from the same data, but the bias should be small as long as the data set is reasonably large and the model for \(\mu\) is sufficiently parsimonious [Cressie, 1993]. This will cause slight underestimation of covariances but should have little impact on ranges. I have not found errors in estimating \(\Theta\) to be of importance in applications discussed in this paper, but for short data records, universal kriging and IUK may not be suitable.

2.1.3. Step 3 (\(\epsilon_M\)). Here unknown residuals \(\epsilon_M\) are estimated from known residuals \(\epsilon_A\) using SI. The previous \(\epsilon_M\) must not be used, since the proper convergence of EM requires that each missing datum be replaced by its conditional expectation value given \(\Theta, a\), and the available data [Dempster et al., 1977]. Interpolations are only necessary at missing \((s, t)\) and need not be done on a domain-filling grid. The implementation of SI differs from the standard technique described by Sherwood [2000b] in two respects, aside from the use of asynoptic data mentioned above.

First, the minimum-variance estimator is used here. More often, the so-called “unbiased estimator” (in which weights are forced to sum to unity) would be used to ensure that interpolations are globally unbiased: in effect, estimating a constant \(\mu\) without bias. Since we already treat \(\mu\) explicitly, this would be inappropriate here. Cressie [1993] gives a good description of the distinctions between these two estimators.

The other difference is due to computational realities. In applying SI, one must make an important choice between (1) solving the relevant equations exactly, which requires inverting an \(N \times N\) matrix where \(N\) is the number of available observations or (2) using a neighborhood restriction, where only \(k\) nearby neighbors are used for interpolation, and a separate \(k \times k\) matrix is inverted for each prediction. The latter expedient (commonly used in the “optimal interpolation” (OI) procedure performed until recently in data assimilation) can lead to poor results [Cohn et al., 1998] but is sometimes the only tractable way to proceed if \(N\) is large since the expense of matrix inversion increases so rapidly with size (for space-time interpolation, \(N\) is the total number of observations over all time). On the other hand, limiting the number of neighbors can sometimes reduce errors if \(\sigma_e\) is poorly known [Lorenc, 1986]. In IUK, step 3 will typically require limited neighborhoods since the interpolation is done in at least three dimensions. This limitation does not appear problematic, for reasons illustrated in section 4.4.

2.2. Relation to Other Methods

It is widely recognized that plain SI techniques can be poorly suited for reproducing both large- and small-scale variability simultaneously. When SI is used outside of the data-assimilation context, seasonal cycles and other known variations are usually removed from the data to ameliorate this problem, though estimating these signals itself can be problematic, especially when the data are incomplete, as we shall see below. IUK synergistically estimates such cycles while predicting missing data values.

A clever and fast technique for interpolating multiscale data accurately in two dimensions was given by Fieguth et al. [1995], though their technique does not generalize as well to more than two dimensions and does not have the ML parameter estimation property of IUK.

The IUK method is conceptually similar to that used by...
Carroll et al. [1997] to analyze ozone exposure in Texas, except that these authors used EM only to fill in the data without involving parameter estimation in the EM loop. Here we are interested in formulating the model so that there is a direct relationship between the \( \alpha \) and the desired scientific conclusions, and the focus is on estimation of the \( \alpha \). This also distinguishes IUK from a variety of ad hoc iterative methods that have proven reasonably successful in performing simple interpolation (reviewed by Lorenc, 1992).

IUK is analogous in some ways to Kalman filtering but simpler and not designed for forecasting. Most mixed-model treatments have, as Kalman filtering, focused on forecasting, often using autoregressive models [e.g., Angulo et al., 1998]. Iteration as used here is inappropriate for forecasting but advantageous for posterior estimation.

3. Test I, Toy Problem

As an illustrative test, consider the following “toy problem.” Two instruments S1 and S2, separated by unit distance, measure a spatiotemporally coherent variable \( Z(x, t) \) at unit time intervals but not every time. An investigator wishes to estimate the mean difference in \( Z \) (\( \Delta Z \)) between the 2 locations as accurately as possible. (Alternatively, one could think of S1 and S2 as two times when data are available from a variety of locations, from which the investigator wants to determine the spatially averaged temporal change \( \Delta Z \).)

The problem was set up by generating pink (in time and space) noise at the stations having Gaussian statistics (assumed known to the investigator). A slow cycle of frequency \( \omega \) and amplitude \( A \) was added at each station (with \( A = 0 \) or 5). The investigator may or may not be aware of the cycle; if so, we assume only its frequency (the real part) is known. A constant was finally added to the simulated data at each station as necessary to make the true means 0.0 and 1.0, for a desired signal \( \Delta Z = 1.0 \).

Missing data were established by generating a second, random process at the stations and flagging observations as missing when this process exceeded a threshold value. This missing-data process was coherent temporally, so missing observations would tend to be clustered as in real life but not spatially. A sample time series at each station is shown in Figure 2. Alternate tests were done with the entire first half of the record at S1 missing, representing the appearance of a new station in a data set.

Six different estimation methods were tested by generating a large number of simulated data sets (each with a different missing-data pattern) and presenting them to each method. The methods were evaluated by their ability to estimate the desired statistic \( \alpha = \Delta Z \) (discussed first), then \( \alpha = A \) (discussed subsequently). The retrieved and true \( \Delta Z \) were computed only over times when at least one station reported an observation. The methods are as follows: (1) Naive means, compute \( \alpha \) directly from the data; (2) Naive difference, same but use only times when both instruments report; (3) OI\(_1\), filling in missing values by SI using only cotemporal observation (if any), then compute \( \alpha \); (4) OI\(_2\), SI using any observations, up to \( k = 10 \) points; (5) IUK\(_1\), IUK using only cotemporal observations; (6) IUK\(_2\), IUK using any observations, \( k = 10 \).

The “naive” methods estimate \( \Delta Z \) without taking coherent variations into account in any way. For observations that are independent and identically distributed (iid), they are ML methods given all the data (“naive means”) or given the complete data records only (“naive difference”).

The OI methods are a direct implementation of SI with \( k = 10 \). “OI\(_1\)” is a synoptic data-mapping procedure for computing \( \alpha \), while OI\(_2\) is a semisynoptic procedure that also uses information at nearby times. The \( k = 10 \) limitation is irrelevant in OI\(_2\) since there are only two stations but limits the amount of information available at long time lags in OI\(_2\).

The IUK methods fit the data to a four-parameter model for \( \mu \) which includes the mean at each station and the amplitude and phase of the cycle. Estimated \( \Delta Z \) is just the difference between the fitted means at S1 and S2.

The amplitude \( A \) was also estimated in these tests, by projecting the available anomalies onto \( \exp(\mu \omega t) \) using linear regression. The anomalies had to be taken from some mean value, which for the naive mean method was just the naive means. For the naive difference method the SI mean was set to the sample mean there, and the S2 mean set to that at S1.

Figure 2. A sample simulated data series from two instruments (offset for clarity).
plus the naive difference estimate. For the OI methods the regression was performed using the available and filled data.

3.1. Results

3.1.1. Spatial gradient. The most notable result from the \(\Delta Z\) tests (Figure 3) is the sizeable bias shown by the OI methods, especially \(\text{OI}_1\), in estimating the spatial gradient. This is because information on the mean and anomalies is borrowed from other locations to fill in missing observations, biasing the mean at each station toward that of the other station. The bias is reduced in \(\text{OI}_2\) since temporal as well as spatial information is available.

When a large-scale cycle was present, the naive mean estimates deteriorated dramatically due to incidental interaction between this cycle and the missing data process. Methods that fill in missing data or ignore incomplete records were little affected.

In the challenging case where this large-scale variation was strongly correlated with the missing-data process (Figure 3c), the naive mean suffered from a very large bias (\(\sim 3\) times the signal). The OI procedures also worsened due to their increasing reliance on information from S1 in filling in the long gap at S2. Note that this problem has nothing to do with limited neighborhoods; it is inherent in the difference between the fundamental goal of SI (minimum squared error of individual interpolants) and that of the investigator (accurate statistics). The naive difference and IUK methods were only slightly degraded by this missing-data problem and remained unbiased.

3.1.2. Temporal cycle. Figure 4 shows the errors in estimating \(A\) for the two \(A = 5.0\) cases above. When the missing-data process was stationary and homogeneous (Figure 4a) all methods performed well. With only a partial record at S1, however, the naive mean and OI methods consistently overestimated \(A\) (Figure 4b). This is because estimation bias at S1 in the first half of the record aliases onto the temporal behavior. The naive method performed well only when the naive difference was used to estimate \(\Delta Z\) but was again the worst method when the anomalies were taken from the individual station means as would typically be done in practice.

Figure 3. Absolute errors in estimating \(\Delta Z\). Boxes show 25th-75th percentile range; thin bar inside box shows median absolute error; hatched bars show 1st, 99th percentile range. Bias (mean error) is shown by thick dashed bar. The methods are (from left) Naive means, naive difference, \(\text{OI}_1\), \(\text{OI}_2\), IUK1, IUK2 (see text). (a) \(A = 0\); (b) \(A = 5\); (c) \(A = 5\) and first half of record missing from S1. Note the logarithmic scale in Figure 3(c).

Figure 4. (a, b) as in Figure 3b and 3c except estimation errors for \(A\) are shown.
3.2. Discussion

These simple tests illustrate how interactions among the missing data process, climatology, and natural variations can conspire to introduce biases into simple estimates of climatology or climate signals. The problems may have been exaggerated here by the small number of instruments but, in principle, will occur regardless of this number.

Interestingly, the “naive difference” estimator of $\Delta Z$ worked surprisingly well. Thus if a reasonable fraction of the observation times in a data set are complete (i.e., all instruments reporting), then a good procedure is simply to discard the incomplete observation times. Sherwood [2000b] employed essentially this method to estimate the atmospheric tide. Unfortunately, as the number of observables increases much beyond 2, the number of complete observations can quickly approach zero, making this technique hopelessly inefficient.

4. Empirical Determination of $g$ and a Test With Rawinsonde Data

In the toy example, our functions $\{f, g\}$ consisted of the climatology and known waveforms. However, an important advantage of IUK is the ability to use standard pattern recognition techniques to determine from the data useful functions $g$ which compactly represent modes of natural variation that are not part of the signal $f$. This empirical feat is possible because a complete data set (essential for most pattern recognition techniques) is synthesized by the procedure. Though it would be better to know $g$ independently of the available data (to reduce the total number of parameters that must be estimated from a data set of fixed size), in practice this requires a dynamical model with sufficient fidelity to simulate modes of variability better than they can be obtained from data, a dubious proposition in many cases. By estimating $g$ the method becomes self-contained, also avoiding the less efficient option of dividing the data into separate parts for determining $g$ and $a$. The catch is that estimating $g$ in this manner involves subtle issues, requires care on the part of the investigator, and is not guaranteed to work well.

4.1. Basic Principles

Variability across a wide range of scales is well known to be the rule rather than the exception in chaotic dynamical systems. Unfortunately, SI is poorly suited to represent this; large-scale behavior is sensitive to the tails of $\sigma_z$, which are hard to estimate and require complicated parameterizations of $\sigma_z$ to represent adequately. Also, large-scale variations tend to occur in particular ways determined by the system dynamics or forcing, rather than stochastically, and thus tend not to be statistically homogeneous. There should be much better ways of representing these variations than by a GRF, provided that a sufficiently long data set is available to learn what they look like, hence the rationale for dividing the “noise” into $\mu$ (i.e., $g$) and $\epsilon$.

Since this distinction is essentially spectral, it will be important to ensure that $g$ functions are smooth at scales comparable to or less than the decay scales of $\sigma_z$ in each dimension; rough variation in $g$ between adjacent stations, or spectral overlap between $\epsilon$ and $g$, is a sign of overfitting or attempting to learn too complicated or numerous functions $g$. This leads to poor interpolation and performance [see Daley, 1991]. If it occurs, the analysis should be repeated with a smaller basis $g$. Overfitting is likely unless $m + n \ll N$ (though if any of the individual $a_i$ are constrained by too few data points, $m + n$ will have to be reduced further).

4.2. Functions $g$ From EOFs

We now need a pattern recognition technique for obtaining the pattern (basis) functions $g$ from the noise and a way of deciding how many patterns to use. Here EOF analysis is suggested. This technique represents a time-space matrix of data by a superposition of modes that captures maximal variance in a minimal number of EOF modes, which is ideal here since it maximizes model parsimony. Also, the modes are conveniently orthogonal. One other option would be principal oscillation patterns or POPs, which capture time dependence in a possibly useful way but are neither orthogonal nor as parsimonious as EOFs [e.g., Von Storch et al., 1995], or the data may simply be used to choose empirically a manageable number of standard (e.g., harmonic) functions from some large basis.

Since spatial EOFs do not use time information, it is desirable to prefilter the data with a time smoother before computing EOFs so that the algorithm will only recognize patterns that are coherent beyond a certain timescale. This also helps keep the modes smooth in space by filtering out small-scale, random fluctuations. The shape of the filtering kernel affects that of $\sigma_z$ and should be similar to that of the unfiltered autocovariance of $Z$ (though decaying more rapidly) in order to preserve good shape in $\sigma_z$. If the kernel is too fat, $\sigma_z$ will dip below zero before decaying back, behavior that cannot be represented by our simple parametrization of this function. Exponential smoothing is likely to be a good choice in geophysical applications; boxcar filtering (for example) is not, tending to produce an oscillatory $\sigma_z$ [e.g., Kendall and Ord, 1990, chapter 5].

Another benefit of prefiltering the data relates to parametric complexity. Each retained mode adds $n$ additional basis functions to (2), where $n$ is the number of times $t$ in the data set, since the mode amplitude must be estimated at
each time. If the data were, say, filtered by time averaging in blocks of 100, then \( n \) would decrease by a factor of 100. With smoothly decaying kernels, \( n \) does not strictly change, but the serial correlation in the amplitudes \( b \) similarly reduces the effective number of independent parameters that must be estimated. Thus more modes can be retained without overfitting; recall that the unfiltered data set with all its information content is used to calculate \( \varepsilon \).

A second preparation detail is that prior to EOF analysis, the observing sites should be weighted according to area by rescaling the data at each site by the square root of the area closest to that site before analysis, then unscaling the EOFs afterward. This procedure correctly compensates for variations in station sampling density \([\text{Buell, 1971}]\), preventing clusters of nearby stations from dominating the determination of modes.

Two ways are available for choosing the number of EOFs to use in (2). First, each mode can be examined in turn starting with the leading mode, keeping modes as long as they continue to look reasonable as described above. Alternately, modes can be retained as long as the eigenvalues are dropping faster than would be expected with modes obtained from pure noise.

### 4.3. A Wind Example

Here IUK is tested on a real data set, with \( g \) determined empirically from EOFs. The data are 70 hPa wind observations from operational rawinsondes in the tropics. Two data sets are analyzed: a “full” set covering December-February and June-August of the years 1985-1998 which includes 145 stations throughout the tropics, and a “reduced” set for method-testing purposes covering only 1987-1992 which includes 35 stations from the Indo-Pacific region. The full data set contains twice-daily sondes with tides (as estimated by \textit{Sherwood} [2000b]) removed, while the reduced data set contains only 00 UT observations. The percentage of observations available at each station in the data sets varies from 5.9\% to 83.0\%, with 43.8\% of observations available overall. Many stations are completely devoid of data for multiple years.

The field \( Z \) here is thus the vector wind \((u,v)\). For this test, the desired signals \( a_i \) are chosen to be the mean \( u \) and \( v \) at each of the 35 stations (i.e., \( 70 f_i \), each a time-independent unit vector). Time filtering of the noise field employed an exponential kernel with bidirectional e-folding decay lag of 4 days. Area rescaling was based on the area closest to each station, out to a maximum distance of 7000 km. Both strategies mentioned in section 4.2 suggested that four or five EOFs should be retained. The main tests were performed with three modes, and comparisons are shown subsequently using other numbers.

In IUK the residuals \( \varepsilon \) were screened to identify bad data. Points any more than six “pseudo standard deviations” (4.45 times the interquartile range) away from zero in either the meridional or zonal components were considered bad and removed. At the end of the first iteration the central 99\% of the distribution of \( \varepsilon \) was already nearly Gaussian (Figure 5). The number of samples from such a Gaussian expected to lie outside the six pseudo-\( \sigma \) cutoff was less than 1, making this a very conservative test that only identifies truly bad points. If the same criterion is used to identify bad points from \( Z \), only half the number is identified, showing the power of IUK to identify bad data more powerfully yet reliably.

After screening for bad data, the missing residuals were interpolated using up to \( k = 25 \) “best” neighbors. The best neighbors were those whose \( \varepsilon \) was most correlated with the target value based on \( \sigma_\varepsilon \) (e.g., the closest). Results were not sensitive to \( k \).

#### 4.3.1. IUK computational details.

Since tests showed that \( g \) and \( \theta \) changed negligibly after three iterations of the algorithm, these were fixed for subsequent iterations and no further bad data removed. This facilitated an important computational shortcut, since the weights for interpolating the missing \( \varepsilon \) also became constant and could be stored in an \( N_{\text{had}} \times k \) matrix, \( N_{\text{had}} \) the number of missing data points and \( k = 25 \). Looking up these weights in subsequent iterations avoids costly matrix inversions (though requiring substantially more memory). In these tests the algorithm sped up by 1 to 2 orders of magnitude as a result.

Another strategy to speed convergence was to accel-
4.3.2. Test results. The same methods used in the toy problem are employed in this test. In SI, limited neighborhoods were required as in IUK since exact solution using all data in time and space required far more memory than available resources would allow, even with the reduced data set.

These tests focus on the $u$ component in the reduced data set. Figure 6 shows the reconstructed time series at a station with significant missing-data problems, using unbiased OI and IUK. There is clearly a high bias in the OI reconstruction, most visible in the third panel (third year): the interpolated curve passes through the observations (crosses) but jumps up between them when no data are available. There is no obvious bias in the IUK reconstruction.

To quantify the skill of each method, any station $i$ is degraded by blacking out four out of the six years of its data (leaving only 1988-1989). Then each method is used to estimate $\mathbf{a}$ from the degraded data, and $a_i$ is compared to the “control” estimate obtained with all the data, a challenging test due to interannual variability in $u$. This is repeated for each station that is missing no more than half its data (so that the “control” estimates themselves will be accurate and the test meaningful). All three methods agree to within 0.3 m/s in their control estimates at these stations.

Figure 7 shows the resulting discrepancies (tests during other seasons and with different blackout years produced similar results). IUK is the most accurate overall, with a maximum error of $1.4 \text{ m s}^{-1}$ compared with $\sim 4 \text{ m s}^{-1}$ for the other two methods. The naive mean is the least accurate, usually deviating in the test by at least $1 \text{ m s}^{-1}$ from the control value. OI is accurate at most stations but makes substantial errors at a few of them.

Note that the true OI error is underestimated here because our test only captures random error and not biases, discovered in section 3, which will equally affect the control and degraded runs (this is also fundamentally why signal error analysis is impractical from mapped data). At the best reporting stations here, the error underestimate is negligible. At the opposite extreme (a hypothetical station with no actual data), IUK and the naive method would be unable to provide any estimate, whereas OI would happily provide one (the same one in the control and degraded runs) based on information from other stations. These factors account for the trends in Figure 7 toward greater IUK error and, apparently, less OI error the less data a station has. Recall, however, that once IUK is complete, $\mathbf{a}$ itself can be mapped which achieves the same benefit as OI near the less reporting sites [Sherwood, 2000b].

Figure 8 shows how the performance of IUK varies with the number of EOFs retained in this test. The performance in this case is not very sensitive to the exact number used, with even 1 EOF giving substantial improvement over OI. With no EOFs, performance was slightly worse than OI. Tests with data sets of different sizes (not shown) suggest that the IUK performance improves faster than OIs with the number of observation times $t$, but more slowly than OIs with the number of stations $s$.

4.4. Interpolation Characteristics: Standard Versus IUK

Here I demonstrate some technical advantages of IUK over standard kriging or SI. The SI approach is based on the autocovariance $\sigma_{Z_i}$ of $u$ and a constant $\mu$. Figure 9 shows empirical covariances between $u$ at station pairs plotted against distance from the full data set. Clearly, the statistics of $u$ are not homogeneous, since the scatter is far greater than...
can be explained by sampling errors (a structure function plot, not shown, shows similar or greater scatter). Since field homogeneity is a basic assumption of SI, the results may be expected to suffer from this problem, regardless of any further approximations made in fitting $\sigma_Z$ or applying SI.

A similar plot made using the residuals $\epsilon$ (Figure 10) after completion of IUK looks much better. The scatter is much less in absolute terms and is less even by comparison with the remaining total variability. Thus the homogeneity assump-

Figure 7. Difference between estimated mean wind with and without a data blackout at that station during the years 1987 and 1990-1992, listed in decreasing order of data available without the blackout. Only those stations missing fewer than half their data are included in the test.

![Graph showing difference between estimated mean wind with and without a data blackout.](image)

Figure 8. Maximum absolute error among the first six stations from the station degradation test, as a function of how many empirical orthogonal functions (EOFs) are retained.

![Graph showing maximum absolute error among the first six stations.](image)

Figure 9. Covariance of $u$ between station pairs, plotted versus transformed distance $\sqrt{(d\bar{x})^2 + (d\bar{y})^2}$ between pairs, where $d$ is the ratio between the best fit decorrelation scales in the $x$ and $y$ directions.

![Graph showing covariance of $u$ between station pairs.](image)

5. Application to Tropical Winds

Here a few results are briefly presented to give a flavor of what signals can and cannot be obtained from the operational rawinsonde network. These results are derived from the full data set, implementing IUK as before but retaining six EOFs. More results will be presented elsewhere.
First, Figure 11 shows the leading two EOFs obtained for the 50 hPa wind field in JJA. The first of these is clearly the quasi-biennial oscillation, which comes into phase with the seasonal cycle through most of the record (this is also true in DJF, not shown). The second mode is part of the seasonal cycle, in which the climatological easterlies near 20°N accelerate during June and early July. This acceleration appears robust from year to year. The spatial consistency of these modes (and the others, not shown) reassures us that they constitute reasonably smooth basis functions.

Previous work [Pawson and Fiorino, 1999] has shown that operational reanalyses of meridional velocities in the lower stratosphere are inconsistent. Here we may obtain estimates of meridional velocity that are unaffected by either forecast model biases or clumsy assimilation of satellite data. A sample of this is shown in Figure 12 at two levels, along the 20°N parallel (chosen because a good quantity of stations are available near this latitude).

Clear problems are evident in both 50 hPa analyses, especially those of the United Kingdom Meteorological Office (UKMO). Interestingly, this analysis agrees well with Hawaii and Wake island data points (indicated on either side of the date line) but swings wildly between them like a curve fit gone awry. Agreement is poor near islands that report less often, suggesting that these stations are undervalued in climate signals estimated from sequential maps (“interpolation bias”). The National Centers for Environmental Prediction (NCEP) analysis also shows excessive variability relative to the data and significant disagreements. The 100 hPa analyses are performing much better and generally agree with each other and the data except for an ~1 m s⁻¹ underestimate of the westerlies over Indonesia.

The cross-equatorial velocity was specifically examined by Pawson and Fiorino [1999, Figure 9] and found to be highly problematic even at 100 hPa. This is because $v$ becomes small near the equator, and there are relatively few radiosonde stations there. An equivalent to their figure appears in Figure 13. It demonstrates two things. First, each analysis shows discrepancies from radiosonde “truth” which are significant at some longitudes (though only marginally so in the case of UKMO), and each also appears to show an offset (further examination shows that the European Centre for Medium-Range Weather Forecasting (ECMWF) offset varies dramatically from year to year). The NCEP $v$ is almost certainly too high near Indonesia, supporting the suggestion of Pawson and Fiorino [1999] that in this respect at least the NCEP product was more accurate before it began assimilating satellite retrievals. Second, the uncertainty of our direct estimates becomes too large in much of the tropics to be of any use, and accuracy of this signal is clearly much more limited by spatial than by temporal sampling. Accurate zonal means of $v$ near the equator must therefore inevitably rely on the quality of forecasts from a physical model (or new instruments) since there is simply not enough coverage from the present radiosonde network to determine the behavior.
Evidently, however, in data-rich regions, IUK analysis can produce better estimates of at least some climate signals than can be obtained from present operational products. Sherwood [2000a] inferred downward motion at the tropopause over Indonesia, not present in the NCEP reanalyses, by applying the kinematic method to IUK wind estimates.

6. Instrument Changes

Thus far we have concentrated mainly on data availability problems, except to note that bad data could be more effectively screened through the use of a model $\mu$ than without (in principle, operational centers should be able to do even better if analysis increments are much smaller than natural variability). However, in the introduction the importance of instrument heterogeneity problems was noted.

In fact, one Australian station used here was relocated slightly during the period, and its World Meteorological Organisation (WMO) station number changed. It was treated as two separate stations. Even though they do not overlap, information is available on their relative behavior through other nearby stations that overlap with both of them in time. IUK thereby produces ML estimates of the signal shifts due to relocation, given all the available data. The same goes for stations that change instruments or sensors (a problem particularly for thermodynamic variables, Elliot and Gaffen [1991]). The only requirement is that the analyst know about these changes in advance so that the data can be appropriately treated.

If such changes are undocumented, objective methods for detecting “change points” or “level shifts” in a time series are required. A discussion of these is beyond our scope here. However, any such method should be more powerful when applied to $\epsilon$ than to $Z$, for the same reason that outlier points can be detected more easily. A subjective look at the present data set did not reveal any obvious examples of level shifts.

7. Conclusion

Uneven data availability patterns can introduce errors into statistical climate signals or climatologies if the former are correlated with natural variations in the field, or if instruments differ. Correlations can arise either accidentally or due to systematic sampling preferences. Standard data analysis techniques can only go part way toward eliminating this problem, hints of which are found even in operational analyses. The IUK method outlined here can considerably reduce the problem relative to other approaches and (even more importantly) allow quantitative assessment of the remaining uncertainty in a given climate signal. Tests on simulated and actual data sets showed that inaccurate estimation of the mean (which is assumed known in many approaches) can also contaminate estimates of variability. IUK signal estimation proved superior to methods based on synoptic statistical analysis of data (which often produced biased results) or “naive” methods (which suffered from much larger random

---

**Figure 12.** Mean meridional wind near 20°N from National Centers for Environmental Prediction (NCEP) reanalyses (dashed), United Kingdom Meteorological Office (UKMO) analyses (solid), and sondes (symbols) in DJF at (a) 50 and (b) 100 hPa. Error bars are determined using bootstrap sample variance as described by Sherwood [2000b].

---

**Figure 13.** Mean meridional wind on the equator from NCEP (dashed), UKMO (solid), ECMWF (European Centre for Medium-Range Weather Forecasting, small diamonds), and sonde analysis (dotted) in DJF at 100 hPa. One-sigma uncertainty envelope of sonde analysis indicated by shaded area.
errors than IUK).

IUK achieves its accuracy in three ways: (1) it relieves mapping techniques of the burden of representing large-scale variations and those with heterogeneous statistics, which are poorly captured by autocorrelation models, instead mapping only a residual field which is much better behaved; (2) it exploits temporal and spatial coherence in the data; and (3) it iteratively converges to maximum-likelihood estimates. The method permits the “signal- mapping” approach [Sherwood, 2000b] to be implemented for a wide class of signals, regardless of the amount of data missing and without having to fit the data to geometric or model-generated functions (though this is possible if preferred). Unlike synoptic or binning methods, IUK exploits statistically the sampling pattern of the observing network, using historical patterns to fill in large data gaps.

Instrumentation changes or site relocations can be handled by treatment as two separate stations that only observe a portion of the record each. This resolves, in principle, a serious handicap to detecting interannual and longer-term climate changes given changing instrumentation (although further work will be required to accommodate satellite platforms). Inability to address this problem is a major weakness of current operational reanalyses with regard to long-term climate variations.

IUK estimation of tropical meridional winds directly from radiosondes demonstrated both strengths and weaknesses relative to operational products. The former were clearly hampered by sparse coverage, to an extent that good operational-style analyses will ultimately be mandatory for an accurate knowledge of equatorial mean $\eta$. On the other hand, the presently available operational products evidently suffer from biases (including those related to data availability patterns) which are prohibitively large for accurate knowledge of this signal, and larger than the IUK uncertainties in well-instrumented areas.

In summary, the IUK method offers data analysts one relatively straightforward way (on those occasions where operational products may be inadequate or unavailable) of using data directly and of knowing the limits of the data. When data coverage is lumpy, IUK or some other method is essential to avoid signal contamination. An area for future work is to develop an analogous approach for moving instrumentation such as satellites and ships of opportunity.

Acknowledgments. Many thanks to Steve Pawson for providing the ECMWF data and helpful discussions, and to Stephen Cohn for comments on a preliminary version of the manuscript. This work was performed under USRA contract NAS5-32484.

References


S. Sherwood, Yale University, P. O. Box 208109, New Haven, CT 06520-8109 (ssherwood@alum.mit.edu)
February 8, 2000; revised August 8, 2000; accepted September 9, 2000.

This preprint was prepared with AGU’s IdTEX macros v4, with the extension package ’AGU++’ by P. W. Daly, version 1.5e from 1997/11/18.