Simultaneous detection of climate change and observing biases in a network with incomplete sampling

Steven C. Sherwood

Yale University, New Haven, CT 06520 (ssherwood@alum.mit.edu)

ABSTRACT

All instrumental climate records are affected by instrumentation changes and variations in sampling over time. While much attention has been paid to the problem of detecting "change points" in time series, little has been paid to the statistical properties of climate signals one ends up with after adjusting ("homogenizing") the data—or to the effects of the irregular sampling and serial correlation exhibited by real climate records. These issues were examined here by simulating multi-station datasets. Simple homogenization methods, which remove apparent artifacts and then calculate trends, tended to remove some of the real signal. That problem became severe when change-point times were not known *a priori*, leading to significant underestimation of real and/or artificial trends. A key cause is false detection of change points, even with nominally strict significance testing, due to serial correlation in the data. One conclusion is that trends in previously homogenized radiosonde datasets should be viewed with caution.

Two-phase regression reduced, but did not resolve this problem. A new approach is proposed in which trends, change-points, and natural variability are estimated simultaneously. This is accomplished here for the case of incomplete data from a fixed station network by an adaptation of the "Iterative Universal Kriging" method of Sherwood (2000b), which converges to maximum-likelihood parameters by iterative imputation of missing values. With careful implementation this method's trend estimates had low random errors and were nearly unbiased in these tests. It is argued that error-free detection of change points is neither realistic nor necessary, and that success should be measured instead by the integrity of climate signals.

1. Introduction

Some of the most important observing systems for detecting climate changes are networks of fixed observing stations, including surface meteorological and radiosonde stations and buoy networks. Climate changes in such networks can be affected by at least three types of spurious variation. First, instruments (or the way their readings are interpreted or stored) can change over time, altering the relationship between reported and actual values of the observable. Second, the spatial sampling pattern, and therefore the sampling bias, can change over time due to the new arrival, termination, temporary suspensions, or low reporting rate of some stations. Finally, the station network may inadequately span the desired domain (for example leaving high mountain ranges, remote rain forests, or parts of the Southern Ocean under- or unrepresented) adding uncertainty to estimates of global mean quantities. Only the first two problems are addressed here (a strategy for mitigating the last problem was given by Sherwood 2000a).

The attempted removal from a climate record of instrument-change and sampling problems is known as homogenization (see Free et al. 2002; Peterson 1998). While surface networks appear to have been homogenized relatively successfully at least for purposes of signals on the largest length and time scales (Parker 2004; Peterson 2003), the radiosonde record appears to be more problematic. Sampling is poorer than at the surface. Multiple instrument design and other changes have occurred at most if not all stations, leading to changes in observing bias for both temperature and humidity (Lanzante et al. 2003; Parker and Cox 1995). Each of these problems appears to be worst in the Tropics, which also happens to be where reported trends disagree most strongly with those expected from surface data and physical arguments (e.g. Santer et al. 2005). Reported trends among individual stations show unrealistic scatter, suggesting that differences are dominated by sampling and/or instrumental artifacts (Free and Seidel 2005).

Several temporal sampling issues may affect trends in any fixed network. All three networks cited above started small and grew (though recently radiosonde station numbers have been dropping again). These changes pose the problem of how to make use of the higher data density in "good" times without introducing spurious changes relative to other times. This problem is often addressed either by ignoring it or discarding stations spanning too short a time period. A more sophisticated approach is to determine basis

Corresponding author address:
S. Sherwood, Yale University, New Haven, CT 06520 (ssherwood@alum.mit.edu)

functions (orthogonal spatial patterns that capture the natural variability) from a portion of the record (or "reference period") when data are complete and project the more limited data at other times onto them (e.g., Mann et al. 1999), but this requires that such a period exist. Many radiosonde stations, particularly those in the Tropics, report only sporadically and/or experience long hiatuses in operation. In some cases sampling is even weather-dependent. These irregularities cause any temporally smoothed climate statistic to behave heteroscedastically (possess non-stationary variance), posing a data-analysis challenge and interfering with attempts to remove the instrument-related changes.

The present work is motivated by the goal of attaining reliable trends at as many station sites as possible in a fixed network, in a manner resistant to the above problems. Tests here are motivated by difficulties known to afflict the radiosonde record, whose climate signals are particularly uncertain and currently debated (e.g. Christy and Spencer 2005; Sherwood et al. 2005). Specifically, we assume that only a limited network of stations is available with no additional "buddy-check" opportunities outside the network; that none of these stations can safely be assumed free of heterogeneities; that biases of similar sign may be ubiquitous across the network; that at many stations a large percentage of the nominal sampling opportunities are unrealized; and that different stations span different time periods. A method is presented which, though motivated by the radiosonde situation, is applicable to any fixed network that samples regularly and suffers any of the same problems.

A considerable climatic and statistical literature exists on the problem of detecting undocumented "change points," or discontinuities in the statistics of a time series (see Menne and Williams 2005, and references therein). Climate-relevant changes are usually modeled as simple step discontinuities in observing bias, due e.g., to changed sensor design, relocation of the sensor, etc., and are thereby distinguishable—at least in principle—from the relatively smooth variation of the underlying observable. Detection of the change point is followed by estimation (and ultimately, removal) of its associated level shift.

These studies have left key issues unresolved. First, detection methods typically assume that the observations possess little or no serial correlation, but real climate records contain variability on all time scales. This makes false detections more likely since the natural variability begins to resemble the artifacts. Second, the goal is usually not detection *per se* but accurate climate signals, yet previous studies have not carefully investigated to what extent that actually occurs. A tendency has been noted for radiosonde temperature trends to disappear upon homogenization (Free et al. 2002). Finally, while the value of using data from neighboring sites is well-recognized for level-shift estimation (e.g. Karl and Williams 1987), detection studies have dwelled on the case of an isolated time series; the use of neighbor information remains ad-hoc in practice, and its efficacy untested.

Several groups have recently tried to produce homogenized radiosonde records. Trends in the unhomogenized tropospheric data were small since 1979, and the studies found little increase upon homogenization. All used external information—from other stations, a separate observing system, or a model forecast—to help anticipate genuine fluctuations at a particular station and thereby aid detection of artifacts. Lanzante et al. (2003) embarked on a painstaking, subjective analysis of 87 individual station time series through 1997, often using neighbor time series and/or climate indices such as the ENSO index to help identify natural variability. The resulting records were used by Thorne et al. (2005) as a backbone to quantify natural variability and aid detection at a much larger number of stations. Haimberger (2005) made similar use of forecasts from a forecast model driven by many observing systems. Others (e.g., Christy et al. Submitted 2005) have used satellite data. No objective, end-to-end tests of the methods used have yet been published documenting their estimation properties.

It is likely that pervasive artifacts remain in these records (Sherwood et al. 2005). One problem is that reference information used for anticipating genuine fluctuations was itself not free of artifacts. Also, the above studies first looked for suspicious changes, made adjustments, and then calculated trends or other signals from the adjusted data. This process is analogous to sequential, univariate regressions for elucidating multiple effects, which produces optimal results only if the effects are orthogonal.

I will demonstrate here that each of these problems is significant. Clear benefits accrue from approaches in which the data are fitted simultaneously, rather than sequentially, to a model that includes real and artificial effects. Techniques that try to do this are available in the literature, but have not been widely adopted, and anyway do not solve the problem. I will present a more successful (though more complicated) approach: an extension of a method called Iterative Universal Kriging (IUK), originally proposed for coping with the incomplete sampling problem, to cases where data are also heterogeneous. Its performance on simulated data will be compared to that of two simple methods proposed as benchmarks, and its advantages for the radiosonde problem will be discussed. Deployment of the method on genuine data will be described in a subsequent publication.

2. Homogenization by Iterative Method

Two similar iterative methods were recently and independently developed (Schneider 2001; Sherwood 2000b) to analyze incomplete data without relying on a "reference period." Each is an implementation of the EM (Expectation-Maximization) algorithm (Dempster et al. 1977), in which missing values are imputed based on the currently estimated model parameters, model parameters are re-estimated based on the imputed plus observed values, and the procedure is iterated to the desired accuracy. Both approaches employed empirical orthogonal modes determined from the data as the basic model of natural variability. Schneider (2001) estimated a maximum-likelihood covariance matrix directly fol-

lowing Beale and Little (1975), but reduced noise more elegantly by regularizing the matrix, while Sherwood (2000b, hereafter, S00) chose a more standard truncation that was not connected to a maximum-likelihood matrix (a drawback discussed in Section c). On the other hand, S00 did not rely completely on a normal model but instead employed Universal Kriging, in which field variability is decomposed into a spatially coherent part (represented by the principal components of a normal model) plus locally random fluctuations with limited spatio-temporal coherence represented by a Gaussian random field.

One advantage of the latter approach is that it makes use of serial correlation to help impute missing values rather than treating each discrete time as an independent realization, which is helpful in datasets with sporadic reporting. S00 was able to apply this approach to instantaneous instrumental data rather than averages from spatiotemporal bins (e.g., 2 $\times$ 2 degree by one month). This confers an advantage since individual observations can reasonably be regarded as homoscedastic, but binned averages are heteroscedastic when the sampling varies from one bin to the next, a challenging problem not confronted by Schneider (2001) nor considered by most recent data analyses. The assumption in S00 of spatially homogeneous Gaussian random field statistics could prove problematic for surface data from heterogeneous regions. In summary, the methods of Sherwood and of Schneider were similar but optimized for different situations. Here, we adopt the IUK approach.

a. Review of the "IUK" method

The IUK method was based on "universal kriging" (e.g., Cressie 1993), or the representation of the data as a parametric model plus a random process,

$$Z = \mu + \epsilon, \tag{1}$$

where each of these quantities is a function of space and time. The parametric model is a linear superposition of basis functions:

$$\mu(s,t) = \sum_{i=1}^{m} a_i f_i(s,t) + \sum_{i=1}^{n} b_i g_i(s,t). \tag{2}$$

The variables s and t in (2) 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 $\mathbf{f}$ represent the desired signal patterns and should be specified 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 additional functions $\mathbf{g}$, which may be determined empirically or a priori, represent variability patterns that are coherent over multiple stations and presumably change relatively smoothly in time. S00 and the present work use principal components ("PC"s) or "empirical orthogonal functions" of time for $\mathbf{g}$. The signal amplitudes $\mathbf{a}$ and others $\mathbf{b}$ are then determined from the data. In S00's application of the method to seasonal wind data in the tropical upper troposphere and lower stratosphere, six empirical functions $\mathbf{g}$ were retained; the first corresponded to the quasi-biennial oscillation, the second to a residual seasonal cycle.

Less coherent variability is represented in (1) by the field ϵ, which was modeled as a Gaussian random field having a heterogeneous and stationary autocovariance function $\sigma_\epsilon(d\mathbf{x}, dt)$ (e.g., Daley 1991) characterized by four independent parameters. This partitioning is a generalization of both standard regression approaches (such as Schneider (2001) or Mann et al. (1999)) where the second term is simply treated as white noise, and statistical interpolation techniques (e.g., Vinnikov et al. 1990) where the first term is just the mean (time-average) field and the second term models all other variability. S00 showed that the ϵ term obeyed the assumption of homogeneous, Gaussian statistics far better than did the original field Z in a radiosonde wind dataset. Also, bad data points can be identified more easily through extreme ϵ values than through extreme Z values.

As with any EM implementation, S00 assumed that the analyst has available a straightforward method (e.g., linear regression) for obtaining the ML estimate of $\mathbf{a}$ from a complete set of data. They also assumed that the data are not in fact complete, with only a subset of the variables $\mathbf{Z}$ actually observed. The EM process iterates between imputation of the missing values and estimation of $\mathbf{a}$ as follows:

1. from a simple initial guess for μ (which can be evaluated for all s and t), obtain residuals ϵ for all observations using (1);
2. Obtain kriging parameters (ranges in each dimension, "nugget" variance and "sill" variance) describing ϵ by fitting to observed residuals;
3. Impute unobserved ϵ using kriging and Z computed from (1);
4. Refit the model μ by regressing the complete set of Z (including imputed values). Go back to step 1.

This leads to maximum-likelihood estimates of $\mathbf{a}$,$\mathbf{b}$ given the incomplete data, the basis functions, and the four kriging parameters.

b. Representing change points in IUK

As S00 noted briefly, extension of μ to represent instrument heterogeneities is straightforward in principle if the times of change are known *a priori*. One may specify a separate basis function f for each known change point at each station, equal to zero everywhere except at that station prior to the change point where it equals -1. When IUK is complete, the function's corresponding fitted coefficient a gives the best estimate of the change in mean value or observation bias, and the retrieved mean value for Z is valid for the final homogeneous data segment at that station. A reconstruction

of the data with the step components of μ omitted yields the "homogenized" dataset with all means adjusted to match those at the end of the analysis period. Note however that if the relationship of a station to others is expected to change significantly—for example if the station is moved to a distant site with different meteorology—then it may be better instead to consider the records before and after as two separate stations. This is easy to do but increases the number of model parameters, decreases the usefulness of the station, and increases the number of missing data points that must be imputed, so it should be done sparingly.

Our chief goal is to estimate long-term changes correctly in the presence of both change points and natural variability. To be successful, we must represent each effect in the model (1). Thus, in addition to the mean value and step functions I include among **f** a linear function of time for each station. Note that some of the trend at a station will be contributed by that of **g**; the inclusion of a linear term in **f** merely ensures that all of the trend will be represented whether or not it projects onto the natural variability. The "true" trend is recovered by adding the contributions from all basis functions.

I have thus represented each mode of natural variability as a time series (analogous to the ENSO index) specified by one basis function g per station, with independent loadings b (such a mode series may, but need not, be identical at each station). This enables the regression to be performed separately at each station during step 4 (the "E" step) of the algorithm, since each fitted parameter affects only one station. This is not a formal requirement—one could, for example, fit temporal loadings to spatial patterns rather than spatial loadings to temporal ones—but without this separability the regressions would have to be done simultaneously at all stations at times, which would be impractical for large datasets.

c. Issues in empirically estimating **g**

A possible weakness of this procedure is that the maximum-likelihood property of the resulting solution **a** holds only with respect to known or assumed natural basis **g** and Kriging parameters. If these must also be estimated from the data, a good overall model is not guaranteed. Tests indicate that while the Kriging parameters are not a problem, estimation of **g** is more delicate. Uncontaminated representation of genuine variability is indeed the core challenge faced by all methods.

Following S00 I obtain **g** by principal components analysis of the infilled dataset. One would like to begin with a dataset already free of artifacts, since if the basis functions capture artificial behavior this will compromise the separation effort. One thus has a chicken-and-egg problem. As with the missing-data imputation problem, this can be resolved iteratively, in particular by inserting an extra step into the IUK procedural loop.

I initialize by fitting the mean, shift, and trend components of μ (but without any **g**) to the available data via multiple linear regression. The first three IUK steps, described in Section a, are then executed. Because the initial μ had only the **f** components, ϵ initially assumes most of the variability, giving it much greater variance and longer "range" parameters than it will have on subsequent iterations.

An extra "basis estimation" step, having two parts, is now added after Step 3. First I refit the model μ with only the **f** basis—as was done initially, but now to the infilled data. Second, I perform PC analysis on the "natural variability" obtained by subtracting this fitted, **f**-only μ from the infilled data. The leading principal component time series are retained as **g**. This extra step is repeated only for a specified number of iterations, after which the model for μ is "frozen" to allow proper convergence to the solution associated with that basis. We retain the number of basis iterations as a parameter, N_g. While the method did not prove absolutely convergent, measures were devised to prevent divergent behavior as described in the Appendix.

d. Initial tests and modifications

Preliminary tests and modifications to the IUK procedure are described in the Appendix. It was determined that the basis functions used for IUK should be computed separately for each station using only the data from other stations, rather than using universal functions of time. A stopping criterion was also developed for deciding when to stop re-estimating this basis. Tests indicated, first, that the inability of the method to converge to a true max-likelihood fit to the data when natural variability basis functions are not known did not appear to be a significant shortcoming; and second, that iteration becomes increasingly beneficial as more artifacts are present in the data.

e. Application to radiosonde data

As mentioned above, Z can be a scalar or vector. The radiosondes motivating the present study measure four vectors: temperature, humidity and two wind components each as a function of pressure. Wind data show promise in constraining slow temperature changes, since the two are geostrophically balanced (Allen and Sherwood submitted 2006). This would not work well for stations in isolation, since winds are not related to temperature locally but to its horizontal gradient. But IUK can exploit winds by incorporating them into the basis **g** across the network—using them to help identify the real variability from which artifacts must be differentiated. Similar benefits could apply to surface temperature data through use of surface pressure. In principle Z could also represent information from all levels in a radiosonde ascent, although this would probably not be advantageous since in the vertical direction (unlike the horizontal) inhomogeneities are more coherent than are natural variations.

For simplicity, the tests below will be restricted to performance on a scalar quantity. A subsequent paper will describe more details on application of IUK to radiosonde data, and results.

3. Testing homogenization methods

a. Design of simulated datasets

For testing, I generated a large ensemble (500 trials) of simulated radiosonde datasets with known properties. Each trial consisted of five stations collecting data over a fixed period. The (dimensionless) station longitudes were 5,4,0,5,5, and latitudes were 10,9,4,4,0, giving the stations a spatial arrangement like the numeral "4."

The time period contained 365 sample opportunities. One can think of this as a year of daily data or approximately 30 years of monthly data; for discussion purposes I will call the sample interval a "day." The synthetic Z (also unitless) were generated by adding a series of terms: a mean value equal to the station number (1-5), a linear trend, two modes of large-scale variability, and small-scale variability. The linear trend component was chosen randomly for each station and trial, so as to contribute a total change between the beginning and end of the time series whose absolute value was uniformly distributed between zero and 1.0. In some tests the sign was always positive, while in others it had a 50/50 chance of either sign.

The two large-scale variability modes were for simplicity made proportional to longitude and latitude respectively, and the amplitude of each mode varied (independently) in time according to a smoothed random process (Gaussian white noise smoothed with a 7-day running boxcar mean). Because all station latitudes and longitudes were nonnegative, both modes produce variability that is positively correlated among all stations, but with unequal variance among the stations. The small-scale variability was first generated on a uniform grid by smoothing spatiotemporal white noise by a uniform 3x3x3 (longitude/latitude/day) averaging kernel, then sampled at the station locations. The standard deviation of the variability associated with each of the two modes, over all stations, was about 0.35. That of the small-scale variability was about the same.

This fully established the simulated "true" values of the observable, from which I then generated "observations" (see Fig. 1). First, some data points were randomly tagged as being missing. The percentage missing varied linearly from 10% at Station 1 to 90% at Station 5. Next, step discontinuities of amplitude 1.0 (twice the size of the average total change due to the linear trend) were added at randomly selected times; in some tests these were all upward, while in others each step had a 50/50 chance of being upward or downward. The standard case included one change point per station (five in total); an easier case was included with change points only at stations #2 and 4, (two in total), and a harder one with two at each of these stations and one at each of the others (seven in total). The time of each change point was chosen randomly, but was not allowed to be less than 30 days away from another change point or from the beginning or end of the record. No additional "noise" per se was added to the observations, but the added small-scale variability can be thought of as the combination of instrumental and "natural" noise. Tests showed that this noise has very little impact on results.

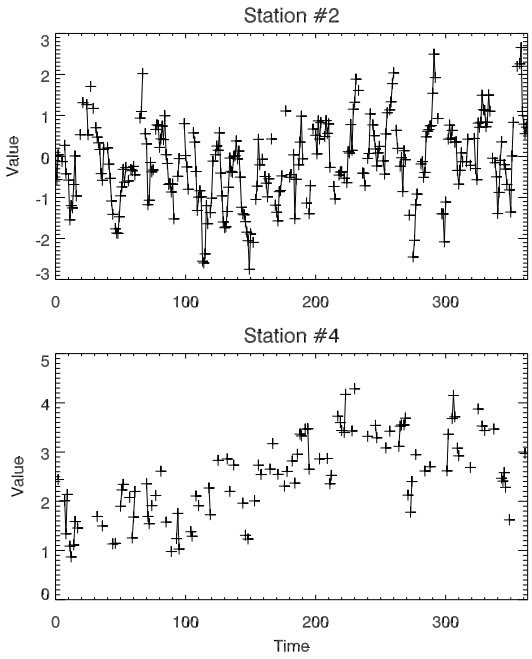

FIG. 1. Simulated observations at Stations 2 and 4 for the first randomly generated trial out of 500. Artificial step increases of 1.0 are present at times 328 and 121, respectively.

This scheme does not mimic all possible problems. For example, no attempt was made to produce lengthy dropouts of data at stations (which S00 showed were treated relatively well by IUK). I did not include blunders (large errors) because testing showed that reasonable blunder rates did not effect results with even crude quality control measures. I did not allow change points to be too close together or non-step-like even though both complications will occur in reality. It is unlikely that any statistical approach could detect such difficult events and would overfit the data if it tried to. We must hope that more pernicious artifacts are not too common in actual data.

In assuming only linear trends plus red noise, no explicit attempt was made to build "climate shifts" into the tests. Nonetheless the simulated time series often showed them—the change evident about halfway through the record at Station #4 (Fig. 1), for example. More vivid examples occurred frequently enough. In most observed datasets, apart from clear volcanic impacts or other events, it would be difficult or impossible to reject the linear-trend-plus-red-noise model used here, although the required model parameters would clearly vary. One could criticize our red-noise process for having too little long-term memory, but this is compensated by the short duration of the test series.

In summary, we now have a set of "true" values and an incomplete, "observed" subset that includes spurious level shifts. The performance of any method can be assessed by comparing climate changes estimated from the simulated

observations to those of the simulated "truth."

b. Benchmark methods

In addition to IUK I tested two simple benchmarks, each implementable with different parameters. These are not meant to be state-of-the-art, but rather, methods that are easy to understand and reproduce, and are similar to those used in past work. Their behavior will help illustrate potential pitfalls, but may not be quantitatively representative of past efforts.

1) SINGLE-STATION METHOD (SS)

This method estimates level shifts at each station by considering only the station's own data. Three weighting strategies were tried. In the "global" version (SS-G), all data before or after a change point (but not going past another change point) were used with equal weight; the shift was estimated as the difference between means before and after. Usually, however, analysts use only those data within a certain time interval of the change point, which is also tested here as the standard version (SS). The weighting function was maximal at the change point and decreased linearly to zero by 50 days before or after this. Though other window widths could be considered, comparison of this example and the "global" alternative illustrates the basic properties and likely range of results. Finally, we consider applying a restricted two-phase linear regression to the data following Wang (2003), which allows for an overall trend and a step at the change point, designated (SS-2). The standard SS tested here is close to the methods used by Lanzante et al. (2003), Haimberger (2005), and Christy et al. (Submitted 2005) except that in the latter two cases, shifts were detected in a difference series between radiosonde and some other source of temperature information.

2) REFERENCE SERIES METHOD (R1, R2)

Several previous radiosonde homogenization efforts (Haimberger 2005; Lanzante et al. 2003; Thorne et al. 2005) have used information from neighboring stations to help identify and quantify discontinuities, typically by generating a "reference" time series for that station with which to compare its own data. It would not be possible to explore all ways of doing this, but I tested one simple procedure. First, each station's sample mean was subtracted from all its observations to generate anomalies. A reference anomaly series was then generated for each station by averaging the available anomalies from other stations at each time (on rare occasions when no other stations collected data, the reference anomaly was set to zero). Level shifts were estimated as with SS (using standard and "global" weighting) but using differences between station and reference anomalies. Since the reference series themselves would be contaminated by heterogeneities, I also tried a second iteration initiated using the homogeneity-adjusted data from the first round, repeating both change point detection (when not known *a priori*) and level shift estimation. The results from the first and second iterations are denoted R1 and R2.

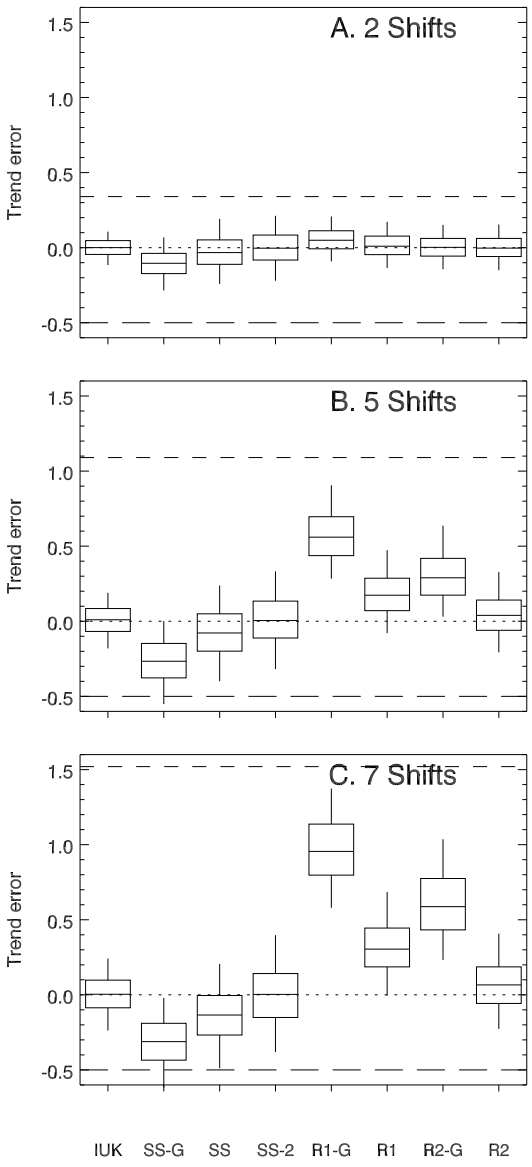

FIG. 2. Box/whisker plot showing median, 25th/75th percentile (box), and 5th/95th percentile (whisker) error in trends estimated by each method. The short- and long-dashed lines show results with no homogenization, and with all trend estimates equal to zero, respectively. IUK method includes two retained PC's, unique to each station, with $N_g = 4$.

4. Scenario I: Known change points

We first explore performance when change point times are all known in advance. In these tests trends and level shifts are always upward, to reveal biases associated with the imperfect separation of the two effects. Section 5 will explore the more important situation where change points must be identified from the data.

Fig. 2 compares the performance of all methods in estimating trends. For each method, the median (shown) is very nearly equal to the mean (not shown), and indicates the bias (if any) of the method in the test situation. All benchmark methods except SS-2 are biased, and all have random errors greater than that of IUK. However, R2 is nearly as good on both counts as IUK.

SS and SS-G trends were biased low for a straightforward reason: true underlying trends were aliased onto estimated level shifts, causing the latter to be overestimated. Not surprisingly this bias was worst with the "global" version. Recent efforts by the Hadley Centre have also encountered this (P. Thorne, personal communication).

The bias disappeared completely using two-phase regression, illustrating the importance of simultaneously fitting the data to a model that includes both real and artificial effects. However, while biases were reduced going from SS-G to SS to SS-2, random errors increased; they were significantly higher for SS-2 than for IUK. This shows the benefit of fitting (but not overfitting) the data to a more complete model of real effects.

The reference methods exhibited low bias for a more subtle reason. Tests (not shown) confirmed that the bias originated from the systematic nature of the level shifts. On average, reference time series were contaminated by spurious increases over time, reducing the apparent increases detected from difference series. This bias was again much worse for the global versions. It rapidly increased with the number of change points, indicating that reference methods must be implemented very carefully on large networks. These biases would presumably disappear if separate reference series were generated for each change point at a station using only homogeneous portions of neighbor records, following Karl and Williams (1987). That strategy, however, becomes difficult to implement for radiosonde data where many far-flung neighbors are often required for generating useful reference series (Thorne et al. 2005), and is unhelpful for detecting change points in the first place (see next section).

We may conclude from these tests, however, that simultaneously fitting data to a step discontinuity and a trend is necessary for unbiased trend estimation, and that if not doing this, one should at least estimate level shifts using data from a restricted neighborhood around the change point (as is, in fact, commonly done) to minimize biases. The same presumably goes for other variations besides linear trends. Given this, performance was not too bad overall. "Global" methods will not be considered further here.

5. Scenario II: Unknown change points

For this, more challenging and realistic scenario we need two performance tests. In the "detection test," simulated trends are half upward and half downward among the trials but level shifts are always upward; in the "confounding test," level shifts are half $+1$ and half -1 while trends are always upward. In the detection test, estimation bias indicates over- or under-correction of systematic artifacts, and will be called "detection bias"[1]. Any bias in the confounding test reveals that genuine trends are being mistaken as artifacts, and will be called "confounding bias." Both biases must be small if a method is to be useful in practice.

a. Detection method

1) BASIC ALGORITHM

A number of methods have been proposed for detecting change points. Their relative skill depends on one's measure of success and overall winners do not clearly emerge (DeGaetano 2006); performance is probably best if consensus is sought among more than one method (Menne and Williams 2005). Multiple change points were detected here using two methods. First, we applied the iterative nonparametric method of Lanzante (1996, hereafter L96), based on the Mann-Wilcoxon-Whitney test, to individual station time series. L96 reported favorable performance compared to previous methods including that of Easterling and Peterson (1995), which was based on two-phase regression. One advantage of a nonparametric method for our tests is that it yields a more realistic assessment, since real data may not obey the assumptions of parametric methods.

On the other hand, L96 did not control for background trends. Insofar as change-point detection requires quantifying a putative level shift, one may expect the issues raised in the previous section to apply also to detection. The reference methods and IUK provide opportunities for natural variability and trends to be removed from the record prior to detection: for the reference methods shifts were be detected in the anomaly series, while for IUK the method was run once with single-station shifts, then shifts were detected from data with the natural variability and trend removed (leaving in the previously estimated shifts plus ϵ), and the method was run again.

Two-phase regression may also be used to detect change points in the presence of linear trends, and can be combined with the above strategies. This method relies on an F-statistic that compares the goodness of fit of a model including both a step and trend, with a reduced model considering only a step (Lund and Reeves 2002). I consider here only the restricted version where trends are assumed stationary (Wang 2003), implemented in a simple heirarchical scheme for multiple detections such as that used by L96. While heirarchical schemes are not optimal (Menne and Williams 2005), this matters only when single change point detection is more successful than we will find here.

2) ADDITIONAL TESTS

Previous studies have often found uncomfortable rates of false detection. The above tests each have one adjustable

[1]Note that the chosen terminology is with respect to estimation of artifacts rather than trends, an arbitrary choice. Any overall bias in one implies an equavalent, opposite bias in the other.

parameter, a confidence threshold for deciding whether a potential change point is significant, for which I considered a wide range of values. As noted by previous authors, the assumptions underlying this confidence test are violated when more than one change point is detected or when the data exhibit serial correlation. The challenge of detecting multiple change points has received attention from the statistics community for some time (see Chib 1998; Sullivan 2002).

L96 recommended two parametric tests as expedients for weeding out false detections. The first was a "signal-to-noise" (S/N) test, where S/N is the ratio of the variance in data around the change point explained by a step function to that not explained. This tests whether a change point is sufficiently "important." L96 advocated minimum S/N thresholds in the range 0.05-0.2, but acknowledged that this depends on the dataset and cannot be rigorously established *a priori*.

The S/N test does not address the fact that sustained, natural changes can mimic a change point, and only rejects change points that will probably have little impact anyway. L96 suggested also a "step vs. trend" (S/T) test based on the ratio of variance explained by a step function to that explained by a linear trend. If this ratio falls below unity, the variability looks more like a trend than a step and the change point should be discarded. Both the S/N and S/T tests require the arbitrary specification of a time window around the candidate change point within which to apply the test.

Experimentation here indicated that the both ratios were helpful in separating "hits" from false detections, so I adopted a dual-ratio filter using both of them. A threshold value of 0.2 and time interval of 50 data points worked well for the S/N test and was used for all methods. The S/T test (whose threshold is 1 *a priori*) worked somewhat better with intervals encompassing the entire homogeneous segments on either side of the candidate point. Points not meeting both tests were filtered out. An S/N threshold > 0.2 improved IUK further, but was too strict for other methods. In practice the S/N test must be applied very conservatively, since there is no *a priori* way to establish the threshold, so it is definitely the less valuable of the two. Fortunately the S/T test was almost as good by itself as the two together.

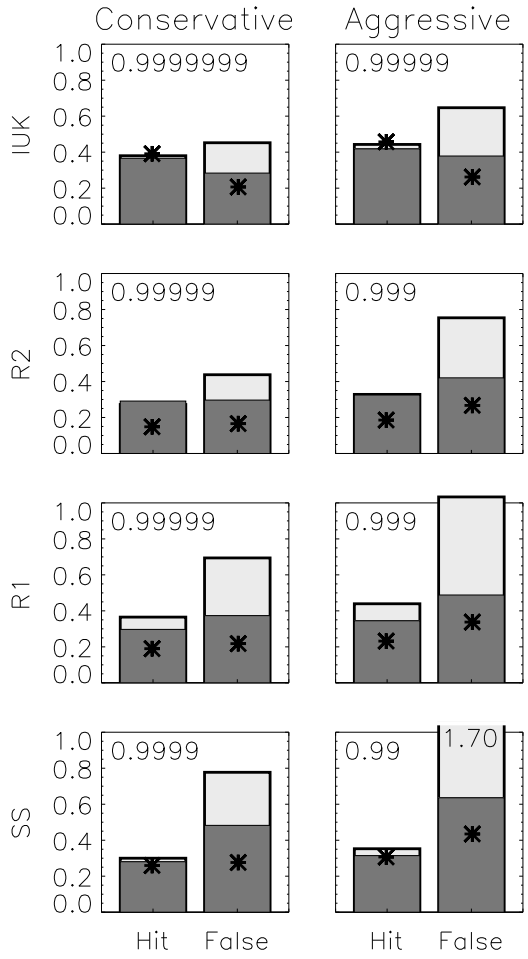

FIG. 3. The hit and false-detection rates for "aggressive" and "conservative" settings of the confidence threshold (quoted in upper left corners) using L96 with no filtering (thick outline), L96 with ratio-based filtering described in text (solid bar), and two-phase regression with ratio filtering (asterisk symbols). Results are for five change points with all trends positive; results for all shifts positive are nearly identical. The detection rates are defined as the ratio of true or false detections to the total number of discontinuities actually present. The value for one bar that goes off scale is printed.

b. Hit and false detection rates

Change-point detection success can be measured by hit and false detection (type I error) rates, defined as the number of correctly or incorrectly identified change points respectively divided by the total number of actual change points[2]. As detection becomes more conservative (confidence threshold increases toward unity), both rates decrease monotonically. It turns out that detection rates—hence method "aggressiveness"—vary considerably among methods using the same nominal confidence threshold. To compare methods at similar aggressiveness levels we choose different confidence thresholds for each, indicated in Fig. 3.

Performance was not particularly encouraging. No method achieved a hit rate above 45%, yet false detections were rampant even with conservative thresholds (as high as 0.9999999). Only the IUK method managed (with filtering) to push the false detection rate below the hit rate, although R2 came close if a conservative threshold was used. Counterintuitively, ratio filtering actually increased the hit rate for R2, as discarding bad change points on the first iteration enhanced subsequent detection. False detections always well exceeded hits when ratio filtering was not applied. High rates of false detection have also been implied by previous radiosonde homogenization comparisons (Free et al. 2002).

[2]This definition of false detection rate differs from the traditional "false alarm rate," but is more convenient for present purposes.

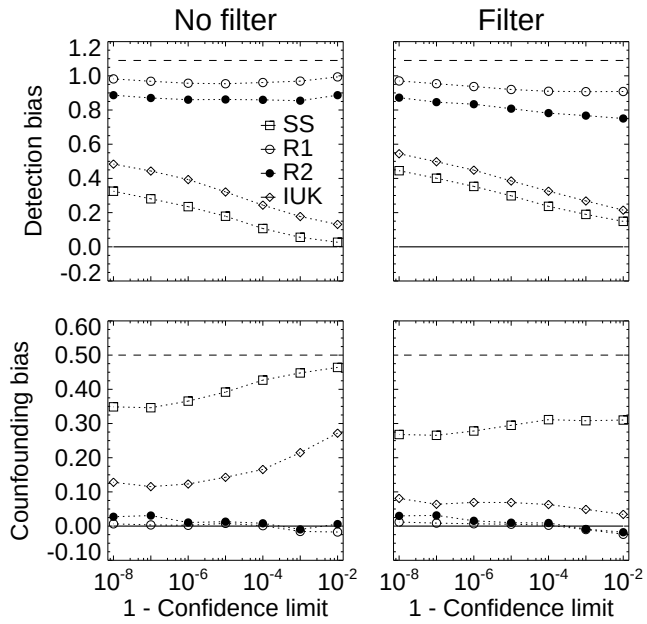

FIG. 4. Detection and confounding bias (upper and lower panels respectively) using L96 detection without and with dual-ratio filtering (left and right panels), for four methods indicated in legend, as a function of detection confidence threshold. Confounding bias has been multiplied by -1 to make it positive. Long dashes show perfect behavior, short dashes show the worst-case bias (that with no detections or with all trend removed, for detection and confounding cases respectively).

Two-phase regression produced better performance than L96 for the SS and IUK methods, reducing the number of false detections without significantly affecting hits. With reference methods, two-phase regression produced significantly fewer detections but with little impact on the percentage of these that were successful.

c. Trend errors

We now examine how accurately trends can be estimated. This may be acceptable even with many false and/or missed detections, if the accompanying level shift errors "average out." This is tested for change points detected by the L96 method in Fig. 4.

Biases appeared in all methods, including IUK, and in general were of comparable magnitude to random errors. Detection biases were uniformly positive, indicating that no method fully removed the impact of heterogeneities regardless of confidence setting. As change-point detection was made more aggressive (to the right in the plot), detection bias decreased but confounding bias increased for both the SS and IUK methods. The SS and IUK methods removed most of the artificial trend, and tests confirmed that their remaining biases were due to undetected change points, which decreased in number with increasing aggressiveness explaining this trend.

The reference method removed only about 20% of the detection bias, a result much worse than expected. Detailed inspection revealed that false detections led to large, spurious adjustments that, on average, nearly canceled out the legitimate adjustments. This was due to the corrupting influence of undetected shifts in the reference series. Iteration of the reference method (R2) did little to ameliorate this serious problem. As successful and false detections both increased with increasing detection aggressiveness, the latter had little net impact on the bias.

Confounding biases led to underestimation of trends (the degree of underestimation shown in the figure as a positive number) by SS and IUK. Tests confirmed that these biases were due mainly to false detections, onto which the genuine upward trend tended to alias and be adjusted out of the record. This explains why the biases grew worse with increasing aggressiveness, and why dual-ratio filtering reduced them (compare lower right and left panels). Confounding biases were not significantly improved by using two-phase regression to estimate level shifts for SS (SS-2), except for the "global" case which we already abandoned.

The reference methods showed no significant confounding bias. In essence, they are optimized for the confounding test because of their implicit assumption that real variability is common to all stations but level shifts are limited to the station under investigation.

For all methods, the best overall performance seems to result from adopting the most aggressive confidence threshold (0.99) and applying filtering. Performance at this threshold is summarized for both detection algorithms in Fig. 5. One might expect the above biases to abate significantly when two-phase regression is used to detect change points rather than L96, since the former produced fewer false alarms and is based on a model that includes an explicit trend effect. Surprisingly, there was no improvement overall, only a small shift from confounding to detection bias, and a slight increase in error variability compared to L96.

The reasons for this are uncertain but I would speculate as follows. Any time series will possess abrupt natural variations that, when fitted to a parametric model, look artificial even though they are not. Such "parametric spoofs" will, if identified as change points, lead to large and erroneous parametric adjustments. It is likely that the two-phase method would detect them more often than would L96, since the former employs a parametric test statistic similar to that used to adjust the data. This may counteract the benefit of a lower overall false detection rate. In any case, the result underlines the limitations of using hit/false detection rate as a performance measure.

While not vastly superior to SS, IUK did perform the best overall in terms of bias and random error. It was the only method that correctly estimated the signs of the true trend and of the artifacts each at least 90% of the time (the whiskers stayed within the two dashed lines in Fig. 5). With L96 filtered detection it successfully removed 81% of the artificial trend due to discontinuities, while sacrificing only about 7% of the genuine trend in the record; this despite

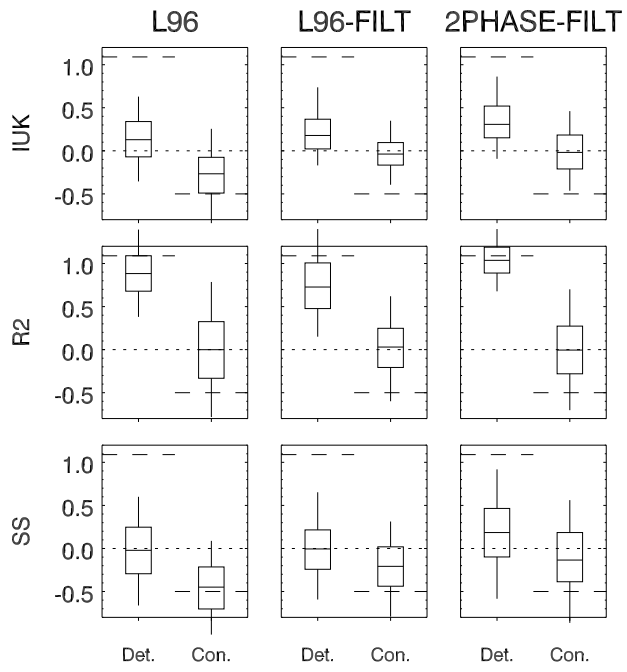

FIG. 5. Detection and confounding error distribution (left and right bars respectively) for (left to right) L96 without and with dual-ratio filtering and two-phase with, for (top to bottom) three methods. Short dashes are as in Fig. 4; boxes and whiskers are as in Fig. 2. All results employ 0.99 confidence threshold and dual ratio test.

false and missed detection rates that, while better than those of the other methods, were still high. The single-station methods tended to remove much of the real trend along with the artificial, while the reference methods removed neither. Reference-type methods could improve with more sophisticated use of neighboring stations, though the benefits of this are likely to be limited if change-point detection is as poor as in these tests.

6. Discussion and conclusions

Unlike previous studies, this investigation focused on the statistical properties of trends estimated from inhomogeneous, serially correlated and incomplete datasets with various homogenization strategies. Performance when change point times are known *a priori* was compared to that when they were not. When change point times are not known, they must be detected in the data; very agressive methods for doing this will find enough change points so that data homogenization removes real as well as artificial variability, while very cautious methods will leave most of the artifacts in place along with the genuine variations. Separation of the two is much more difficult for time series with realistic serial correlation than for the weakly or uncorrelated series typically used in tests.

I investigated several methods of quantifying level shifts—including a newly proposed method based on iterative universal kriging (IUK)—and two methods for detecting them. Results were assessed on the basis of the mean error (bias) and the spread of errors about this mean (noise). The investigation led to the following conclusions:

- True trends are biased toward zero if artificial level shifts are quantified using differences of means before and after the change point. This bias is reduced, though noise is increased, by taking means from shorter segments. Two-phase regression and IUK eliminate the bias by fitting the data to a model that accounts for underlying variability, but IUK was less noisy.

- When change points aren't known *a priori*, all methods become biased due to detection errors. Estimation biases were often comparable to the signals being sought and/or to the estimation noise. Two-phase regression improved detection statistics but not trend estimation. IUK was best able, of the methods tested, to eliminate artifacts while keeping most of the genuine trend.

- Methods where nearby stations are used to generate "reference" series can easily fail to remove artifical trends if inhomogeneities at the neighboring stations are not treated carefully. Lack of *a priori* knowledge of change point times may represent a serious impediment for this strategy.

- Confidence limits based on the null hypothesis of uncorrelated data can be extremely misleading; most change-point detections here were false, even when nominal confidence levels as high as 0.99999999 were demanded.

- Additional tests designed to differentiate between step-like and gradual changes were found to be very helpful in eliminating false detections. Even so, false detections outnumbered hits in these (fairly challenging) tests unless IUK was used in detection.

- Poor performance in detecting change points does not imply poor performance in estimating climate signals. The best results occurred from liberal detection combined with methods (primarily, IUK) that were able to assign small adjustments (on average) to those that were detected falsely.

These findings have important implications for previous homogenization efforts. Such efforts have often used nominal confidence levels to argue the legitimacy of change points, but false detections are much harder to avoid than is commonly assumed. Such false detections are a serious problem because they tend to occur at precisely the times when natural variability shows large changes, such that the spurious adjustment to the time series will also be large. IUK did not prevent false detections but, by explicitly representing real and artificial behavior, limited their damage.

No previous radiosonde homogenization effort has employed change-point test statistics or level-shift estimators

that control for background trends, such as two-phase regression. This probably caused climate trends to be underestimated. Some have used satellite data (Christy et al. Submitted 2005; Free et al. 2002) or reanalysis forecasts (Haimberger 2005) to establish reference series at station locations; in these cases, the tendency is instead to regress toward the trends of the reference dataset (which, as it happens, were also small and of uncertain accuracy). The idealized, generic tests here cannot quantitatively address the accuracy of previous studies, but would seem to indicate caution in interpreting their results.

The IUK procedure introduced here was able to overcome much of the problem. Its chief advantage lies in the fact that the data are regressed onto a model that includes both natural and artificial effects, while previous efforts have identified artifacts and true variability in separate steps. A few modifications were found necessary, the most important of which was to ensure that basis functions at each station are independent of that station's data. One should ensure that spatial loadings of these functions are well distributed over stations and appear physically reasonable. If a cluster of stations (for example, from one country) has change points all at the same time, it may be important to eliminate all stations from the cluster when computing their variability basis. Some care is also necessary to avoid overfitting but tests indicate good robustness to parameter choices that affect this. Two other benefits of IUK noted by Sherwood (2000b) include the fact that fitted trend parameters have independent errors across stations (given the basis), which is crucial if spatial smoothness of the final trend is to be used as a measure of success; and the ability of the method to cope well with difficult missing data patterns including large blocks of missing data at different stations.

A final benefit is that one can easily bring non-temperature information to bear. For example, wind fluctuations are closely related to those of temperature and their observations appear to be less inhomogeneous (Allen and Sherwood submitted 2006). Results here indicate that any independent information that helps constrain the natural variability will improve results, especially when change points are not known *a priori*.

Acknowledgments. Thanks to J. Lanzante, P. Thorne and M. McCarthy for helpful comments. This work was funded by the NOAA Climate Change Detection Program grant NA03OAR4320153.

APPENDIX

The IUK procedure was run for values of N_g (the number of basis iterations) from 1-19. For comparison, it was also run with the large-scale variability basis $\mathbf{g}$ that had actually been used to generate the true data provided *a priori*. The mean error of estimated level shifts (bias) was zero in all cases, and is not shown. Errors were approximately normally distributed and performance is measured here by the mean-squared error (MSE) of the shift estimates, shown in the top panels of Fig. A1. The bottom panels show the variance of ϵ, a diagnostic for the quality of fit.

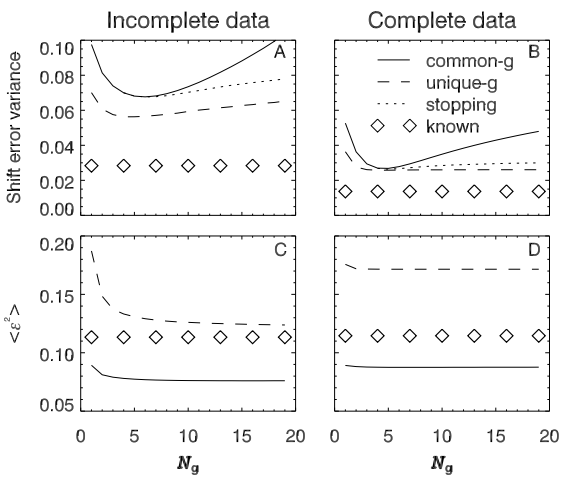

FIG. A1. The mean-squared value of the level-shift estimation error among shifts (A,B), and of ϵ among observed data (C,D), as a function of the number of iterations N_g used in estimating $\mathbf{g}$, with five change points. Results shown for incomplete data (A,C) and complete data (B,D). Symbols indicate the results when both $\mathbf{g}$ are known *a priori*. The solid and dashed curves correspond respectively to the "common-$\mathbf{g}$" and "unique-$\mathbf{g}$" procedures (see Section 4d), while the dotted lines in (a,b) show common-$\mathbf{g}$ results with the likelihood-based stopping criterion in force (stopping criterion has no significant impact on unique-$\mathbf{g}$ results). All tests assume two basis functions in fitting.

The MSE for the standard test (Panel A) ranged from 0.07 to 0.1 (a RMS error of 25% to 32% of the shift amplitude itself) depending on N_g. The MSE was roughly halved with no data missing (Panel B), which is consistent with sampling error behavior since twice as many data were available in this case. MSE was also cut by a factor of two or more when "correct" $\mathbf{g}$ were supplied *a priori*, showing that performance is significantly affected by the imperfect ability to determine natural variability empirically.

The most worrisome feature of the results is that MSE did not stabilize but increased monotonically with N_g after attaining a minimum around $N_g = 5$. This occurred even when imputation of missing values was unnecessary (Panel B). Detailed examination revealed that instead of showing a general mild tendency to increase, errors exploded in a few "problem" trials. In these, a basis function g became highly loaded at one station and nearly degenerate with a level-shift step function at the same station, with the two effects growing in opposite directions as the method iterated.

1) IMPROVING CONVERGENCE AND ROBUSTNESS

Two strategies were investigated to make the algorithm more robust. The first was based on model likelihood (the logarithm of the "likelihood" of the data given the model is inversely proportional to the sample variance of ϵ). Likeli-

hood increases with each iteration after N_g(as it must due to the max-likelihood property of EM) but in problem trials it often began decreasing before reaching N_g. This suggests that the basis re-estimation should be stopped if and when the RMS of ϵ increases. This to some extent obviates the ad-hoc choice of N_g, though a maximum number of basis iterations must still be specified *a priori* and this was often reached.

Results with this stopping criterion in force are shown by the dotted extensions in Fig. A1A,B (for this curve N_g is taken as the upper limit of the number of iterations). The stopping condition made the results significantly less sensitive to the choice of N_g, though not fully eliminating the runaway error problem. The remaining deterioration even while likelihood (fit) is improving is a sign of overfitting, discussed further below.

The second strategy was to modify the IUK method to give each station a unique basis **g** independent of the station's own data. Recall that each station has its own set of basis functions, which need not be identical among the stations. Making them different does not formally add any model complexity, in that the total number of estimated parameters is identical. I repeated the analysis with principal components calculated separately for each station based on all data but its own, a modification designated "unique-**g**" (with the old way denoted "common-**g**"). This makes IUK more like the reference methods and that of Thorne et al. (2005).

Results with unique-**g** are shown by the dashed curves in Fig. A1A,B. With incomplete data (Panel A) the nonconvergence problem was again mitigated but not eliminated. Unlike the stopping criterion, however, unique-**g** reduced MSE by at least 15% for all values of N_g, a significant improvement. When data were complete (Panel B) the method became fully convergent. Application of the stopping criterion to unique-**g** runs yielded no further gains and is not shown. Unique-**g** was adopted for the remainder of this study.

The above results may be understood on the basis of overfitting. Fig. A1C shows that, with common-**g**, the empirical models for μ fitted the data better than did the correct model! This is a clear sign of overfitting. But there is good news, on two counts. First, model likelihood continued to increase with N_g, assuaging concerns (Section c) that the empirically-determined **g** might not fit the data well. Second, the unique-**g** version of IUK clearly mitigates the overfitting problem. It may be concluded that excess model complexity (i.e., number of modes n) is a greater concern than failure to fully maximize likelihood at the chosen complexity level.

2) FURTHER TESTS

Based on the above we might expect performance to improve if only one PC were retained, even though two modes were used to generate the data. This turned out to be true for common-**g** (not shown), but for unique-**g** (Fig. A2) it was only true when $N_g > 18$, well beyond the optimal value. In the optimal N_g range, $n = 2$ performed better. Going to $n = 3$ produced no additional gains and, not surprisingly, exacerbated nonconvergence at high N_g. Wiggles in the $n = 3$ curve also show some sensitivity to small changes in the model, another symptom of overfitting. Note that the affordable value of n depends on the size of the dataset. Standard methods are available for choosing a truncation. S00 found that a broad range of values (∼2-6) produced reasonable results on a radiosonde wind dataset. The small magnitude of differences in (Fig. A2), about 10% range from best to worst for $2 < N_g < 15$, indicates robustness to the value of this parameter.

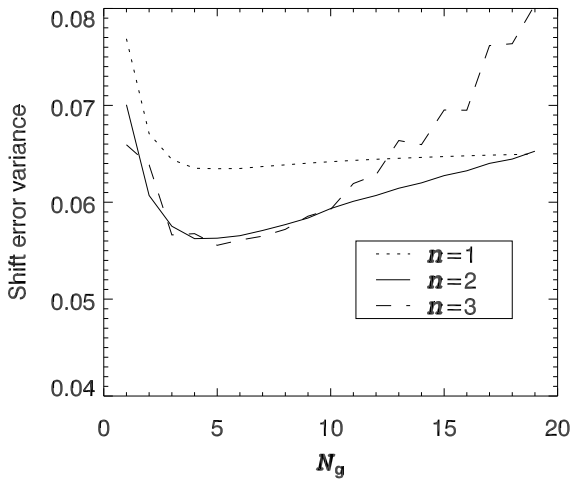

FIG. A2. Mean-squared level shift error for unique- IUK, with five change points, for between one and three retained PC's per station. Diamonds denote results with **g** known *a priori* (which are the same for two or five shifts).

Finally I examined the performance on problems of varying difficulty. Fig. A3 shows MSE for two, five, or seven change points. Error increases with more change points, since information from neighboring stations is more likely to be contaminated. Interestingly, more basis iterations N_g were necessary to minimize error when there are more change points: only 2-4 were needed with two change points, but 6-7 were needed with seven. Also, the improvement relative to $N_g = 1$ was substantially greater with more change points. Evidently, multiple iterations become more important on more difficult problems.

REFERENCES

Allen, R. J., and S. C. Sherwood, submitted 2006: Utility of radiosonde wind data in representing climatological variations of tropospheric temperature and baroclinicity in the western tropical Pacific. *J. Climate*.

Beale, E. M. L., and R. J. A. Little, 1975: Missing values in multivariate analysis. *J. R. Statist. Soc.*, **37**, 129–145.

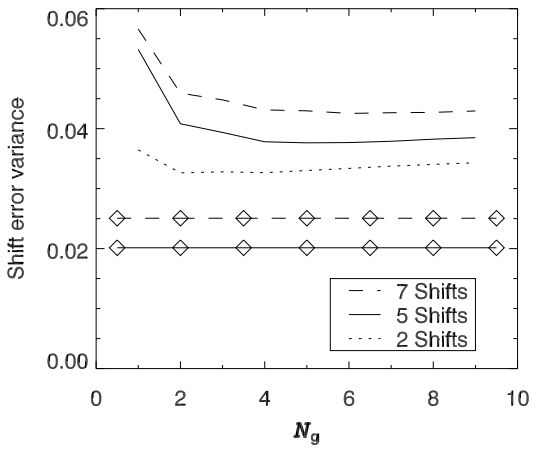

FIG. A3. Mean-squared level shift error at stations 2 and 4 for unique- IUK with two, five, or seven change points.

Chib, S., 1998: Estimation and comparison of multiple change-point models. *J. Econometrics*, **86**, 221–241.

Christy, J. R., and R. W. Spencer, 2005: Correcting temperature data sets. *Science*, **310**, 5750.

Christy, J. R., W. B. Norris, and R. W. Spencer, Submitted 2005: Tropospheric temperature change since 1979 from tropical radiosonde and satellite measurements. *J. Geophys. Res.*.

Cressie, N. A. C., 1993: *Statistics for Spatial Data*. revised ed., John Wiley.

Daley, R., 1991: *Atmospheric Data Analysis*. Cambridge.

DeGaetano, A. T., 2006: Attributes of several methods for detecting discontinuities in mean temperature series. *J. Climate*, **19**, 838–853.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977: Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc., ser. B*, **140**, 1–22.

Easterling, D. R., and T. C. Peterson, 1995: A new method for detecting undocumented discontinuities in climatological time-series. *Int. J. Climatol.*, **15**, 369–377.

Free, M., and D. J. Seidel, 2005: Causes of differing temperature trends in radiosonde upper air data sets. *J. Geophys. Res.*, **110**, D07,101.

Free, M., et al., 2002: Creating climate reference datasets: CARDS workshop on adjusting radiosonde temperature data for climate monitoring. *Bull. Amer. Meteor. Soc.*, **83**, 891–899.

Gaffen, D. J., M. A. Sargent, R. E. Habermann, and J. R. Lanzante, 2000: Sensitivity of tropospheric and stratospheric temperature trends to radiosonde data quality. *J. Climate*, **13**, 1776–1796.

Haimberger, L., 2005: Homogenization of radiosonde temperature time series using ERA-40 analysis feedback information. ECMWF, Tech. rep., ERA-40 Project Report Series #23, 68 pp.

Karl, T. R., and C. N. J. Williams, 1987: An approach to adjusting climatological time series for discontinuous inhomogeneities. *J. Clim. App. Meteor.*, **26**, 1744–1763.

Lanzante, J. R., 1996: Resistant, robust and nonparametric techniques for the analysis of climate data: Theory and examples, including applications to historical radiosonde station data. *Int. J. Climatol.*, **16**, 1197–1226.

Lanzante, J. R., S. A. Klein, and D. J. Seidel, 2003: Temporal homogenization of monthly radiosonde temperature data. Part I: Methodology. *J. Climate*, **16**, 224–240.

Lund, R., and J. Reeves, 2002: Detection of undocumented change-points: A revision of the two-phase regression model. *J. Climate*, **15**, 2547–2554.

Mann, M. E., R. S. Bradley, and M. K. Hughes, 1999: Northern hemisphere temperatures during the past millennium: Inferences, uncertainties, and limitations. *Geophys. Res. Lett.*, **26**, 759–762.

Menne, M. J., and C. N. Williams, 2005: Detection of undocumented changepoints using multiple test statistics and composite reference series. *J. Climate*, **18**, 4271–4286.

Parker, D. E., 2004: Climate—large-scale warming is not urban. *Nature*, **432**, 290–290.

Parker, D. E., and D. I. Cox, 1995: Towards a consistent global climatological rawinsonde database. *Inter. J. Climatol.*, **15**, 473–496.

Peterson, T. C., 2003: Assessment of urban versus rural in situ surface temperatures in the contiguous United States: No difference found. *J. Climate*, **16**, 2941–2959.

Peterson, T. C. *et al.*., 1998: Homogeneity adjustments of *in situ* atmospheric climate data: A review. *Int. J. Climatol.*, **18**, 1493–1517.

Santer, B. D., et al., 2005: Amplification of surface temperature trends and variability in the tropical atmosphere. *Science*, **309**, 1551–1556.

Schneider, T., 2001: Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *J. Climate*, **14**, 853–871.

Sherwood, S. C., 2000a: Climate signal mapping and an application to atmospheric tides. *Geophys. Res. Lett.*, **27**, 3525–3528.

Sherwood, S. C., 2000b: Climate signals from station arrays with missing data, and an application to winds. *J. Geophys. Res.*, **105**, 29,489–29,500.

Sherwood, S. C., and A. E. Dessler, 2001: A model for transport across the tropical tropopause. *J. Atmos. Sci.*, **58**, 765–779.

Sherwood, S. C., J. R. Lanzante, and C. L. Meyer, 2005: Radiosonde daytime biases and late-20th century warming. *Science*, **309**, 1556–1559.

Sullivan, J. H., 2002: Estimating the locations of multiple change points in the mean. *Comput. Stat.*, **17**, 289–296.

Thorne, P. W., D. E. Parker, S. F. B. Tett, P. D. Jones, M. McCarthy, H. Coleman, and P. Brohan, 2005: Revisiting radiosonde upper-air temperatures from 1958-2002. *J. Geophys. Res.*, **110**, D18,105.

Vinnikov, K. Y., P. Y. Groisman, and K. M. Lugina, 1990: Empirical-data on contemporary global climate changes (temperature and precipitation). *J. Climate*, **3**, 662–677.

Wang, X. L., 2003: Comments on "Detection of undocumented changepoints: A revision of the two-phase regression model". *J. Climate*, **16**, 3383–3385.

Printed December 6, 2006.