OUTLIER DETECTION IN GRIDDED SHIP DATA SETS

Pascal Terray, Laboratoire d'Océanographie Dynamique et de Climatologie, and Université Paris 7, Paris, France

1. INTRODUCTION

This is the second of two papers attempting to develop robust statistical methods to deal with gridded ship data sets. The earlier study (Terray, 1999) focused on an extension of the traditional empirical orthogonal function (EOF) analysis which allows arbitrary positive weights to be assigned to each entry of the data matrix. If these weights are constructed in a responsible manner (for example, as a smooth function of the number of ship reports used to compute a particular raw monthly mean in the data set), it was demonstrated that this method allows us to analyse the natural variability exhibited by gridded ship data sets by directly taking into account the irregular space-time sampling of marine observations. In particular, the method takes care of missing values by assigning zero weights to such data entries.

In the current study, we discuss another robust statistical method to detect 'local errors' in gridded ship data sets. More precisely, we tackle the problem of outlying areal averages in gridded ship data sets such as the Comprehensive Ocean-Atmosphere Data Set (COADS; Woodruff *et al.*, 1987) 2° lat × 2° long monthly summaries and how to test their statistical significance. Since the majority of climate researchers use gridded ship data sets instead of individual ship reports, we suggest that these data sets must be checked for the presence of doubtful raw monthly means in the same manner as individual ship reports are quality controlled before being integrated in ship reports databases. Moreover, it should be noted that such an approach may be a solution to the trimming problems which are apparent in COADS monthly summaries (Wolter, 1997).

The rest of this paper is organized as follows. First, we present some elements of outlier detection theory and the basic statistical tests we have used. Next, we discuss how these statistical tests may be adapted to ship data sets and integrated as building blocks in a fully computerized procedure for detecting many outliers in such data sets. Finally, this new approach has been experimented on a marine product in order to show how it works in practice. As a conclusion, we suggest that the two procedures, namely outlier detection and weighted EOF analysis, may be combined to obtain a truly robust statistical method particularly well suited to gridded ship data sets.

2. STATISTICAL THEORY OF OUTLIER DETECTION

In the context of gridded ship data sets, an outlying observation, or 'outlier', is a raw monthly mean in a 2° lat × 2° long box (depending on the resolution of the data set) that appears to deviate markedly from adjacent or neighbouring grid-points in area or/and in time. Outliers in gridded ship data sets may be generated by three basic mechanisms (Wolter, 1997):

- An outlying raw monthly mean may be merely an extreme manifestation of the sampling inherent in the data, since some raw monthly means in 2° lat × 2° long boxes are computed with very few marine observations for a given date while adjacent boxes may be well sampled.
- Outlying raw monthly means in some 2° lat × 2° long boxes may also be the results of potential biases due to the origin of the 'source-decks' merged into the gridded ship data set or processing errors. For example, biases in sea surface temperature (SST) associated with different methods of measurements (bucket or intake) may well introduce errors in gridded ship data sets in particular atmospheric conditions and along some ship tracks.

- Finally, an outlying areal average may be the result of errors relating to instrumental readings or coding mistakes. But, most of these types of outliers must be discovered during basic quality controls which are automatically applied to individual ship reports merged into any reasonable marine product.

The problem of detecting outliers in a random sample has been extensively researched by statisticians in recent years and a number of test statistics are available for both the single outlier case and the many outlier case for testing a specified number k of outliers (Barnett and Lewis, 1978). In particular, the detection of outliers in a normal sample has received considerable attention. It is far beyond the scope of this paper to give a review of this vast subject. Suffice to say here, that the problem of outlier detection is generally treated as the statistical testing of a hypothesis. The null hypothesis, as usually stated, is that all the observations are drawn from the same (normal) population; the alternative hypothesis is that at least one of the observations has been drawn from another distribution. To discriminate between these two hypotheses, a sample criterion T which uses the doubtful observation(s) is calculated. This statistic is then compared with a critical value λ_α based on the theory of random sampling to determine whether the doubtful observation is to be retained or rejected. This critical value is the value of the chosen sample criterion which would be exceeded by chance with some specified and small probability α (say 0.01 or 0.05), which is the so-called significance level of the test, if the null hypothesis is true. Intuitively, this significance level is the risk of erroneously rejecting a good observation (statistical type I error). More precisely, statistical tests for outliers are the following:

(1) Find λ_α such that $Pr(T>\lambda_\alpha)= \alpha$ if the null hypothesis is true for some statistic T;

(2) Reject the null hypothesis and declare an outlier present if $T>\lambda_\alpha$, or accept the null hypothesis and declare the sample is clean if $T\leq\lambda_\alpha$.

In this statistical framework, outlier detection procedures differ by:

- The form of the underlying parent population (normal, gamma, etc.);
- The form of the test criterion T which has to be computed on the sample: among these test criteria, we can distinguish those which clearly identify particular observations as possible outliers from those which test the hypothesis that the random sample as a whole did indeed come from the specified parent distribution;
- The number of suspected outliers in the sample;
- The fact that the doubtful observations may be to one side of the bulk of the data or that some are too large and some are too small.

Several hundreds of statistical tests of this type are described in the book written by Barnett and Lewis (1978) which is a kind of 'bible' on the subject. In the context of gridded ship data sets, the problem is then to decide which tests to apply, and how to use them in order to obtain a fully computerized procedure for detecting outliers which may be applied to any ship data set. In this way, one can hope to trap anomalous cases and so ensure the integrity of most of the ship data sets currently in use.

We have used here a simple model, which is well documented in the statistical literature: when the data with the possible exception of any outlier form a sample from a normal distribution with unknown mean μ and unknown variance σ^2. We recognize that this model is certainly not perfect in the context of samples of adjacent raw monthly means in 2° lat × 2° long boxes extracted from gridded ship data sets. However, as we will show below, this model works 'reasonably' well as implemented in our computerized procedure on the basis of the spatial coherence of neighbouring 2° lat × 2° long area values for many meteorological parameters. Several reasonably powerful statistical tests exist to detect one outlier in a normal sample, and our approach involves the following classical statistical criteria:

Let $x_1, x_2, \ldots, x_n$ be the observations of a random sample. Order the observations according to increasing magnitude and denote the i_{th} largest by y_i; thus, $y_1\leq y_2\leq \ldots \leq y_n$ is the ordered set of observations. Suppose the largest observation y_n is suspect. To test for discordancy in this single upper observation in a normal sample, a reasonable test statistic is:

$$T = \frac{y_n - \bar{x}}{s}$$

where:

$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$ is the sample mean, and

$s^2 = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{n} \right)^2$ is the sample variance calculated with n degrees of freedom.

If y_1, the lower observation, rather than y_n, is the doubtful value, the criterion is as follows:

$$T' = \frac{\bar{x} - y_1}{s}$$

and the rest of the statistical procedure will be unchanged on the basis of the symmetry of the normal distribution. Finally, when it is not known a priori whether the contaminant is the lower or the upper observation in the sample, we should compute:

$$T^* = \max(T, T')$$

But, in this last case, we must use a critical value corresponding to the $\alpha/2$ significance level if we want the true significance level to be 0.05.

The rationale behind these tests may be found in Hawkins (1980) or Barnett and Lewis (1978). The null hypothesis that we are testing in every case is that all the observations in the sample come from the same normal population. It may be shown that these statistics are optimal in the sense of maximizing the probability of correct identification of an outlier when one is present. It should be noted, however, that these statistics may produce quite misleading results in the presence of many outliers, especially when suspected values are closer to each other than to the bulk of the other observations. This inability of a testing procedure to identify even a single outlier in the presence of several suspected values is called the masking effect. We will discuss this point further in the next section when we describe our computerized procedure for detecting outliers.

Before using these test statistics in outlier checks, we must know the significance probability attached to an observed value t of the statistic T (or T',T*). That is to say, the probability that, on the null hypothesis of no contamination, T takes values more discordant than t. For this purpose, we need to find the null distribution of T or at least some fractiles λ_α of this distribution corresponding to specified significance levels α, say 0.01 or 0.05. The null distribution of T is available as a recursion relationship (Barnett and Lewis, 1978) or as a complicated multiple integral (Grubbs, 1950), and tables containing critical values for some standard significance levels have been published (Grubbs and Beck, 1972; Hawkins, 1980). However, we will show how approximate critical values for a given significance level α can be computed since our computerized procedure may involve a number of observations outside the range of these published tables.

Without loss of generality, we consider only the case of an upper outlier; approximate fractiles for T' or T* may be derived similarly. We may compute some fractiles of the test distribution of T as follows:

Under the null hypothesis of no contamination, $x_1, x_2, \ldots, x_n$ are observations of random variables $X_1, X_2, \ldots, X_n$ which are independent and identically distributed as $N(\mu, \sigma^2)$. In this case, if x_i is an observation selected arbitrarily from the random sample of n items, it may be shown that if:

$$T_i = \frac{x_i - \bar{x}}{s}$$

then the probability density function of:

$$t_i = \frac{T_i\sqrt{n-2}}{\sqrt{n-1-T_i^2}}$$

is given by the 'student's' t-distribution with n-2 degrees of freedom. This is easily verified because t_i is the test statistic of the classical student's two sample t-test, where one sample consists of x_i and the second sample of the n-1 other observations. From this result, we are able to find the probability that an arbitrary observation i will be outlying since:

$$\Pr[T_i > \lambda] = \Pr\left[t_{(n-2)} > \frac{\lambda\sqrt{n-2}}{\sqrt{n-1-\lambda^2}}\right]$$

where λ is an arbitrary value in the range $\left]-\sqrt{n-1}, \sqrt{n-1}\right[$ and $t_{(n-2)}$ follows a student's t-distribution with n-2 degrees of freedom. However, this result does not yet give us an exact test for one outlier, because this probability is different from the probability that a particular observation (the lowest or the largest) will be greater than λ. More precisely, we need the distribution not of an arbitrary T_i, but of T, the greatest of the quantities T_i for i=1 to n.

Now, note that the event ($T>\lambda$) is the union of the n events ($Ti>\lambda$). Thus:

$$\Pr[T > \lambda] = \Pr\left[\bigcup_{i=1}^{n}(T_i > \lambda)\right]$$

In other words, the probability of the event ($T>\lambda$) is the probability that at least one of the n events ($T_i>\lambda$) is true. Bounds on $\Pr[T>\lambda]$ may then be obtained in terms of the component events ($T_i>\lambda$) through the use of the so-called Bonferroni inequality (Feller, 1968):

$$\sum_i \Pr[T_i > \lambda] - \sum_{i<j} \Pr\left[(T_i > \lambda) \cap (T_j > \lambda)\right] \le \Pr[T > \lambda] \le \sum_i \Pr[T_i > \lambda]$$

Since the events ($T_i>\lambda$) are equiprobable, and likewise the events ($T_i>\lambda$)$\cap$(($T_j>\lambda$), we have the following inequality for arbitrary i and j:

$$n\Pr[T_i > \lambda] - \frac{n(n-1)}{2}\Pr\left[(T_i > \lambda) \cap (T_j > \lambda)\right] \le \Pr[T > \lambda] \le n\Pr[T_i > \lambda]$$

Now, by using the fact that for arbitrary i and j (Doornbos, 1966):

$$\Pr\left[(T_i > \lambda) \cap (T_j > \lambda)\right] < (\Pr[T_i > \lambda])^2$$

we finally obtain:

$$n\Pr[T_i > \lambda] - \frac{n-1}{2n}\left(n\Pr[T_i > \lambda]\right)^2 < \Pr[T > \lambda] \le n\Pr[T_i > \lambda]$$

for an arbitrary i. Thus, if:

$$\frac{\lambda\sqrt{n-2}}{\sqrt{n-1-\lambda^2}}$$

is the 1 - (α/n) fractile of the student's t-distribution with n-2 degrees of freedom, the last equation shows that:

$$\alpha - \frac{n-1}{2n}(\alpha^2) < \Pr[T > \lambda] \le \alpha$$

A result indicating that λ is a good and conservative approximation of the true critical value λ_α of the distribution of T under the null hypothesis of no contamination for any reasonable significance level α, say 0.01 or 0.05. Moreover, it can be shown that this method gives the exact critical value λ_α of T if:

$$\lambda \ge \sqrt{\frac{n-1}{2}}$$

(for example, the 0.05 critical value for any n<15) since in this case we have:

$$\Pr\left[(T_i > \lambda) \cap (T_j > \lambda)\right] = 0$$

for arbitrary i and j. Following the same procedure, we may approximate the true critical value λ_α of T* on the null hypothesis of no contamination by λ^*, if:

$$\frac{\lambda^* \sqrt{n-2}}{\sqrt{n-1-(\lambda^*)^2}}$$

is the 1 - $\alpha/(2n)$ fractile of the student's t-distribution with n-2 degrees of freedom.

3. OUTLIER DETECTION IN GRIDDED SHIP DATA SETS

Suppose now that we want to check the 'local' consistency of a given ship data set with, say, a 2° lat × 2° long resolution. This data set may contain raw data or anomaly fields after removal of the annual cycle with a climatology. In both cases, the same algorithm is used and the preceding theoretical results are then used as follows:

(1) First, we specify upper and lower limits for detecting doubtful monthly mean or anomaly values in the gridded ship data set. These limits may vary depending on calendar month and area. Any value which exceeds the upper limit, or is less than the lower limit, is considered a priori doubtful and will be tested for compatibility with monthly mean or anomaly values in adjacent or neighboring 2° lat × 2° long boxes. These upper or lower limits determine the number of values which will be tested in the detection procedure for a given gridded data set. Thus, if we want to test nearly all the data values for compatibility, we just have to specify a very low upper limit and/or a very high lower limit in the algorithm. Such a choice means that the algorithm will use more computer time since a lot of data values will be tested; but in any case, a data value will be declared an outlier only on the basis of the probabilities of rare events as outlined in section 2 (see below).

(2) For any date, doubtful monthly mean or anomaly values identified in step 1 are arranged from the most outlying to the most inlying compared to the bulk of the data. For this purpose, absolute values of residuals of these doubtful values from the overall mean of the observed data for this date are sorted in descending order, and the doubtful values are ranked accordingly.

(3) These doubtful values are then considered consecutively, from the most outlying to the most inlying, and a sample is constructed from adjacent or neighbouring 2° lat × 2° long area values for any of these possible outliers. The number of 2° lat × 2° long boxes in the vicinity of each doubtful value which are scanned, in order to construct a sample, may be chosen by the user before running the procedure. It should be noted that the number of items in this sample may vary depending on the date and the area. However, the significance level α of the test will be the same for any suspected raw monthly mean value, as we will see below.

(4) At this stage, several different possibilities exist:

(*a*) First, we need to consider the case when it is not possible to pick up a sample to test the doubtful value because none of the surrounding boxes contain data. Frequently, this means that the doubtful raw monthly mean is calculated from very few ship reports. In such a case, the user may decide, before running the procedure, to flag or reject all these unrepresentative values.

(*b*) Second, suppose that there is only one doubtful value in the collected sample, the one we want to test. If this value is at the upper end of the sample, we use T as a test criterion; if it is at the lower end of the sample, the statistic T' is considered instead. In both cases, the doubtful value is declared an outlier if the statistic exceeds the critical value λ_α corresponding to a specified significance level α. In this case, the suspected value is rejected or flagged (a user choice) and the next most outlying doubtful value is processed.

(*c*) Finally, imagine that there is more than one doubtful value in the constructed sample of n items, according to the upper and lower limits specified in step 1. Let K be the number of such doubtful values and x be the

suspected value that we are currently processing. In order to take into account the possibility that the sample contains more than one outlier, a consecutive procedure is applied. One naive approach is to use repeated applications of the single outlier statistical test T* described above, deleting the 'outlier' detected at each step and applying the test again to the reduced sample until an insignificant result is obtained or the suspected value x is tested for compatibility with the remaining observations. However, this 'forward selection' approach may be quite misleading in practice (Hawkins, 1980). The problem is the so-called masking effect discussed in the preceding section, namely that the presence of two or more outliers may produce an insignificant result in the initial single outlier test. In view of this defect, the following variant is recommended: remove the K most extreme values of the sample (absolute values of residuals from the sample mean of the successively reduced sample are used to rank the observations). If the current doubtful value x is not thrown away in this process, declare x as 'clean' and process the next most outlying doubtful value for the current date. Otherwise, apply the following 'backward selection' algorithm: starting with the n-K 'clean' observations, test the most inlying of the K extreme values for compatibility with the clean observations by the statistic T* at a nominal significance level α. If it is compatible, then include it with the clean observations and repeat the procedure with the next most outlying suspected value in the sample until the current doubtful value x is processed and declared as compatible. This sequence of tests is immediately stopped when an observation is rejected by the statistical test T* since all the subsequent outlying raw monthly means, including x, are then incompatible with the clean observations. In this case, the 2° lat $\times$ 2° long area mean value corresponding to x is rejected or flagged, and the next doubtful value for the current date is processed. Note, however, that the other rejected values in the sample are not set to missing at this stage.

It is fair to say that, while the backward consecutive algorithm described in 4(c) is immune to masking (providing that the actual number of outliers in the sample does not exceed the number of suspected values K in the test procedure), it provides important distributional difficulties associated with finding suitable fractiles λ_α if we require (as we do) an actual significance level α for each of the successive null hypotheses which are tested in the backward selection algorithm. A comprehensive discussion of this problem is given by Hawkins (1980), and we omit the details owing to the lack of space and the difficulty of the problem. Suffice to say here, that it is necessary to resort to simulation if we require exact fractiles, but that there is little error by approximating these fractiles, as outlined in the preceding section, excepted for small n, say $n<15$. The latter solution was adopted in this study. Consequently, the sequence of tests used in the backward consecutive algorithm described in 4(c) may have actual significance levels in excess of 25 per cent of the specified nominal significance level α according to Hawkins (1980). We will try to correct this deficiency in a future version of our outlier detection procedure by carrying out the required simulations.

4. EXAMPLE

The outlier detection algorithm has been applied to several ship data sets and various examples were presented during the workshop. In particular, an experiment was undertaken on a pre-COADS marine product with known systematic errors, in order to show the benefit of this type of procedure in the context of marine climatology.

An extensive description of the ship data set used in this experiment may be found in Terray (1994). Briefly, SST data are presented as raw monthly means in 2° lat $\times$ 2° long boxes in a domain extending from 30° to 100°E longitude and from 30°S to 30°N latitude. The period of analysis extends from 1900 to 1986. Figure 1 documents the irregular space-time sampling associated with this gridded ship data set.

Many well-known deficiencies were observed in this data set before and around the Second World War (Terray, 1994). In addition, a suspicious warming

trend is apparent on the SST time series during 1954-1976, and it was anticipated that this trend may be linked to important changes in the origin of the 'source-decks' merged into this marine product, or to the presence of a large amount of erroneous ship reports that were not rejected during basic quality control of the ship reports. Suspect raw monthly means were mainly confined along the shipping routes from Madagascar to Sumatra and from Sumatra to the Northern Arabian Sea for the 1968-1974 period.

In view of this, the outlier detection algorithm of the preceding section has been applied to this SST gridded ship data set in a two-step procedure:

- First, the algorithm was applied to all the raw monthly SST fields with 15°C as a lower limit and 35°C as an upper limit to identify doubtful 2° lat × 2° long monthly means which must be tested by the algorithm. A nominal significance level of 0.05 was chosen for all the tests. This first step was only intended as a check on 'evident' outliers far away from the bulk of the data. In this first step, 481 raw monthly values were tested for all the monthly fields of 1900-1986 and, among them, 361 were identified as outliers by the statistical tests (this number includes isolated monthly mean values) and rejected.
- The second step is designed to remove outliers with respect to anomaly fields. For this purpose, the raw monthly mean SST fields were expressed as monthly anomaly fields by using a monthly climatology obtained from a weighted EOF analysis on COADS SST data (Terray, 1998). The outlier detection algorithm was applied to these anomaly fields with -3°C as a lower limit and 3°C as an upper limit. Again, a nominal significance level of 0.05 was chosen for all the tests. In this second step, 10 917 anomaly values were tested; among them, 2 826 were identified as outliers and the corresponding raw monthly mean values were rejected.

The 15°C–35°C limit in the first step and -3°C–3°C limit in the second step were choosen as a compromise between a good use of computer ressources and the quality of the final product. Lower upper limits or higher lower limits in both steps of the algorithm give roughly the same final results, but are more expensive with respect to CPU time.

On average, five to ten grid squares have been removed for each date from the beginning of this century until the Second World War. During the 1940-1968 period, the number of outliers is quite low, being less than five for each date on average. Finally, recent decades have witnessed a substantial increase in outlier losses. The number of outlier rejections may be as high as one hundred during 1968-1974 and outliers are still very common after this period. It is interesting to note that, except for the 1968-1974 period, outlier losses are very similar to the trimming losses observed for computing COADS trimmed monthly summaries for

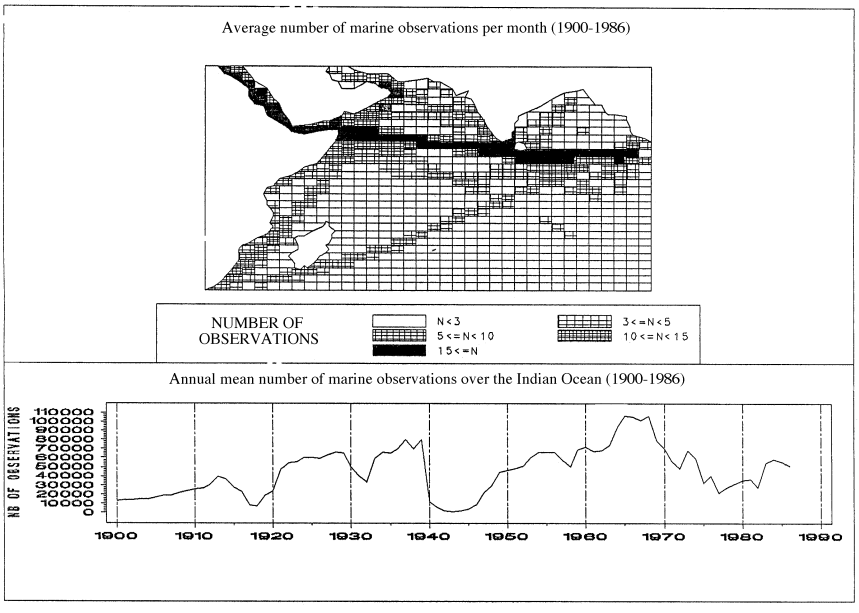

Figure 1—Space-time sampling of the ship reports associated with the gridded data set.

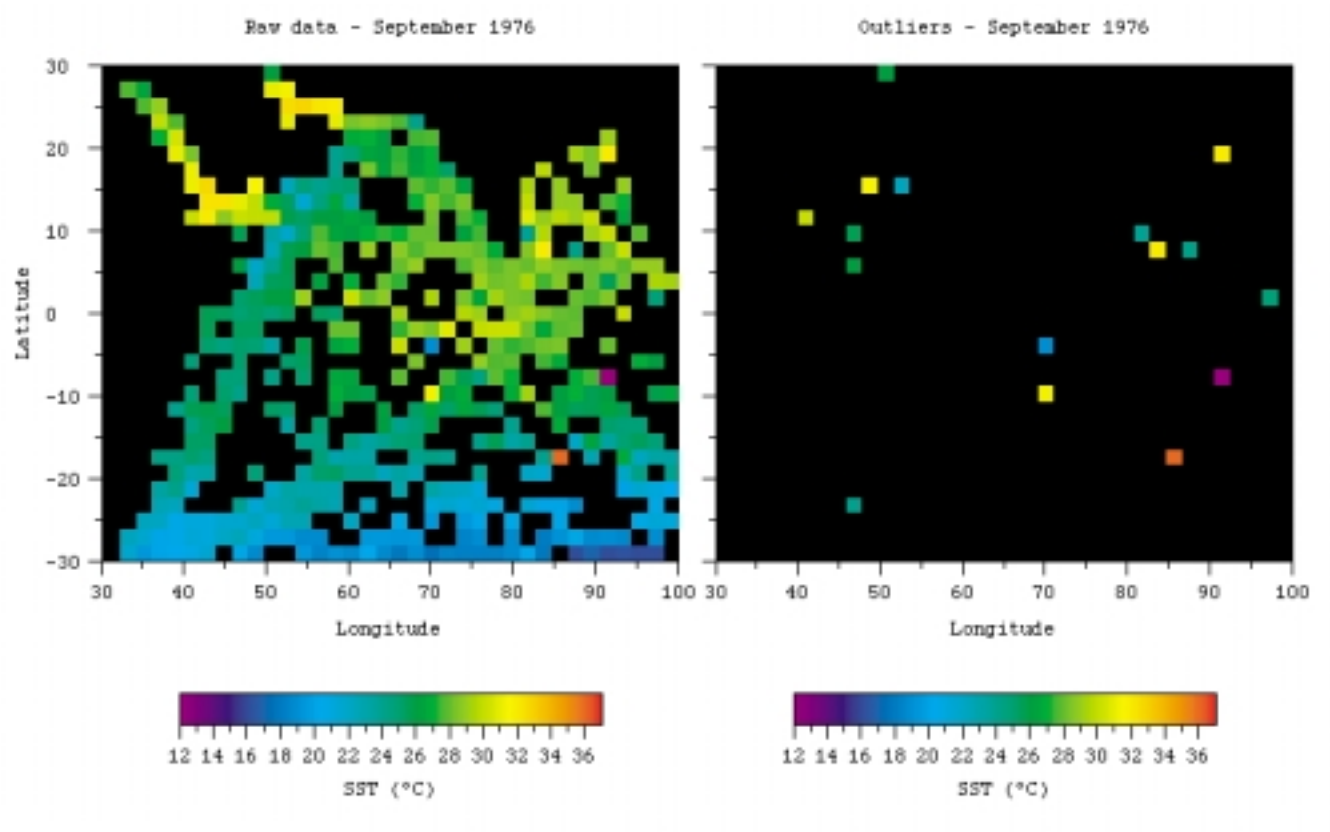

Figure 2—Results of the two-pass outlier detection algorithm for September 1976.

the global ocean (Wolter, 1997). Figure 2 presents the results of the two-pass outlier detection algorithm for September 1976 and may be used to obtain some understanding of the grid squares which were rejected by the outlier detection process.

To investigate the impacts of the outlier detection algorithm, the following computations were also undertaken on the SST ship data set both before and after the 'cleaning' of the data:

(i) First, the 1954-1976 interval was used as a reference period for calculating a climatology for each calendar month and each 2° box, provided that data for at least 10 years with more than 5 observations per month were available in the period. The monthly means for each i grid point and j month were computed as a weighted average:

$$\overline{X}_{ij} = \left(\sum_{k=1954}^{1976} W_{ijk} X_{ijk}\right) / \left(\sum_{k=1954}^{1976} W_{ijk}\right)$$

where $W_{ijk} = 1 - \exp(-N_{ijk}/5)$

Here X_{ijk} is the value computed for the i^{th} box, j^{th} month and k^{th} year. N_{ijk} is the number of ship observations used in computing X_{ijk}. W_{ijk} is in the neighbourhood of 1 if $N_{ijk}>10$ and near 0.5 if N_{ijk} equals 5.

(ii) After this first step, time monthly anomaly series for each 2° box during the 1900-1986 period were computed by simply subtracting this climatology from each value, provided that neither the datum nor the climatology was missing. These anomalies were then subsequently spatially averaged over the whole Indian Ocean with the same weighting scheme (e.g., W_{ijk}) as used in the computation of the climatology.

The two SST anomaly time series computed, respectively, before and after the 'cleaning' of the data, were then subjected to the X11 monthly additive scheme (Terray, 1994), a powerful technique for describing a time series, to assess their consistency. In the X11 procedure, the analysed X_t monthly time series is decomposed into three terms:

$$X_t = T_t + A_t + I_t$$

The T_t term is used to quantify the trend and low-frequency variations in the time series. The A_t term describes the annual cycle and I_t can be used to assess the level of noise in the data, though this term can also contain some signal in a climatological sense. All the terms are estimated with specific moving averages of various lengths.

Figures 3 and 4 give the results of the analysis for the SST time series computed before and after outliers were rejected, respectively. The monthly number of observations is also plotted on the bottom of each figure as an aid for interpreting the results and detecting accurately any change in the composition of the 'source-decks' contributing to the time series. While the two series and their associated X11 components are similar in many aspects, an important discrepancy may be noted during 1968-1974: the unlikely warm anomalies observed in the data before running the outlier detection procedure (Figure 3) are considerably reduced on the time series computed after outliers were rejected (Figure 4). As a consequence, the trend components of the two series are different during 1968-1974. This difference is consistent with the hypothesis of the artificial nature of the warming trend observed during 1968-1974 over the Indian Ocean. Finally, it may be noted that the 'clean' series is less noisy, as demonstrated by the irregular components.

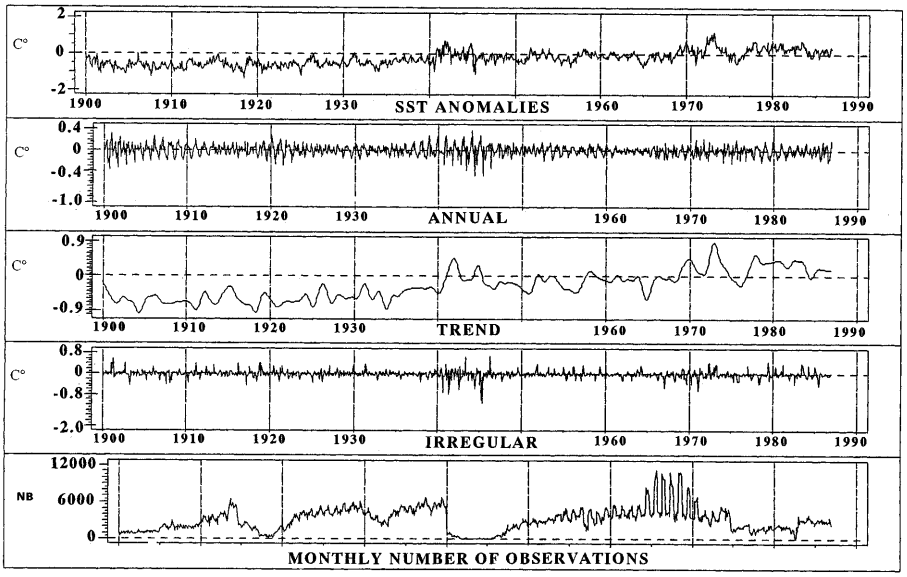

Figure 3—SST monthly anomalies relative to 1954-1976 for the whole Indian Ocean before outlier detection. The series has been broken down into annual, trend and irregular components by the X11 procedure. The monthly number of ship reports used to construct the series is given at the bottom of the figure.

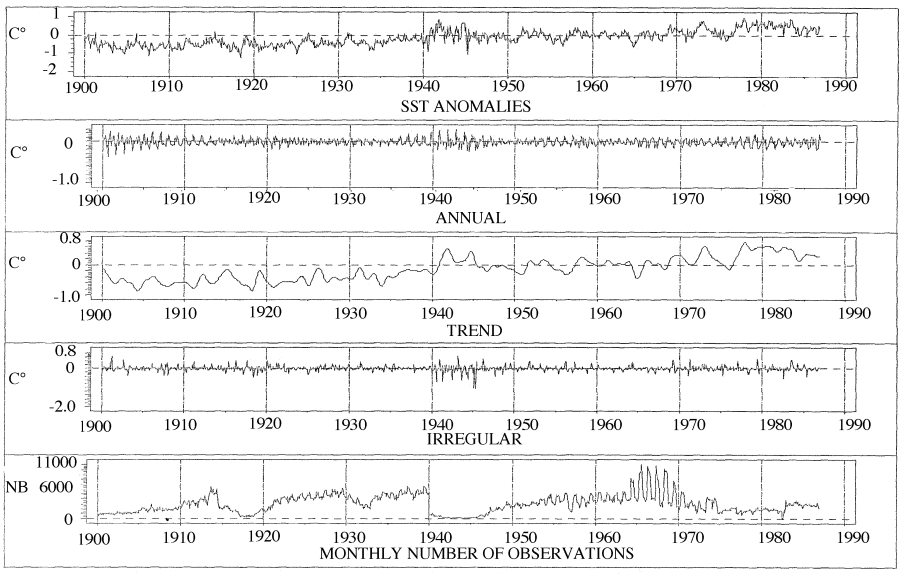

Figure 4—SST monthly anomalies relative to 1954-1976 for the whole Indian Ocean after outlier detection. The series has been broken down into annual, trend and irregular components by the X11 procedure. The monthly number of ship reports used to construct the series is given at the bottom of the figure.

5. CONCLUSIONS

A recurring problem in the creation and maintenance of large gridded ship data sets is the accuracy of the information entering these products. The fact that large volumes of data are involved suggests that, as far as possible, the reliability of such data sets should be assessed through a computerized screening procedure. For this purpose, a new method for detecting outliers in gridded ship data sets has been proposed. It is our hope that this approach will aid climate scientists in determining which, if any, of the raw monthly area values included in a particular ship data set may be outliers.

Once potential outliers have been identified, it is suggested that these values may be flagged or, more drastically, rejected. In any case, the impact of these doubtful values in a particular data analysis may be easily assessed by comparing the results obtained before and after these potential outliers are rejected. In this way, it may be possible to obtain more reliable results in marine climatology.

The proposed approach may also be considered as a valuable alternative to trimming procedures which are applied to ship reports before computing monthly mean summaries for 2° lat × 2° long boxes in order to reduce erroneous data losses.

REFERENCES

Barnett, V. and T. Lewis, 1978: *Outliers in Statistical Data*. John Wiley & Sons, Inc., New York.

Doornbos, R., 1966: *Slippage Tests*. 1st ed. Mathematical Centre Tracts, No. 15, Mathematisch Centrum, Amsterdam.

Feller, W., 1968: *An Introduction to Probability Theory and its Applications*. vol. 1, 3rd ed., John Wiley & Sons, Inc., New York.

Grubbs, F.E., 1950 : Sample criteria for testing outlying observations. *Ann. Math. Statist.*, 21, 27-58.

Grubbs, F.E. and G. Beck, 1972: Extension of sample sizes and percentage points for significance tests of outlying observations. *Technometrics*, 14, 847-854.

Hawkins, D.M., 1980: *Identification of Outliers*. Chapman and Hall, London.

Terray, P., 1994: An evaluation of climatological data in the Indian Ocean area. *J. Meteor. Soc.* Japan. 72, 359-386.

Terray, P., 1999: Detecting climatic signals from ship's datasets. *Proceedings of International Workshop on Digitization and Preparation of Historical Surface Marine Data and Metadata* (15-17 September 1997, Toledo, Spain). H.F. Diaz and S.D. Woodruff Eds., WMO/TD-No. 957.

Wolter, K., 1997: Trimming problems and remedies in COADS. *J. Climate*, 10, 1980-1997.

Woodruff, S.D., R.J. Slutz, R.L. Jenne and P.M. Steurer, 1987: A Comprehensive Ocean-Atmosphere Data Set. *Bull. Amer. Meteor. Soc.*, 68, 1239-1250.