Theor Appl Climatol (2009) 96:235–248
DOI 10.1007/s00704-008-0027-3

ORIGINAL PAPER

A theoretical framework for the sampling error variance for three-dimensional climate averages of ICOADS monthly ship data

Mark L. Morrissey · J. Scott Greene

Received: 5 September 2007 / Accepted: 12 March 2008 / Published online: 29 May 2008
© Springer-Verlag 2008

Abstract Meteorological and oceanographic data from ships of opportunity are the largest contributor to the world's ocean surface database and thus are extensively used to estimate the change in climatic properties over the world's oceans during the previous 150 years. The importance of these data for climate change studies underscores the need to fully understand the error associated with averages of these data. The sampling error problem is especially acute for ship data due to the fact that ships are moving platforms and, thus, report observations from constantly varying locations with time. This paper develops a theoretical framework for assessing the averaged sampling error associated with monthly, 1°×1° latitude-longitude box averaged ship data. It should be noted that the time-space distribution of ships within the averaging domain strongly affects the sampling error. This is shown in our derivation. The framework developed here can be used to improve upon existing methods for estimating the sampling error associated with three-dimensional box averages of meteorological and oceanographic data obtained from ship records. The framework is complimentary to existing methods of assessing biases and random error due to instrumentation, recording, etc. It is demonstrated mathematically that the uncertainty due to incomplete sampling is primarily a trade off between of the number of observations and their relative locations within the box as well as the inherent time-space correlation structure of the variable of interest. This work differs from other studies in that the three-dimensional interdependence of data is taken into account in deriving an expression for the sampling error.

M. L. Morrissey (✉)
School of Meteorology, University of Oklahoma,
120 David L. Boren Blvd., Suite 5900,
Norman, OU 73072, USA
e-mail: mmorriss@ou.edu

J. S. Greene
Environmental Verification and Analysis Center,
University of Oklahoma,
100 E. Boyd, SEC 410,
Norman, OU 73019, USA

1 Introduction

Knowledge of the present and future states of our climate requires that in-situ observations be taken over large time and space domains. As the oceans make up about 70% of the earth's surface area, they are of particular interest to climatologists. While satellites provide excellent spatial coverage of the earth, they have only been operational since the early 1970s. Therefore, it is essential to collect data from as far back in history as possible to assess variations in climate longer than the current satellite record can discern. In addition, satellites measure meteorological and oceanographic variables only through indirect conversion of received irradiance using statistically and physically based algorithms. Thus, in-situ data are required to check the validity of satellite estimates.

For these reasons, a cooperative project to collect and digitize meteorological and oceanographic observations taken from ships, especially merchant mariners, was begun during the 1980s (Woodruft et al. 1987). This project, referred to as the Comprehensive Ocean-Atmosphere Data Set (COADS, now called International COADS or ICOADS), is a cooperative international project to collect global weather and ocean observations primarily from ships of opportunity from 1854 through present. The year 1854 was significant in that the Brussels Maritime Conference of 1853 concluded that ship weather records should be

assimilated as much as possible and made available for the study of weather and climate. This effort has continued to this day.

The primary variables measured from ships include wind speed and direction, atmospheric pressure, sea surface temperature, and present weather. The ICOADS data set now includes data from other oceanic sources in addition to ships of opportunity (Woodruff 2001). Data collected from relatively recently installed oceanic platforms such as buoys, are currently being included in ICOADS.

While there are many different averaging techniques available, the simplest by far, and that used by the developers of the ICOADS project to produce the basic 1°×1° latitude-longitude, monthly box, is a non-weighted, arithmetic time-space averaging scheme. This scheme is used for its computational simplicity, as the number of ship reports contained in the data set is in the millions. The disadvantage of using this method is that it does not weight the data according to their relative locations in time and space and, thus, does not produce a value which minimizes the inherent sampling error in the average.

While merchant ships travel all of the world's oceans and provide important measurements from these regions, the ships also tend to travel along specific shipping lanes for speed, safety and economic reasons, rather than to optimize the sampling of environmental data. Thus, from a sampling standpoint, data collected from these ships suffer from extremely irregular time and space sampling. This calls into question the uncertainty of box averages constructed from ICOADS data. For example, ships traveling in convoys would produce a quasi-linear sampling scheme as observed in a three-dimensional sense. Ships close to each other in time and space tend to have similar data values and, thus, would be considered somewhat statistically dependent. This tends to reduce the degrees of freedom of data making up the average, thereby increasing the sampling error. In other words the 'effective number of independent number of observations' is reduced (refer to Jones et al. 1997).

It is well known that sampling error is only one contributor to the uncertainty of climate averages constructed from ship data. One ship may contribute a substantial number of reports to a specific monthly box average as it moves through the box during a month. If this ship's data contain a significant amount of systematic error, the resulting average will also contain a large amount of systematic error. Systematic and other non-sampling related random errors can arise from a multitude of sources, including instrumentation biases, transmission and transcription errors. In addition, the instrumentation itself has changed throughout the years, creating a degree of inhomogeneity in the data. It is quite difficult to quantify these types of errors and much work has been done to do just that. For example, Kent et al. (1999) focused on the random errors of individual measurements taken onboard voluntary observing ships. An extensive list of publications relevant to ship-measured observation errors, both biased and random, can found at the James Rennell Division (JRD) component of Southampton Oceanography Centre in Southhampton, United Kingdom[1]. A sampling of studies relevant to ICOADS instrumentation errors and inhomogeneities are given by Jones (1994) and Folland and Parker (1995).

Others have derived various relationships to ascertain the sampling error of climate data averaged in time, space and/or time-space. A sub-set of these studies include Parker (1984), Wigley et al. (1984), Trenberth (1984a), Trenberth (1984b), Briffa and Jones (1990), Kagan (1997, reprinted from earlier Russian manuscript), and Jones et al. (1997). Trenberth (1984a, b), in a rather complete treatment of this subject, demonstrated the importance of carefully computing the uncertainty in climate averages and developed a method of separating out climate signal from noise in time. Trenberth's development termed 'climatic noise' as error which arose from incomplete sampling. He also fully realized that there are many other contributions to 'noise' in climate averages such as instrumentation error, recording errors, etc.

The most recent method for accessing the sampling error associated with climatic time-space box averages was first developed by Briffa and Jones (1990) and later re-derived and applied to global air temperature data by Jones et al. (1997). Hereafter, we will simply refer to Jones et al. (1997), since this is the most often-referenced work in climate related journals, e.g., refer to Parker and Horton (2005) Smith and Reynolds (2005) and Brohan et al. (2006). While most of their constructs are based upon practical considerations, their derivation was incomplete. Not incorporated in their method is the contribution to the sampling error of a three-dimensional time-space average from incomplete sampling in the temporal dimension. They assumed, advertently or inadvertently, that measurements from stations from a fixed location contained zero temporal sampling error. This can be shown through their use of a time-averaged 'spatial' correlation function in their development. The effects on the sampling error from varying temporal sampling schemes or simple discrete temporal sampling were not incorporated into their scheme. They utilized an exponential decay function to represent the spatial correlation between observations. This function contained a single parameter representing only the spatial separation distance between data points within a time-space box. Since most climate variables (excepting rainfall from

[1] http://www.soc.soton.ac.uk/JRD/MET/met_pubs_all.php

accumulating rain gauges) are measured discretely in time as well as space, error is introduced into climate box averages from incomplete sampling in time.

Wolter (1997) notes that "marine air temperature has been shown to have such high daily persistence (Parker 1984) that effectively only about five independent samples can be drawn in any given month". Jones et al. (1997)'s method would lead to an overestimate of the number of independent samples taken in time which in turn leads to an underestimate of the sampling error associated with a climate box average. Thus, the dependence among data in the temporal dimension should be included in any technique for estimating sampling error in climate box averages. In the method of Jones et al. (1997), the dependence among daily air temperature data would seriously reduce the 'effective number of degrees of freedom' which is the central parameter around which their method is developed (refer to the discussion of this parameter, N_{eff}, in Jones et al. 1997).

The inclusion of the dependence effects among data separated in time considerably complicates the development of a practical method for the determination of the sampling error in time-space boxes. The equations derived in this paper incorporate quite a few parameters which would in practice be estimated using statistics calculated from data (i.e., statistical estimates). Since this is a theoretical development no assumptions need be made at this stage about the homogeneity or stationarity of the statistics that would be used to estimate the parameter values (since no statistical estimates are used in the construct). An example utilizing ship data is shown only to illustrate theoretical relationships among parameters and to clarify the meaning of the resulting sampling error expression. In this manner, this paper differs from that of Jones et al. (1997).

Thus, it is not the main purpose of this paper to develop a practical method which researchers can readily to ICOADS. Rather, due to the complexity of the issues involved with three-dimensional sampling, a framework is constructed whereby the theoretical relationships among relevant parameters are demonstrated. It is hoped that researchers will use the resulting equations and theoretical constructs to develop a practical method for operational use with three-dimensional averages which now incorporates the effect of discrete temporal sampling.

Another complication treated in this paper is the fact that ships are generally traveling and reporting data at specified, discrete time intervals, usually 6 h apart. Thus, they present ever changing sampling structure within climate boxes, say from month to month. In the past, it has been common practice to estimate the uncertainty of three-dimensional ship data averages using indirect methods such as Monte Carlo sub-sampling schemes (Legler 1991, Cayan 1992, Gulev and Hasse 1998). While relatively easy to apply, these methods assume that the ship reports are randomly distributed within the averaging domain, which in most situations they are not. One can easily imagine a likely sampling scheme associated with moving ships. Observed in a three-dimensional sense, such a 'network' in a time-space coordinate system would likely be linear or clustered. For example, one ship traveling from east to west through an ICOADS box during a month would produce a linear network of data reports. A convoy of ships traveling northwest to southeast through one corner of an ICOADS box could be likened to a clustered network. Thus, given a set of fixed parameter values (e.g., the point variance, σ_p^2), the different sampling 'networks' from one ICOADS box to another would produce a different value for the sample error for different boxes simply due to the varying relative time-space 'locations' of the ship reports within each box.

In developing an expression for the uncertainty associated with three-dimensional 'networks', a review of the theoretical work done with various fixed two-dimensional structured rain-gauge networks eases the process (e.g., Journel and Huijbregts (1989); Rodriguez-Iturbe and Mejia 1974; Morrissey 1991, Morrissey and Greene 1993, 1998, Krajewski et al. 2000, and Gebremichael et al. 2003). The developments in these studies allow one to make the rather large leap from two to three dimensions. These studies have all utilized random function theory which we will utilize as well.

One very important, but complex, notion is the definition of sampling error. We define the sampling error as the additional variance contributing to the temporal box variance which is due to incomplete sampling. Ideally, with ICOADS one would like to determine the sampling error associated with each value of a set of climatic box averages. This is impractical since it requires accurate estimates for the required parameters made from statistics constructed from very small sample sizes (i.e., the data within a given 'box'). Thus, the best that can be done is to estimate the sampling error associated with ICOADS boxes having a given sampling structure given specified assumptions about the statistics used to estimate the parameter values. For example, suppose we want to estimate the sampling error associated with a specific ICOADS box centered on the equator and the dateline for January 1959. This box has a given sampling structure in time and space which most likely differs from most other boxes. There are not enough data within this single box to accurately estimate the parameter values that are required in the computation of the sampling error for this one box. If we utilized data outside this box in time and/or space to estimate the required statistics, as did Jones et al. (1997), then we can determine the sampling error for this ICOADS box. However, the validity of the assumption that the

statistics computed from regional or long-term data are representative of parameter values for that box (e.g., the true variance of the observations within the box) depends upon the bounds of the domain where data are gathered to compute these statistics and the sample size used. Therefore, the best we can say is that for a given time-space sampling structure, the computed sampling error is accurate only as far as the statistics accurately represent the box parameters. Another way to state this problem is to say that the estimated sampling error is representative of boxes having a set time-space sampling structure over the domain where the statistics are stationary and homogeneous. The difficulty, of course, is defining this domain.

Perhaps the most important parameter used in our derivation is the box mean. A box mean estimated using only the data available from a single box cannot be used since the difference between the estimated mean (using the data) and the parameter value representing the mean is due to the sampling error itself. Thus, we will take the tack of using the long-term mean for the spatial location of the box in our development. We do this since the long-term mean parameter is likely to be accurately estimated from the long-term data from the box spatial location, assuming the data has been de-trended. Using the long-term mean parameter thereby forces us to re-define the sampling error as the time-averaged sampling error associated with the selected box's sampling structure.

2 Application of random function theory to geophysical data

Perhaps the best way to understand the theoretical approach taken in this paper is to take an analogy from geostatistics (Journel and Huijbregts 1989) applied to geology and mining. In these fields, it is common to work with three-dimensional spatial volumes to assess the amount of ore or mineral content within each spatial volume. In the case presented in this work, the 'time' dimension is analogous to the 'vertical' spatial dimension. To obtain a basic understanding of random function theory one assumes that $z(\mathbf{x})$ represents a random variable at a point, $\mathbf{x}$, in three-dimensional space. It is assumed that $z(\mathbf{x})$ has a probability distribution function or more properly a probability density function (PDF). Quoting Journel and Huijbregts (1989), "The problem is to represent the variability of the function $z(\mathbf{x})$ in space" (or space-time in the case of the deviations presented in this paper). There may or may not be observations of $z(\mathbf{x}_0)$ at location $\mathbf{x}_0$. However, there usually are observations at nearby locations $(z(\mathbf{x}_i), i \neq 0)$ all of which are assumed to have the same PDF. If nearby locations having observations are correlated with each other, the problem then becomes the determination of the function $Z(\mathbf{x})$ where position vector $\mathbf{x}$ varies throughout the three-dimensional domain, i.e., D^3. The set of auto-correlated random variables is the random function $Z(x)$ (i.e. $z(x_0), z(x_1), \ldots \in Z(x)$). The problem then reduces to one of determining the correlation function between the different random variables in $Z(x)\ \forall\ \mathbf{x} \in \mathbf{D}^3$.

It should be noted that an observation is considered as one realization of the random variable $z(\mathbf{x})$, the set of which constitutes the random function $Z(\mathbf{x})$. Given that the random variable $z(\mathbf{x})$ has a certain PDF the question becomes, how to estimate this PDF? Obviously, the PDF cannot be determined from one observation. However, given a certain domain, it is often appropriate to assume that the probability density of $z(\mathbf{x})$ is the same throughout the domain (i.e., homogeneous and stationary in D^3) at least up to the second order. Thus, within this domain, all the available observations can be considered realizations from the same probability distribution which then can be used to estimate the PDF. The question then becomes whether or not the assumption of homogeneity/stationarity is appropriate for a given domain under study.

The meaning of 'second order' is that the mathematical expectation of the second central moment of the random variable $z(\mathbf{x})$ is independent of location in the averaging domain and that the covariance between any two random variables is a function of distance only. Statistics computed from the data are used to estimate the value of certain 'parameters' (e.g., s^2 is used to estimate σ^2). In reality, physical processes vary over time and space. Thus, the parameters are only homogeneous/stationary within certain dimensional boundaries and then only approximately. In any use of the expression developed here, it must assume that the expectation of the different moments of the random variables (i.e., the parameters) are at least quasi-homogeneous/stationary within the domain of interest only and that errors resulting from this assumption will be small enough for our results to be useful.

3 Domain descriptions

While the technique described in this paper can be applied to any space-time averaging domain, it simplifies matters to select a specified domain within which we define a linear, three-dimensional coordinate system. For clarity, the derivations in this paper will utilize two distinct domains over which the expectation of the required moments of different random variables will be taken. The first domain is a cubic domain D_1^3 bounded in time and space representing month, M. This domain represents a single ICOADS 1°×1°, monthly box. A cubic space-time coordinate system (x, y, t) is defined within this domain. A visual representation of this domain is shown in Fig. 1. The variables X, Y are

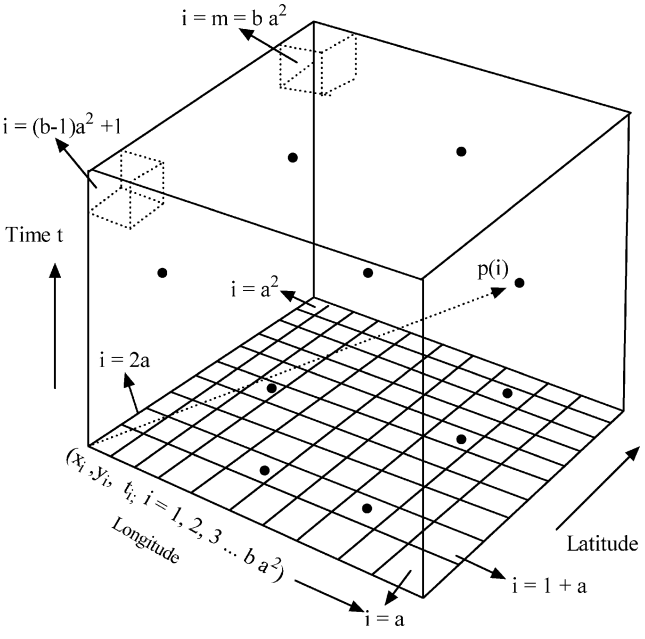

Fig. 1 The ICOADS 1°×1°×1 month box domain, . Hypothetical ship observations in time and space are shown by the *dots*. Examples of two 'sub-boxes' are shown by the *two cubes within the ICOADS box*

longitude and latitude, respectively, of the centroid of the ICOADS box and M is time or 'month' the box represents. The highest reporting resolution in ICOADS for individual ships is to the nearest 0.1°×0.1° latitude, longitude with each ship reporting approximately at a 6-h time interval. Thus, the box domain can be divided into 10×10×120 (i.e., 12,000) sub-boxes with each sub-box representing a 0.1° latitude, longitude, 6-h cube (assuming a 30-day month). The vector $\mathbf{p}(i)$ is a location vector from the domain origin (e.g., the southwest corner at the beginning of the month) to the centroid of a given sub-box 'i' within the box domain $\left(\mathbf{p}(\mathrm{i}) \in D_1^3\right)$. Thus, $\mathbf{p}(i)$ is pointing to the sub-box centered at longitude x_i, latitude y_i, and the 6-h period, t_i (i=1 to m, where m is the total number of sub-boxes within D_1^3 with m=12,000). The random variable $R(\mathbf{p}(i), M)$ represents any ship-measured variable such as air temperature, sea surface temperature, wind speed, etc. within the sub-box located at vector $\mathbf{p}(i)$. The 'length' of $\mathbf{p}(i)$ with respect to the origin is

$$\mathbf{p}(i) = \sqrt{x_i^2 + y_i^2 + t_i^2} \tag{1}$$

We have little choice but to assume that a single ship observation within a sub-box adequately represents the sub-box's true volume average and that the average of more than one ship value in a sub-box represents the value of that sub-box. A second domain is now defined as that bounded in space but unbounded in time. This domain, which we refer to as D_2^3 is defined so that it would be clear to the reader over which domain the statistical expectation operator is taken. A visual representation of this domain is shown in Fig. 2. Note that domain D_2^3 encompasses domain D_1^3, i.e., $D_1^3 \in D_2^3$. We will also use two different expectation operators, one for the expectation over D_1^3 (i.e., E_1) and one for that taken over D_2^3 (i.e., E_2).

4 Error variance due to incomplete sampling in time and space

Given a fixed time-space distribution of ship reports within domain D_1^3, it is of interest to determine the time-averaged sampling error variance within an ICOADS box. For example, a user may want to know what the time-averaged sampling error variance is for an ICOADS box with the distribution of ship reports like that shown in Fig. 1. In the derivation presented in this paper, the delta symbol (e.g., Parker 1984; Morrissey et al. 1995) is used as a binary variable indicating whether a given sub-box contains one or more observations. If more than one ship report is in the same sub-box, then the observations in that sub-box are averaged to produce a single value for that sub-box. Thus, the delta symbol $\delta(\mathbf{p}(i), M)$ refers to the presence (i.e.,

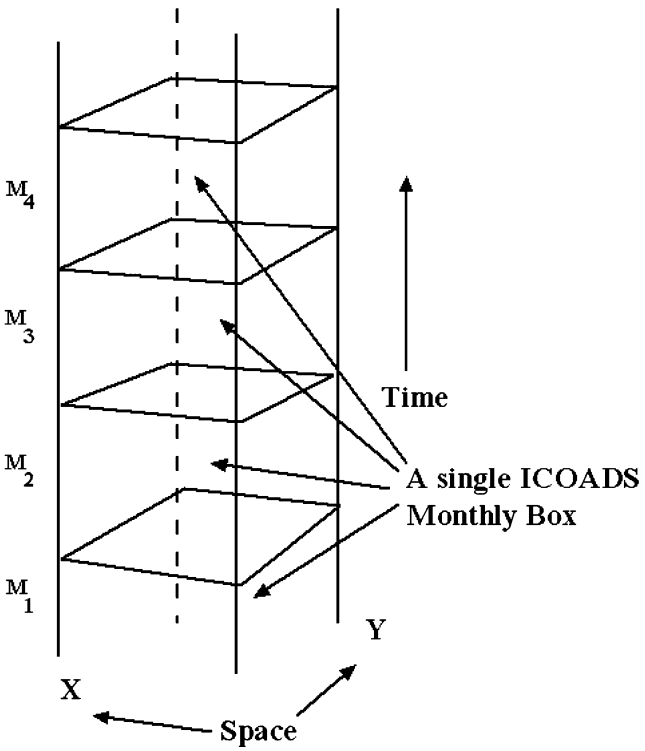

Fig. 2 A visual representation of domain D_2^3. With $M = -\infty$ to $+\infty$ representing different months. This domain can be visualized as one ICOADS box fixed in space but for different months

$\delta(\mathbf{p}(i),M) = 1$) or absence (i.e., $\delta(\mathbf{p}(i),M) = 0$) of an observation within a sub-box defined by a vector, $\mathbf{p}(i)$., i.e.,

- $\delta(\mathbf{p}(i),M) = 0 \rightarrow$No ship observation in a sub-box located at vector position $\mathbf{p}$(i) and month M within D_1^3 and
- $\delta(\mathbf{p}(i),M) = 1 \rightarrow$At least one ship observation in sub-box located at vector position $\mathbf{p}(i)$ and month M within D_1^3.

Within domain D_1^3 there are 'n' ($n \in \{0,1,2,\ldots m\}$) number of sub-boxes containing at least one ship observation and m total sub-boxes ($i = 1,2, \ldots m$). The total number of sub-boxes must be greater than or equal to the number of occupied sub-boxes, n,

$$n(M) = \sum_{i=1}^{m} \delta(\mathbf{p}(i),M) \quad \forall \quad n \leq m \tag{2}$$

For the following derivations we will utilize the long-term box mean parameter defined as the expectation of $R(\mathbf{p}(i),M)$ over D_2^3. In other words, the long-term mean is defined as the expectation of the variable of interest, R, within the domain shown in Fig. 2, i.e., over all M,

$$\mu = E_2[R(\mathbf{p}(i),M)], \quad \mathbf{p}(i) \in D_1^3; \quad M \in D_2^3 \tag{3}$$

An estimate of the long-term mean parameter can be easily computed from the available data using

$$\mu \approx \overline{R} = \sum_{M=1}^{M_{\max}} \sum_{i=1}^{n(M)} \frac{R(\mathbf{p}(i),M)\delta(\mathbf{p}(i),M)}{n(M)}, \quad \mathbf{p}(i) \in D_1^3; \; M \in D_2^3 \tag{4}$$

where M=1 and $M_{\max}$ redefine as the first and last months in the record, respectively. The total number of available observations within that ICOADS box for month M is defined as n. Generally, a very large sample size is available to estimate this value which gives the estimate of μ (i.e.$\overline{R}$) a relatively small error variance.

The estimated mean value of a random variable for a specific ICOADS ship box (i.e., D_1^3), is defined using

$$\overline{R}(M) = \sum_{i=1}^{m} \frac{R(\mathbf{p}(i),M)\ \delta(\mathbf{p}(i),M)}{n(M)} \tag{5}$$

The mean parameter value for the same random variable is defined as the expectation over the domain shown in Fig. 1, i.e.;

$$\mu(M) = E_1[R(\mathbf{p}(i),M)], \quad \mathbf{p}(i) \in D_1^3 \tag{6}$$

Note that $\overline{R}(M)$ is a statistic and $\mu(M)$ is a parameter. Also note that while $\mu(M)$ is a constant for domain D_1^3 (i.e., the month M is fixed), it is a random variable within domain D_2^3 since within this domain the month varies (i.e., M=1, 2, 3 ….). Note that the difference between $\overline{R}(M)$ and $\mu(M)$ is due to sampling error only.

We now will develop an expression for the sampling error variance, given a fixed sampling distribution within D_3^2. We will begin by finding the total variance (i.e., variance in time) of $\overline{R}(M)$ about the long-term mean parameter, μ. This is done by taking the expectation of the square of the box mean estimate about the long-term mean parameter, i.e.;

$$\sigma_T^2 = E_2\left[\left(\overline{R}(M) - \mu\right)^2\right] \tag{7}$$

By taking the expectation over D_2^3 (using E_2) we are including all possible realizations of $\overline{R}(M)$ within D_2^3. The total variance, σ_T^2 , includes variance contributed by the sampling error (i.e., σ_e^2, $n < m$) and the box to box temporal signal variance (i.e., σ_S^2). The signal variance is mathematically expressed as,

$$\sigma_S^2 = E_2\left[(\mu(M) - \mu)^2\right] \tag{8}$$

From the analysis of variance relationship the variances can be partitioned into

$$\sigma_T^2 = \sigma_e^2 + \sigma_S^2\ ; \quad n(M) \leq m; \quad \sigma_e^2, \sigma_s^2 \geq 0 \tag{9}$$

The sampling error variance can now be found by rearranging the above equation, i.e.,

$$\sigma_{\mathrm{e}}^2 = \sigma_{\mathrm{T}}^2 - \sigma_{\mathrm{S}}^2 \tag{10}$$

In the case of a completely sampled box (i.e., n=m with $\delta(\mathbf{p}(i),\ M) = 1\ \forall\ i$) the sampling error variance, σ_{e}^2 , becomes zero and, thus, the expression in Eq. (10) becomes

$$\sigma_{\mathrm{T}}^2 = \sigma_{\mathrm{S}}^2; \quad (n(M) = m)\ \forall\ M \tag{11}$$

since the difference between the total variance and the signal variance is the sampling error variance (this was also demonstrated by Jones et al. (1997, e.g., 2) in a slightly different manner). Below we will develop an expression for the difference between σ_S^2 and σ_T^2, which is, of course, the sampling error variance, σ_e^2.

Before developing this expression two additional expressions need to be defined. The point variance parameter is defined as the expectation of the squared deviation of the values in all sub-boxes within D_2^3 about the long-term mean,

$$\sigma_{\mathrm{p}}^2 = E_2[(R(\mathbf{p}(i),M) - \mu)^2] \tag{12}$$

Likewise the covariance is also defined using the long-term mean and is equal to

$$\begin{aligned} Cov[R(\mathbf{p}(i),M),R(\mathbf{p}(k),M)] &= E_2[(R(\mathbf{p}(i),M)-\mu)(R(\mathbf{p}(k),M)-\mu)] = \\ E_2[R(\mathbf{p}(i),M)R(\mathbf{p}(k),M)] &- \mu E_2[R(\mathbf{p}(i),M)] - \mu E_2[R(\mathbf{p}(k),M)] + \mu^2 \\ = E_2[R(\mathbf{p}(i),M)R(\mathbf{p}(k),M)] &- \mu^2 \\ \because\ E_2[R(\mathbf{p}(i),M)] &= E_2[R(\mathbf{p}(k),M)] = E_2[\mu] = \mu \end{aligned} \tag{13}$$

5 Use of the anomaly in determining the standard error

To simplify our expressions we will use the anomaly defined as

$$r(\mathbf{p}(i),M) = R(\mathbf{p}(i),M) - \mu \tag{14}$$

where its expectation over D_2^3 is zero ($\overline{r}(M) = E_1[r(\mathbf{p}(i),M)]$). Substituting Eq. (14) into Eq. (7), the total variance parameter using the anomaly now becomes

$$\sigma_T^2 = E_2[\overline{r}(M)^2] \tag{15}$$

The anomalous box parameter mean can be defined as $\mu_r(M) = E_1[r(\mathbf{p}(i),M)]$ and is estimated from

$$\overline{r}(M) = \sum_{i=1}^{m} \frac{r(\mathbf{p}(i),M)\ \delta(\mathbf{p}(i),M)}{n(M)} \tag{16}$$

If we substitute Eq. (16) into Eq. (15) and expand we get

$$\sigma_T^2 = E_2\left[\begin{array}{l} \dfrac{1}{n(M)^2}\displaystyle\sum_{i=1}^{m} r(\mathbf{p}(i),M)^2\delta(\mathbf{p}(i),M)^2 + \\ \dfrac{2}{n(M)^2}\displaystyle\sum_{i=1}^{m-1}\sum_{k=i+1}^{m} r(\mathbf{p}(i),M)r(\mathbf{p}(\mathrm{k}),M) \\ \qquad \delta(\mathbf{p}(i),M)\delta(\mathbf{p}(\mathrm{k}),M) \end{array}\right] \tag{17}$$

The point variance parameter using the anomaly is now,

$$\begin{aligned} \sigma_{\mathbf{p}}^2 = E_2[(R(\mathbf{p}(i),M)-\mu)^2] &= E_2[(r(\mathbf{p}(i),M)+\mu-\mu)^2] \\ &= E_2[r(\mathbf{p}(i),M)^2] \end{aligned} \tag{18}$$

In a similar manner, the anomaly can be used to simplify the covariance expression[2],

$$\begin{aligned} \mathrm{Cov}[R(\mathbf{p}(i),M) - R(\mathbf{p}(k),M)] &= E_2[(R(\mathbf{p}(i),M)-\mu)(R(\mathbf{p}(k),M)-\mu)] \\ &= E_2[r(\mathbf{p}(i),M)\,r(\mathbf{p}(k),M)] \end{aligned} \tag{19}$$

An important note is needed here. Although we are taking the expectation of the covariance over domain D_2^3 we only use the covariance at lags within the confines of D_1^3 as shown by the limit, m, in Eq. (17). This means that the largest lag is the largest diagonal distance within D_1^3. For an ICOADS box near the equator this is roughly 157 km–120 6-h periods. Estimates of the covariance can be made using regional and/or long-term data as long as the homogeneity/stationarity assumption is met. The correlation is defined as

$$\rho[\mathbf{p}(i)-\mathbf{p}(k)] = \frac{\mathrm{Cov}[\mathbf{p}(i)-\mathbf{p}(k)]}{\sigma_p^2} \tag{20}$$

Taking the expectation of the two terms in Eq. (17) and substituting the expressions for the variance and correlation we arrive at

$$\sigma_T^2 = \frac{\sigma_{\mathbf{p}}^2}{n(M)^2}\left[\sum_{i=1}^{m}\delta(\mathbf{p}(i),M)^2 + 2\sum_{i=1}^{m-1}\sum_{k=i+1}^{m}\rho[\mathbf{p}(i)-\mathbf{p}(k)]\,\delta(\mathbf{p}(i),M)\delta(\mathbf{p}(k),M)\right] \tag{21}$$

This expression is similar to that found for a two-dimensional rain-gauge network by Rodriguez-Iturbe and Mejia (1974).

Now that we have an expression for σ_T^2 we need one for the signal variance σ_S^2 to finally arrive at an expression for the sampling error variance, σ_e^2. One way to find σ_S^2 is to use the total variance expression Eq. (21) and assume that we have complete sampling by allowing the sample size, n to equal m and, by consequence, letting $\delta(i,M) = 1\ \forall\ i \in D_1^3$. This reduces the total variance Eq. (21) to an expression for the signal variance, σ_S^2,

$$\sigma_S^2 = \frac{\sigma_{\mathbf{p}}^2}{m}(1 + (m-1)\overline{\rho}) \tag{22}$$

where $\overline{\rho}$ results from taking the average of $\rho[|\mathbf{p}(i) - \mathbf{p}(k)|]$ over all locations, $\mathbf{p}(i)$, $\mathbf{p}(k)$ within D_1^3. Note that with m sufficiently large, Eq. (22) reduces to

$$\sigma_S^2 = \sigma_p^2\overline{\rho} \tag{23}$$

Using the same principles, Jones et al. (1997) found the same expression (Eq. 2 in Jones et al. 1997). Thus the signal of box averages with complete sampling (and no systematic error) is completely determined by the time-space averaged covariance only since $\sigma_{\mathbf{p}}^2\overline{\rho} = \overline{Cov}$.

By substituting Eqs. (21) and (23) into Eq. (10), we arrive at an expression for the sampling error variance,

$$\sigma_e^2 = \tfrac{\sigma_{\mathbf{p}}^2}{n(M)^2} \times \left[\sum_{i=1}^{m}\delta(\mathbf{p}(i),M) + 2\sum_{i=1}^{m-1}\sum_{k=i+1}^{m}\rho[\mathbf{p}(i)-\mathbf{p}(k)]\,\delta(\mathbf{p}(i),M)\delta(\mathbf{p}(k),M)\right] - \sigma_{\mathbf{p}}^2\overline{\rho} \tag{24}$$

or

$$\sigma_e^2 = \sigma_{\mathbf{p}}^2\left[\frac{1}{n(M)} + \frac{2}{n(M)^2}\sum_{i=1}^{m-1}\sum_{k=i+1}^{m}\rho[\mathbf{p}(i)-\mathbf{p}(k)]\,[\delta(\mathbf{p}(i),M)\delta(\mathbf{p}(k),M)] - \overline{\rho}\right] \tag{25}$$

[2] For simplicity we will hereafter allow $\mathrm{Cov}[R(\mathbf{p}(\mathbf{i}),M),R(\mathbf{p}(\mathbf{k}),M)]$ to equal $\mathrm{Cov}[|\mathbf{p}(\mathrm{i}) - \mathbf{p}(\mathrm{k})|]$

Note that $\delta(\mathbf{p}(i),\ M)^2$ is equivalent to $\delta(\mathbf{p}(i),\ M)$. We now have an expression for the error variance as a function of sample size, *n(M)*, the point variance, $\sigma_{\mathbf{p}}^2$, the time-space correlation, $\rho[|\mathbf{p}(i)-\mathbf{p}(k)|]$weighted by the relative distribution in time and space of ship reports within D_1^3 and true average correlation, $\overline{\rho}$ within D_1^3. The square root of Eq. (25) is the commonly known as the standard error equation. For a given ICOADS box the values for *n(M)*, *m* and $\delta(\mathbf{p}(i),M)\ \forall\ i \in D_1^3$ are constants. Values for the parameters, $\sigma_{\mathbf{p}}^2$, $\rho[\mathbf{p}(i)-\mathbf{p}(k)]$ and $\overline{\rho}$ must be estimated from data.

Note that with uncorrelated data expression Eq. (25) reduces to the common expression for the error variance associated with averages of independent data, $\frac{\sigma_{\mathbf{p}}^2}{n(M)}$. Also note that the last term in the expression is constant for a given three-dimensional physical correlation structure and is not affected by the time-space location of ship reports. Thus, it is the second term which determines the effect of different sampling structures on the sampling error.

6 Examination of the standard error equation

6.1 The point variance

An estimate of the point variance, $\sigma_{\mathbf{p}}^2$ can be easily computed using data within the averaging domain from the following equation,

$$\sigma_{\mathbf{p}}^2 \approx s_{\mathbf{p}}^2 = \frac{1}{T}\sum_{M=1}^{T}\sum_{i=1}^{m}\frac{r^2(\mathbf{p}(i),M)\delta(\mathbf{p}(i),M)}{n(M)} \tag{26}$$

where T is the total number of months available in the data record (i.e., j=1, 2, 3, …T). However, the questions arises, how closely does the statistic, $s_{\mathbf{p}}^2$, approximate the parameter $\sigma_{\mathbf{p}}^2$? Since $\sigma_{\mathbf{p}}^2$ is defined about the long-term mean in our derivation (refer to Eq. 12), the answer will depends solely upon the sample size, $\sum_{M=1}^{T} n(M)$. If the sample size is relatively small refer to Jones et al. (1997) for several other ways to estimate $\sigma_{\mathbf{p}}^2$.

6.2 Issues concerning the space-time covariance and correlation function

Most of the information in Eq. (25) lies in the correlation function representing the domain, D_2^3. Computing a representative time-space correlation function is difficult at best. Meteorologists and hydrologists have traditionally assumed that time-space estimates of environmental quantities have 'separable' covariance functions (Rodriguez-Iturbe and Mejia 1974). In other words, it was assumed that the correlation function could be constructed from purely independent spatial and temporal covariance functions, such that

$$\mathrm{Cov}(\mathbf{d},t) = Cov(\mathbf{d})\mathrm{Cov}(t) \tag{27}$$

where $\mathbf{d}$ is the spatial component of the vector distance between two points, and t is the time component between those two points. While the computation of this function was relatively simple, its practical use implies several unrealistic assumptions (Kyriakidis and Journel 1999). For one, the product of two independent covariance functions adequately representing the time-space covariance is usually hard to justify physically. Also, functions of this type do not allow for time-space interactions (i.e., the time and space covariance are simply inversely proportional to each other). It has also been noted by Gneiting et al. (2007) that a space-time covariance process may not be 'fully symmetric', that is

$$\mathrm{Cov}[R(s_1,t_1)R(s_2,t_2)] = \mathrm{Cov}[\mathrm{R}(\mathrm{s}_1,\mathrm{t}_2)\mathrm{R}(\mathrm{s}_1,\mathrm{t}_1)] \tag{28}$$

where s_1,t_1 and s_2,t_2 are two different spatial and temporal coordinates, respectively. With atmospheric and oceanic properties, it is noted that the fields are generally not fully symmetric due to transport effects by winds or ocean currents. Thus, the concept of symmetry of the covariance function may be more appropriately applied using a Lagrangian reference frame. (May and Julian 1998). Gneiting et al. (2007) pointed out that others (Gneiting 2002 Stien 2005; de Luna and Genton 2005) have shown the effects of the transport of environmental properties on the symmetry. Tests for non-separability have been developed (Scaccia and Martin 2005), Lu and Zimmerman 2005). For a complete description of these issues regarding the time-space covariance function refer to Gneiting et al. (2007).

Gneiting et al. (2007) notes that while these issues concerning the covariance function require further research, they do not rule out the use of a covariance function as a tool for describing the statistical characteristics of three-dimensional physical data. They go on to mention that "one needs to be cognizant of the effects that violations of certain assumptions may cause on the resulting statistics". Gneiting et al. (2007) also demonstrate that these assumptions are not always invalid and suggest that "the relevancy of a given covariance function to a given study will vary on a case by case basis".

6.3 A practical set of parametric functions for a non-separable time-space covariance

Given the reservations of assuming a separable time-space covariance relationship, Cressie and Huang (1999) and

Gneiting (2002) formulated a general set of parametric functions for non-separable time-space covariance that may be applicable to ICOADS data. We will summarize their formulation below.

Gneiting (2002 refer to Eq. 14) provides a family of potential covariance functions where one relevant example is

$$\mathrm{Cov}(\mathbf{d},t) = \frac{\sigma_{\mathbf{p}}^2}{(at^{2\alpha}+1)^{\tau}} \exp\left(-\frac{c|\mathbf{d}|^{2\gamma}}{\left(a|\mathbf{t}|^{2\alpha}+1\right)^{\beta\gamma}}\right); \quad (29)$$

$$(\mathbf{d};t) \in D_1^3$$

where symbols a and c are scaling parameters for time and space, respectively. The smoothness and shape of this class of functions are controlled by the parameters, $\alpha, \gamma \in (0,1]$. The symbol $\tau \geq 1$ is another parameter which can be adjusted to improve the fit of the data. The corresponding correlation function is simply

$$\rho(\mathbf{d},t) = \frac{1}{(at^{2\alpha}+1)^{\tau}} \exp\left(-\frac{c|\mathbf{d}|^{2\gamma}}{\left(a|\mathbf{t}|^{2\alpha}+1\right)^{\beta\gamma}}\right); \quad (30)$$

$$(\mathbf{d};t) \in D_1^3$$

The interesting, but not surprising, aspect of this class of functions is that they were constructed from familiar correlation functions for environmental time and space processes (i.e., the exponential and power law functions; refer to Gneiting 2002). The parameter $\beta \in [0,1]$ controls the interaction of the covariance between time and space. For example, with β=0, we have a separable model where time and space covariance is proportional to each other. As $\beta \rightarrow 1$ the strength of the interaction between time and space covariance increases. Also, as $\beta \rightarrow 1$ the correlation values at non-zero lags fall off less and less, as compared with a separable model (Gneiting 2002).

If Eq. (30) is to be used to represent the time-space correlation structure of a variable, an appropriate set of parameters must be found which provide an adequate fit to the data. To fit this three-dimensional function to data we must first find the function parameters associated with time—(i.e., Cov(0,t))—and then do the same with space (i.e., Cov(**d**,0)). In the case with Eq. (30) with **d**=0 the temporal correlation function would be

$$\rho(0,t) = \frac{1}{(at^{2\alpha}+1)^{\tau}} \quad (31)$$

With t=0 the spatial correlation function becomes

$$\rho(\boldsymbol{d},0) = \exp\left(-c\boldsymbol{d}^{2\lambda}\right) \quad (32)$$

If parameters can be adequately estimated which allow each function to fit the data separately in time and then space then, after Cressie and Huang (1999) and Gneiting (2002), the following function can be minimized to obtain a value for the time-space interaction parameter, β, over $\beta \in [0,1]$

$$W(\beta) = \sum_{i,j}\sum_{t}\left(\frac{\widehat{\rho}(d_{i,j},t) - \rho(d_{i,j},t|\beta)}{1-\rho(d_{i,j},t|\beta)}\right)^2 \quad (33)$$

where $\mathbf{d}_{i,j}$ is the spatial lag between stations i and j. Note that $\widehat{\rho}$ is the estimated correlation value found from the data and ρ is the functional value at the corresponding lag values.

The steps in fitting such a model to data involve first assuming that β=0 (i.e., a separable model) and fitting the spatial (with t=0) and temporal (with $d_{i,j}$=0$\forall i,j$) correlation functions separately to the data. Once the various parameters are found which provide adequate time and space fits, they can be substituted into Eq. (30) with β=0 and Eq. (30) can be used in Eq. (33) to find a value for β which minimizes equation $W(\beta)$ using numerical methods. Finally, the resulting β value is used to reconstruct Eq. (30) with parameter values which produce a three-dimensional correlation function with a reasonable fit to the data.

6.4 Testing the fit of the correlation function

To find a set of parameters for Eq. (30) which may be appropriate for use in examining our values using the standard error equation, we experimented with a sample of ship air-temperature data taken from a subset of ICOADS data. The data were taken from the region bounded by latitude 0 to 10°N and longitude 140–150°E from 1961 to 2004. We further define a smaller ICOADS box-sized domain D_1^3, which is bounded by 0°×1°N and 145°×146°E and 120 consecutive 6-h periods (i.e., the size of a standard ICOADS 1°×1° , monthly box). The estimated correlation values at different space-time distances within D_1^3 were computed using data over the length of the record. Only lags out to approximately 157 km-120 6-h periods (the diagonal distance in a monthly, 1°×1° box) were required in the computation. However, extending the lags somewhat further out allowed a more precise inspection of the goodness of fit. Using a least-squares approach, both Eqs. (31) and (32) were fit to the resulting correlation estimates and best-fit parameter values were found for each equation. The results are shown graphically in Figs. 3 and 4 for space and time, respectively. The best-fit estimates for the

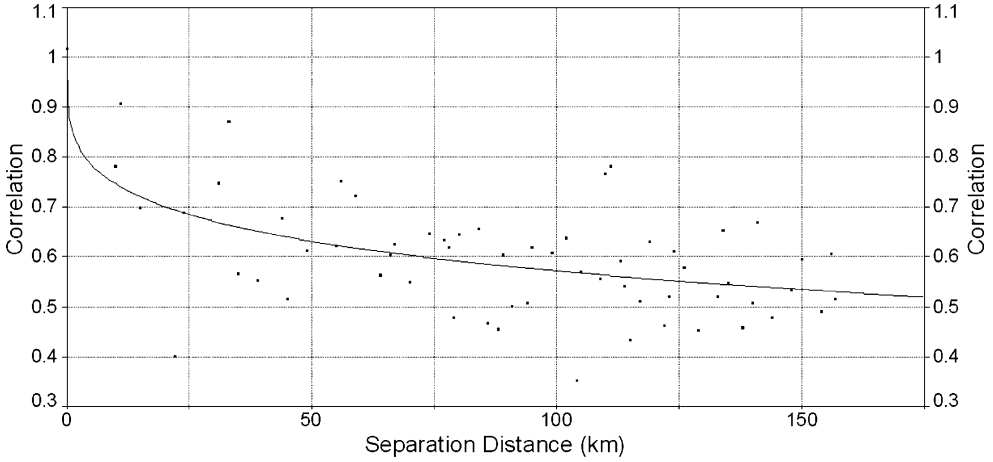

Fig. 3 The least-squares fitted spatial correlation function

parameters for each equation are shown in Table 1. From the figures, it can be observed that both the spatial and temporal functions to fit the data reasonably well (at least for illustrative purposes). Note that the correlation values fall off quite rapidly with time lag with an approximately *e*-folding time separation of 2.5 days. However, after this, the correlation levels off at approximately 0.1 throughout the month. This is not inconsistent with the results found by Wolter (1997) mentioned earlier.

To demonstrate the procedure of determining the magnitude of separability, a value for the interaction parameter, β in Eq. (33) was found which minimized $W(\beta)$ between $0 \leq \beta \leq 1$. By using a least-squares method a value of β equal to 0.1 was found to minimize $W(\beta)$. This indicates that degree of separability in our example is quite high, but not zero. Thus, assuming $\beta = 0.1$, the resulting time-space correlation function representing our selected region is

$$\rho(\mathbf{d},t) = \frac{1}{(1+27.5t^{0.999})^{0.184}} \; e^{\frac{-0.153\mathbf{d}^{0.282}}{(t^{1.998}+1)^{.014}}}$$

$$\forall\ t > 0,\ \ t \in \{0,\ 120;\ \ \text{six hour periods}\},\ \mathbf{d} \in \{0,\ 157\text{km}\} \tag{34}$$

To incorporate this correlation function into Eq. (25) note that $\mathbf{d}$ and t are simply the spatial and temporal components of the separation vector, $|\mathbf{p}(i) - \mathbf{p}(k)|$. It should be stressed that this is an isotropic correlation structure in space and in reality a more complicated expression taking into account anisotropy may be warranted. However, for simplicity in illustrating the use of the sampling error expression, we will assume an isotropic spatial correlation structure. The resulting three-dimensional correlation function is shown in Fig. 5

7 Finding a value for $\overline{\rho}$

The last term in Eq. (25) involves the estimation of the averaged correlation within the averaging domain D_1^3, i.e. $\overline{\rho}$. Now that a representative correlation function has been found for the domain, the correlation averaged among all pair of points within the domain D_1^3 can be computed.

For ICOADS data and the coordinate system define in this paper, it is relatively simple to compute $\overline{\rho}$ for ICOADS boxes directly. One can use the determined correlation function and numerically compute $\overline{\rho}$ by averaging the correlation over all pairs of sub-boxes within D_1^3.

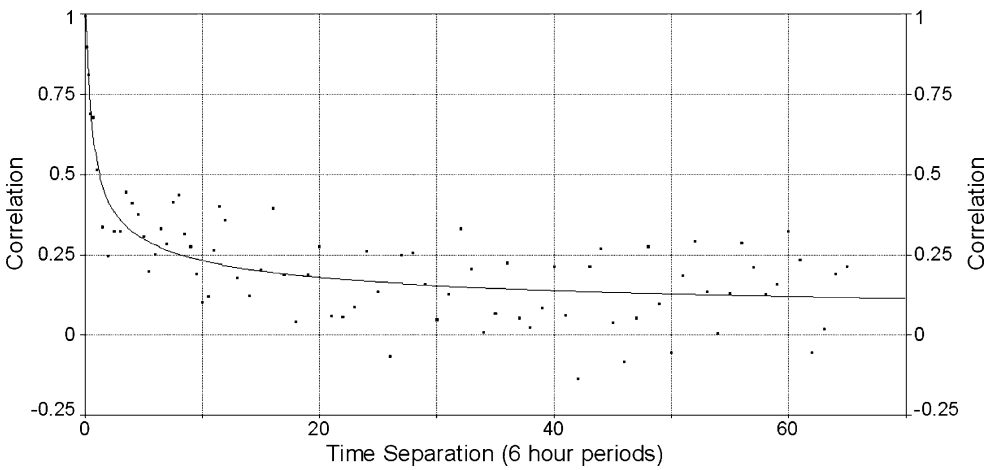

Fig. 4 The least-squares fitted temporal correlation function

Table 1 Estimated values of the parameters found for the fitted space–time correlation function

	Spatial	Temporal
A	NA	27.5
α	NA	0.999
τ	NA	0.184
C	0.153	NA
γ	0.141	NA

However, given the large number of sub-boxes (i.e., 10×10×120=12,000) it becomes computationally intensive when considering the number of pairs of sub-boxes (i.e., 71,989,050 pairs). A much more efficient way to estimate $\overline{\rho}$ is to reduce the dimensions of the ICOADS box somewhat. For ICOADS boxes a logical transformation would be to convert the 10×10×120 system to a 10×10×10 system. This can be done by defining a new 'spatial' variable as

$$d' \; sub-boxes = \frac{d}{11.1} \; \frac{sub-boxes}{km}$$

since each side of a sub-box represents approximately 11.1 km (near the equator). Also, since there are 120 6-h periods per month this means that there are 12 $\frac{6\ hour\ periods}{sub-box}$. To transform the temporal dimension from 1–120 sub-boxes to 1–10 sub-boxes, a new temporal coordinate system is defined as

$$t' \; sub-boxes = \frac{t}{12} \; \frac{sub-box}{6\ hour\ period}$$

Thus, using these new coordinates we now have a 10×10×10 domain which gives only 494,550 pairs of sub-boxes. The correlation function in the new coordinate system is now

$$\rho(\mathbf{d}', t') = \frac{1}{(1+27.5(12t')^{0.999})^{0.184}} \; e^{\frac{-0.153(11\mathbf{d}')^{0.282}}{((12t')^{1.998}+1)^{.014}}}$$
$$\forall \; \{t', \mathbf{d}'\} > 0, \;\; t' \in \{1,\ 10;\ \text{sub}-\text{boxes}\}, \;\; \mathbf{d}' \in \{1, 10;\ \text{sub}-\text{boxes}\} \tag{35}$$

Using the estimated parameters values shown in Table 1 with β=0.1 and Eq. (35), simple averaging was used to find the average correlation $\overline{\rho}$ within an ICOADS box. This value turned out to be $\overline{\rho} = 0.102$. This relatively small value is reasonable when one considers the very small correlation values with lag in the temporal dimension.

8 An examination of the variance reduction factor, VF

Different forms of the expression within the large brackets in Eq. (25) have found by many authors for two dimensions (e.g., Rodriguez-Iturbe and Mejia 1974, Morrissey et al. 1995, Jones et al. 1997, Kagan 1997). This expression in its various forms has been referred to as the 'variance reduction factor' or VF by Rodriguez-Iturbe and Mejia (1974). The primary effect of this factor is to account for the effect of sample size, the correlation structure and the relative spacing of reports in D_1^3 on the sampling error variance. Higher values of VF correspond to larger the sampling error values. The VF is equivalent to the reciprocal of the effective number of degrees of freedom described by Jones et al. (1997; i.e. $\frac{1}{N_{\text{eff}}}$). An examination of Eq. (25) shows that it consists of three terms. The first term, which is the reciprocal of the number of reports in an ICOADS box, acts to reduce the error variance with increasing sample size. The second term acts to increase the error variance from correlation or 'dependence' among the data points. This is due to a reduction in the degrees of freedom from ships that are close to each other in either time and/or space. Note that the magnitude of this term is weighted by the points' location relative to each other within D_1^3 only. In essence, this term represents the averaged time-space correlation within D_1^3 as estimated by the existing data. The third term, which is not a function of a particular boxes sampling structure, is the true correlation averaged within D_1^3. Note that this is simply a statistical expression of the signal variance (when multiplied by the point variance). This can be observed through a comparison with Eq. (23). This term is a constant for a given correlation structure and is independent of sampling structure or sample size.

For our experiment, it is of interest to examine VF by varying the sample size n to its extremes. If, for example, n equalled the total number of sub-boxes, m, and we assume m to be very large, then the VF reduces to $\sigma_p^2\overline{\rho} - \sigma_p^2\overline{\rho} = 0$. This can be seen in Eq. (25) by noting that as $n \to m \to \infty$ the first term becomes zero and the second term becomes $\overline{\rho}$. Thus, the sampling error variance equals zero as it should with complete sampling. If n equals one, VF would reduces to $1-\rho$. In this case the total

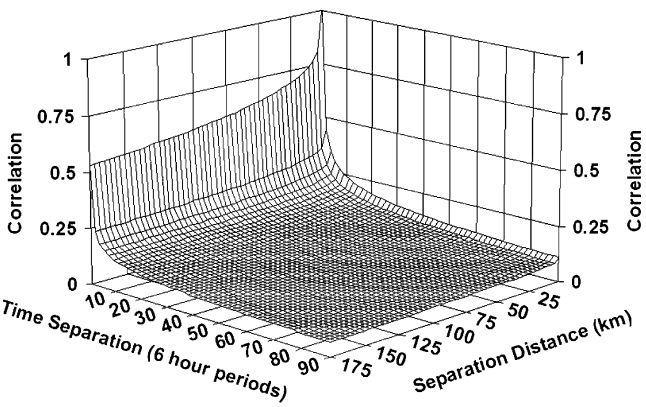

Fig. 5 The least-squares fitted time-space correlation function

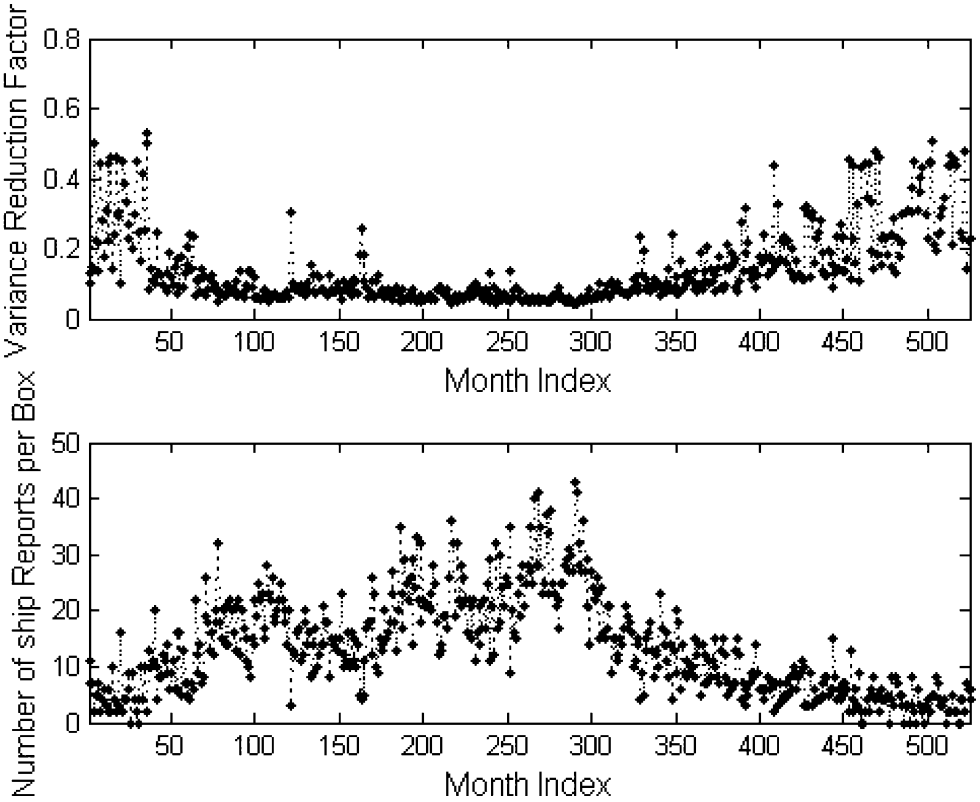

Fig. 6 The variance reduction factor, VF, for ICOADS monthly, 1 degree boxes bounded by 7°N, 8°N and 146–147°E from the months spanning January 1961 through December 2004 (*top*). The number of observations per box for that location from the same period. Note that VF values representing months when there were zero observations are not plotted (*bottom*)

variance (Eq. 21) of the boxes is entirely a function of the point variance, i.e.,

$$\sigma_T^2 = \sigma_e^2 + \sigma_s^2 = \sigma_{\mathbf{p}}^2(1-\rho) + \sigma_{\mathbf{p}}^2\rho = \sigma_{\mathbf{p}}^2 \tag{36}$$

This should be obvious since the box averages are now simply constructed from a single point and the total variance of the boxes is the same as the variance of the individual points.

If the correlation was zero everywhere in D_1^3 the error variance will reduce to the standard expression for independent data, i.e. $\frac{\sigma_p^2}{n}$, since term 2 and term 3 in Eq. (25) would reduce to zero leaving term 1.

9 An illustrative application of the sampling error variance, $\overline{\sigma_e^2}$

To observe how the standard error derivation above may be utilized in a practical sense, ship air-temperature data within a 1°×1° location was selected within a time-space domain bounded by 7°N, 84°N and 146–147°E and from 1961 through 2004. These data were used to estimate the parameters in Eq. (25). The variance reduction factor, VF, was then computed for each month. Note that, given a fixed set of parameters, only the sample size and the relative ship report locations varied from month to month. Equation (35) was used for the representative space-time correlation function.

The results of this experiment are shown in the top plot of Fig. 6. The VF values are relatively high and erratic during the beginning and end of the period. During the middle of the period the VF values slowly decrease, reach a minimum near the 250 to 200th month of the period and then start to increase again. This indicates that the sampling error is quite high during the beginning and end of the period relative to the middle portion. To see what may have contributed to the variation in VF, we also plotted the number of observations per month throughout the period (bottom Fig. 6). While the number of observations per month is relatively low at the beginning and the end of the period, they cannot completely explain the lack of variation in VF during the middle of the period. The number of observations reaches a maximum near the 300th month but fluctuate considerably from month to month, while VF is relatively constant. One might expect the large monthly fluctuations in the number of observation to produce similar fluctuations in VF, but this is not seen. The most likely explanation for this is that, given less than approximately ten observations per month, the sample size is the controlling factor in VF (i.e., term 1). With greater than ten observations per month, the relative time-space positions of the observations (i.e., term 2) control the magnitude of VF. Thus, given the estimated time-space correlation structure and months having more than ten observations, the individual observations become somewhat redundant due to the higher correlation values resulting from ship

reports located (in time and space) close together (i.e., N_{eff} becomes constant). Possible reasons for the variations in number of observations for this particular location throughout the record may be the altering ship travel lanes, data not updated near the end of the period and a fall off of observations near the beginning of the period, etc.

Nevertheless, what is clear from this experiment is that the uncertainly of the estimated box averages is considerably less near the beginning and end of the period. Note that the sampling error throughout the middle portion of the period for this location is approximately constant at 15% of the point variance compared to about 50% for beginning and end of the period. The results also indicate that a hypothetical addition of more ship reports during the middle portion of the period would not substantially alter the magnitude of the sampling error since the data are already quite statistically interdependent.

10 Conclusions

A mathematical expression for the sampling error variance of standard ICOADS 1.0°×1.0° latitude/longitude monthly boxes has been derived. This study differs from similar past studies in that the resulting expression incorporates the effects of incomplete sampling in both time and space on the sampling error.

This study is unique in climate research in that it was realized that the correlation function used in the computation of the sampling error must represent the interdependence among observations in not only space, but time as well, if observations are taken discretely in all three dimensions. Using ship data from the equatorial western Pacific, the resulting time-space correlation function was found to have a relatively large amount of time-space separability as indicated by the estimated value of β which was 0.1 which ranges from zero to one. It would be interesting to determine if this is true for mid-latitude regions as well.

The illustrative example using data from the 1.0°×1.0° latitude/longitude location, indicated that, for that particular area, the sampling error was quite high during the beginning and end of the record compared with the middle of the period (approximately 50% compared to 15%). The effect of ship-report locations on the sampling error variance within a standard ICOADS box was shown to be a strong function of the time-space correlation structure and the sample size itself. In the particular case of ship-recorded air temperature from the ICOADS box in the western equatorial Pacific, it was shown that 1/VF, or the effective degrees of freedom, is strongly affected only when ship reports are less than approximately ten observations per month. Several reasons for this were hypothesized. However, the point of this exercise was to demonstrate the potential usefulness of the derived sampling error expression and questions into why the results of the expression behaved the way they did requires further research.

Acknowledgements The authors would like to thank those who attended the Brussels CLIMAR-II Conference held in 2003 for their encouragement to pursue this work. We also would like to acknowledge the generous help from NOAA's Climate and Global Change Program's Climate Observations Element contract number NA17RJ1227. Thanks also go out to M. Klatt for his help in organizing the data.

References

Briffa KR, Jones PD (1990) Basic chronology statistics and assessment. In: Cook E, Kairiukstis L (eds) Methods of dendrochronology: applications in the environmental sciences. Kluwer, Dordrecht, The Netherlands, pp 137–152

Brohan P, Kennedy JJ, Harris I, Tett SFB, Jones PD (2006) Uncertainty estimates in regional and global observed temperature changes: a new data set from 1850. J Geophys Res 111:12,106–12,127

Cayan D (1992) Variability of latent and sensible heat fluxes estimated using bulk formulae. Atmos Ocean Techn 301–42

Cressie N, Huang H (1999) Class of nonseparable, spatio-temporal stationary covariance functions. J Am Stat Assoc 95:1330–1340

De Luna X, Genton MG (2005) Predictive spatio-temporal models for spatially sparse environmental data. Stat Sin 15:547–568

Folland CK, Parker DE (1995) Correction of instrumental biases in historical sea-surface temperature data. Q J R Meteorol Soc 121:319–367

Grebremichael M, Krajewski WF, Morrissey M, Langerud D, Huffman GJ, Adler R (2003) Error uncertainty analysis of GPCP monthly rainfall products: a data based simulation study. J Appl Meteorol 42(12):1837–1848

Gneiting T (2002) Nonseparable, stationary covariance function for space-time data. J Am Stat Assoc 97:590–600

Gneiting T, Genton MG, Guttorp P (2007) Geostatistical space-time models, stationarity, separability and full symmetry. In: Finkenstadt B, Held L, Isham V (eds) Statistical methods for spatio-temporal systems, CRC, Boca Raton, FL, USA, pp 151–175

Gulev SK, Hasse L (1998) North Atlantic wind waves and wind stress fields from voluntary observing ship data. J Phys Oceanogr 28:1107–1130

Jones PD (1994) Hemispheric surface air temperature variations: a reanalysis and an update to 1993. J Climate 7:1794–1802

Jones PD, Osbborn TJ, Briffa KR (1997) Estimating sampling errors in large-scale temperature averages. J Climate 10:2548–2568

Journel AG, Huijbregts ChJ (1989) Mining geostatistics. Academic, San Diego, CA, USA, 600 pp

Kagan RL (1997) Averaging of meteorological fields. Kluwer, Dordrecht, The Netherlands, 279 pp

Kent KC, Challenor PG, Taylor PK (1999) A statistical determination of the random observational errors present in voluntary observing ships meteorological reports. J Atmos Oceanic Tech 16:905–914

Krajewski WF, Ciach GJ, McCollum JR, Bacotiu C (2000) Initial verification of the Global Precipitation Climatology Project monthly rainfall over the United States. J Appl Meteor 39:1071–1086

Kyriakidis PC, Journel AG (1999) Geostatistical space–time models: a review. Math Geol 31:6651–684

Legler D (1991) Errors in five-day mean surface wind and temperature conditions due to inadequate sampling. J Atmos Oceanic Technol 8:705–712

Lu N, Zimmerman DL (2005) Testing for directional symmetry in spatial dependence using the periodogram. J Stat Plan Infer 129:369–385

May DR, Julian PY (1998) Eulerian and Lagrangian correlation structures of convective rainstorms. Water Resour Res 34:2671–2683

Morrissey ML (1991) Using sparse raingauges to calibrate satellite-based rainfall algorithms. J Geophys Res Atmos 96:18561–18571

Morrissey ML, Greene JS (1993) A comparison of two satellite-based rainfall algorithms using Pacific atoll raingauge data. J Appl Meteorol 32(2):411–425

Morrissey ML, Greene JS (1998) Uncertainty of satellite rainfall algorithms over the tropical Pacific. J Geophys Res 103:19,569–19,576

Morrissey ML, Maliekal JA, Greene JS, Wang J (1995) The uncertainty of simple spatial averages using rain gauge networks. Water Resour Res 31:2011–2017

Parker DE (1984) The statistical effects of incomplete sampling of coherent data series. J Climatol 4:445–449

Parker DE, Horton B (2005) Uncertainties in central England 1878-2003 and some improvements to the maximum and minimum series. Int J Climatol 25:1173–1188

Rodríguez-Iturbe I, Mejía JM (1974) The design of rainfall networks in time and space. Water Resour Res 10(4):713–728

Scaccia L, Martin RJ (2005) Testing axial symmetry and separability of lattice processes. J Plann Infer 131:19–39

Smith TM, Reynolds RW (2005) A global merged land-air-sea surface temperature reconstruction based on historical observation (1880-1997). J Climate 18:2021–2036

Stein ML (2005) Space-time covariance functions. J Am Stat Assoc 100:310–321

Trenberth KE (1984a) Some effects of finite sample size and persistence on meteorological statistics, part I: autocorrelations. Mon Weather Rev 112(12):2359–2368

Trenberth KE (1984b) Some effects of finite sample size and persistence on meteorological statistics, part II: potential predictability. Mon Weather Rev 112(12):2369–2379

Wigley TML, Briffa KR, Jones PD (1984) On the average value of correlated time series, with applications in dendroclimatology and hydrometeorology. J Clim Appl Meteorol 23:201–213

Wolter K (1997) Trimming problems and remedies in COADS. J Climate 10:1980–1997

Woodruff SD (2001) COADS updates including newly digitized data and the blend with the UK meteorological office marine data bank and quality control in recent COADS updates. Proceedings of Workshop on Preparation, Processing and Use of Historical Marine Meteorological Data, Tokyo, Japan, 28–29 November 2000, Japan Meteorological Agency and the Ship and Ocean Foundation, Tokyo, pp 9–53

Woodruff SD, Slutz RJ, Jenne RL, Steurer PM (1987) A comprehensive ocean-atmosphere data set. Bull Am Meteorol Soc 68:1239–1250