Standard Error Estimation of COADS Monthly Mean Winds

Mark L. Morrissey

Oklahoma Climatological Survey
University of Oklahoma
Norman, Oklahoma 73019

Jose A. Maliekal

Dept. of the Earth Sciences
SUNY Brockport
Brockport, New York 14420

Introduction

The Comprehensive Ocean-Atmosphere Data Set (COADS) contains ship observations from 1854 to 1990. Several versions of this data set are available including one containing individual ship reports of various meteorological and oceanographic measurements. The extremely large data volume of this version necessitated another version containing spatially and temporally averaged reports representing monthly averages over $2° \times 2°$ areas (i.e. boxes). The highly variable distribution of ships, especially in tropical regions, produces considerable uncertainty in these box averages. One measure of the random error associated with the monthly averages is the standard error of the mean. Unfortunately, the small sample sizes and the difficulties of incorporating changing ship locations has hindered an early development of a robust standard error equation for COADS box averages.

By defining a grid system based upon the minimum spatial resolution of COADS individual ship reports (i.e. $0.1° \times 0.1°$ latitude-longitude), a practical standard error relationship is presented which can be applied to the monthly averages. By developing the equation using the long-term mean, sample statistics, such as the point variance and the lagged correlation, are relatively unbiased making the standard error equation quite robust. The equation was initially developed for two dimensional fields by Morrissey et al. (1994), but is easily applied to three dimensional box averages.

Equation Development

A grid system is used whereby each $0.1° \times 0.1°$grid location is numbered systematically from 1 to 400 (i.e. 20×20) for hour 1 in the month, from 401 to 800 for hour 2 and so on. The sample time-space average for month *m* can be defined by,

$$\bar{X}_m = \frac{\sum_{i=1}^{N} x(i)\delta(i)}{n}$$

where *x(i)* is an individual ship report (e.g. the u component) located at grid i, N is the total number of grid points and *n* is the number of ship reports in the box. An indicator variable, ∂(i) is one if a ship is present at grid 1 and is zero otherwise. This variable is used to incorporate the ship locations into the equation. Also, the overall mean of the ship reports is removed from each *x(i)* value. The field mean for month *m* is defined by,

$$\mu_m = \frac{\sum_{i=1}^{N} x(i)}{N}$$

Morrissey et al. (1994) derived a practical form of the standard error equation by substituting these two expressions into

$$\sigma_{\bar{x}}^2 = E\left[\bar{X}_M - \mu_m\right]^2$$

and expanding to arrive at

$$\sigma_{\bar{x}} = \sigma\left[\frac{1}{n} + \frac{2}{n^2}\sum_{L=1}^{N-1}\rho(L)w(L) - \frac{2}{N^2}\sum_{L=1}^{N-1}(N-L)\rho(L)\right]^{\frac{1}{2}}$$

where σ = *point variance*

$$w(L) = \sum_{iu=1}^{N-L} \delta(i)\delta(i+L)$$

ρ(L) = *lagged auto correlation*

where *w*(*L*) is a weight factor which is a function of the network configuration. The quantity in the large brackets is the variance factor which accounts for the sample size and the dependence among the reports in both time and space. The second term in the variance factor accounts for the variance of the sample mean about the long-term mean and the third term accounts for the variance of the population monthly field about the long-term mean. The second term is an estimate of the average correlation within the times-pace volume. A clustered network will generally provide overestimates of the average correlation since w(L) will be large when the correlation is large (i.e. L is small). Thus, the difference between terms 2 and 3 should be rather large for a clustered network.

Examining the Standard Error Equation

By assuming a two dimensional anisotropic exponential spatial correlation function (Fig. 1), the behavior of the standard error given specific grid configurations can be observed. Four grid configurations are shown in Fig. 2, a random, a clustered, a uniform and a linear network (i.e. linear network #1). A fifth network (not shown) is a simple 90 degree rotation of the linear network #1 (i.e. linear network #2). By multiplying the denominator in the exponent of the correlation function by a constant, the e-folding distance along the major axis can be varied. It can be observed (Fig. 3) for different e-folding distances, the variance factor, and hence the standard error, generally decreases with increasing correlation. This results from the increased areal representation of a given ship report (i.e. increased dependence among ship reports). Also, linear network #2, which is aligned along the major axis of the spatial correlation function provides higher variance factor values per e-folding distance than does linear network #1. This is due to a larger amount of redundant information measured by linear network #2 (i.e. w(L) is large when the correlation is high). It can also be observed that for all of the networks except the random and uniform networks, the variance factor initially increases with increasing enfolding distances. This results from the increasing difference between terms two and three in the variance factor with these networks. This behavior is dependent upon, not only the network configuration, but the correlation function as well.

Relevance to COADS

The use of the long-term mean in the sample statistics means that the sample statistics should be relatively unbiased. Thus, the standard error equation is fairly robust. By estimating a representative time-space correlation function for different oceanic regions, the standard error of monthly box averaged wind components can be found given different COADS ship distributions. Thus, standard error estimates can now be produced for COADS $2° \times 2°$ monthly averages.

References

Morrissey, M.L., J.A. Maliekal, J.S. Greene, and J. Wang, 1994: The uncertainty of simple spatial averages: The standard error equation, submitted to the *J. Climate*.

Figure 1: The two dimensional exponential correlation function used to test the standard error equation.

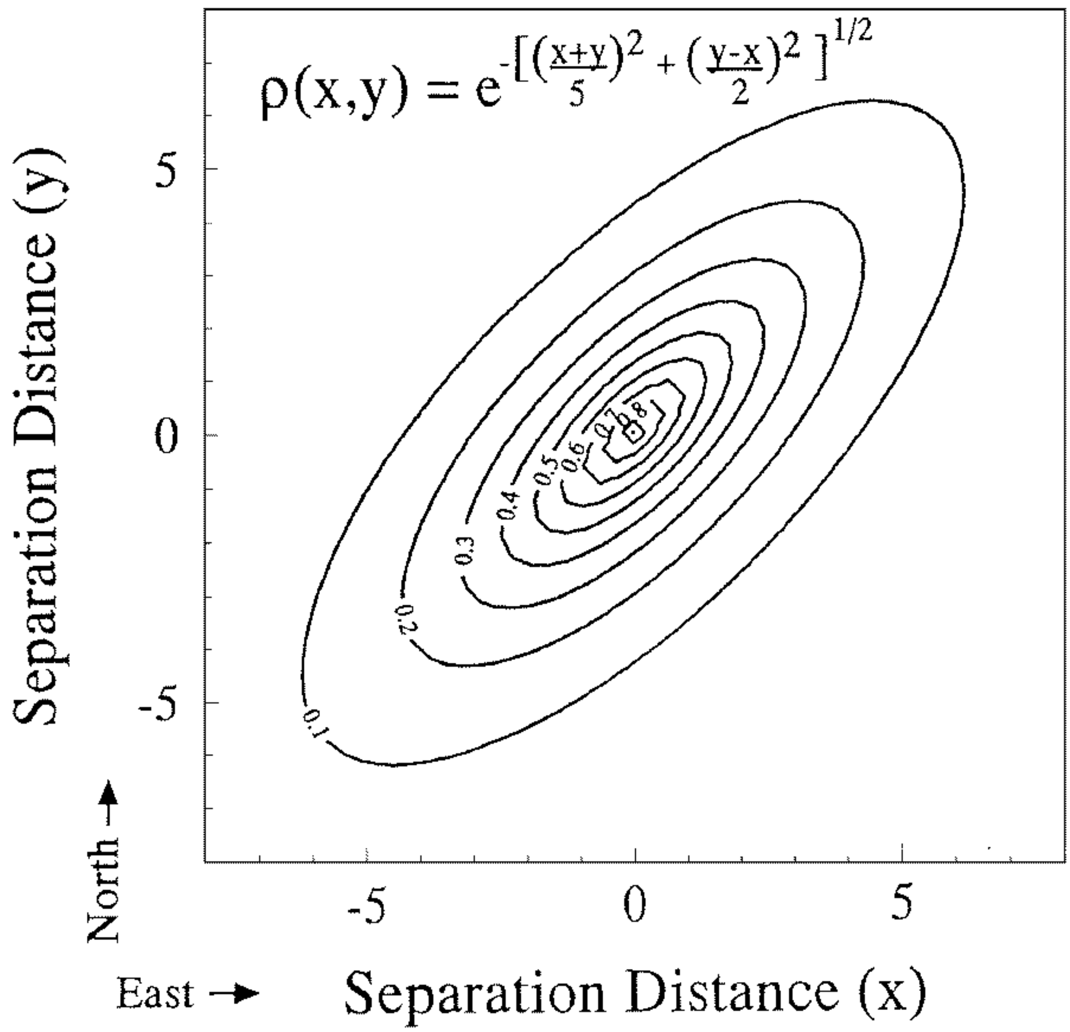

Figure 2: Four sample network configurations overlaid on a 100×100 grid.

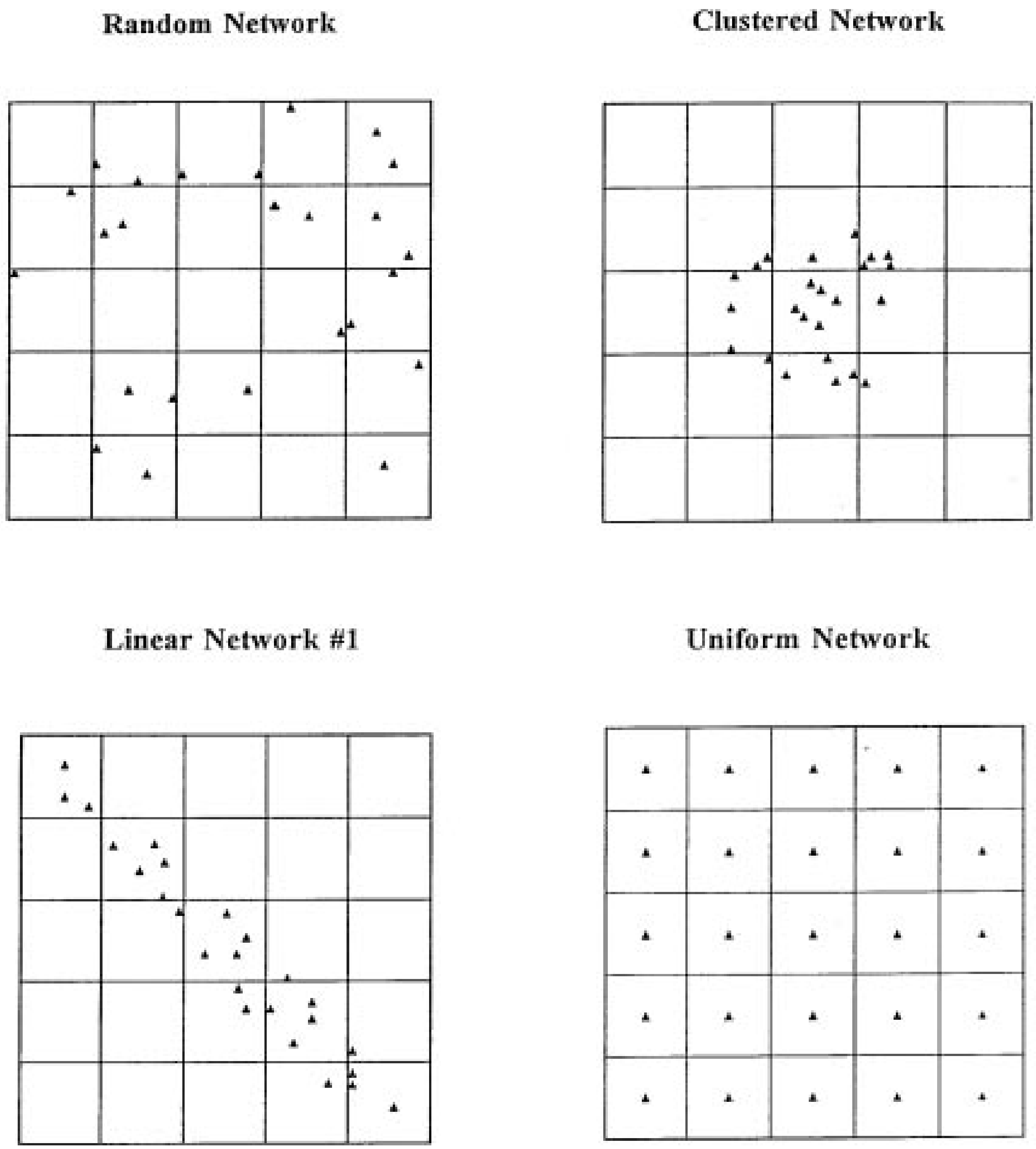

Figure 3: The variation of the variance factor defined in the text as a function of the e-folding distance using the anistropic spatial correlation function shown in Fig. 1.