The uncertainty of climatological values

Edgar G. Pavia

Centro de Investigación Científica y de Educación Superior de Ensenada, Ensenada, Baja California, Mexico

Received 14 May 2004; revised 10 June 2004; accepted 2 July 2004; published 30 July 2004.

[1] The uncertainty of climatological values is calculated with a method using the average absolute deviation from the mean, or absolute error. This method is related to the mean of the magnitude of the anomalies and its applicability is readily verified with a proposed approximate numerical technique, in contrast with the standard error method whose applicability is cumbersome to verify. Different data sets were tested with this technique showing that the method applies in a wide variety of geophysical cases, including non-Gaussian cases. When it applies, the absolute error method provides additional information on the mean, such as the error due to the sample size. *INDEX TERMS:* 1620 Global Change: Climate dynamics (3309); 3309 Meteorology and Atmospheric Dynamics: Climatology (1620); 4215 Oceanography: General: Climate and interannual variability (3309). **Citation:** Pavia, E. G. (2004), The uncertainty of climatological values, *Geophys. Res. Lett.*, *31*, L14206, doi:10.1029/2004GL020526.

1. Introduction

[2] In the simplest and schematic way we estimate a climatological value (or simply climatology), μ, by averaging a record of N available data,

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{1}$$

and similarly we obtain a measure of its uncertainty by calculating the standard error (or standard deviation of the sample mean),

$$\varepsilon_s = \sigma(x)/\sqrt{N}, \tag{2}$$

where

$$\sigma(x) = \left[\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^{1/2}, \tag{3}$$

is the standard deviation of x. This is done in order to state the range in which we are confident our climatology lies, $\bar{x} - \varepsilon_s \leq \mu \leq \bar{x} + \varepsilon_s$ or $\mu = \bar{x} \pm \varepsilon_s$. A common use of this result is to discern anomalies compared to ε_s, that is if $|\delta x_i = x_i - \bar{x}| \leq \varepsilon_s$ then $\delta\mu_i = 0$, else we reduce the magnitude of the anomalies by ε_s to find $\delta\mu_i$ ($|\delta\mu_i| = |\delta x_i| - \varepsilon_s$). In other words, if the uncertainty is greater than $|\delta x_i|$ the climatological anomaly, $\delta\mu_i$, is null and otherwise its magnitude is reduced by ε_s.

[3] When the data are Gaussian and uncorrelated the above method is valid and statistically robust; but in many geophysical situations the data do not clearly meet these requirements, and thus we would like to know if our measure of uncertainty is valid or not. In these situations ε_s is sometimes called the "lowest limit" of the standard error [see, e.g., *Pavia and Graef*, 2002], strictly meaning that the uncertainty of the climatological value is greater than ε_s. Thus the purpose of this work is to propose a new method whose applicability we can test in more general situations. Of course, there are many works on different aspects of this subject, and in this short note we do not intend to make a thorough review of it since we deal only with a small part of this problem, namely the validity test of the method to estimate the uncertainty of climatological values. Nevertheless, as a background the reader is referred to the works of *Leith* [1973], for an example of early work on this subject, *Guttman* [1989], for a review of climatological normals; *Taylor* [1997], for a more general statistical approach to error analysis; and *von Storch and Navarra* [1999] and *von Storch and Zwiers* [1999], for a discussion of applications of statistical techniques to climate research. We begin by introducing an alternative to equation (2) as a measure of uncertainty.

2. The Absolute Error

[4] Climatological data might have large and non-Gaussian deviations basically because each outcome x_i is not an independent measure of $\bar{x}$. A measure of the mean deviation may be

$$\alpha(x) = \frac{1}{N}\sum_{i=1}^{N}|x_i - \bar{x}|, \tag{4}$$

or absolute deviation. This leads to another measure of uncertainty called the absolute error

$$\varepsilon_a = \alpha(x)/\sqrt{N}, \tag{5}$$

which, in some cases, may be preferable to equation (2). For example: i) To compare climatological values, because $\alpha(x) \leq \sigma(x)$ and the smaller the measure of uncertainty the stricter the comparing criterion; ii) To obtain additional information about $\bar{x}$, because $\alpha(x)$ is equivalent to the mean error of the poorest climatological estimate of $\bar{x}$ [$\bar{x}(N = 1)$], that is if we select only one outcome to represent the climatology and repeat this process for every outcome, the mean error will be given by $\alpha(x)$; and iii) To test the applicability of the uncertainty method, because it is straightforward to verify if an expression equivalent to

0094-8276/04/2004GL020526$05.00

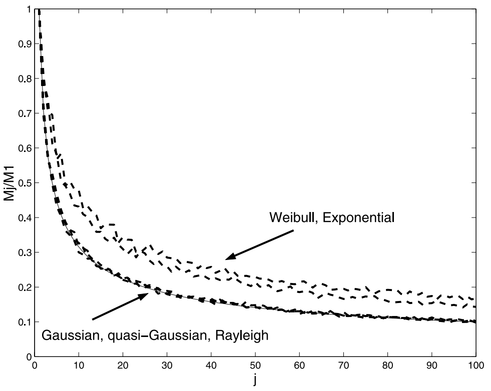

Figure 1. M_j/M_1 for artificial data. The corresponding curves for Gaussian, quasi-Gaussian and Rayleigh distributed data approximate the theoretical curve ($j^{-1/2}$).

equation (5) applies or not for a particular data set. This last point is explained in more detail in the next section.

3. The Equivalent Absolute Error

[5] Let us define the sample-size dependent absolute deviation from the mean as the error due to the sample size

$$Z(j) \equiv |\overline{x_j} - \mu|, \qquad (6)$$

where

$$\overline{x_j} \equiv \frac{1}{j}\sum_{n=1}^{j} x_n, \quad j = 1, 2, ..., J, \qquad (7)$$

in this case j (not N) is the sample size, J its greatest value (to be determined a priori), and μ is the expected value of x, $\mu = E[x]$. Indeed we are interested in the expected value of equation (6)$M_j = E[Z(j)]$, $j = 1, 2,\ldots, J$; that is the mean error of different estimates of $\overline{x_j}$ with the same sample size j. Since equation (6) is non-analytical calculating M_j is usually rather difficult but if we assume that it is related to equation (5) then $M_j \sim j^{-1/2}$, or $M_j = C/\sqrt{j}$, and by inspection for $j = 1$, $C = M_1 = E[Z(1)]$, thus we get

$$M_j \equiv E[Z(j)] = \frac{E[Z(1)]}{\sqrt{j}}. \qquad (8)$$

[6] This is the sample-size dependent average absolute deviation of the mean, which for $\mu = \overline{x}$ and $j = N$ would be equivalent to the absolute error, that is $M_N \approx \varepsilon_a$. But for equation (8) to apply to a particular data set estimates of M_j with different j should behave like the right-hand side of equation (8). If they do the above assumption is correct and M_N is an equivalent measure of uncertainty. One way to estimate M_j in order to check our method is suggested in the next section.

4. Numerical Verification

[7] As an example consider a very simple non-Gaussian but symmetric ideal case. Suppose x_j are drawn from a 0 to 1 uniform distribution, then $\mu = 0.5$ and $E[Z(1)] = 0.25$. Obviously μ represents the climatology and $E[Z(1)]$ represents the average of its poorest estimation with a sample size $j = 1$. To reduce M_j in equation (8) by one order of magnitude we need $j = 100$, therefore in this case $M_{100} = 0.025$.

[8] To test our method we use a linearized least-squares technique by rewriting equation (8) as $M_j = aj^b$, and $\ln(M_j) = \ln a + b \ln j$, or $Y_j = A + BX_j$, where $Y_j = \ln(M_j)$, $A = \ln a$, $B = b$, $X_j = \ln j$. That is we minimize an expression of the form

$$\sum_{j=1}^{J} \epsilon_j^2 = \sum_{j=1}^{J} \left[(A + BX_j) - Y_j \right]^2,$$

where from the above we choose $J = 100$. To calculate Y_j, $j = 1, 2,\ldots, J$ we approximate M_j by averaging 1000 random samples

$$M_j \sim \frac{1}{1000}\sum_{k=1}^{1000} Z(j)_k, \qquad (9)$$

where $Z(j)_k = |\overline{(x_j)}_k - \mu|$; that is the x_j are taken at random with replacement from the uniform distribution. For this ideal case this procedure yields $a = e^A = 0.23$, $M_{100} = 0.0232$ and $b = B = -0.50$, indicating that equation (8) satisfactorily applies in this uniform case. Other examples with artificial data of known distribution gave similarly acceptable results in Gaussian, quasi-Gaussian and Rayleigh cases, and non-acceptable results in exponential and Weibull (with shape parameter $c < 1$) cases (see Figure 1).

5. Real Data Examples

[9] As with artificial data our method does not always apply to real data (see Figure 2), which can be attributed to different reasons. For example for mean annual records of sea surface temperature (SST) data in Vizcaine Bay, Mexico, we find a climatological value of $\mu = 20.7 \pm 0.1$ C for the period 1982–1999, but the validity test yields $b \sim -0.2$ (perhaps due to internal correlations or small N) and thus $\varepsilon_s = 0.1$ is just the "lowest limit" of the uncertainty because for autocorrelated data the effective N (N_{eff}) in equation (2) is smaller than N ($N_{eff} < N$). Similarly, wind speed data with large percentage of calms (zero values) yield $-b \ll 1/2$ and thus in these cases neither equation (8) applies nor a real measure of uncertainty can be obtained because the data are exponentially distributed as the artificial data example given in the previous section.

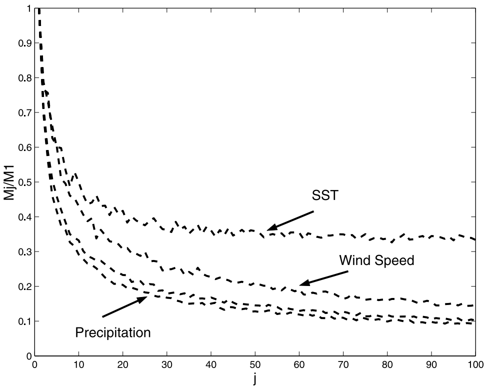

Figure 2. M_j/M_1 for real data. The corresponding curves for total annual precipitation data in Ensenada and San Diego approximate the theoretical curve ($j^{-1/2}$).

[10] Next we compare two climatologies where our method is found to be valid; the first case is the 1948–2003 annual precipitation climatology in Ensenada, Baja California, Mexico, and the second case is the 1850–2003 annual precipitation climatology in San Diego, California. Annual precipitation data are much more closely Gaussian than higher frequency data so in this sense our method is expected to apply. The precipitation in these two stations is highly cross-correlated for coincident periods [*Pavia and Badan*, 1998], but it has been suggested that the Ensenada data may be overestimated (the station is located in a dam between two small hills) and that the San Diego data set may be more reliable [*Pavia and Graef*, 2002] since it is subject to a more rigorous quality control than the Ensenada data. Thus we would like to know if there is a statistically significant difference between the two climatologies. In the first case $N = 56$, $\bar{x}(N) = 260$ mm, $\sigma(x) = 125$ mm, $\varepsilon_s = 17$ mm, and we get $M_1 \sim a = e^A = 95$ mm (equivalent to $\alpha(x)$), $b = B = -0.4882$, and $M_N \sim 13$ mm (equivalent to ε_a). Thus in terms of the absolute error the Ensenada climatology should be written $\mu_{ens} = 260 \pm 13$ mm. In the second case $N = 154$, $\bar{x}(N) = 253$ mm, $\sigma(x) = 103$ mm, $\varepsilon_s = 8$ mm, and we get $M_1 \sim a = e^A = 83$ mm (equivalent to $\alpha(x)$), $b = B = -0.4958$, and $M_N \sim 7$ mm (equivalent to ε_a). Thus in terms of the absolute error the San Diego climatology should be written $\mu_{sdi} = 253 \pm 7$ mm. Therefore equation (8) applies, the measures of uncertainty are valid, and we cannot say that there is a statistically significant difference between the two climatologies (at $1 - 1/28 = 0.9643$ confidence).

6. Discussion and Conclusions

[11] The absolute error and the standard error are both practical measures of the uncertainty of the mean in most climatological cases, but the former is more conservative than the latter, which might be preferable in situations where a more strict criterion than the one obtained with the standard error is needed. Furthermore from the absolute error concept we can derive an equivalent expression equation (8) whose validity for each particular case can be verified by a numerical technique. The validation is done by comparing the obtained value of b with $b = -1/2$, and the application of this technique suggests that the method generally applies in symmetric distributed cases and some moderate asymmetric distributed cases, but it fails for extreme asymmetric cases. When equation (8) is valid the method provides additional information, such as the approximate size of the largest error and how this error decreases ($\sim j^b$, $b \sim -1/2$), when equation (8) fails it was found in all cases that $-b \ll 1/2$. This paper responds to the call for additional descriptors of symmetric and asymmetric distributed data besides the measure of central tendency [*Guttman*, 1989]; in other words methods for the analysis of entire data sets which perhaps should be used before and in addition to advanced time series techniques for climate research (see, e.g., *Godtliebsen et al.* [2003], for analysis in the time domain, and *Ghil et al.* [2002], for analysis in the frequency domain).

[12] **Acknowledgments.** I thank Pepe Ochoa, Guido Marinone and Julio Sheinbaum for numerous brief but insightful comments. The help of two anonymous reviewers to improve this work is sincerely acknowledged. Special thanks go to Prof. James J. O'Brien who twenty years ago introduced me to the study of climate and statistics. My research is funded by the Mexican CONACYT system.

References

Ghil, M., et al. (2002), Advanced spectral methods for climate time series, *Rev. Geophys.*, *40*(1), 1003, doi:10.1029/2000RG000092.

Godtliebsen, F., L. R. Olsen, and J.-G. Winther (2003), Recent developments in statistical time series analysis: Examples of use in climate research, *Geophys. Res. Lett.*, *30*(12), 1654, doi:10.1029/2003GL017229.

Guttman, N. B. (1989), Statistical descriptors of climate, *Bull. Am. Meteorol. Soc.*, *70*, 602–607.

Leith, C. E. (1973), The standard error of time-averaged estimates of climatic means, *J. Appl. Meteorol.*, *12*, 1066–1069.

Pavia, E. G., and A. Badan (1998), ENSO modulation of rainfall in the Mediterranean Californias, *Geophys. Res. Lett.*, *25*, 3855–3858.

Pavia, E. G., and F. Graef (2002), The recent rainfall climatology of the Mediterranean Californias, *J. Clim.*, *15*, 2697–2701.

Taylor, J. R. (1997), *An Introduction to Error Analysis*, 327 pp., Univ. Sci., Sausalito, Calif.

von Storch, H., and A. Navarra (Eds.) (1999), *Analysis of Climate Variability*, 342 pp., Springer-Verlag, New York.

von Storch, H., and F. Zwiers (1999), *Statistical Analysis in Climate Research*, 484 pp., Cambridge Univ. Press, New York.

E. G. Pavia, CICESE, P.O. Box 434844, San Diego, CA 92143, USA. (epavia@cicese.mx)