NOTES

Selection of a Map Grid for Data Analysis and Archival

WILLIAM B. ROSSOW

NASA Goddard Space Flight Center, Institute for Space Studies, New York, NY 10025

LEONID GARDER

Department of Geological Sciences, Columbia University, New York, NY 10025

22 March 1984 and 4 June 1984

ABSTRACT

Aggregation of atmospheric data using the common equal-angle (latitude–longitude) map grid is shown to introduce an unnecessary degradation of data quality compared with aggregation using an equal-area grid. Analysis of this problem shows that the analysis grids can be different from the archival grids for convenience.

1. Introduction

Study of atmospheric phenomena on Earth (and other planets) generally requires analysis of observations and consideration of theoretical calculations in terms of the spatial distribution of some quantity. Higher order moments of the spatial distribution and statistics of variations in time are also useful for describing phenomena. The advent of satellite remote sensing has created high volume data sets which need to be reduced for understanding to a few descriptors of the spatial distribution of physical quantities. An example of this reduction is averaging a quantity in time and space and calculating a standard deviation. Consequently, a common step in data analysis is collection of observations into arrays representing some map grid. These arrays are used to aggregate data (to reduce volume) and to archive the quantities obtained from the analysis of the observations.

Selection of a map grid for use in data analysis can be made on the basis of three criteria: data quality, volume and convenience. Quality refers to preservation of the original statistical properties of the data and a proper representation of the spatial distribution of quantities. Volume refers to concern with the difficulty of storing or accessing large volume data sets. Convenience refers to issues concerning the ease of retrieval and manipulation of the data. Much discussion may precede the selection of a map grid for a particular data analysis project, but heavy weight is usually given to convenience and simplicity. The problem is that spatial distributions are best described on a closed spherical surface rather than a rectangular, flat surface. The latter is obviously considered more convenient than the former, since the most commonly used map grid for both aggregation and archival is a regular array with equal increments in latitude and longitude.

This brief note reiterates the scientific arguments for selection of a map grid by illustrating the quantitative effects on data quality caused by using different grids. Although the effects shown here are well known, the wide use of the rectangular latitude–longitude grid indicates that its (apparent) convenience is given disproportionate weight in the selection process. We show that the use of this particular map grid actually degrades data quality, increases stored data volume and increases the complexity of data manipulation. Use of an equal-area grid is shown to be a proper way to aggregate data, but such equal-area grids are thought to be inconvenient. We argue that equal-area grids can be convenient but propose a resolution to this dilemma by showing that the *analysis* and *archival* map grids need not be the same. The results of a proper analysis of the data on an equal-area grid can be remapped to the more "convenient" rectangular latitude–longitude grid without loss of quality.

2. Test of different map grids

There are two separate consequences of using different map grids for data analysis: the effects of variable resolution and the effects of changing statistical significance. The former will not be considered here since the principal criterion is straightforward: the map projection should not *degrade* spatial resolution differentially, i.e., the spatial resolution should not be a function of location. Most map grids do not have constant spatial resolution over a spherical surface; however, this criterion can also be met by making the lowest resolution portion of the map equivalent to the intrinsic resolution of the observa-

tions. Thus, the higher resolution portions of the map can be filled by data replication, thereby avoiding the introduction of spurious spatial structure. The drawback to this approach is that the mapped data volume is now greater than the original data volume. Furthermore, degradation of data resolution at a later time must be done using proper area weighting, otherwise the quality of information on spatial structure will not be uniform over the map. Any map grid, which has grid boxes representing (nearly) equal spatial dimensions at all locations, will avoid these problems.

Here we consider the statistical effect of different analysis grids. The specific issue is the consequence of varying statistical weights produced by map grids which have grid boxes representing varying surface area. If the observations are distributed over the earth with nearly constant density, but the grid box areas vary, this may affect the quantities calculated using that particular map grid and introduce spurious spatial variations. We explore this effect by using a synthetic data set obtained by "observing" values from a specified probability distribution with a known mean and standard deviation. Two distributions were tested, a symmetric Gaussian and an asymmetric distribution; but only the results from the asymmetric distribution are presented (Fig. 1). This synthetic data set is equivalent to measuring some spatially varying quantity which has no fixed structure but varies from point-to-point with the probability shown in Fig. 1. Alternately, the data set could represent measurement of some constant quantity with high uncertainty (instrumental noise or analysis uncertainties). The results shown simulate one observation of the asymmetric distribution from the NOAA polar orbiting meteorological satellite, representing, e.g., a single measurement of the spatial distribution of IR radiance in a single orbit with a spatial resolution varying between 24 and 48 km.

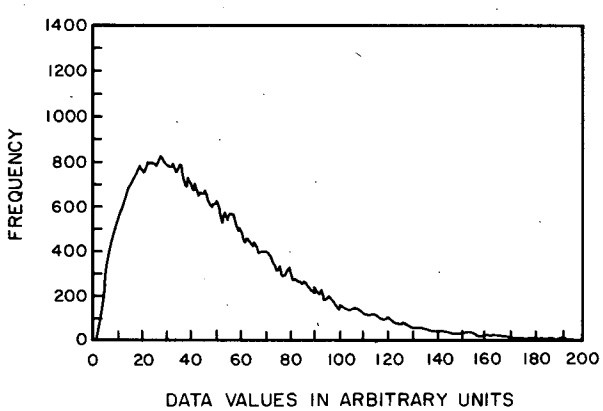

FIG. 1. Asymmetric probability distribution used to produce synthetic observations. Note arbitrary units of magnitude on the horizontal axis.

The observations are aggregated in two map grids, the commonly used rectangular latitude–longitude grid with 2.5° resolution and an equal-area grid with 2.5° latitude increments ($\Delta\theta = 2.5°$) and a longitude increment proportional to $\cos^{-1}\theta$. (The last latitude increment at each pole is divided into three triangular boxes.) The mean value and the standard deviation are calculated for each grid box. The mean and standard deviation represent two common quantities calculated to reduce data volume by reducing spatial resolution (spatial average) while maintaining some information about the higher resolution spatial distribution (standard deviation). These two quantities also represent examples of a linear and nonlinear product, respectively, derived from higher volume data.

Figures 2 and 3 show the frequency distribution of mean and standard deviation values obtained for grid boxes in several latitude zones. Since the number of observations in each box can vary with longitude and latitude (as illustrated in Figs. 2c and 3c), the variation of values about the correct (statistical) answers, indicated by solid vertical lines, is simply a consequence of sampling. The point is that this sampling problem is always present in actual observations so that the width of the histograms in Figs. 2 and 3 can be interpreted to represent uncertainties in determining some quantity from observations. Increasing the number of observations of the fixed distribution reduces the magnitude of this "dispersion," but does not alter the relative variation with latitude. Since most observables vary with time, examination of a single observation is a fair test of the effect.

Figures 2 and 3 make clear the problems with the rectangular latitude–longitude map grid: uncertainty in both the mean value and the standard deviation increases with latitude (Figs. 2a and 2b) as a direct consequence of the decreasing number of observations in each grid box (Fig. 2c). More importantly, the standard deviation is biased with latitude (Fig. 2b); even the zonal mean of this nonlinear quantity is biased on this grid. Both of these spurious latitudinal dependencies are introduced by the variation of statistical weight, given by the number of observations, over the map. Hence the data quality, which was originally nearly uniform with location, is no longer uniform over the globe. Figure 3 shows that analysis of the data on the equal-area grid removes these deficiencies.

3. Discussion

Selection of an *analysis* grid should be made to maintain uniform spatial resolution and statistical significance. Use of the "more convenient" rectangular latitude–longitude grid degrades data quality by introducing greater uncertainty for some locations (usually near the poles) and nonuniformity in the properties of the analysis products over the map. This latter

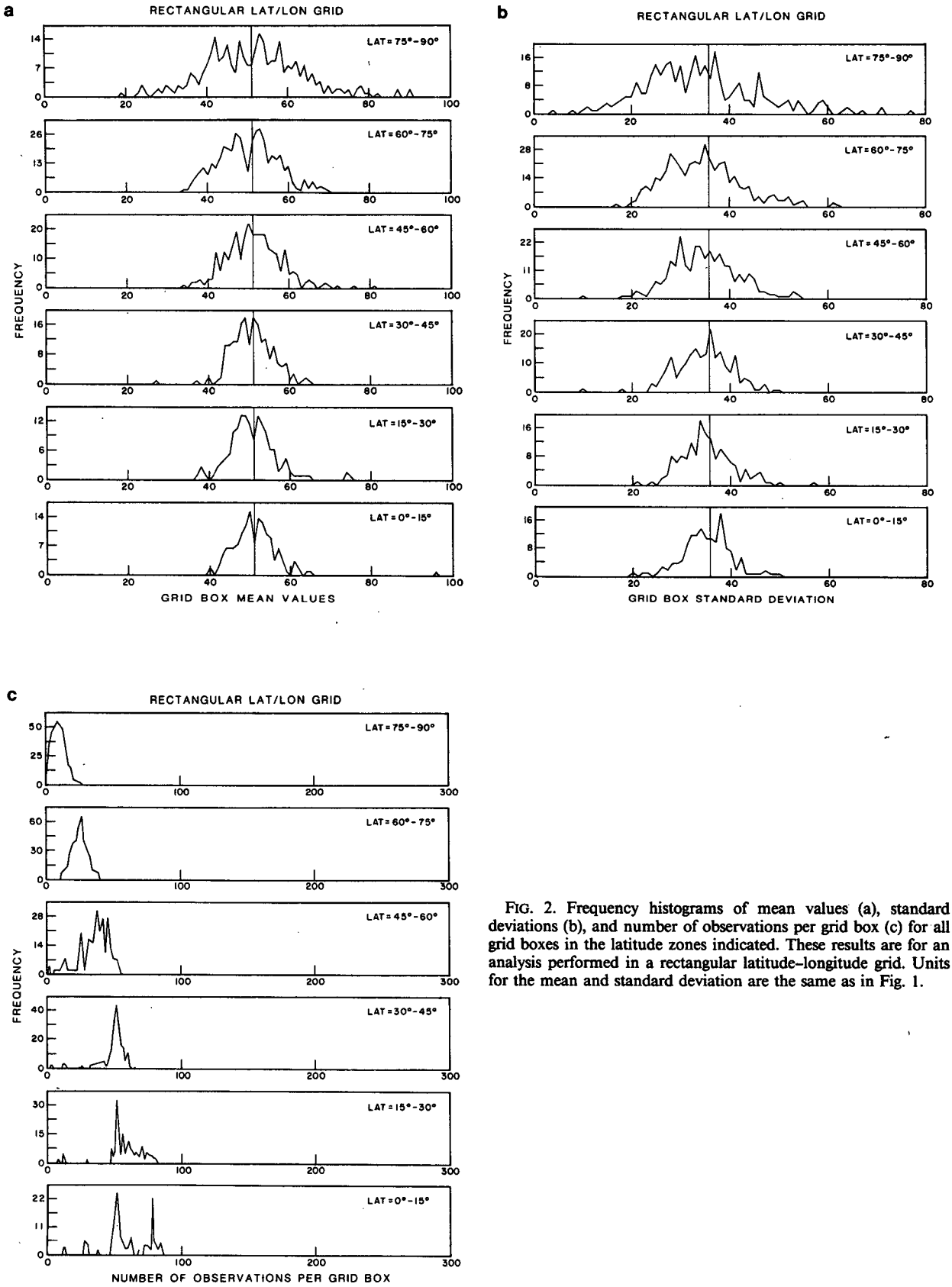

FIG. 2. Frequency histograms of mean values (a), standard deviations (b), and number of observations per grid box (c) for all grid boxes in the latitude zones indicated. These results are for an analysis performed in a rectangular latitude–longitude grid. Units for the mean and standard deviation are the same as in Fig. 1.

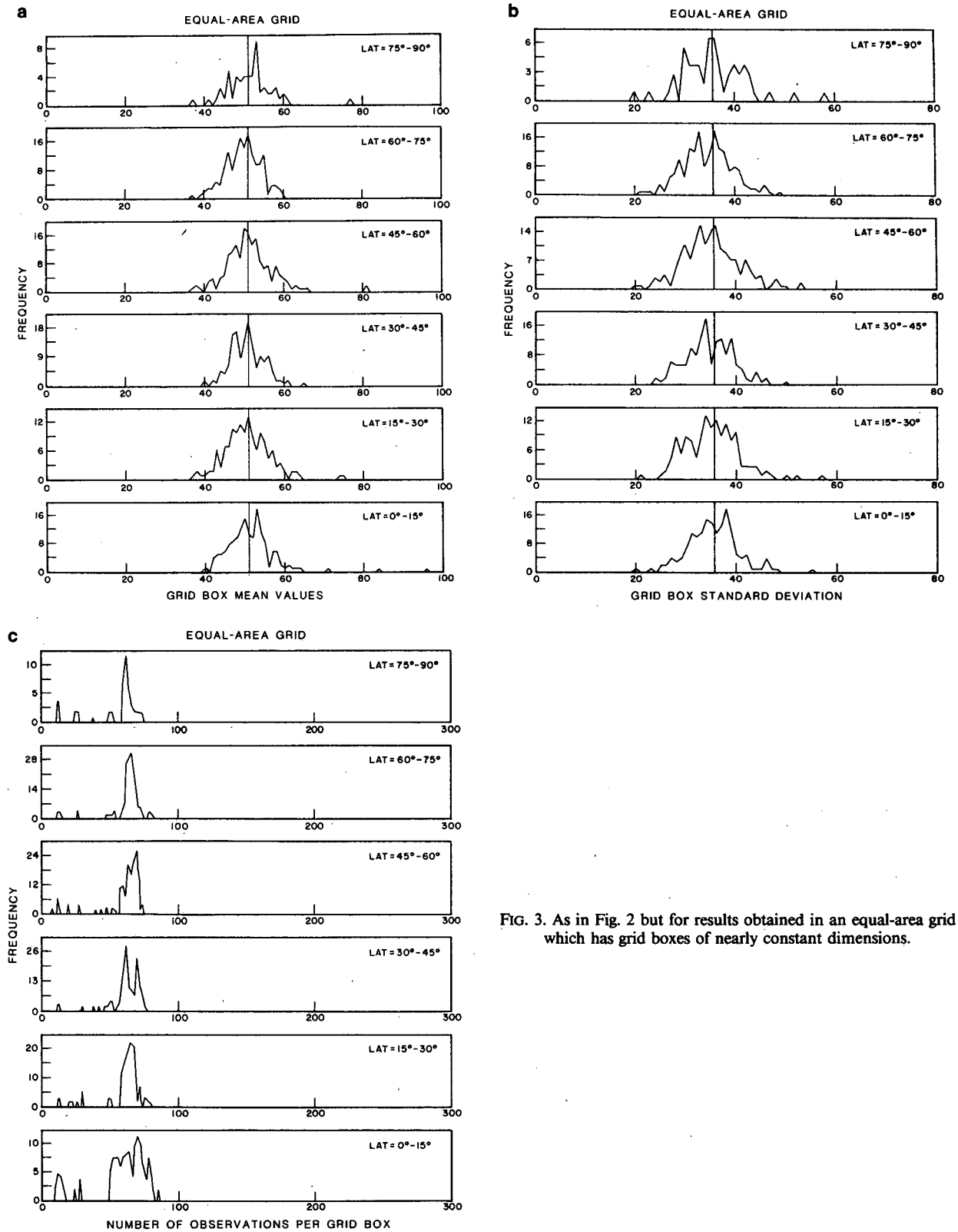

FIG. 3. As in Fig. 2 but for results obtained in an equal-area grid which has grid boxes of nearly constant dimensions.

problem is serious when higher order statistics (e.g., the variance) of the spatial distribution are calculated; furthermore, variation of the grid box area with location alters the meaning of these statistics with location. These statistical effects, together with the requirement for uniform spatial resolution, provide strong arguments for use of an analysis grid, such as the equal-area grid used here, that has grid boxes or elements with nearly constant area and linear dimensions.

The "convenience" of the rectangular latitude–longitude grid over other alternatives is actually more apparent than real. Since computers and data storage devices actually store numerical "arrays" as linear lists with addresses, the use of equal-area grids introduces only a trivial modification of the address indices to make one a function of the other. In other words, an array $A(I, J)$ becomes $A[I, J(I)]$. Manipulation of data stored in an equal-area grid is consequently no more difficult than for any other map grid; indeed, remapping or spatial averaging of data are easier because each grid box has the same statistical and area weight, unlike other grids. The volume of the equal-area grid is also about 25% smaller than the equal-angle grid for the same resolution at the equator.

The results of the analysis on an equal-area grid can easily be remapped to a rectangular latitude–longitude grid using simple linear area weights. (Reprojection of the data from the rectangular latitude–longitude grid to another grid is a more complex operation because the area weights are not constant.) Comparison of the reprojected data analysis with the original equal-area analysis shows no difference. That is, analysis results properly calculated on an equal-area grid can be projected onto another grid without loss of quality, as long as spatial resolution is not decreased. The criterion of "convenience" could be met by *archiving* the data in a rectangular latitude–longitude grid, but *analysis* on an equal-area grid is necessary for maintaining data quality.

The effects of different analysis grids on data properties are not surprising; they result from the uniformity, or lack of uniformity, of data density over the map. What is surprising is that most data sets are still analyzed (and archived) in map grids which alter the character of data in undesirable ways. Use of an equal-area grid would seem to be more appropriate for aggregation of atmospheric data. However, if the rectangular grid is to be retained for the archived data product, then use of a separate analysis grid and archival grid would overcome these problems.

Acknowledgments. The ideas in this paper have benefited from discussions with B. Barkstrom, E. Harrison and R. Schiffer. Computations were carried out by L. Kanganis; figures were drafted by L. DelValle.