Conf-9404237--1

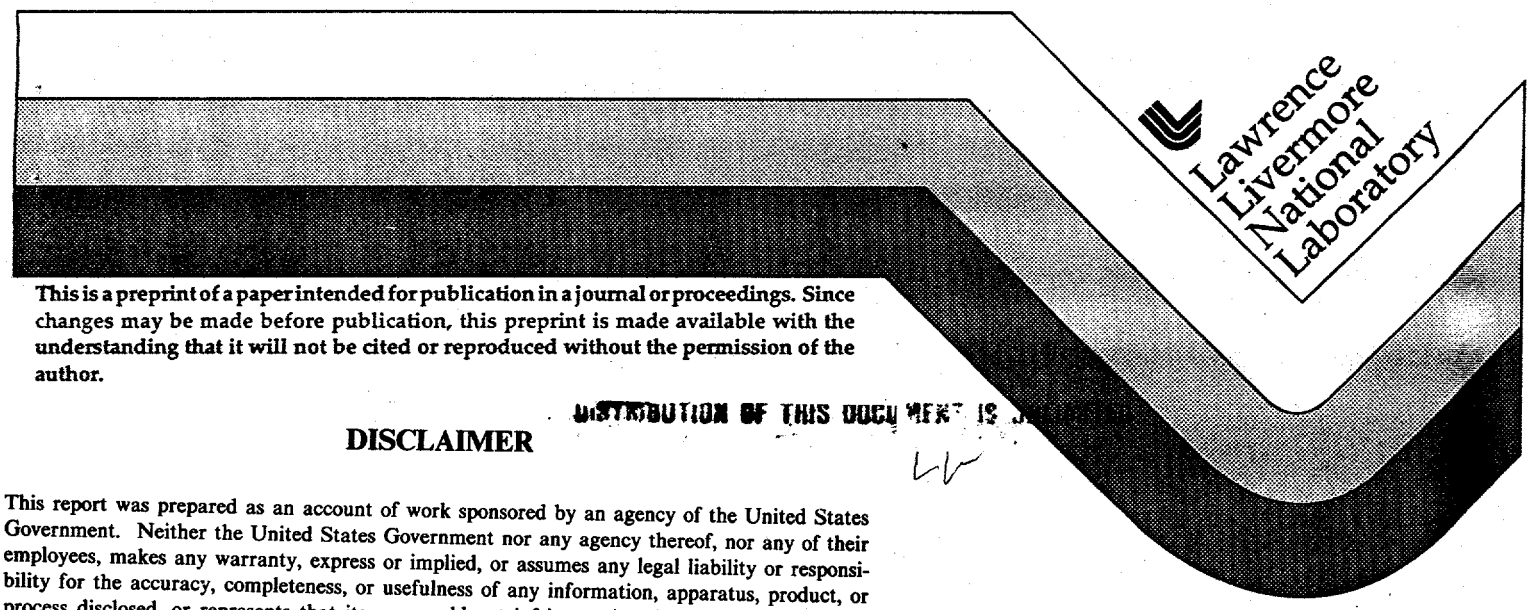
UCRL-JC-116492
PREPRINT

Scientific Statistics and Graphics on the Macintosh

Stanley L. Grotch
Lawrence Livermore National Laboratory
Livermore, CA 94550

Submitted for Publication in
Proceedings of the AGU/Paclim Workshop
Asilomar, CA
April 19-22, 1994

September 1994



DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Scientific Statistics and Graphics on the Macintosh

Stanley L. Grotch
Laboratory Associate
Lawrence Livermore National Laboratory
Livermore, California

Introduction

In many organizations scientists have ready access to more than one computer, often both a workstation (e.g., SUN, HP, SGI) as well as a Macintosh or other PC. The scientist commonly uses the work station for "number-crunching" and data analysis whereas the Macintosh is relegated to either word processing or serves as a "dumb terminal" to a larger main-frame computer. In an informal poll of my colleagues, very few of them used their Macintoshes for either statistical analysis or for graphical data display.

I believe that this state of affairs is particularly unfortunate because over the last few years both the computational capability, and even more so, the software availability for the Macintosh have become quite formidable. In some instances, very powerful tools are now available on the Macintosh that may not exist (or be far too costly) on the so-called "high end" workstations. Many scientists are simply unaware of the wealth of extremely useful, "off-the-shelf" software that already exists on the Macintosh for scientific graphical and statistical analysis.

This paper is a very personal view illustrating several such software packages that have proved valuable in the author's own work in the analysis and display of climatic datasets. It is not meant to be either an all-inclusive enumeration, nor is it to be taken as an endorsement of these products as the "best" of their class. Rather, it has been found, through extensive use that these few packages were generally capable of satisfying my particular needs for both statistical analysis and graphical data display. In the limited space available, the focus will be on some of the more novel features found to be of value.

The following discussion is divided into three sections, the first two illustrating Macintosh software for statistical data analysis and for graphical data display. The final section will summarize the work and offer some comments regarding the future.

Statistical Analysis Software for the Macintosh

A number of general purpose statistical software packages are now available for the Macintosh. For a detailed review intercomparing their capabilities see Best and Morganstein 1991. Review articles frequently appearing in popular journals such as MacWorld are of great benefit in keeping abreast of the more recent developments. In my

own work, two statistical packages have proved to be of particular value: (1) Data Desk and (2) StatView. (To obtain further information regarding the software discussed here see Appendix A.)

Invariably, no single package offers *all* of the features that one would desire. Generally, each has its peculiar strengths and weaknesses. This is both good and bad for the scientist. Good in that feature duplication is minimized, but bad in that multiple packages must be purchased and subsequently mastered. It is this latter point, the continual intellectual demand placed on the scientist that, I believe, has largely contributed to the lack of small computer usage noted above. Unfortunately, the scientist feels so overwhelmed with day-to-day responsibilities that "makes us rather bear those ills we have, than fly to others that we know not of". This is particularly so with the more sophisticated software packages which require frequent usage to maintain the necessary skills for effective operation.

Both Data Desk and StatView provide the user with a veritable arsenal of the most important statistical analysis tools: standard summary statistics (means, variances, non-parametrics, etc.), inference testing (equivalence of means), correlation, regression, analysis of variance. Data Desk, particularly, provides excellent instruction manuals and both have competent telephone technical support. With networking becoming commonplace, both programs will readily accept data matrices generated on other computers in a range of formats. Tab-delimited ASCII matrices are easily read without user intervention.

The two packages differ, however, in their basic philosophy regarding graphical data display. Data Desk is far more interactive, but produces cruder graphics. On the other hand, although StatView is typically slower, it is capable of generating truly presentation-quality graphics. To achieve speed and high interactivity, Data Desk has few (or no) controls for user-determined plot limits, gridlines, fonts, annotation, etc. , features nicely implemented in StatView. On the other hand, Data Desk can generate and rapidly rotate three dimensional point clouds, a feature still not implemented in StatView.

One of the most useful and innovative features implemented in Data Desk (but not in StatView) is the concept known as "linked plots". Here, any points highlighted in one display are also correspondingly highlighted in all others. As an illustration of the power of this technique consider a very common problem encountered in data analysis. We wish to compare two datasets and spatially locate those points which show the greatest similarities and the greatest differences.

Assume we have available on a common grid observations of precipitation from two different sources. A histogram of the gridpoint differences in precipitation is displayed in the top panel of Fig. 1 If a binary indicator of land(=0) or water (=1) is available at every gridpoint, a second plot showing the continental land masses can also

readily be produced (lower panel of Fig. 1). To generate the lower display, all grid points are first plotted as a scatter plot, yielding a simple rectangular grid. The land grid points are then selected using the land/water index and instantly only the land grid points are highlighted. It should be emphasized that in the static displays presented here several *extremely important* distinguishing features present with a TV monitor are absent: (1) color, (2) intensity, and (3) temporal on/off flashing. The color and shape chosen to differentiate a given group of points is maintained in all displayed plots.

To select a region of the histogram (here the lower tail) the user merely touches the vertical bars of the histogram with a program tool, and the selected bars instantly darken on the histogram display. Simultaneously, those grid points captured in the highlighted ranges will also glow on the lower map. These map-selected points can either be preserved using color and/or shape or the user can choose another set of ranges in the histogram and the initial selection will automatically disappear. To spatially locate and differentiate the three regions: (1) the lower tail, (2) the upper tail, and (3) the central region (best agreement), each grouping is chosen, in turn on the histogram. As the relevant points are automatically selected on the lower map, they can be differentiated using different colors and/or shapes. The remarkable interactivity of this process has to be experienced to be fully appreciated.

Linked plots also function in the opposite direction. If, for example, one wanted to determine what the histogram of differences was for only the tropical region, the same tool would be moved along the latitudinal axis of the lower map, capturing the desired range of latitudes (see lower panel of Fig. 2). At the same moment, a sub-histogram would darken (upper panel of Fig. 2), showing the histogram relevant to only the selected points. Similarly, if the histogram for a specific area such as the continent of Africa was required, another tool (appropriately, a lasso) would be used to encircle the appropriate area on the map, and again the captured points would yield a darkened sub-histogram.

Additionally, any selected points are automatically highlighted on *all other* displays shown on the screen using the same colors and symbols. Thus, in the example here, if data for say, temperature vs. cloudiness at these gridpoints were available, when the African points were selected with the lasso tool, these points would also glow on the temperature /cloudiness plot. The truly extraordinary potential of this technique for interactively analyzing multivariate datasets has simply not been exploited by most scientists, largely due to their ignorance of the existence of such tools on the Macintosh.

Graphical Software

The author has found three software packages of particular value for scientific data display: (1) Kaleidagraph, (2) Delta Graph, and (3) Spyglass Transform and Dicer. Once again, each package has its own virtues and disadvantages which will be briefly discussed here.

Many excellent software packages exist for producing the "bread and butter" plots of the scientist: two dimensional scatter and line plots. Thus, any "recommended" choice among these is particularly subjective. In my own experience, Kaleidagraph has proved to be particularly easy to use and versatile in permitting the user considerable control in the embellishment of two dimensional graphics. In Kaleidagraph the user has considerable latitude over setting axis limits, axis direction, the appearance of grid lines, fonts, symbols, colors, arrows, lines, background color. These can be quickly and interactively added, changed, and moved, as desired. While such capabilities might seem superfluous, I strongly believe that in scientific graphics these capabilities are *absolutely essential* to produce both aesthetically pleasing and scientifically informative graphics.

As might be expected, the number of software packages which can produce truly effective 3D scientific plots is much less than for 2D. Unfortunately, Kaleidagraph has no 3D plotting capability. The packages Data Desk, Delta Graph and Spyglass all have three dimensional capabilities, but each has important advantages and limitations.

Data Desk can display point clouds in space, and it is highly interactive in rapidly rotating these points to produce a very realistic 3D effect. (The linked plot feature described above also functions with these 3D plots). However, although the three dimensional effect using parallax motion is visually excellent on the screen, when produced as a hard copy, the resultant is often quite disappointing. Delta Graph and Spyglass, on the other hand, are both too slow to produce motion interactively, but both do generate presentation-quality 3D graphics.

The Spyglass suite of software (currently, Transform, View, Plot and Dicer) is perhaps the most innovative and most impressive in its capability for scientific data display. Although the software has virtually no numerical statistical capability, the packages are remarkable in their ability to produce both two and three dimensional false color images and animations. Several illustrative examples will be presented using Spyglass Transform and Dicer. Unfortunately, in the black and white figures presented here, much of the visual impact and differentiation produced by using color is lost.

Transform is the primary false color plotting package of the Spyglass suite of programs. Any data matrix, say a meteorological field, expressed as a function of latitude and longitude, is quickly rendered in either two or three dimensions using color to code the magnitude of the variable presented. A broad range of built-in color mappings is available and these can be quickly changed in a very interactive manner. Continental outlines, vector and contour overlays can be also added to any two dimensional display as in Fig. 3 which shows the temperature distribution predicted by a global GCM. Unfortunately, the subtle gradations in the color original are lost in this black and white reproduction.

With Spyglass Transform these same data can also be quickly rendered as a wiremesh surface in three dimensions as is shown in Fig. 4. Transform permits the user to interactively change viewpoint and the aspect ratio used as well as the color mapping selected. To better permit spatial orientation in these displays the continental areas can also be directly shaded on the data surface. Once again, the importance of color in such displays cannot be ignored. Once a satisfactory plot is generated in either 2D or 3D, the user may save all of the instructions as a macro. This greatly facilitates reproducing similarly scaled plots for intercomparison of different datasets or for generating animations.

For many years the Macintosh has been the computer of choice for work in the graphic arts. The wealth of capabilities which has developed in this area is truly impressive. These have been largely ignored by most scientists. The ability to "cut and paste" graphics from disparate sources can prove to be of considerable value in scientific graphics. No longer does the scientist have to generate all of the components of each graphic in a single program.

In Fig. 5, for example, a 3D wiremesh rendition of the temperature data of Figs. 3-4 from Spyglass Transform is merged seamlessly with a world map using Adobe Photoshop. Alternatively, the 2D map of Fig. 3 could be easily inserted in this lower plane. With the many tools provided in existing graphics programs, the user can quickly resize and overlay different plots, add textual and arrow annotations. To further enhance the final appearance of the printed output, colors can be interactively altered and the brightness and contrast quickly changed. For presentations on the printed page or as view graphs, these capabilities can change a rather dull plot into one that is both far more visually appealing and ultimately much more informative.

Dicer is probably the most novel of the suite of Spyglass programs. As the name implies, Dicer permits the user to consider three dimensional data as a solid piece of food through which slices or blocks can be cut out and the results displayed using a variety of user-specified color tables. In Fig. 6 are displayed the monthly average climatologies for 850 mb temperature. The horizontal cut planes show the January and July temperature distributions. Vertical cuts in the remaining planes show the corresponding temporal-spatial distributions. After selection, the user can quickly move or delete any of these slices.

Blocks can also be interactively cut out of the dataset or isosurfaces produced highlighting other aspects of the data. In Dicer the user can also generate a series of parallel slices to produce very effective animations. For example, in the example of Fig. 6, the user can very simply produce twelve horizontal slices to obtain a 2D animation of monthly climatology. These frames can be readily converted to Apple's Quick Time system, permitting easy viewing on virtually any Macintosh.

The range of features and the degree of interactivity incorporated into this software is truly remarkable. These capabilities are particularly valuable in the geophysical sciences where highly dimensional data are common.

Conclusions

There can be little doubt that scientific data analysis is currently undergoing explosive growth. Hardware advances on all fronts permit the examination of datasets undreamed of even a few years ago. The intense competition among software vendors has proven an extraordinary blessing to the scientist. The constant improvements in capabilities and ease of operation that have occurred in only the last few years are truly remarkable.

These software enhancements are very likely to continue at a rapid pace. One area of practical importance to the scientist is the development of scripts or macros to facilitate reproducing the same analyses or displays using different datasets. This will be of particular importance in intercomparing and contrasting different results and in producing animations of data where hundreds of individual frames must often be produced.

The future infusion of techniques coming from the field of graphic arts and the world of multimedia should exert considerable influence in scientific graphics. Today, most scientists simply do not think of combining graphics generated by different programs. With software such as Adobe Photoshop, it becomes a simple matter to interactively overlay bits and pieces of plots obtained from entirely different sources to produce considerably enhanced products. The application of such procedures is now in its infancy.

Animation (or, more generally, "multimedia") is currently another major growth area in graphics. Until very recently, scientists had to have the assistance of graphics specialists to produce effective movies. This was typically very costly, and severely limited the number of animations attempted and also greatly restricted experimentation. This situation is now rapidly changing. Desktop movie making made directly by the scientist is now quite feasible with packages such as Spyglass. Movie editing software for merging, titling, adding sound, etc., is becoming much easier to master and much more available on the Macintosh. Finally, standard movie playing software such as Apple's QuickTime will greatly facilitate the process of sharing animations among colleagues.

Unfortunately, my experience has shown that, in practice, many scientists still are unwilling to make the time and/or intellectual commitment needed to master these many new techniques on the small computer. There is little question that many do

require a significant investment of time and near-constant usage to be of value. Some, such as animations, may be unnecessary diversions in many applications. Hopefully, the next generation of scientists, for whom these technologies may be more familiar and ultimately less threatening, will be much more receptive to their use. My own strong belief is that these new technologies can provide the scientist with enormously powerful tools in the area broadly characterized as "scientific data analysis".

Acknowledgment

This work was performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

References

Best, A.M., Morganstein, D., 1991. Statistics Programs Designed for the Macintosh: Data Desk, Exstatix, Fastat, JMP, Statview II, and Super Anova. *American Statistician.* 45(4), 318-333

Appendix A-Software Sources

Data Desk, Data Description, Inc., P.O. Box 4555, Ithica, N.Y. : Tel: 607 257 1000.

DeltaGraph: DeltaPoint, Inc., 2 Harris Court, Monterey, Ca. 93940,: Tel. 408-648-4000.

Kaleidagraph: Synergy Software, 2457 Perkiomen Ave., Reading, Pa. 19606: Tel. 215-779-0522.

Spyglass: Spyglass, Inc., P.O. Box 6388, Champaign Ill. 61826 : Tel. 217-355-6000.

StatView, Abacus Concepts, Inc. , 1984 Bonita Avenue, Berkeley, Ca. : Tel. 510-540-1949.

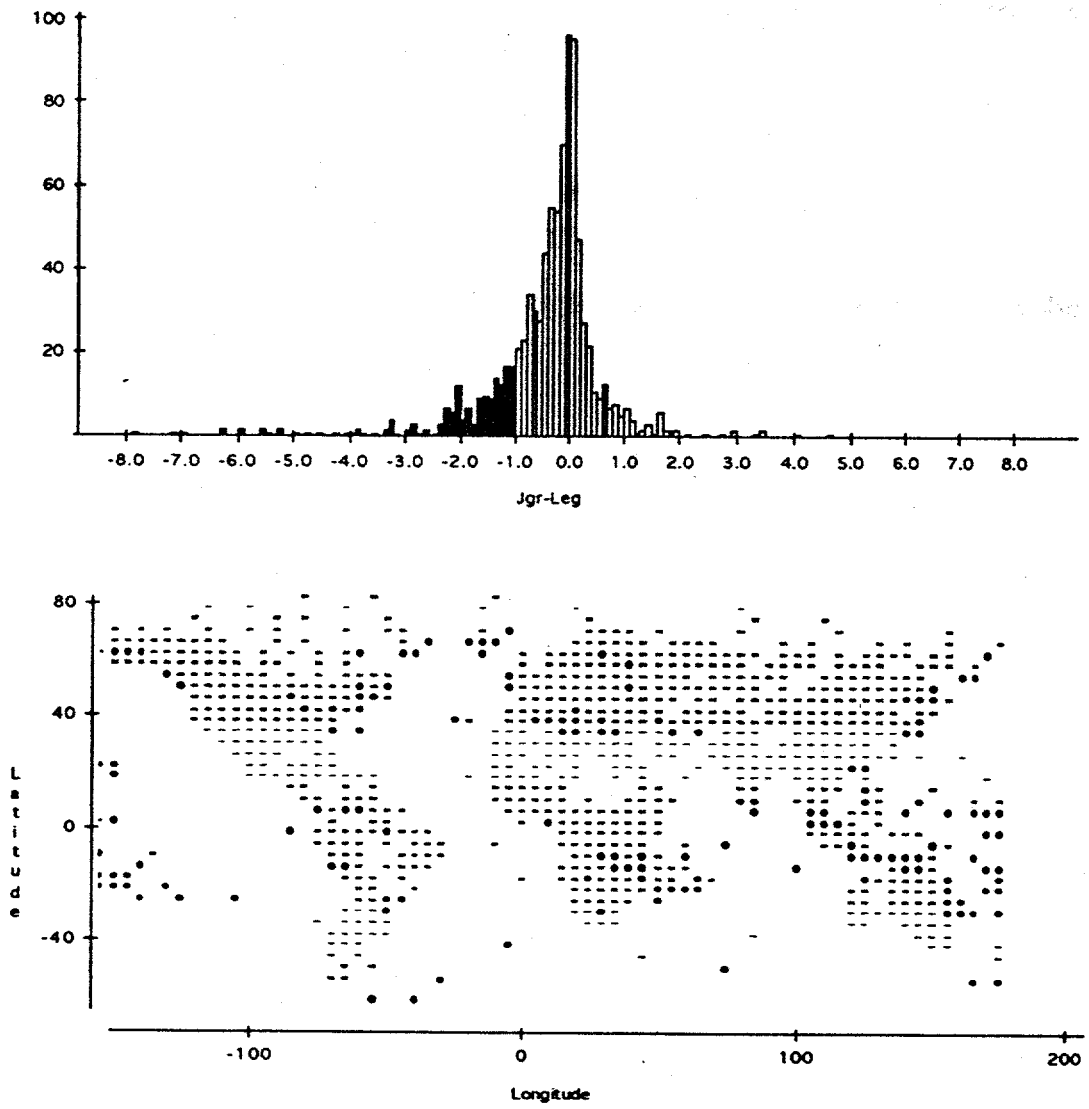


Fig. 1

Displayed in the upper panel is a histogram of the pointwise differences in observed precipitation between two datasets. When the lower tail of the histogram is selected in Data Desk, it darkens. Simultaneously, the points corresponding to these maximal differences glow on a world map indicating where these differences arise.

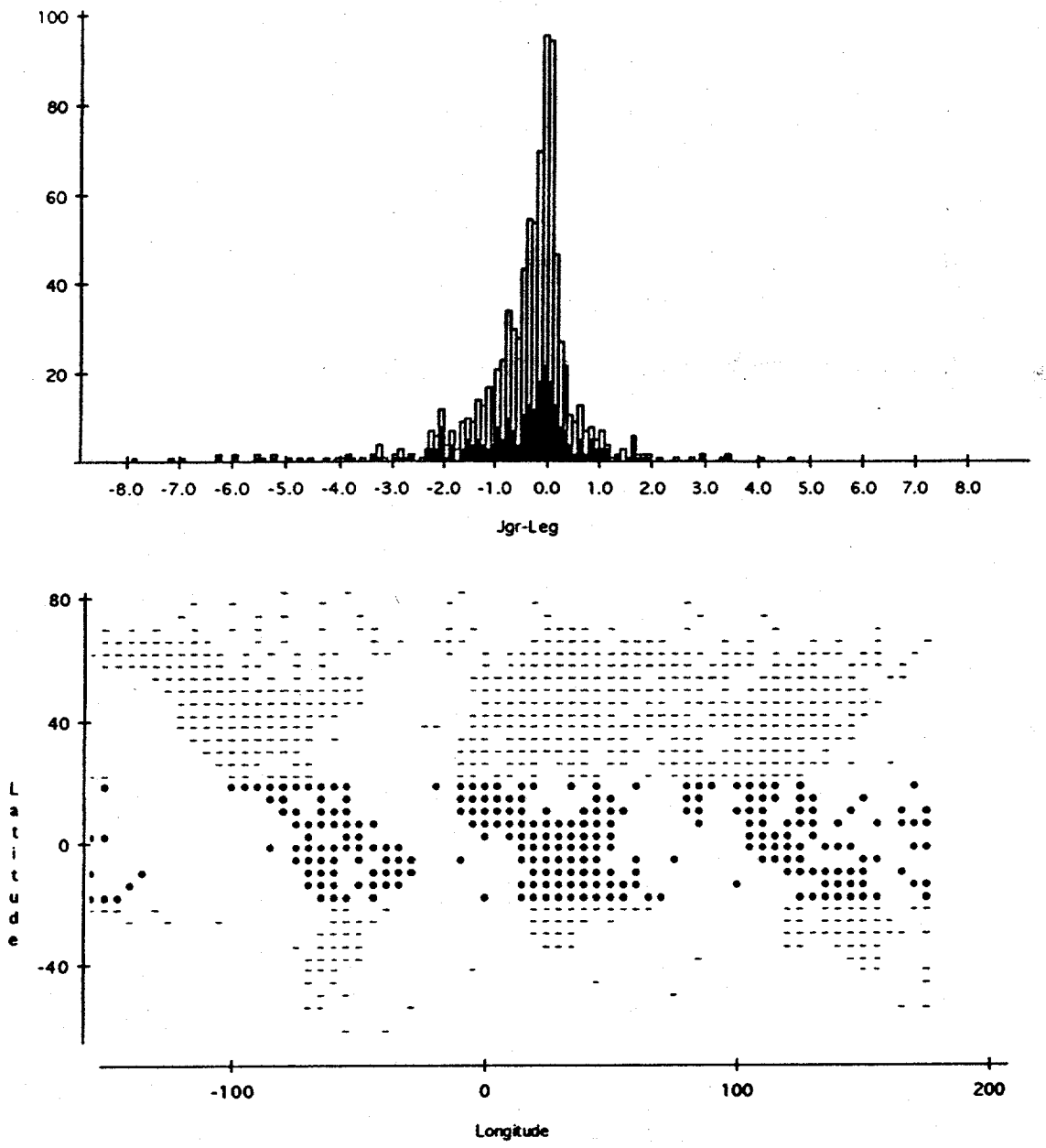


Fig. 2

The histogram of precipitation differences for only the tropics is desired. The user moves a tool along the vertical axis of the lower map until the desired latitudinal range is captured. As the tool is moved the chosen points glow in the map display.

Simultaneously, in the upper histogram a darkened sub-histogram is outlined showing the distribution for only the tropical gridpoints.

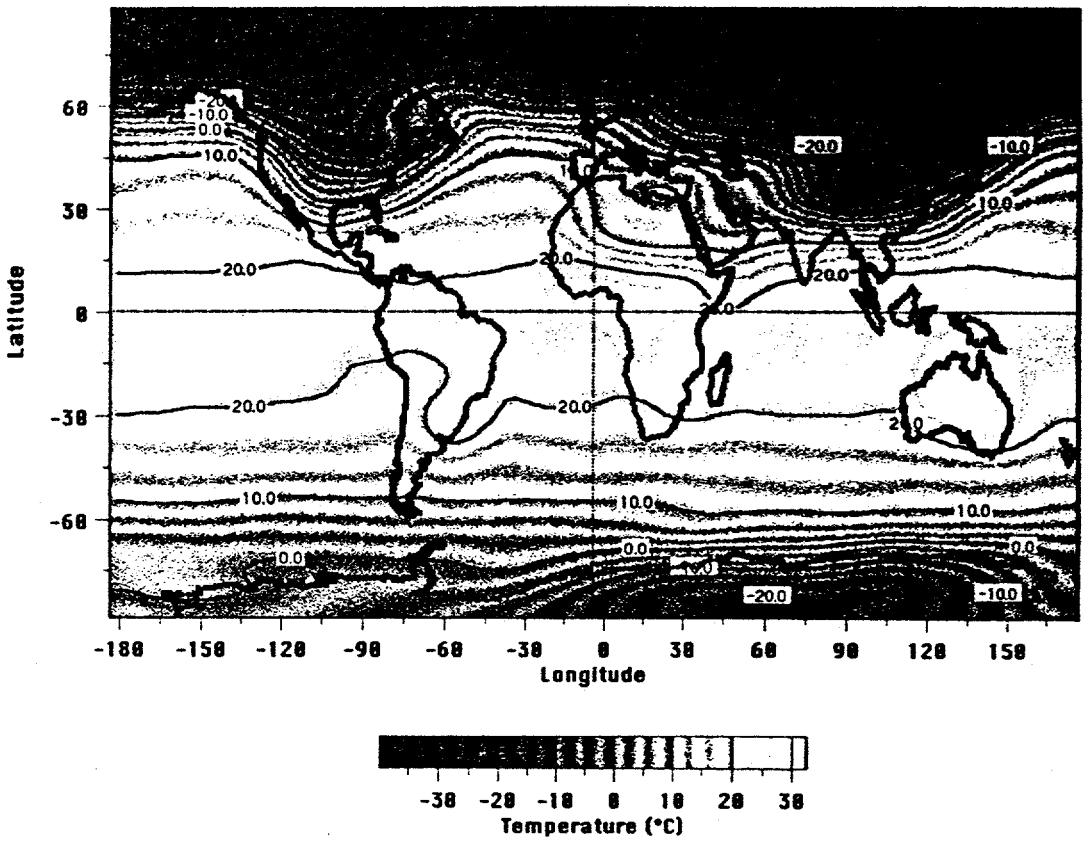


Fig. 3

The global temperature distribution predicted by a GCM is represented as a false color map using Spyglass Transform. To produce this display the underlying colored map is overlaid both with a contour map also generated with Transform and the continental outlines obtained from another source.

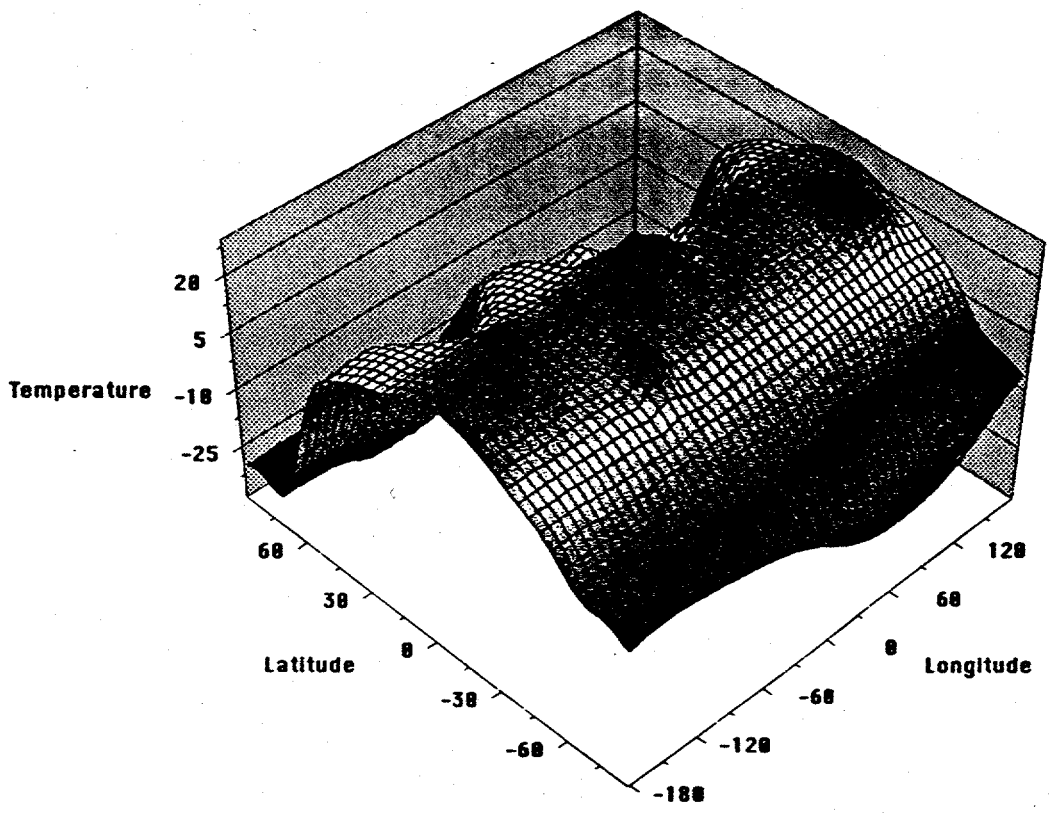


Fig. 4

The global temperature distribution of Fig. 3 is represented as a 3D wiremesh surface using Spyglass Transform. Many features of the plot including: user viewpoint, aspect ratio, axis scaling and labeling, and the colors used can be interactively changed by the user.

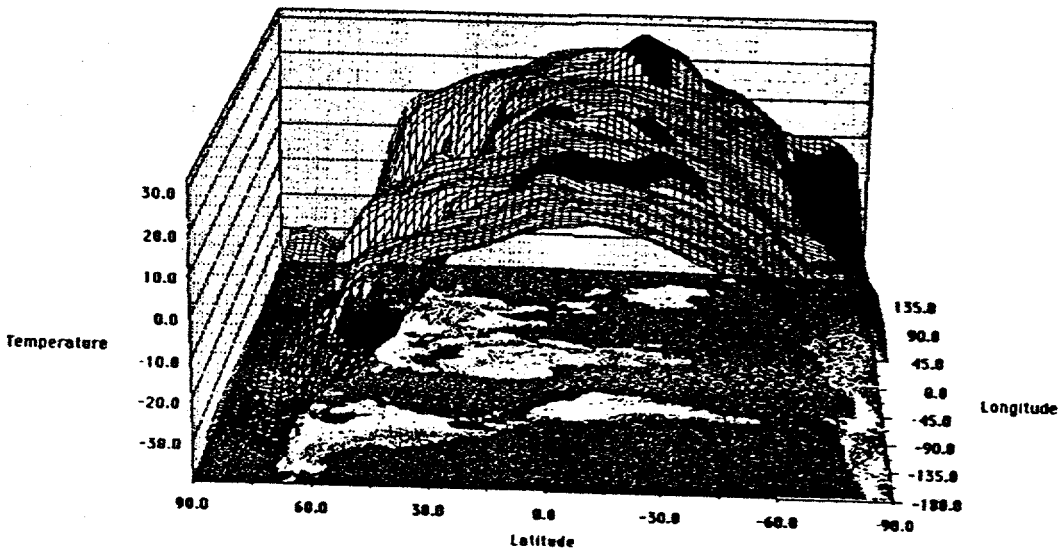


Fig. 5

The power of merging graphics derived from different sources is shown here. The 3D wiremesh surface of global temperature is generated using Spyglass Transform. The lower world map, showing topography was obtained from the Macldas group at Wisconsin via Internet. The two graphics were seamlessly merged using Adobe Photoshop. To obtain better spatial location, the continents may be also shaded in the upper surface using other capabilities in Transform.

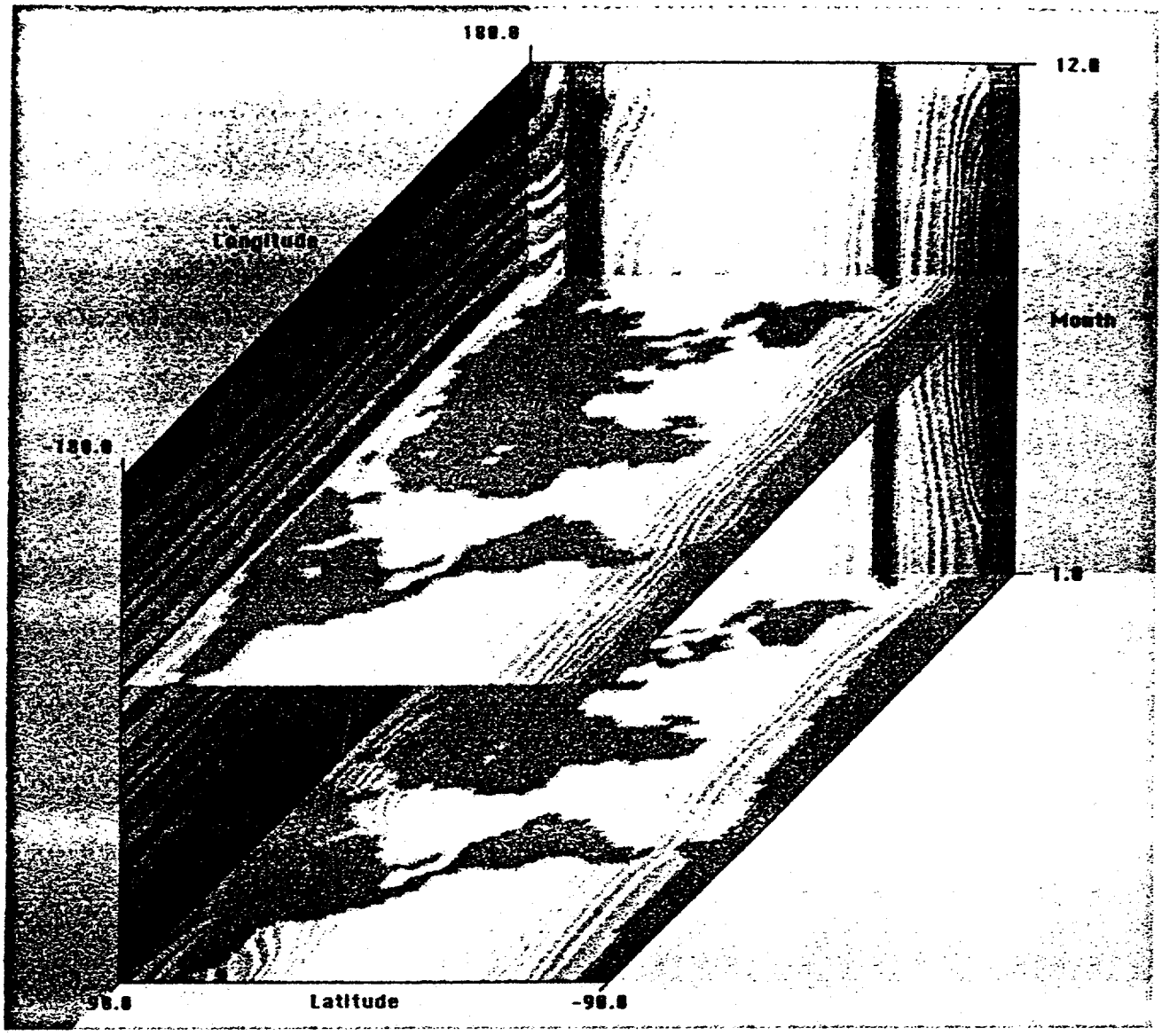


Fig. 6
The global distribution of 850 mb monthly temperature is shown using Spyglass Dicer. The two horizontal cuts in this plot show the spatial distribution of the January and July average climatologies. Dicer permits cuts to be made interactively along any of the axes as well as obliquely through the data. Thus one can examine two dimensional views through the three dimensional data set. Blocks and isosurfaces through the data can also be represented. The effects are particularly useful when displayed in color.