Scientific Statistics and Graphics on the Macintosh

Stanley L. Grotch

ABSTRACT: The personal computer has become commonplace on the desk of most scientists. As hardware costs have plummeted, software capabilities have expanded enormously, permitting the scientist to examine extremely large datasets in novel ways. Advances in networking now permit rapid transfer of large datasets, which can often be used unchanged from one machine to the next. In spite of these significant advances, many scientists still use their personal computers only for word processing or e-mail, or as "dumb terminals". Many are simply unaware of the richness of software now available to statistically analyze and display scientific data in highly innovative ways. This paper presents several examples drawn from actual climate data analysis that illustrate some novel and practical features of several widely-used software packages for Macintosh computers.

Introduction

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

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, powerful tools are now available for the Macintosh that may not exist or may be far too costly for the so-called "high end" work stations. Many scientists are unaware of the wealth of extremely useful, off-the-shelf Macintosh software that already exists for scientific graphical and statistical analysis.

This paper is a personal view, illustrating several software packages that have proved valuable in my own work in analysis and display of climatic datasets. It is not meant to be an all-inclusive enumeration, nor is it to be taken as an endorsement of these products as the best of their class. Rather, extensive use has proven these few packages to be generally capable of satisfying my particular needs for statistical analysis and graphical data display. I focus on some of the more novel features found to be of value.

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

In: C.M. Isaacs and V.L. Tharp, Editors. 1995. *Proceedings of the Eleventh Annual Pacific Climate (PACLIM) Workshop, April 19-22, 1994.* Interagency Ecological Program, Technical Report 40. California Department of Water Resources.

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). Reviews frequently appearing in popular journals such as *MacWorld* are of great benefit in keeping abreast of developments. In my own work, two statistical packages have been of particular value: Data Desk and StatView.

No single package seems to offer all the features one needs. Each has its 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. This latter point, the continual intellectual demand placed on the scientist, has largely contributed to the lack of personal computer use. The scientist feels so overwhelmed with day-to-day responsibility 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 use to maintain the necessary skill for effective operation.

Both Data Desk and StatView provide the user with an 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 can generate truly presentation-quality graphics. To achieve speed and high interactivity, Data Desk has few (or no) controls for user-determined plot limits, grid lines, fonts, annotation, *etc*, features nicely implemented in StatView. On the other hand, Data Desk can generate and rapidly rotate 3-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 correspondingly highlighted in all others. As an illustration of the power of this technique, consider a problem commonly encountered in data analysis. We want to compare two datasets and spatially locate those points that show the greatest similarities and the greatest differences.

Assume we have available on a common grid observations of precipitation from two sources. A histogram of the gridpoint differences in precipitation is displayed in Figure 1, top panel. If a binary indicator of land (=0) or water (=1) is available at every gridpoint, a second plot showing the continental land masses can readily be produced (Figure 1, lower panel). 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. Note that in the static displays presented here, several *extremely important* distinguishing features present with a TV monitor are absent: color, intensity, and 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

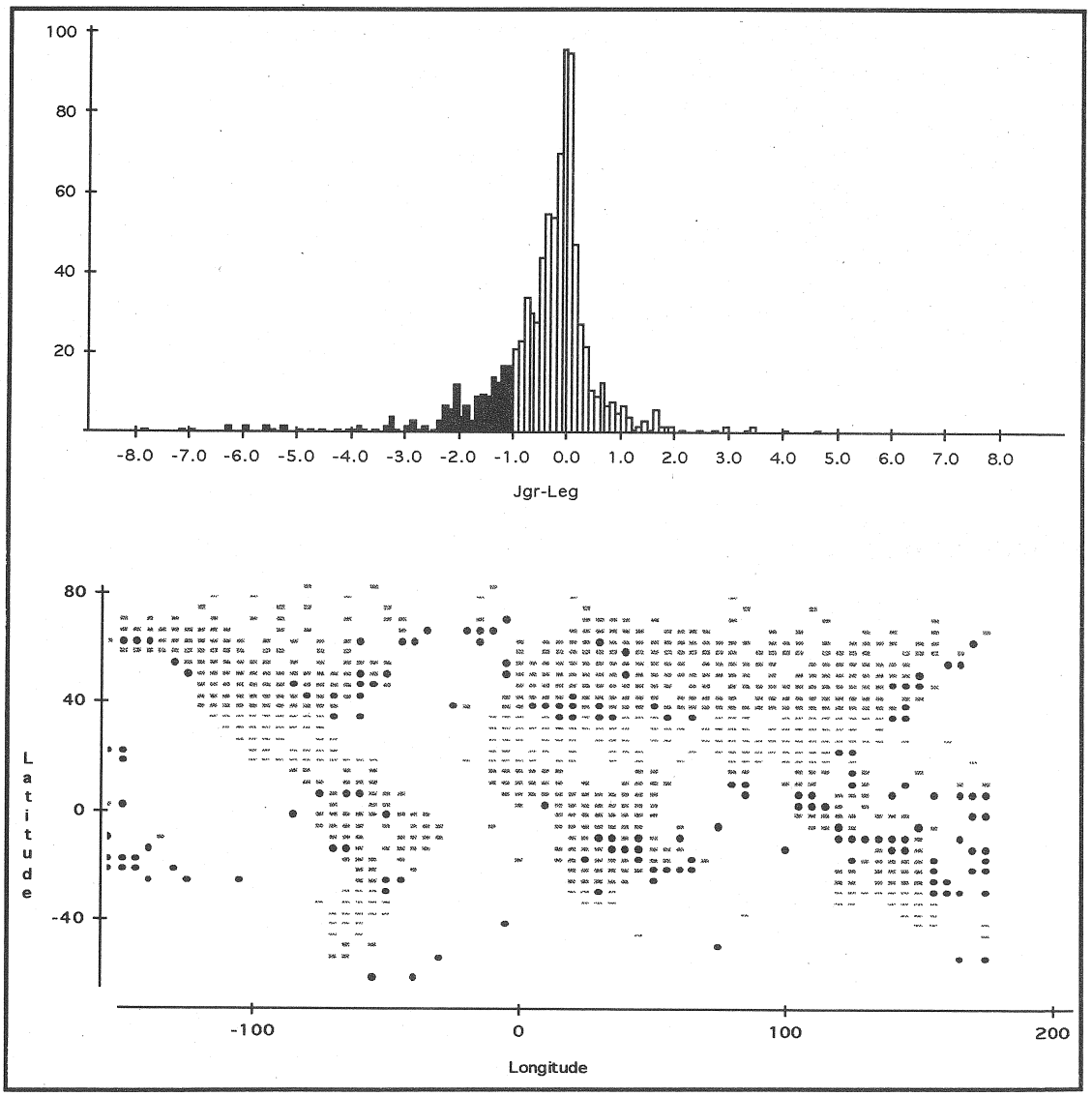

Figure 1. The upper panel shows a histogram of pointwise differences in observed precipitation between two datasets. When the lower tail of the histogram is selected in Data Desk, it darkens. Simultaneously, points corresponding to these maximal differences glow on a world map indicating where these differences arise.

selected bars instantly darken on the histogram display. Simultaneously, those grid points captured in the highlighted ranges also glow on the lower map. These map-selected points can be preserved using color and/or shape, or the user can choose another set of ranges in the histogram and the initial selection will disappear. To spatially locate and differentiate the three regions (lower tail, upper tail, and central region [best agreement]), each grouping is chosen, in turn, on the histogram. As 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 must 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 (Figure 2, lower panel). At the same moment, a sub-histogram would darken (Figure 2, upper panel), 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 (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. 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 extraordinary potential of this technique for interactively analyzing multivariate datasets has not been exploited by most scientists, largely due to ignorance of the existence of such tools for the Macintosh.

Graphical Software

I have found three software packages of particular value for scientific data display: Kaleidagraph, Delta Graph, and Spyglass Transform and Dicer. Once again, each package has virtues and disadvantages.

Many excellent software packages exist for producing the "bread and butter" plots of the scientist: 2-dimensional scatter and line plots. Any "recommended" choice among these is particularly subjective. In my own experience, Kaleidagraph has proven particularly easy to use and versatile in permitting considerable control in the embellishment of 2-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. While such capabilities might seem superfluous, in scientific graphics these capabilities are essential to produce both esthetically pleasing and scientifically informative graphics.

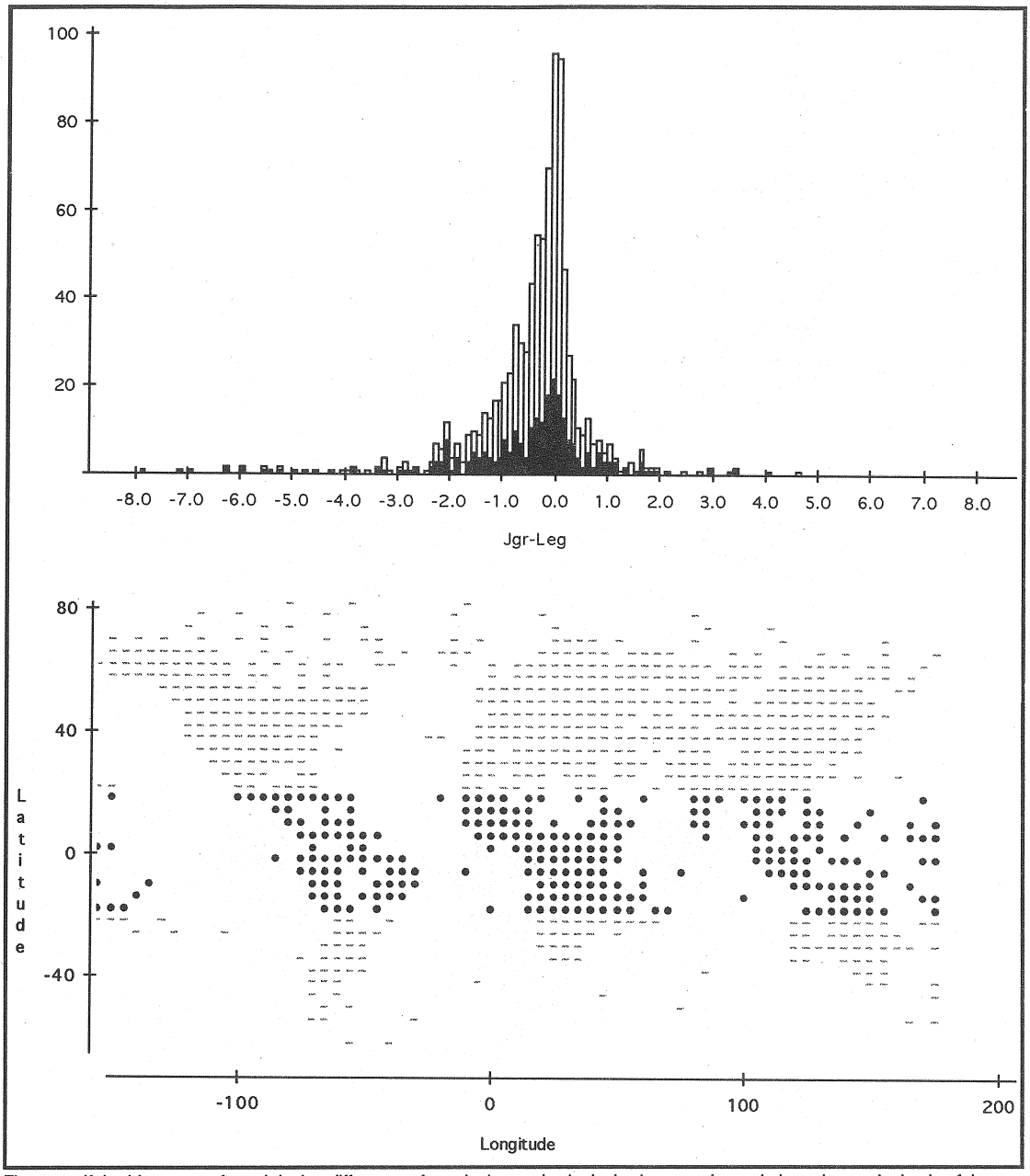

Figure 2. If the histogram of precipitation differences for only the tropics is desired, move the tool along the vertical axis of the lower map until the desired latitudinal range is captured. As the tool is moved, chosen points glow in the map display. Simultaneously, a darkened sub-histogram is outlined in the upper histogram, showing the distribution for only the tropical gridpoints.

As might be expected, the number of software packages which can produce truly effective 3-dimensional scientific plots is much less than for 2-dimensional. Kaleidagraph has no 3D plotting capability. Data Desk, Delta Graph and Spyglass all have 3D 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 realistic 3D effect. (The linked plot feature described above also functions with the 3D plots). However, although the 3D effect using parallax motion is visually

excellent on the screen, the result is often disappointing when produced as a hard copy. Delta Graph and Spyglass, on the screen and are both too slow to produce motion interactively, but both do produce presentation-quality 3D graphics.

The Spyglass suite of software (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 2- and 3-dimensional false color images and contour plots. Several illustrative examples are presented using Spyglass Transform and Dicer. However, much of the visual impact and clarity of detail produced by using color is lost in the figures presented here.

Transform is the primary false color plotting package of the Spyglass suite. 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 may be quickly changed in an interactive manner. Continental outlines, vectors, and contour overlays can be added to any 2-dimensional display. Figure 3 shows the temperature distribution predicted by a global GCM. The subtle gradations in the color original are lost in this reproduction.

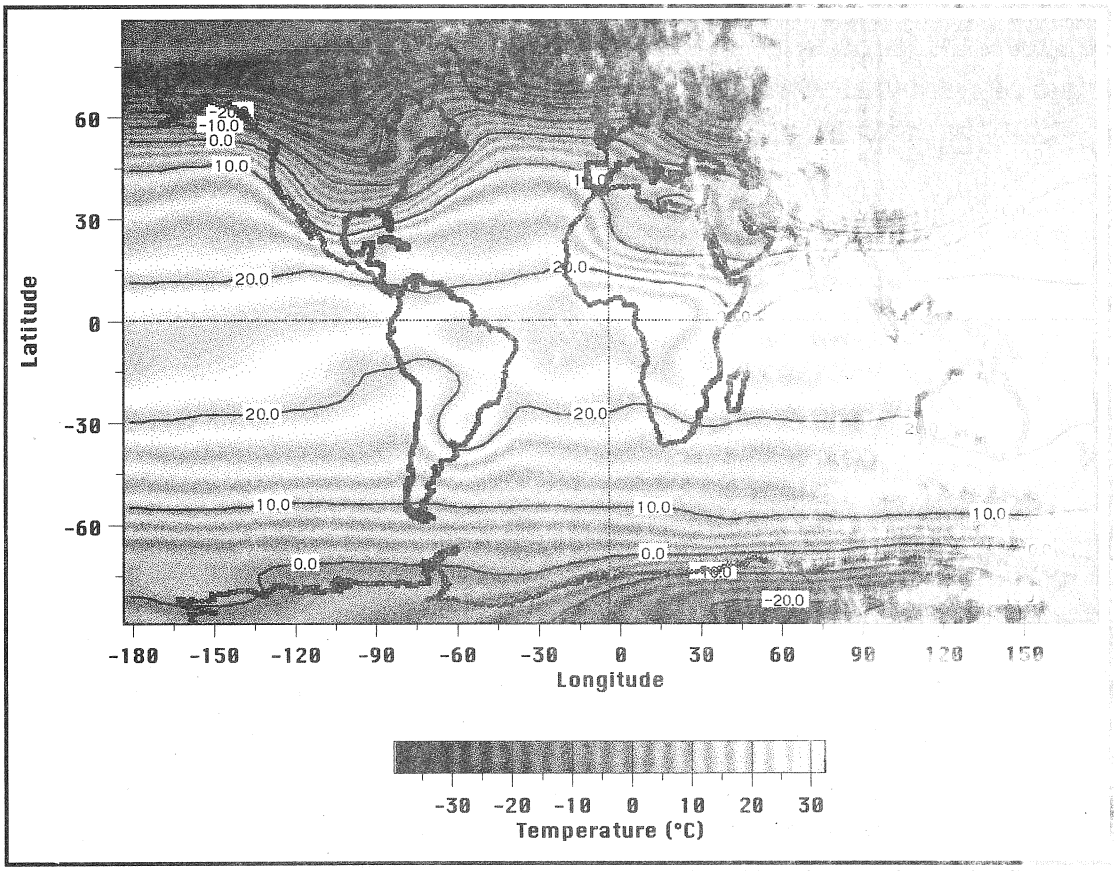

Figure 3 The global temperature distribution predicted by a GCM is represented as a false color map using Spyglass Transform. To produce this display, the underlying colored map is overlaid with a contour map (also generated with Transform) and continental outlines (obtained from another source).

With Spyglass Transform these same data can also be quickly rendered as a wiremesh surface in three dimensions, as shown in Figure 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 2D or 3D plot is generated, the user may save all the instructions as a macro to facilitate reproducing similarly-scaled plots for intercomparison of datasets or for generating animations.

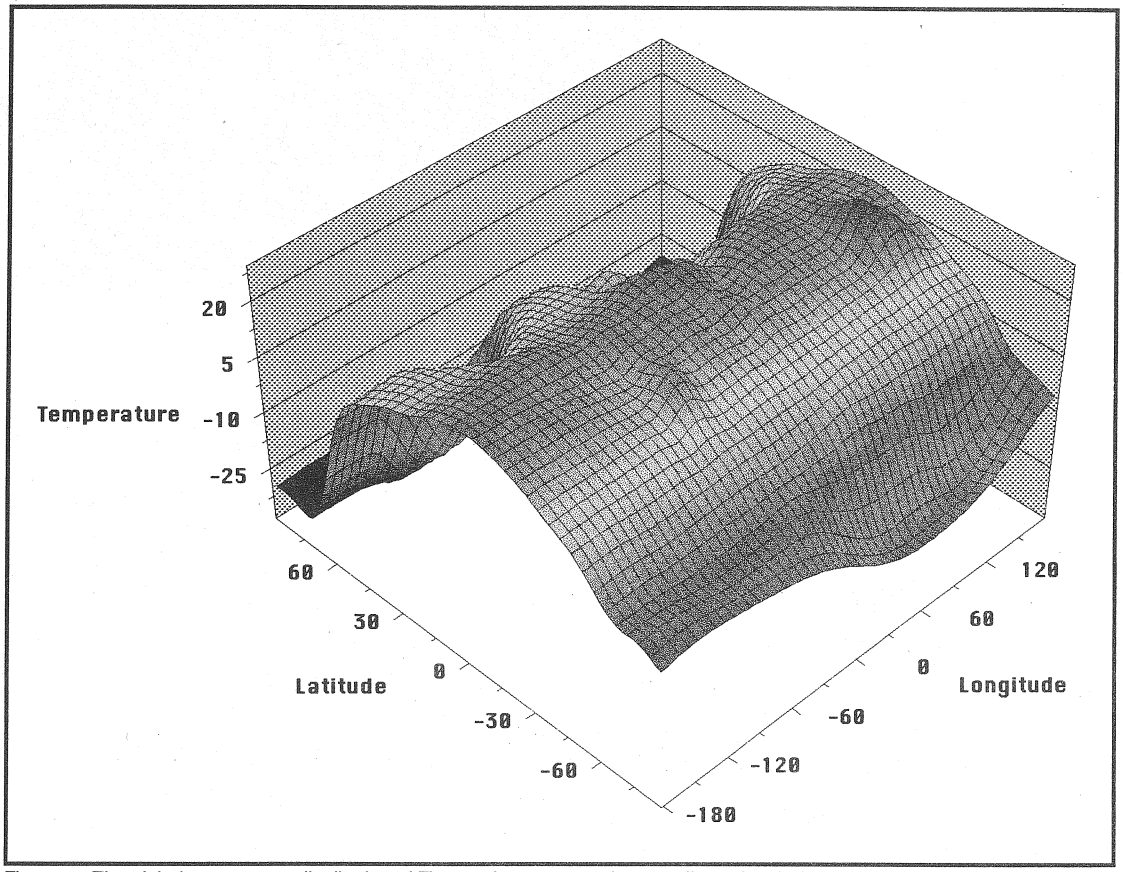

Figure 4 The global temperature distribution of Figure 3 is represented as a 3-dimensional wiremesh surface using Spyglass Transform. The user can interactively change many features of the plot, including user viewpoint, aspect ratio, axis scaling and labeling, and colors.

For many years the Macintosh has been the computer of choice for work in the graphic arts. Capabilities in this area are impressive. The ability to "cut and paste" graphics from disparate sources are of considerable value in scientific graphics. No longer does the scientist have to generate all the components of each graphic in a single program.

For example, in Figure 5 a 3D wiremesh rendition of the temperature data of Figures 3 and 4 from Spyglass Transform is merged seamlessly with a world map using Adobe Photoshop. The 2D map of Figure 3 could just as easily be inserted in this lower plane. With the many tools provided in graphics programs, the user can quickly resize and overlay

different plots and add textual and arrow annotations. To enhance the printed output, colors can be altered interactively, and brightness and contrast can be changed quickly. For presentations on the printed page or as view graphs, these capabilities can change a dull plot into one far more visually appealing and informative.

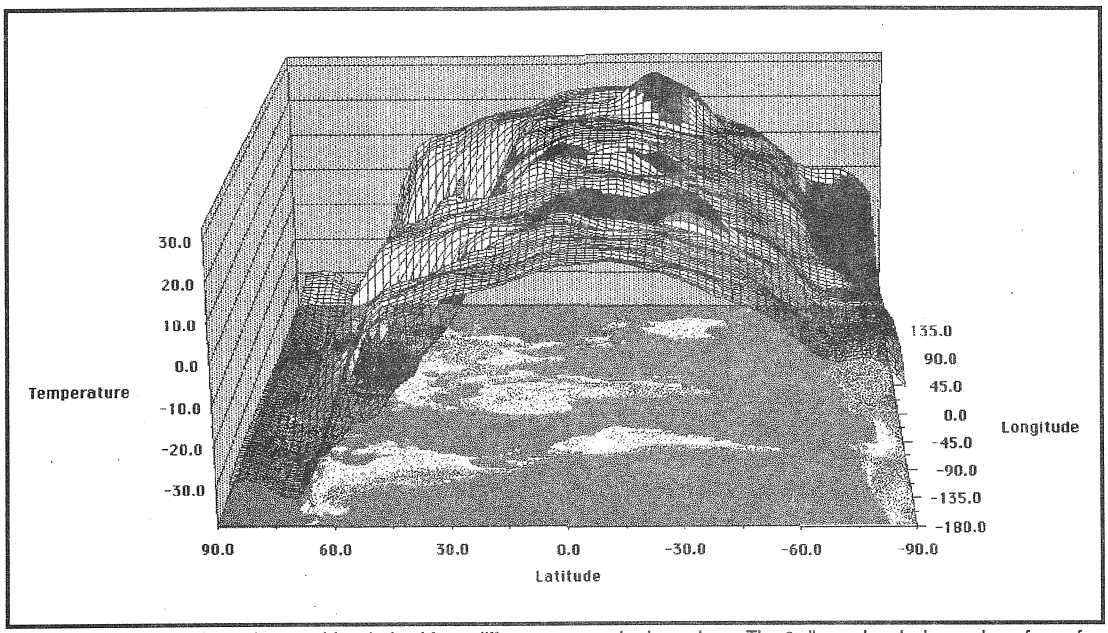

Figure 5. The power of merging graphics derived from different sources is shown here. The 3-dimensional wiremesh surface of global temperature is generated using Spyglass Transform. The lower world map, showing topography, was obtained from the Macldas group at Wisconsin via Internet. The two graphics were seamlessly merged using Adobe Photoshop. To obtain better spatial location, the continents may be shaded in the upper surface as well, using other capabilities in Transform.

Dicer is probably the most novel of the suite of Spyglass programs. As the name implies, Dicer permits the user to consider 3-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. Figure 6 shows 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 can be produced, highlighting other aspects of the data. In Dicer, the user can generate a series of parallel slices to produce very effective animations. For example, in the example of Figure 6, the user can easily produce 12 horizontal slices to obtain a 2D animation of monthly climatology. These frames can be converted readily 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 remarkable. These capabilities are particularly valuable in the geophysical sciences, where highly dimensional data are common.

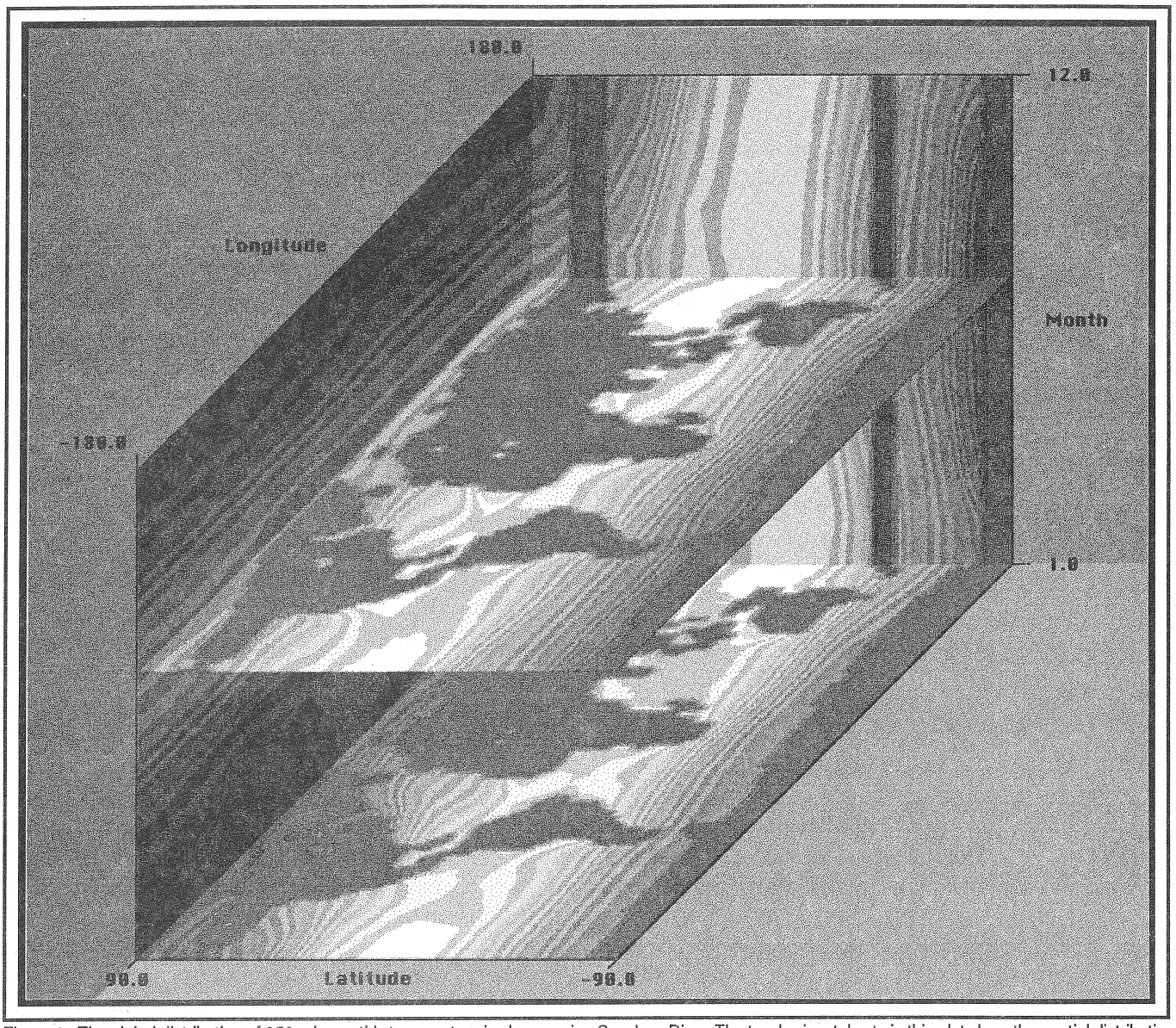

Figure 6, The global distribution of 850-mb monthly temperature is shown using Spyglass Dicer. The two horizontal cuts in this plot show the spatial distribution of the January and July average climatologies. Dicer permits cuts to be made interactively along any of the axes, as well as obliquely through the data. Thus, one can examine 2-dimensional views through the 3-dimensional dataset. Blocks and isosurfaces through the data can also be represented. The effects are particularly useful when displayed in color.

Conclusions

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

Software enhancements are 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 results and in producing animations of data where hundreds of individual frames must often be produced.

Infusion of techniques from the field of graphic arts and the world of multimedia should exert considerable influence in scientific graphics. 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 another major growth area in graphics. Until very recently, scientists had to have help from graphics specialists to produce effective movies. This was typically very costly and severely limited the number of animations attempted. It also restricted experimentation. This situation is changing rapidly. Desktop movie-making 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 for the Macintosh. Finally, standard movie-playing software, such as Apple's QuickTime, will greatly facilitate the process of sharing animations among colleagues.

My experience has shown that many scientists still are unwilling to make the time and intellectual commitment needed to master these techniques on the small computer. There is little question that many techniques 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. The next generation of scientists, for whom these technologies may be more familiar and less threatening, will no doubt be more receptive to their use. I believe 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., and D. Morganstein. 1991. Statistics Programs Designed for the Macintosh: Data Desk, Exstatix, Fastat, JMP, Statview II, and Super Anova. *American Statistician*. 45(4), 318-333

Software Sources

Data Desk: Data Description, Inc., P.O. Box 4555, Ithica, NY (607/257-1000).

DeltaGraph: DeltaPoint, Inc., 2 Harris Court, Monterey, CA 93940 (408/648-4000).

Kaleidagraph: Synergy Software, 2457 Perkiomen Avenue, Reading, PA 19606 (215/779-0522).

Spyglass: Spyglass, Inc., P.O. Box 6388, Champaign, IL 61826 (217/355-6000).

StatView: Abacus Concepts, Inc. , 1984 Bonita Avenue, Berkeley, CA (510/540-1949).