Validating a Large Geophysical Data Set: Experiences with Satellite-Derived Cloud Parameters

Ralph Kahn, Robert D. Haskins, James E. Knighton,

Andrew Pursch, and Stephanie Granger-Gallegos

Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109

Abstract

We are validating the global cloud parameters derived from the satellite-borne HIRS2 and MSU atmospheric sounding instrument measurements, and are using the analysis of these data as one prototype for studying large geophysical data sets in general. The HIRS2/MSU data set contains a total of 40 physical parameters, filling 25 MB/day; raw HIRS2/MSU data are available for a period exceeding 10 years. Validation involves developing a quantitative sense for the physical meaning of the derived parameters over the range of environmental conditions sampled. This is accomplished by comparing the spatial and temporal distributions of the derived quantities with similar measurements made using other techniques, and with model results.

The data handling needed for this work is possible only with the help of a suite of interactive graphical and numerical analysis tools. Level 3 (gridded) data is the common form in which large data sets of this type are distributed for scientific analysis. We find that Level 3 data is inadequate for the data comparisons required for validation. Level 2 data (individual measurements in geophysical units) is needed. A sampling problem arises when individual measurements, which are not uniformly distributed in space or time, are used for the comparisons. Standard 'interpolation' methods involve fitting the measurements for each data set to surfaces, which are then compared. We are experimenting with formal criteria for selecting geographical regions, based upon the spatial frequency and variability of measurements, that allow us to quantify the uncertainty due to sampling. As part of this project, we are also dealing with ways to keep track of constraints placed on the output by assumptions made in the computer code. The need to work with Level 2 data introduces a number of other data handling issues, such as accessing data files across machine types, meeting large data storage requirements, accessing other validated data sets, processing speed and throughput for interactive graphical work, and problems relating to graphical interfaces.

KEY WORDS: large data sets, validation, satellite data analysis

1. Introduction

NASA's Earth Observing System (EOS) will generate vast quantities of data. Hundreds of terabytes of data will be acquired from orbit to characterize the Earth's environment with the kind of spatial and temporal detail needed to study climate change. Such high resolution is required to properly sample the non-linear impact of small-scale phenomena, which can make significant contributions to the global-scale budgets of heat and momentum. It is also expected that the data will be analyzed not just in the traditional manner, concentrating on a single data set at a time, but in new ways that involve routinely comparing data sets from multiple sources. Part of the need to study multiple data sets comes from a growing appreciation for the importance to global conditions of transports across boundaries such as the air-ocean interface (e.g., Earth System Science Committee, 1988).

We are undertaking the validation of cloud parameters derived from the High Resolution Infrared Radiation Sounder 2 (HIRS2) and the Microwave Sounding Unit (MSU) instruments aboard the NOAA polar orbiting meteorological satellites. The instruments provide one of the few global measures of cloud properties extending over many years. They are also capable of obtaining near-simultaneous constraints on the physical characteristics of the atmosphere and surface needed to derive cloud properties. One goal of this work is to learn about analyzing large geophysical data sets in general.

Radiances from the HIRS2 and MSU instruments have been analyzed by Susskind and co-workers using an algorithm that accounts self-consistently for the first-order physical quantities affecting the emergent radiation (Susskind et al., 1984; 1987). The standard data products are (1) monthly mean values for forty meteorological parameters, including effective cloud amount and effective cloud top height, on a grid of boxes 2 degrees in latitude by 2.5 degrees in longitude, and (2) 'daily data' with twice-daily temporal sampling, a spatial resolution of about 125 km, and spacing between points of about 250 km. The monthly mean data are referred to as a 'Level 3' (gridded) product, and the daily

data is called a 'Level 2' product (individual measurements reduced to geophysical units) (Space Science Board, 1982; EOS Data Panel, 1986). The size of the uncompressed Level 3 data is about 4 MB/month, whereas the Level 2 product fills about 25 MB/day (750 MB/month).

By validation we mean 'developing a quantitative sense for the physical meaning of the measured parameters,' for the range of conditions under which they are acquired. Our approach involves: (1) identifying the assumptions made in deriving parameters from the measured radiances, (2) testing the input data and derived parameters for statistical error, sensitivity, and internal consistency, and (3) comparing with similar parameters obtained from other sources using other techniques. A study of this type was performed for sea surface temperature (Njoku, 1985), and our project is one of several parallel efforts currently underway to validate different cloud climatologies (e.g., Rossow et al., 1985; 1990). The validation effort we are undertaking introduces a number of problems that may be of interest to specialists in computational statistics, such as the INTERFACE community, as well as to those involved in research directly related to interpreting large geophysical data sets. This article summarizes the key data handling issues we have encountered.

2. The Need for 'Level 2' Data

Large geophysical data sets, such as cloud climatologies, are often distributed to researchers in gridded (Level 3) form. This can reduce the data volume by orders of magnitude relative to the parameter values for each individual sounding (Level 2), and provides the user with a 'spatially uniform' data product. For example, Figure 1A is the global, monthly-mean cloud amount map for July 1979 from the HIRS2/MSU data, in the original 2 degree by 2.5 degree averaging bins. All accepted cloud amount data from the individual atmospheric soundings that fell within each geographic box were summed, and mean and variance values for each box were calculated.

Several problems occur when using Level 3 products for validation. First, if only the Level 3 parameter values and associated variances are available, there is no way to assess how much of the reported variance is due to inherent non-uniformity of the parameter over the averaging region. Essentially, the instrument resolution is degraded to a scale comparable to the box size, and information originally acquired to measure smaller-scale phenomena in both the spatial and temporal domains is lost. For example, in a 2 by 2.5 degree box, the surface temperature may exhibit random fluctuations of half a degree and may change systematically by several degrees, whereas the box average variance will assign all the variability to random error.

We encountered a second problem when making comparisons among Level 3 products with different gridding schemes. The best concurrent cloud climatology available for comparison with the data in Figure 1A was derived from the Temperature Humidity Infrared Radiometer/Total Ozone Mapping Spectrometer (THIR/TOMS) on the NASA Nimbus 7 satellite (Stowe et al., 1988; 1989). The standard THIR/TOMS Level 3 data product was binned according to a global 500 by 500 km grid that is also used for Earth radiation budget studies. The July 1979 HIRS2/MSU Level 3 data, degraded using area-weighted averaging to the THIR/TOMS spatial grid, is shown in Figure 1B. We then resampled the degraded HIRS2/MSU data back to the 2 by 2.5 degree grid, and subtracted it from the original HIRS2/MSU data (Figure 1C). Note that the differences are nearly as large as the range of the signal, with both positive and negative values. The pattern of differences varies with the location of edges in the original data, and is modulated by the relative position of grid boundaries. Differences are especially large at high latitudes, where the spatial resolution of the THIR/TOMS grid is much lower than that of the HIRS2/MSU grid, and wherever there are sharp edges generated by cloud patterns, such as in the intertropical convergence zone and monsoon areas.

With the Level 2 products, we have access to physical quantities at the full resolution acquired by the instruments, and avoid introducing additional artifacts into the comparison between data sets. Level 2 data are not uniformly distributed over the surface. At low latitudes there are gores in the HIRS2 sampling between orbits, whereas at high latitudes, the surface is heavily oversampled. Data dropouts and calibration lines occur at all latitudes. The sample resolution changes by more than a factor of 2 from nadir to the limits of each scan. As a first step toward making comparisons among Level 2 data sets, surfaces that take account of non-uniform clustering of data points may be fit to the data. We have begun experimenting with locally adaptive surface fitting techniques (e.g., Renka, 1988), and are exploring the use of methods that generate variance surfaces together with each fitted surface (Cresse, 1989, and references therein).

Binning, which is traditionally used to make comparisons among global data sets, is performed as an automatic procedure. In using Level 2 data for validating data sets, geographic sub-regions of the globe must be selected for surface fitting, based upon some criterion that evaluates the density of points relative to the size of local gradients of the parameter field, possibly in several directions. Figure 2 illustrates the role of interactive geographic subset selection a part of the software we are assembling to perform the HIRS2/MSU validation. 'HDF' in this figure refers to Hierarchical Data Format, a transportable file format that eliminates all but an initial file conversion for exchanging data among DEC, Sun, MacIntosh, and other machines used in the validation (NCSA Software Tools Group, 1990).

This allows us to store single copies of data files on centrally located disks, that are accessible across the network to machines with differing architectures. We are currently investigating the criteria for accepting subsets, choice of method for surface fitting, and methods for making formal comparisons among surfaces fitted to data from different sources. The important question of interpolation in the temporal domain we set aside for the present.

To summarize: in spite of the much larger volume of the Level 2 data, relative to Level 3, and the collection of issues related to the spatial and temporal sampling of Level 2 data, we need the ability to access, store, and process Level 2 data for (1) studies of the internal consistency and precision of the data set and (2) comparisons with other cloud climatologies, that are involved in the validation of the HIRS2/MSU cloud parameters. We anticipate that similar needs will arise for interdisciplinary process studies, and in work directed toward using observations to better understand mesoscale climatological phenomena.

3. Tracking Assumptions in the Code

Another issue that bears upon the degree to which we may perform validation, and other scientific analysis on large data sets, is our ability to grasp the collection of constraints imposed on parameter values by the code that generates them. An assumption embedded in a large data handling code may produce results that hide important information in the data, or may produce patterns in the data that could be incorrectly interpreted as scientifically meaningful.

We are experimenting with methods of charting the collection of assumptions, as a way of calling the attention of the user to areas where the code may influence the output parameters. We are using standard charting symbols as much as possible (e.g., Yourdon and Constantine, 1979). An example of this type of chart is Figure 3. This shows the flow of control and the flow of assumptions made in a relatively small part of the HIRS2/MSU analysis code that produces Level 3 data from Level 2 products. This chart made clear the number and complexity of the assumptions involved in generating Level 3 products, and it played a role in our assessment of the value of Level 3 data for the validation exercise.

Charting the flow of control provides a needed context for the constraints placed on the data. These charts take a step in the direction of making it *possible* to keep track of assumptions, but they do not eliminate the work involved in carefully assessing the meaning of derived parameters.

4. Conclusions

The HIRS2/MSU cloud parameter validation effort raises a number of data handling issues that are likely to arise frequently when scientific analysis is attempted on large geophysical data sets. We need Level 2 data (individual measurements in geophysical units) (A) to perform comparisons among data sets with different sampling, and (B) to understand the effects of spatial and temporal sampling on the 'average' values obtained from a single data set. The need for Level 2 data severely complicates data handling. Among the areas where advances would be most helpful are:

1. Surface fitting software for data distributed non-uniformly in 2-dimensional space, and ways to obtain some measure of the associated variances.

2. Software for making formal comparisons among fitted surfaces from several sources, and their associated variance surfaces.

3. Ways of documenting software and data files so they may be exchanged and used by others easily.

4. Ways of documenting the assumptions embedded in retrieval and processing algorithms, so a researcher studying the data products can grasp the collection of constraints placed on the output data by the code.

5. Additional ways of storing data. For a given Level 2 data product, we need readily accessible data storage capacity of between one and two *orders of magnitude* the size of the basic data set, for intermediate and derived products that are created as part of the validation.

Several longer-term needs include:

6. The development of validation procedures that are easy enough to apply so that it will be feasible to generate and access a large number of validated geophysical data sets for interdisciplinary studies of all types.

7. Ways of fitting surfaces to data values distributed non-uniformly in 2-dimensional space and in time, and obtaining a measure of the associated variances.

8. Better ways of discovering patterns and surprises in high-dimensional data sets.

9. Ways of fitting hyper-surfaces to higher dimensional data sets, and techniques for studying them.

We have described our data, the collection of problems we are facing in the validation work, and our approaches to some of these issues. Solutions or partial solutions may exist to some of the problems that are not widely known outside specialized data handling and computational statistics communities. We hope to stimulate experts in these fields to participate in the effort to improve our understanding of Earth through the study of large, geophysical data sets.

Acknowledgments

We thank Paul Tukey for inviting us to participate in the INTERFACE 91 conference, and Daniel Carr, Jeff Dozier, Mike Freilich, Wes Nicholson, Bill Rossow, Victor Zlotnicki, and Richard Zurek for stimulating discussions on many aspects of this work. This project is supported in part by the NASA Earth Sciences Interdisciplinary Program in the Earth Science and Applications Division, and by the Jet Propulsion Laboratory Director's Discretionary Fund. The work was performed at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

References

Cressie, N. (1989), Geostatistics, *The Amer. Statistician*, 43, 197-202.

Earth System Science Committee (1988), "Earth System Science: A Closer View", Report of the Earth System Science Committee, NASA Advisory Council, NASA, Washington, D.C.

EOS Data Panel (1986), The Earth observing system: Report of the EOS data panel, Vol 2a, NASA Tech. Memo. 8777. Washington, D.C.

NCSA Software Tools Group (1990), Hierarchical Data Format, National Center for Supercomputing Applications, Champaign, IL.

Njoku, E.(1985), Satellite-derived sea surface temperature: Workshop comparisons, *Bull. Am. Meteorol. Soc.*, 66, 274-281.

Renka, R.J. (1988), Multivariate interpolation of large sets of scattered data, *ACM Transact. Math. Software*, 14, 139-148.

Rossow, W.B., Mosher, F., Kinsella, E., Arking, A., Desbois, E., Harrison, E., Minnis, P., Ruprecht, E., Seze, G.,Simmer, C., and Smith, E. (1985), ISCCP cloud algorithm intercomparison., *J. Climate Appl. Meteor.*, 24, 877-903.

Rossow, W.B. (1990), Report of the Workshop on Comparison of Cloud Climatology Datasets, NASA Goddard Institute for Space Studies, New York.

Space Science Board (1982), Data management and computation, Vol 1: Issues and recommendations. National Academy of Sciences/ National Academy Press, Washington, D.C.

Stowe, L.L., Wellemeyer, C.G., Eck, T.F., Yeh, H.Y.M., and the NIMBUS 7 Cloud Data Processing Team (1988), NIMBUS 7 global cloud climatology. Part I: Algorithms and validation, *J. Climate*, 1, 445-470.

Stowe, L.L., Yeh, H.Y.M., Eck, T.F., Wellemeyer, C.G., H.L. Kyle, and the NIMBUS 7 Cloud Data Processing Team (1979), NIMBUS 7 global cloud climatology. Part II: First year results, *J. Climate*, 2, 671-709.

Susskind, J., Rosenfield, J., Reuter, D., Chahine, M.T. (1984), Remote sensing of weather and climate parameters from HIRS2/MSU on TIROS-N, *J. Geophys. Res.*, 89, 4677-4697.

Susskind, J., Reuter, D., Chahine, M.T. (1987), Cloud fields retrieved from analysis of HIRS2/MSU sounding data, *J. Geophys. Res.*, 92, 4035-4050.

Yourdon, E., and Constantine, E.E. (1979), Structured Design: Fundamentals of a Discipline of Computer Program and System Design, Yourdon Press, NJ, pp 473.

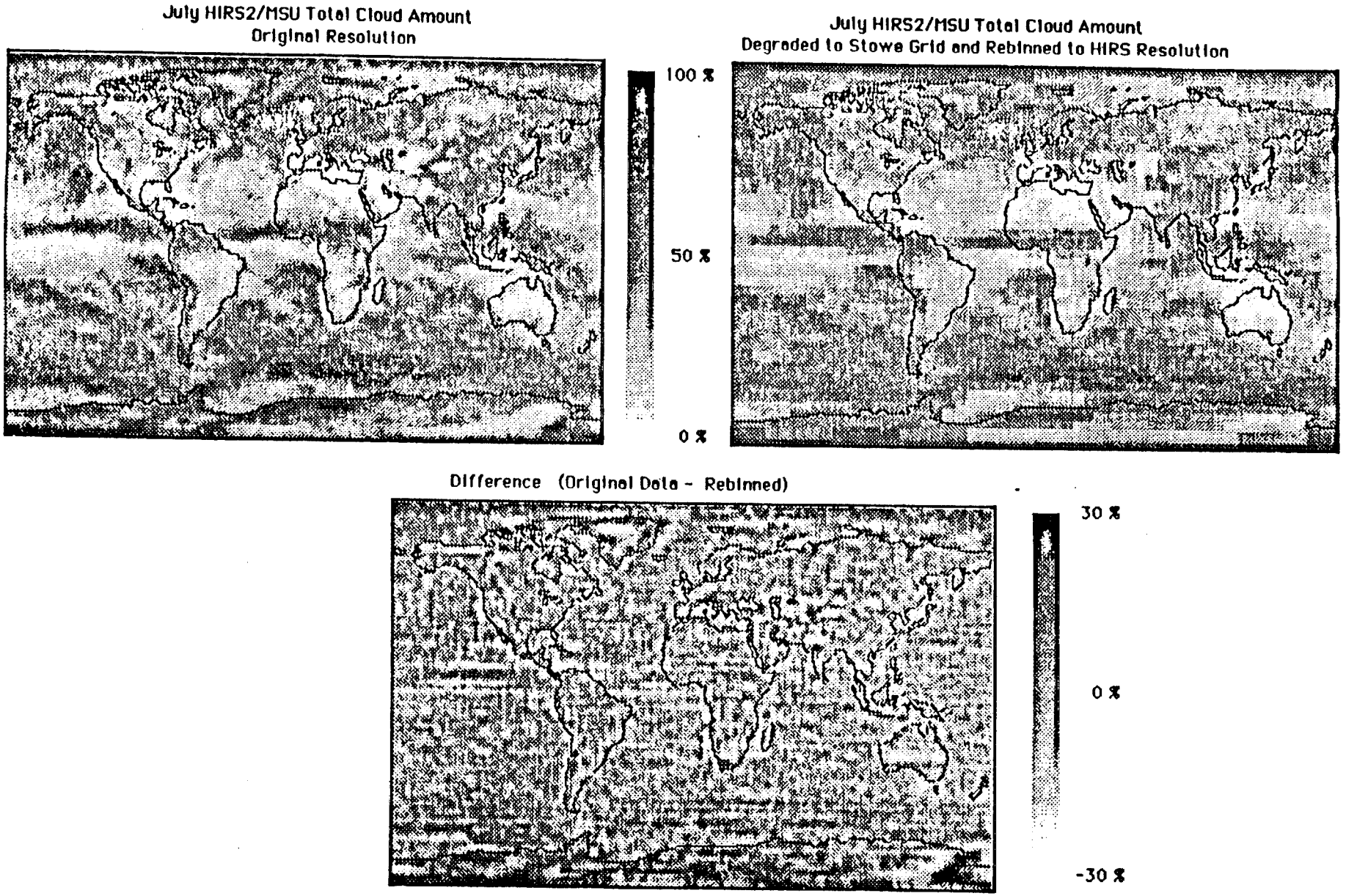

Figure 1. The Effect of Rebinning on Global Cloud Amount

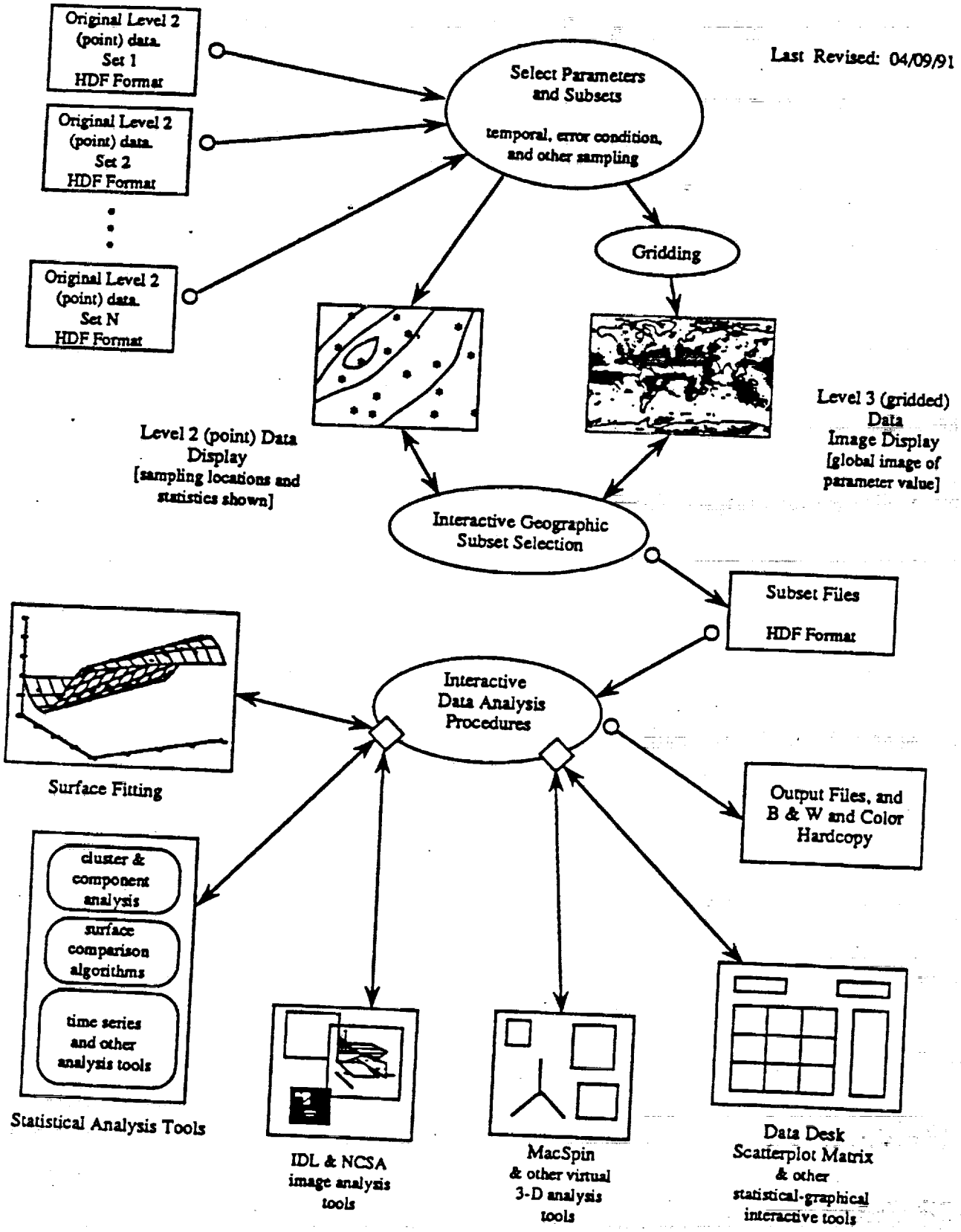

Figure 2. Level 2 Data Analysis Software

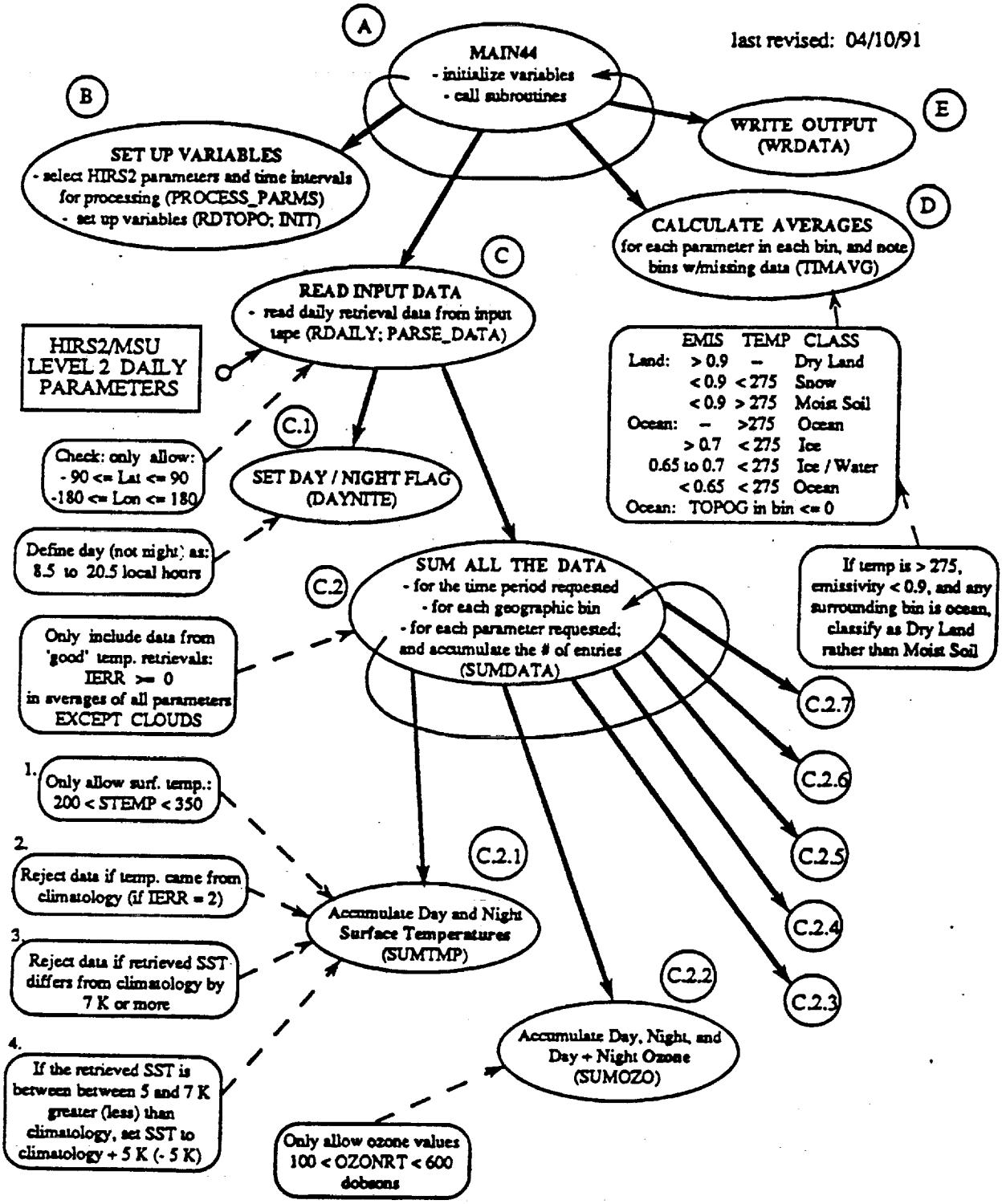

Figure 3. HIRS2 Level 2 to 3 Software Overview / Assumptions

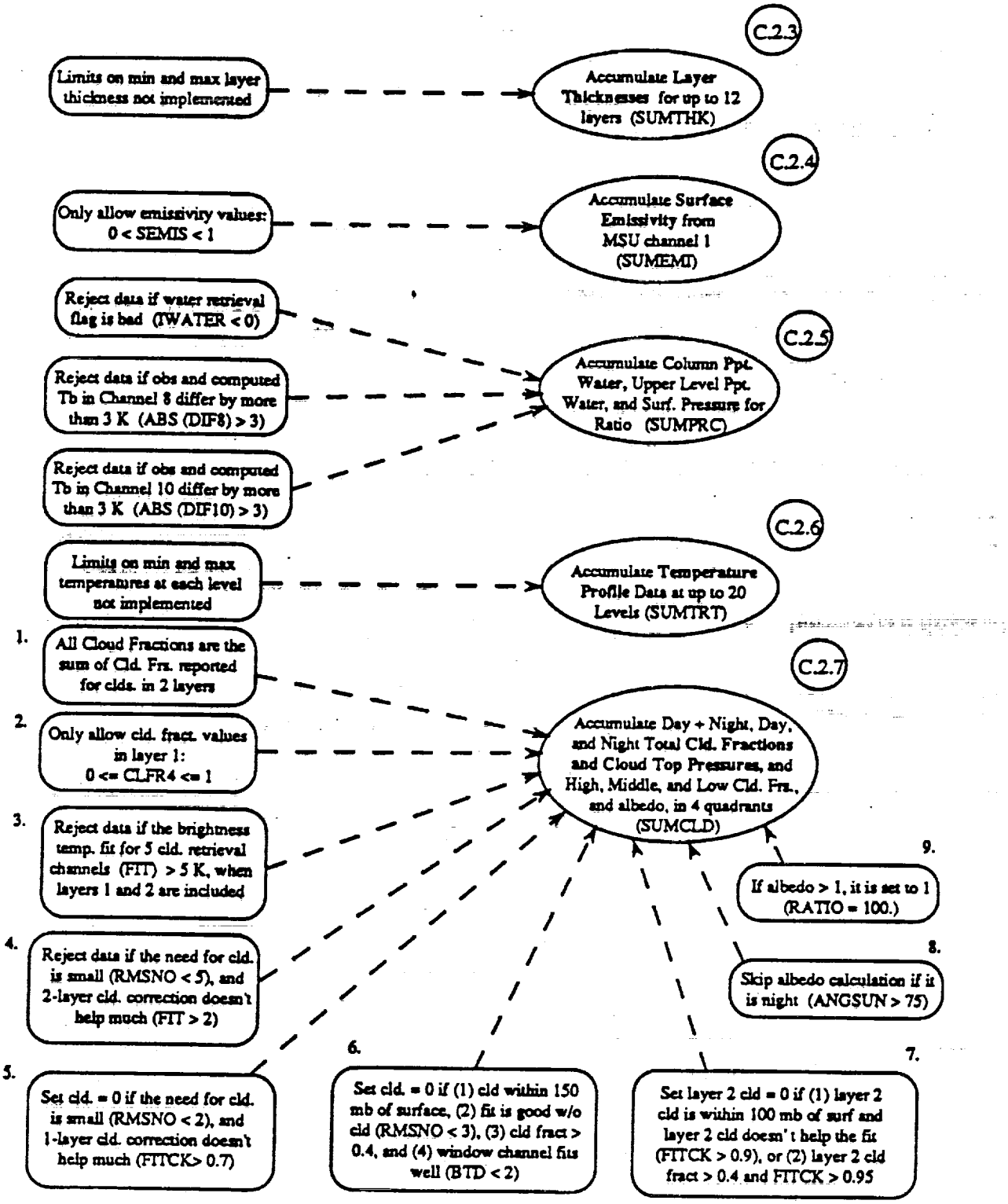

Figure 3. HIRS2 Level 2 to 3 Software Overview (Continued)