For: The Proceedings of the Massive Data Sets Workshop, July 7 & 8, 1995
Committee on Applied and Theoretical Statistics (CATS)
National Academy of Sciences, Washington, DC

last revised: 10/24/95

What Shall We Do With The Data We Are Expecting in 1998?

Ralph Kahn, Jet Propulsion Laboratory /California Institute of Technology, PasadenaCA91109

Abstract

The community of researchers studying global climate change is preparing for the launch of the first Earth Observing System (EOS) satellite. It will generate huge amounts of new data, filling gaps in the information available to address critical questions about the atmosphere and surface of Earth. But many data handling and data analysis problems must to be solved if we are to make best use of the ncw measurements. In key areas, the experience and expertise of the statistics community could be of great help.

1. Introduction

The Earth Observing System (EOS) is scheduled to launch its first platform into polar orbit in June of 1998. The payload includes five remote sensing instruments designed to study the surface and atmosphere of Earth. In a broad sense, the purpose for making these observations is to find indications of how Earth's climate is changing, and to discover clues to the mechanisms that are responsible for these changes. To this end, a 5 to 15 year program of global monitoring is planned, covering many wavelengths, with spatial resolutions as small as **0.25 km** and temporal coverage as frequent as a day. Higher resolution data on regional scales will also be acquired.

The surface area of Earth is about 5 x 10^8 km^2. At 0.25 km resolution, a single instrument acquiring 36 channels of data, such as the Multi-angle Imaging SpectroRadiomctcr (MISR) or the Moderate Resolution Imaging Spectrometer (MODIS) on the EOS platform, will generate upwards of 80 Gbyte/day, or 30 Tbyte/year of basic data. The geophysical quantities arc generally retrieved at lower spatial resolution, but must include quality flags and other ancillary information, resulting in a geophysical data set that will be no smaller than 37 'byte/year for the MISR instrument alone.

The sheer volume of data creates unprecedented challenges for accomplishing basic data handling operations, such as throughput and storage. But there are deeper issues regarding the scientific use of this huge amount of data. The EOS community has adopted a partial framework, and some terminology, for discussing the questions we must face. However, in many areas the development of an approach to the underlying issues is in its infancy. This paper begins with a brief review of the data classification scheme we use to organize our thinking about data handling and analysis. This is followed by discussions of some issues relating to specific classes of data, and a summary of areas to which the statistics community maybe well-equipped to contribute.

2. Data Classification Scheme

The Committee on Data Management And Computing define five general classes of spacecraft data, based on the degree of processing involved (CODMAC, 1982, and subsequent refinements):

- **Level O -- The** raw data stream from the spacecraft, as received at Earth

- **Level 1 --** Measured radiances, geometricall y and radiometricall y calibrated

- **Level** 2-- Geophysical parameters, at the highest resoluti on available
- **Level** 3-- Averaged data, providing spatially and temporally "uniform" coverage
- **Level** 4-- Data produced by a theoretical model, possibly with measurements as inputs

This paper focuses on Level 2 and Level 3 data, which arc the main concerns of most global change research scientists working on EOS instrument teams. Level 2 products are reported on an orbit-by-orbit basis. For a polar-orbiting satellite such as EOS, the Level 2 sampling of Earth is highly non-uniform in space and time, with coverage at high latitudes much more frequent than near the equator. Level 2 data is needed when accuracy at high spatial scale is more important than uniformity of coverage. These situations arise routinely for validation studies of the satellite observations, in the analysis of field campaign data, and when addressing other local- and regional-scale problems with satellite data.

The spatially and temporally uniform Level 3 data arc needed for global-scale budget calculations, and for any problem that involves deriving new quantities from two or more measurements which have different sampling characteristics. To derive a Level 3 product from Level 2 data, spatial and temporal scales must be chosen. It is to this issue that wc turn next,

3. **Grinning and Bidding to Create Level 3 Data**

The creation of **Level 3** data has traditionally involved the selection of a global, 2- or 3-dimensional spatial grid, possibly a time interval as well, and "binning" the Level 2 data into the grid cells. The binning process for large data sets usually entails taking the arithmetic mean and standard deviation of all Level 2 data points falling into a grid cell, with possible trimming of outliers or of measurements flagged as "low quality" for other reasons. Typically, all points included in a grid cell average are given equal weight. Occasionally a median value will be used in place of the mean.

The leading contender for the standard EOS Level 3 grid is a rectangular-based scheme similar to one that has been used by the Earth Radiation Budget Experiment (ERBE) (Green and Wielicki, 1995a). In the proposed implementation for EOS, the Earth is divided zonally into 1.25 degree strips (about 140 km in width). Each strip is then divided into an integral number of quadrilaterals, each approximately 140 km in length, with the origin at the Greenwich meridian. This produces a nearly-equal area grid.

A number of issues arise in using a grid of this sort for the Level 3 data. Anisotropy presents an obstacle for calculating gradients, fluxes, and other quantities based on finite differences. Some neighboring cells sham an edge whereas others share only a point, and there is no general rule as to how the contributions of each should be weighted. only zonal gradients can be calculated in a consistent way on a global scale. Even in the meridional direction, the north-south cell boundaries are aligned only along the prime meridian. Inhomogeneity presents a second set of problems, since the distribution of grid cells varies with latitude, and there are singularities at the poles.

A third set of issues arises from the nesting properties of these grids. Nested grids can be used to relate data sets taken at different spatial resolutions, such as data from ground-based; aircraft, balloon, and satellite instruments. It is often necessary to compare these types of data (particularly for validation work), and to use data from multiple sources to calculate new quantities. To form sub-grids at length scales below 140 km, decisions must be made as to whether the subdivisions will be equi-angular, which are unique and relatively easy to define, or equal area, which has more desirable sampling properties, but requires more complex ccl] boundaries that increase anisotropy. Performing analysis on data sets from non-nested grids introduces errors that may be significant on a global scale (Green and Wielicki, 1995 b), and can be arbitrarily large in regions where the quantities of interest have significant gradients (Kahn et al., 1991).

There are alternative grids, based on triangle or hexagon subdivisions of the spherical surface or a projection thereof, that may alleviate some of these issues (I). Cam and P. Hubcr, personal communication, MDS Workshop, 1995). A considerable body of work exists that explores the characteristics of nested systems of such grids (White ct al., 1992, and references therein).

An effort is being organized to develop such grid schemes into systems that EOS scientists can use (Kiestcr, Kimmcrling, Knighton, Olsen, Sahr, and White, personal communication, 1995). A specific choice of grid system is being made, and its geometric properties characterize, Schemes will be needed to address and store data at different levels within the grid system. If the performance of a triangle or hexagon-based grid is promising, efficient translators to and from commonly used addressing systems, such as latitude-longitude, and conversions to popular map projections would need to be derived and implcmcnted in data processing and GIS software packages widely used by the EOS community,

One would like to embed each data set into a grid within a nested system that is appropriate to its resolution and sampling structure. This raises the related issues of how to select a "native" grid size for a given data set, and how best to calculate the value and associated statistics to be assigned to each grid cell from the Level 2 data for both continuous- and discrete-valued quantities. Once this is done, methods may be developed to aggregate and dis-aggregate grids at various spatial resolutions, calculating the associated error. characteristics along with the data (N. Cressie, personal communication, MDS Workshop, 1995).

Such a system would revolutionize the way the global climate research community works with data.

4. Generating Level 2 Data

The generation of Level 2 geophysical quantities from calibrated radiances introduces a far more diverse set of issues, since the retrieval algorithms vary greatly with the type of measurement made and the retrieval strategy adopted. For specificity, I use the Ml SR aerosol retrieval process as the basis for the discussion in this section (Diner et al., 1994).

Two MISR-rclated issues similar to ones that arisc elsewhere are: how to determine the sensitivity of the instrument to differences in atmospheric aerosol properties, and how to develop climatologies for the retrieved geophysical quantities based on existing constraints.

4.1. Sensitivity Studies

From the point of view of retrieving aerosol properties from MISR observations, the distinctions worth reporting are determined by the sensitivity of the instrument. We use a theoretical model to simulate the measurements at the 4 wavelengths and 9 viewing angles covered by the MISR instrument. We run simulations for a wide range of aerosol size distributions, compositions, and amounts. The full parameter space that must be explored includes mixes of particle size distributions and compositions, atmospheric relative humidity, and surface type.

We designate the one set of simulated rcflectances as the "measured" case, and step through "comparison" models covering a range of alternative size distributions, for example. We usc simple χ^2 statistics to make the comparisons, such as:

$$\chi^2_{abs} = \frac{1}{N\langle m_k\rangle}\sum_{l=1}^{4}\sum_{k=1}^{9}\frac{m_k\left[L_{mes}(1,k) - L_{cmp}(1,k)\right]^2}{\sigma^2_{abs}(1,k)} \qquad (1)$$

where L_{mc}s is the simulated "measured" reflectance, Lcmp is the simulated reflectance for the "comparison" model, l and k arc the indices for wavelength and viewing angle, N is the number of measurements included in the calculation, and σ_{abs} is the absolutc measurement crror in the reflectance. m_k is the weight for terms related to viewing angle k, and $<m_k>$ is the average of the weights for all thc viewing angles included in the sum.

Comparisons made in this way reduce the information content of as many as 36 individual measurements (4 wavelengths x 9 angles) to a single. number. There is more information in the data. Two partly independent ways to compare cases are the maximum deviation of all the measurements used, and a χ^2 statistic weighted by the measurements at the nadir angle:

$$\chi^2_{geom} = \frac{1}{N\langle m_k\rangle}\sum_{l=1}^{4}\sum_{\substack{k=1\\k\neq nadir}}^{9}\frac{k\left[\frac{L_{mes}(1,k)}{L_{mes}(1,nadir)}-\frac{L_{cmp}(1,k)}{L_{cmp}(1,nadir)}\right]^2}{\sigma^2_{rel}(1,k)} \qquad (2)$$

where σ_{rel} is the relative measurement error. We arc experimenting with combinations of these metrics as the criteria to be used for evaluating the comparison cases, both in the sensitivity studies, and in the retrieval algorithm.

Our approach to covering the parameter space is also simple. Wc are planning first to vary particle size distribution and amount for fixed composition, establishing the minimum number of sizes needed to represent the range of expected values within the instrument sensitivity. The discrete sizes will be used to determine sensitivity to composition, which is represented by the particle index of refraction. The sensitivity to mixturcs will then bc tested by a similar process.

These procedures are well-defined and systematic. But they arc empirical, and it is impractical to capture every possible combination of conditions with them. In the absence of new ideas, we will live with these limitations.

4.2. Climatologies

The Level 2 retrieval algorithms for EOS must run in an automatic mode, rapidly processing huge amounts of data at computing facilities far from the purview of the instrument teams. As a first step in understanding the results, we plan to automatically compare them with " the expectations" -- a climatology initially based on the best data available prior to launch.

Consider the aerosol climatology. The quantities of interest are the aerosol column amount and the aerosol "type", which summarizes particle composition, sizc distribution, and shape. There exist global satellite estimates of aerosol amount at 1 km resolution, over oceans only, on a weekly basis for almost seven years. For these observations, particle type is assumed. There arc global models of four of the main particle types, at spatial resolutions ranging from about 100 km to about 1000 km, at monthly or seasonal intervals. Numerous ***in situ*** measurements have also been made, with every conceivable spatial and temporal sampling. Some report aerosol amount, others provide information about aerosol type, and a few include both.

How do wc merge all these data into a "climatology?" Our current approach is to ingest monthly cases of the global satellite data set into our geographic information system (GIS) as the primary constraint on aerosol amount. We will then use thc global models to assign aerosol type, on a region-by-region basis (Figure 1). It is undecided as yet how the mix of particle types will be determined from the models, or how the uncertainty in the results will bc obtained. We plan to use ***in situ*** measurements where available, to improvc the constraints placed by the global data ***sets.*** Again wc are undccidcd as to how to weight the information from different data sources, and how to assign uncertainties. Lastly, we must develop the algorithm that compares the aerosol properties derived from the' satellite data with the climatology, and assigns a measure of "likelihood" to the result.

Wc will develop pragmatic approaches to each of these problems, but a formal procedure for constructing a climatology of this sort is beyond our current capability.

5. Summary of Issues

This paper concentrates on matters of potential interest to the statistics community that relate to the generation of Level 2 and Level 3 data from EOS instruments (Table 1). For Level 3 data, the main issues are: defining an effective system of nested grids, deriving procedures for ingesting Lcvcl 2 data into the system, and developing algorithms for aggregating and translating data that is in the system. Level 2 data presents a more diverse set of issues; we focused on performing sensitivity studies and developing climatologies.

The EOS community is preparing to derive geophysical quantities from measurements that will begin appearing in June 1998. All being well, wc will soon face the challenges of actually studying the data, summarizing the trends, identifying and characterizing the exceptions, and exploring the implications of the results for further data acquisition, and for global climate change... more than enough to keep several large and active communities of researchers very busy.

Acknowledgments

I thank my colleagues on the EOS MISR Team for providing the context for this work. I also thank the participants in Massive Datasets Workshop for their interest in our data handling issues, and their patience with our naive approach to profound problems in statistics. This work is performed at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration, through the EOS office of the Mission to Planet Earth.

References

Committee on Data Management and Computation (CODMAC), Data Management and Computation, olume 1: Issues and Recommendati‘ons, Space Scicncc Board, National Academy of Scicnccs, pp. 167, 1982. ‘

Diner, J., W. Abdou, T. Ackerman, J. Conel, H. Gordon, R. Kahn, J. Martonchik, S. Paradise, M. Wang, and R. West, MISR Level 2 Algorithm Theoretical Basis: Aerosol/Surface Product Part 1 (Aerosol Parameters), EOS Project Documen~NASA/JPL-D11400, 1994.

Green, R., and B. Wielicki, Selection of a nestable Grid for EOS products, NASA Langley Research Center Internal Memorandum, CERES Science Team, September, 1995a.

Green, R., and B. Wielicki, Transfrming data from Grid A to Grid B, NASA Langley Research Center Internal Memorandum, CERES Science Team, September, 1995b.

Kahn, R., R.D. Haskins, J.E. Knighton, A, Pursch, and S. Grangcr-Gallegos, “Validating a large geophysical data set: Experiences with satellite-derived cloud parameters”, Proceedings of the 23rd Symposium on Interface, Computing Science and Statistics, 133-140,1991.

White, D., A.J. Kimmcrling, and W.S. Ovcrton, Cartographic and geometric components of a global sampling design for environmental monitoring, Cartogr. and Gcograph. Inform. Svst. 19, 5-22, 1992.

TABLE 1. SUMMARY OF ISSUES

General Data Processing --

- Throughput
- Storage
- Distribution
- Sorting
- Searching
- Documenting Assumptions, Constraints, Data Quality

Creating Level 3 Data --

- Choice of Nested Grid System and Associated Software
- Binning Algorithm (Continuous- and Discrete-Valued Quantities)
- Measures of Certainty for Comparisons Among Level 3 Products

Sensitivity Studies --

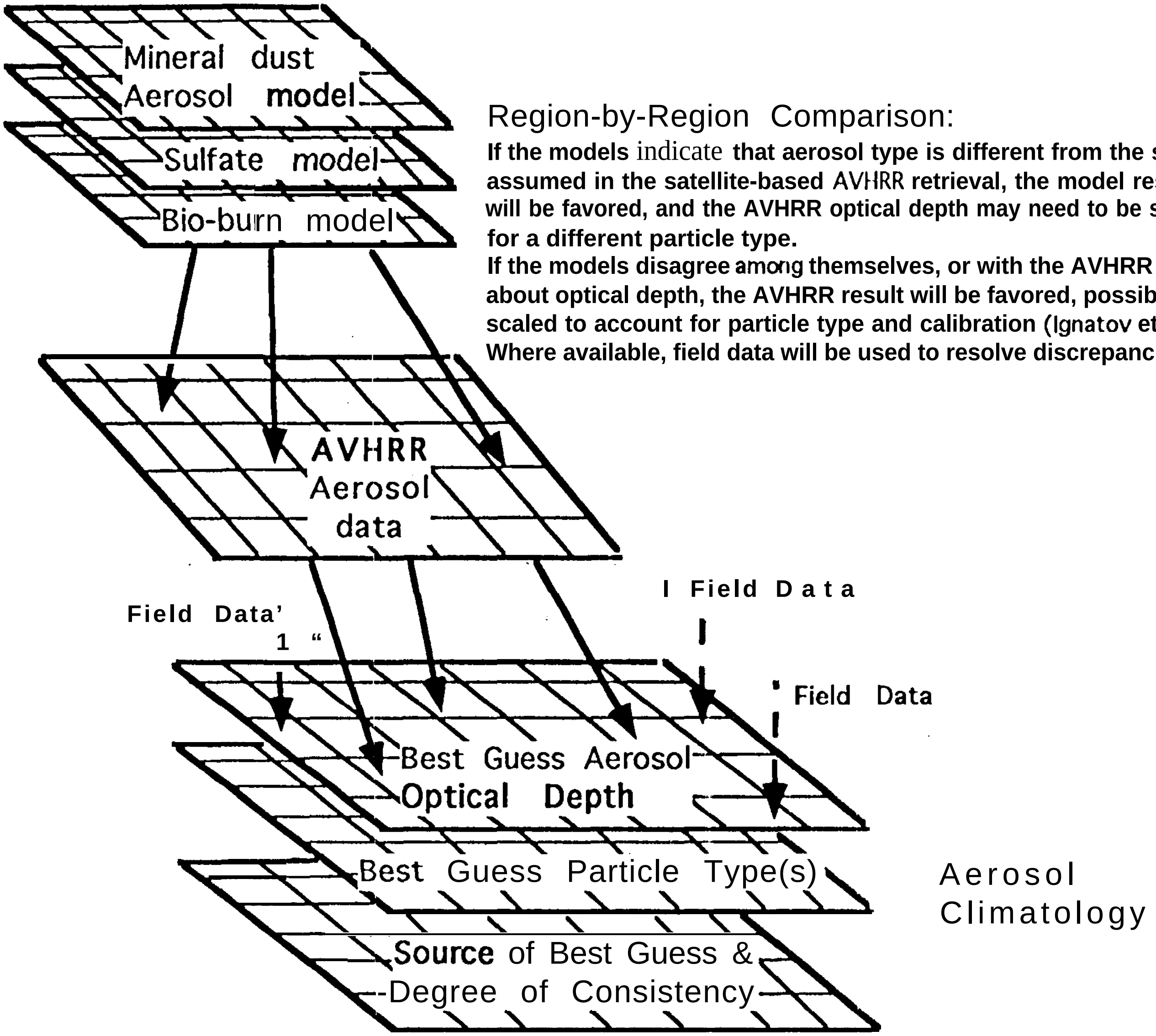
- Choice of Metrics to Distinguish Cases
- Strategy for Running Cases in Multi-Dimensional Space
- Data Visualization Techniques for Analyzing Results

Climatologies --

- Approach to Combining Model-Based and Observational Constraints
- Approach to Applying "Climatological Constraints" in the Retrieval

Studying the Observations --

- Summarizing Trends
- Identifying and Characterizing Exceptions (surprises)


Mineral dust Aerosol model
Sulfate model
Bio-burn model
AVHRR Aerosol data
I Field Data
Field Data’ 1 “
Field Data
Best Guess Aerosol Optical Depth
Best Guess Particle Type(s)
Source of Best Guess & Degree of Consistency
Aerosol Climatology

Region-by-Region Comparison:

If the models indicate **that aerosol type is different from the sulfate assumed in the satellite-based AVHRR retrieval, the model results will be favored, and the AVHRR optical depth may need to be scaled for a different particle type.**

If the models disagree among themselves, or with the AVHRR data, about optical depth, the AVHRR result will be favored, possibly scaled to account for particle type and calibration (Ignatov et al). Where available, field data will be used to resolve discrepancies.

Figure 1. Application of Constraints for Aerosol Climatology