UNCERTAINTY AND ERROR FROM THE STANDPOINT OF EARTH SCIENCE INFORMATICS

1. INTRODUCTION

This brief note is intended to be one of the starting points for a conversation about a Chapman conference on errors and uncertainty in Earth science data – and, particularly, about the role of Earth science informatics in mediating between scientists and data end-users. To keep this initial presentation simple, the core content consists of three figures:

- A "spatio-temporal sampling cube" that may help users identify the kind of error summary they are interested in
- A Data Flow Diagram that shows the relationship between collections of files, the processes that produce these files, and some of the ancillary data that can contribute errors
- An illustration of the calibration error that could arise because contamination on an instrument's optical train lowers the transmission of light from the Earth to the sensor

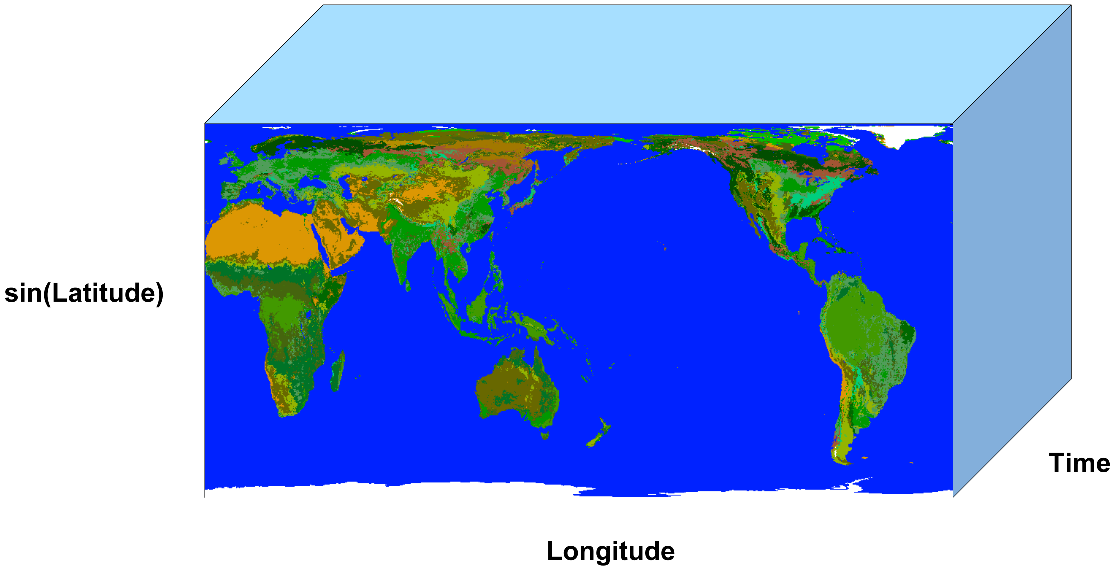

Figure 1. Sampling Cube. *This cube offers a visualization of the horizontal and temporal sampling within which Earth science data are found.*

2. SPATIO-TEMPORAL SAMPLING

All Earth science data takes observations from a four-dimensional space whose coordinates are λ, longitude, Φ, latitude, z, altitude (or a measure of depth within the ocean), and t, time. It is often convenient to collapse the representation down to a cube such as the one shown in fig. 1. Because the Earth is approximately spheroidal, the front surface of this cube cannot represent both spatial outlines and the area of features such as continents accurately. Fig. 1 is an equal-area presentation, in which the left side of the map starts at a longitude of 18° W, thereby keeping the Sahara as a connected region.

The colors on the front surface of the cube correspond to the IGBP types. For example, the white areas are permanently covered by snow and ice, whereas the orange colored areas represent desert. The representation corresponds to 512 bins in sin(Φ) and 1025 bins in λ. This means that near the equator, each pixel covers an area about 40 km by 40 km.

Earth science data come from instruments that have a variety of sampling patterns when they are visualized with this cube as a three-dimensional framework. For example, a single climate network instrument on the Earth's surface would have a spatio-temporal sampling structure like a "needle" inserted into the cube perpendicular to the front map. One end of the "needle" would start when the station started operating; the other end would be located at the date and time at which the station stopped. Sampling from a geostationary satellite would be like a set of flat saucers within a cylinder covering about one-fifth of the map. The saucers would be spaced apart in time by about one-half hour – and might be tilted with respect to the time axis if the scan lines progress from north to south (as in U.S. satellites) or south to north (as in European ones). Low Earth Orbiting satellites usually have some form of "sampling ribbon" as their orbits progress around the Earth.

What is important to us in dealing with errors and uncertainties is that some data users are interested in quantifications for parameters that cover the entire globe. Other users are interested in statistics that describe fields coming from a single ecosystem, which would be represented by an area on the map that has a single IGBP type. Obviously a single ecosystem can be irregular in shape. Also, ecosystems of similar IGBP type can have quite different areas. Climate data users may emphasize variations along the time axis, rather than spatial areas. A critical aspect of Earth science informatics will be to provide the appropriate type of error or uncertainty metric for each kind of data user.

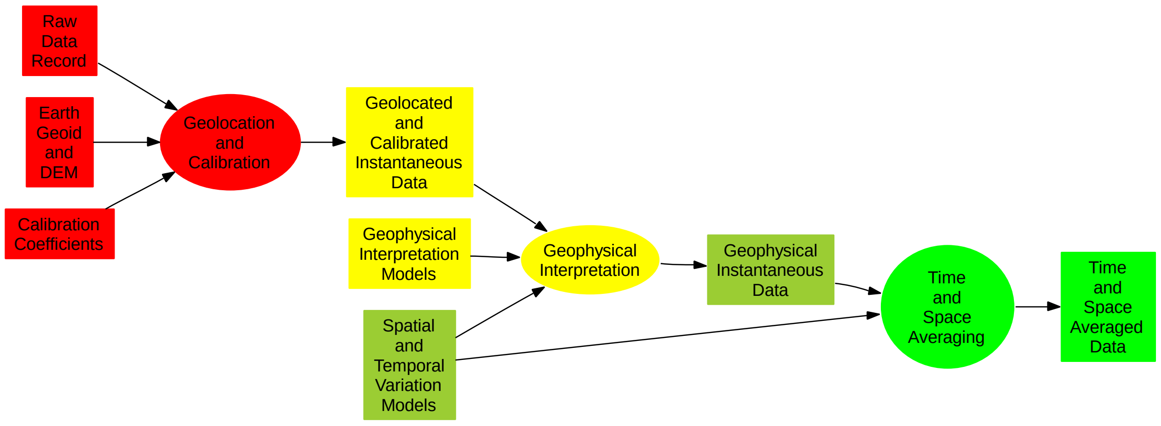
Raw Data Record
Earth Geoid and DEM
Calibration Coefficients
Geolocation and Calibration
Geolocated and Calibrated Instantaneous Data
Geophysical Interpretation Models
Spatial and Temporal Variation Models
Geophysical Interpretation
Geophysical Instantaneous Data
Time and Space Averaging
Time and Space Averaged Data

Figure 2. Generic Data Flow Diagram. *This diagram shows the relationship between various kinds of data and the processes that produce them.*

3. PROVENANCE AND ERROR STATEMENTS

Earth science data has a variety of instruments for measurement and a variety of processes that convert the raw observations into data that is useful for a particular community. A major source of difficulty in dealing with error and uncertainty assessments lies in the variety of data types and data production paradigms. However, the preponderance of Earth science data is in discrete files that are produced in a highlly automated fashion by various scientific and operational groups.

Figure 2 illustrates a generic kind of data production, based primarily on experience with satellite data. The Raw Data from the instruments appears in the upper left of this figure in the red box that indicates a kind of data file. The red ellipse into which the Raw Data enters identifies the location (λ, and Φ) and converts the Raw Data into Calibrated Data. In order to perform this work, the Geolocation process needs an Earth Geoid and a Digital Elevation Map (DEM). The Calibration process needs Calibration Coefficients. The Data Flow Diagram (DFD) assumes that these kinds of ancillary data are also found in files.

The next stage in production is Geophysical Interpretation, in which the Calibrated Data is transformed into quantities other than the ones in the original data stream. Often, this stage in production uses models of the processes (such as radiative transfer models), as well as models of the spatial and temporal structure of the underlying fields. The third stage in this data production model takes the instanteous geophysical quantities and averages them over time and space. Thus, this process might take a large volume of high-resolution spatial measurements at very short time intervals and produce a monthly average of measurements at the resolution of the map in fig. 1.

The rectangles in fig. 2 represent collections of files. Figure 2 not only shows the topology of data production, it also shows the dependence of the data in each collection of files upon both the algorithms that have produced the files and the parameters those algorithms use. Clearly, the error properties of the data in the files depend upon the production provenance of this topology. The Geolocated and Calibrated Instantaneous Data does not depend upon the Geophysical Interpretation Models or on the Spatial and Temporal Variation Models. However, the Time and Space Averaged Data clearly might have errors that depend on all of the algorithms and parameter files shown in fig. 2.

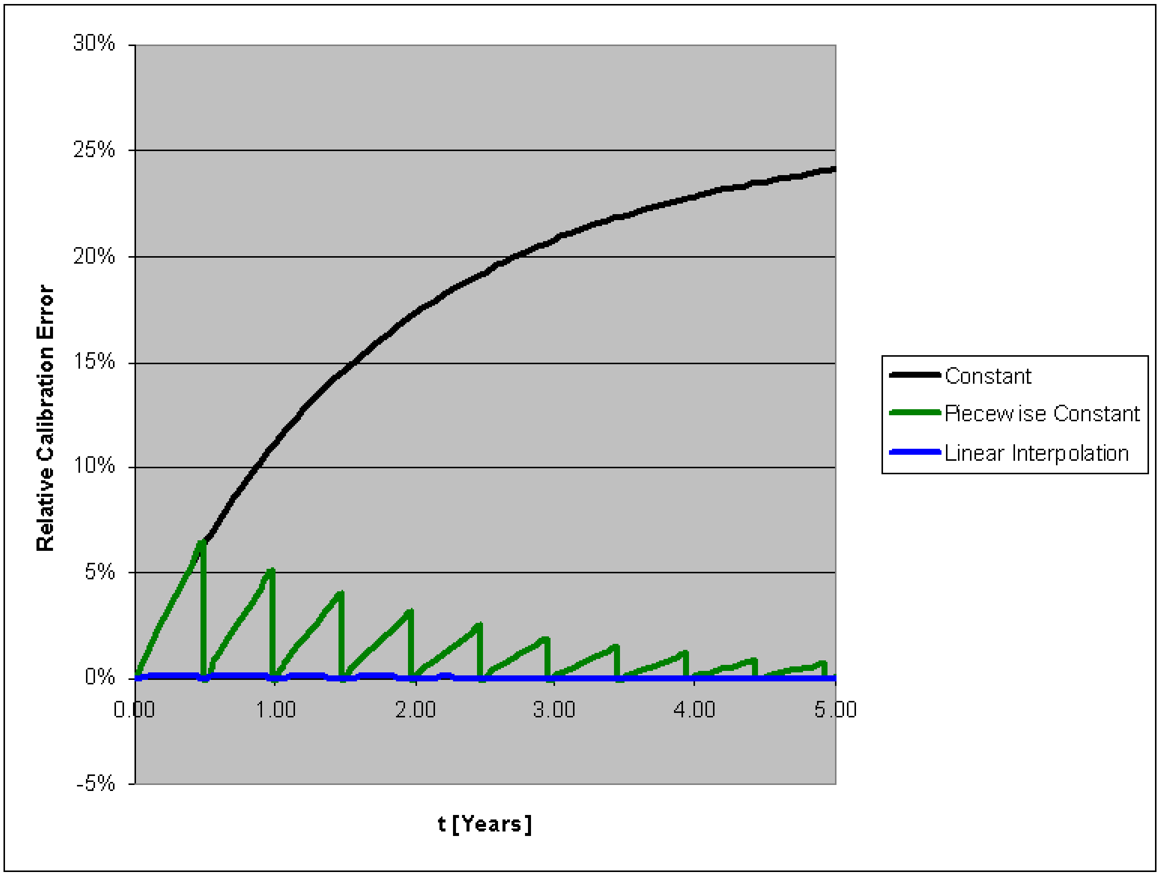

Figure 3. Hypothetical Relative Error in Instantaneous Calibrated Shortwave Radiances with Different Validation and Data Production Algorithms. *This figure shows the relative error in calibrated radiances depending on whether the data producer kept the calibration constant at the pre-operational value, updated the calibration every six months, or performed a linear interpolation between post-operational validation values.*

4. ILLUSTRATION OF CALIBRATION ERRORS

Figure 3 illustrates a time series of relative errors in calibrated radiances from a model of an instrument that has received contamination that darkens over time. The model is a generic one taken from experience with a number of Earth observing instruments, including the VIIRS instrument on the TRMM spacecraft, the MODIS instruments on Terra and Aqua, as well as instruments flown by the Earth Radiation Budget Experiment and by the investigation of Clouds and the Earth's Radiant Energy System (CERES). The important point is that the optical train of the instrument lets through less light as a function of time – the decrease is roughly an exponential decline to an asymptotic value with a time constant of about nine months.

With this kind of instrument model, it is possible to know exactly how the instrument would have behaved without this degradation. Thus, fig. 3 shows the relative error of the calibrated radiances as a function of time for three different approaches to discovering and accounting for the change in the Calibration process. One approach would be to ignore the change and simply keep the Calibration Coefficient constant. In this case, the relative error in fig. 3 absorbs the full error and grows to an unacceptable level after several years. A second approach would be to have a ground site that does observations capable of estimating the radiances seen by the instrument accurately enough to provide a new calibration every six months. If the science team adjusts their calibration every six months, then the relative error has the sawtooth pattern in the figure. A third approach is for the team to interpolate linearly between the vicarious calibration experiments. The team has to wait six months to be able to produce the linear calibrations, but the resulting errors are much more tolerable.

This example is useful in illustrating two important features of real observations:

- Data may have errors that do *NOT* conform to the typical notion that errors are well described by the sum of random errors with stationary statistics and biases with time independent values
- Real statistical distributions of errors – even at a given time – do not necessarily have a normal or Gaussian distribution

5. DATA QUALITY: A CHALLENGE TO EARTH SCIENCE INFORMATICS

The controversy over global warming illustrates the key role uncertainty and error now play in the Earth sciences. Furthermore, we face the substantial challenge of preserving and extending the information contained in existing and future data. Accordingly, practioners of Earth science informatics need to develop new ways of

- Recording existing and new information about error sources – including the context that creates this information (such as the validation experiments suggested in the previous section)
- Organizing the recorded infromation so that error statements and the algorithms that create them are useful across the full spectrum of data users, from members of the general public simply interested in an interesting image to scientific scholars or policy makers who need to assess the certainty of the information
- Ensuring that the recorded information is reliable and has integrity

One strategy for dealing with these requirements is to provide a “graded” disclosure that accommodates data user capabilities. For example, an archive interface might ask a user to choose the level of error and methodology disclosure:

1. No disclosure
2. Disclosure of a standard upper and lower bound, such as a “one-sigma” lower and upper bound or a “three-sigma” pair of bounds
3. Disclosure of a range selected based on the probability of the value not lying outside the bound
4. Disclosure of an error budget for a data product (meaning a collection of files) that would show the quantitative sources and impacts of various error sources
5. An interactive environment that would allow a data scholar to explore the influence of different assumptions on the stated errors

It is particularly important to ensure that statements of data quality and error uncertainty distributions are useful to the end users of data. Members of the user community who do not need a sophisticated quantitative error statement are easy to satisfy. Members of user communities who must make risk-informed decisions can be (and should be) more demanding. At the same time, data providers also need to recognize that end users control the requirements that define the usefulness of error statements. For examples, insurance underwriters and market hedge fund managers can be quite sophisticated in their understanding of probabilistic error or uncertainty distributions.

Error assessments and statements are a critical component of scientific understanding. They are also likely to be highly complex on two grounds:

- the algorithms that process data are complex, as is the dependency of errors on the detailed logic of these algorithms and their input parameters

- the processes that lie in the chain of provenance may have complex procedural elements, particularly when the data producers need to describe the details of the instrument calibration and the complex operations involved in data validation

Furthermore, the problem space is high and has high dimensionality. For example, there are at least three kinds of data quality metrics that are particularly important in climate data:

- Quantification of trends, particularly non-linear ones such as those that we might expect from climate change
- Quantification of changes in extreme value statistics, such as the probability of 100-year or 500-year floods
- Reliable interpretation of cause-and-effect in the presence of statistical correlations of anomalies
- Incorporating non-Gaussian statistical sources and algorithms into error assessments

All of these considerations create a rather challenging situation for Earth science informatics. First, we need to develop a common framework for dealing with production provenance and error propagation. This means that we need to develop tools for organizing the presentation of error sources. Second, we need to provide tools for making the error sources and error propagation understandable over the long term. This suggests that we will need to devote time and energy to creating tools that assist in making large, complex pieces of code and the production environment that used this code understandable. It also means organizing the documents and data sources in a way that allows scientific scholarship to succeed. Third, we need to provide tools for marshalling data sources and algorithms into usable information that increases the efficiency of the Earth science community.