NOAA's Archive, Access, and Assessment of Satellite Data for Climate Applications

John J. Bates, Chief Remote Sensing Applications Division
NOAA/NESDIS National Climatic Data Center, Asheville, North Carolina, United States of America

Abstract

NOAA's National Climatic Data Center (NCDC) is the U.S. Agency records center for weather and climate data. The mission of NCDC is to provide archival and access to the Nation's resource of global climate and weather related data and information, and assess and monitor climate variation and change. As such, NCDC archives all of NOAA's polar orbiting data, selected instruments from the DMSP satellites, and all data from NOAA's geostationary satellites. Stewardship efforts are underway to 'rescue' some of these data sets to ensure their quality. These data can be accessed through the Comprehensive Large-Array data Stewardship System (CLASS) which is under continuing development for the data center. Finally, NOAA's Scientific Data Stewardship Project has been established to produce long-term climate data records to apply to the NCDC mission of assessing and monitoring climate variation and change.

The modern era of U.S. climate records and observations began in 1950 with the Federal Records Act which authorized the General Services Administration to delegate authority for storage, processing and servicing of weather and climate records to a center operated by the head of a Federal Agency. Thus, in 1951, the Weather Bureau Records Center was established per agreement with the General Services Administrator. This agreement established a joint Weather Bureau-Air Force-Navy weather records center in Asheville, North Carolina now known as the National Climatic Data Center (NCDC). NCDC joined the National Oceanic and Atmospheric Administration (NOAA) when several different components related to atmospheres and oceans were merged in 1970. At that time, the mandate of the Data Center was expanded to include environmental satellite data.

Archive – The Comprehensive Large-Array Data Stewardship System

Currently, different categories of data are archived using different mechanisms due to the volume of data. Very large volume data, also referred to as large-array data, are archived using the Comprehensive Large-Array data Stewardship System (CLASS, formerly the satellite active archive). CLASS provides a set of information technology software and hardware productivity tools to contribute effectively to a NOAA-wide system for data ingest, archive, and easy and convenient access by our customers to new and historical national and global original data and synthesized products. It provides important pieces of the infrastructure for increased access to climate and environmental information required by all sectors of the economy and society, including government and academic institutions, major corporations, small businesses, and individual users.

A core set of large array environmental data campaigns are currently, or will soon, be available through CLASS. These campaigns generate huge data volumes for which CLASS will provide information technology hardware and data management tools consistent with an extensible, interoperable, cost effective architecture. CLASS modules address the following eight large-array campaigns:

- NOAA and Department of Defense Polar-orbiting Operational Environmental Satellites (POES) and Defense Meteorological Satellite Program (DMSP)
- NOAA Geostationary-orbiting Operational Environmental Satellites (GOES)
- National Aeronautics and Space Administration (NASA) Earth Observing System (EOS) Moderate-resolution Imaging Spectrometer (MODIS)
- National Polar-orbiting Operational Environmental Satellite System (NPOESS)
- NPOESS Preparatory Program (NPP)
- EUMETSAT Meteorological Operational Satellite (MetOp) Program
- NOAA NEXT generation weather RADAR (NEXRAD) Program and future dual polarized and phased-array radars.
- National Centers for Environmental Prediction Model Datasets, including Reanalysis Products

The POES data sets include all data Earth sensor environmental data, and selected products, from the operational era beginning in 1978 with NOAA-6. DMSP data sets only currently include the SSM/x (where x=I, T1, T2, and IS) from 1997 to date. An effort is under way to provide access to the earlier SSM/x data. A major effort to rescue GOES full resolution imager data, underway for the last 5 years, has just been completed. Imager data, and selected products and sounder data, are now available for all GOES satellites.

To enable CLASS information technology investments and data management tools to fully function within the NCDC in the longer term, a baseline notional architecture has been developed. The target architecture for CLASS (Figure 1) consists of three distinct functional layers: 1) a core CLASS tools layer, 2) a data stewardship layer, and 3) a customer access layer. Within the CLASS tools layer, there are provisions for CLASS to: a) provide the IT hardware and data management software tools to archive original data and synthesized products for the large-array data campaigns, b) ingest and provide information technology hardware and software for metadata storage and access, and c) to enable a NOAA-wide enterprise solution for ingest, archive and access of data from NOAA's centers of data and non-large array data from the NOAA National Data Centers. The access and assessment/stewardship layers are described detail below.

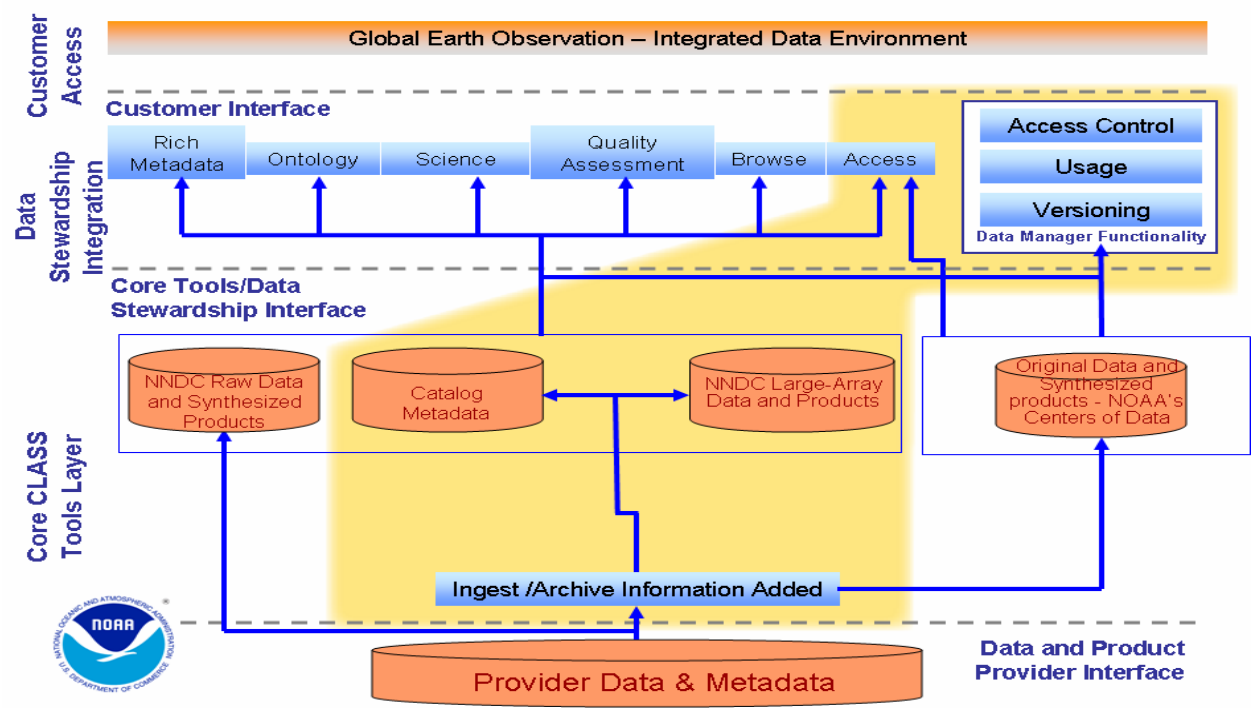


Figure 1. NOAA's notional architecture for data management including CLASS and related services for an integrated data environment.

Access – Current Status and Plans for an Integrated Data Environment

The CLASS provides web access to NOAA satellite data sets including search by satellite, instrument, and location and time. Up to 100 data sets can be ordered in a session and data orders are written to an ftp area for user pickup. Subscription services are also available if a user wants to regularly receive data for a given instrument and area. The CLASS web site can be reached at: WWW.CLASS.NOAA.GOV . Metadata related to a given satellite are also accessible through the CLASS web site by selecting the live links on the data description pages. For large data orders, and for additional user services, users can contact NCDC.SATORDER@NOAA.GOV .

Increasingly, due to the complexity of many environmental problems and to answer questions addressing contemporary societal needs users require data from a variety of different sources. To respond to these needs, NOAA must be able to successfully integrate information from all of its data management systems, including CLASS, and exchange data with partners in the national US-Global Earth Observation System (US-GEO) and the international Global Earth Observation System of Systems (GEOSS). With its Global Earth Observation Integrated Data Environment (GEO-IDE) as its contribution to US-GEO, NOAA will be able to provide easier and more cost-effective access to all of its data and information.

NOAA's GEO-IDE is envisioned as a "system of systems" – a framework that provides effective and efficient integration of NOAA's many quasi-independent systems, which individually address diverse mandates in areas of resource management, weather forecasting, safe navigation, disaster response, and coastal mapping among others. The NOAA GEO-IDE will make NOAA products available in multiple formats and communication protocols, utilizing current information technology standards, where they are mature, and best practices, where accepted standards are still evolving. NOAA data and products will be described by comprehensive metadata that conforms to national and international standards. NOAA observing systems and collection, assimilation, quality control and modeling

centers will provide their data and metadata in accordance with established NOAA GEO-IDE standards.

NOAA GEO-IDE will strive to take full advantage of the opportunities presented by internet technology to make access to environmental data and information as easy and effective as access to digital documents over the Web is today. It will also improve efficiency and reduce costs by bridging the barriers between existing, independent "stove pipe" systems and integrating the data management activities of all NOAA programs. It will do this through a federated approach, where the individual components retain a measure of responsibility and authority within the context of an overarching systematic set of goals, principles and objectives.

GEO-IDE will fundamentally depend upon standards and it is essential that these be thorough, documented, and supported standards with demonstrated benefits. To ensure these standards are embraced and accepted across NOAA, an open and inclusive "standards process" for nominating, evaluating, and implementing NOAA GEO-IDE standards is proposed. The standards process will define what standards are adopted, when they become effective, and how the organization will build up to and support the implementation of those standards.

GEO-IDE aims to retain existing systems as much as possible while building a software infrastructure that links these systems together. This software infrastructure, called a Service Oriented Architecture, is a style of systems design based on using loosely coupled connections among independent programs to create scalable, extensible, interoperable, reliable, and secure systems. Service-based architectures have been proven to solve interoperability problems including integrating systems developed in various programming languages, running on different computing environments and developed by autonomous groups at different times. They make it practical to adapt and connect existing systems quickly for accomplishing new tasks and to benefit from highly evolved and still useful "legacy" applications.

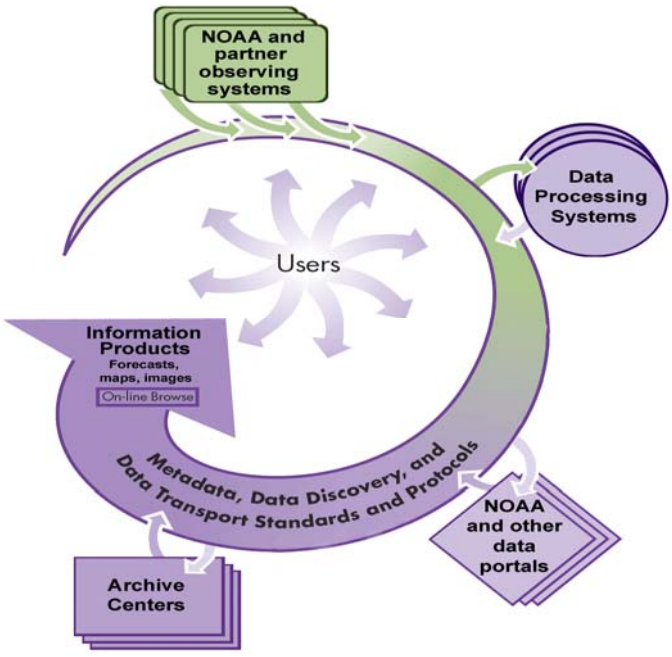


Figure 2. Conceptual view of the GEO-IDE system.

Assess – Scientific Data Stewardship

Scientific Data Stewardship (SDS) is a subset of data management, emphasizing data quality and quantification of uncertainty required for long-term data assessment. These features make NOAA's SDS program particularly well-suited to production of Climate Data Records. A recent National Research Council report (2004) defines a CDR as "a time series of measurements of sufficient length, consistency, and continuity to determine climate variability and change."

That report also differentiates fundamental climate data records (FCDRs), which are calibrated and quality-controlled sensor data that have been improved over time, and thematic climate data records (TCDRs), which are geophysical variables derived from the FCDRs, such as sea surface temperature and cloud fraction. In most senses, it is appropriate to associate FCDRs with Level 1 data (i.e. geolocated and calibrated) and TCDRs with higher level data. Thus, TCDRs may contain instantaneous geophysical fields or spatially gridded and temporally averaged fields. The instantaneous TCDRs usually have the full temporal and spatial resolution of the original data and are thus often described as Level 2 data. The gridded and averaged TCDRs would often be described as Level 3 data.

Characteristics of CDRs

There are several characteristics that distinguish CDRs from other kinds of data products:

- Record Length – Because CDRs are intended to provide reliable data for climate research, it is important that the data record be as long as possible. This characteristic implies that a CDR will need to include data from several different sources, such as similar instruments on a succession of spacecraft.
- Error Structure Homogeneity – Climate investigations usually seek to detect and measure small signals embedded in a highly variable record. Instrument or algorithm artifacts add uncertainty and can substantially reduce the usefulness of data sets for investigations that seek to measure these small climate signals. To achieve error structure homogeneity, a CDR data provider will usually need to carefully check (and often reprocess) data for consistency.
- Assurance of Data Provenance – Because of the importance attached to being able to draw reliable conclusions regarding climate change from these data, CDRs need to ensure that users can retrieve and understand the heritage and chain of custody of the data they want to use.
- Ability to Retain Usefulness over Long Time Periods – Because CDRs are dealing with long data records with high attention to data quality, they may take longer to produce than weather forecast data as they need to be dependable over very long periods of time. This characteristic means that CDRs need documentation of their context and assurance of long-term survivability in the presence of huge changes in their Information Technology (IT) environment and in user access needs.

These characteristics of CDRs place rather stringent requirements on the production configuration management. Figure 3 shows a generic Data Flow Diagram for a family of CDRs. Level 0 data appear on the left of this figure. The processes of calibrating and geolocating the raw data are those that produce the Level 1 data that constitute a FCDR. The instantaneous Level 2 data form one kind of TCDR; the gridded Level 3 data form a second.

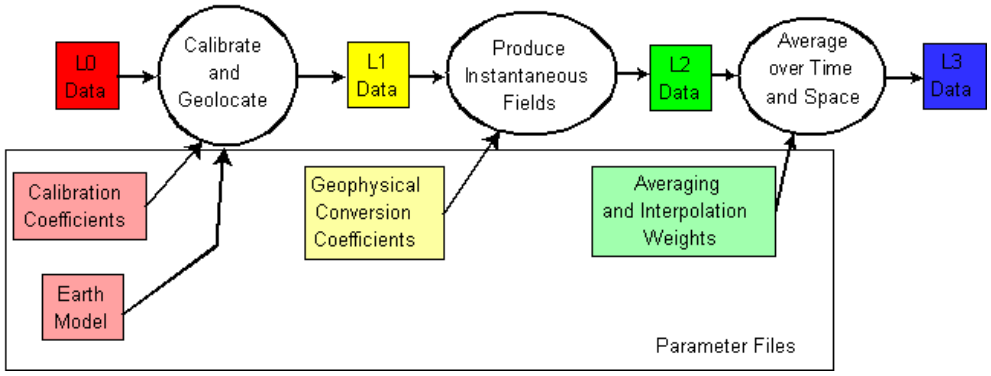


Figure 3. Generic Data Flow Diagram for Climate Data Records.

In addition to illustrating the relationship between FCDRs and TCDRs, figure 3 illustrates how errors propagate between data products and what needs to be tracked in order to maintain a record of data provenance. If we consider the FCDR that appears as Level 1 (or L1) data, for example, it is clear that there are four potential sources of error:

- Level 0 data, which may have gaps or errors introduced by the instrument telemetry or ground receiving systems
- Source code (and algorithms) for the calibration and geolocation process
- Calibration coefficients and Earth model used in the data reduction
- System configuration (computers, operating system, compilers, production scripts) used to actually run the jobs that produce the data

Errors any of these potential sources will propagate into the FCDR. Thus, quality assurance in forming the FCDRs is of the utmost importance.

The same emphasis on production quality also applies for the higher level data products. Figure 1 makes clear that errors or gaps in Level 1 data propagate into Level 2 data. This figure also shows that errors in the Geophysical Conversion Coefficients or in the Level 2 source code (or algorithms) will affect the instantaneous TCDRs. The system configuration may also contribute to error propagation. For gridded and time averaged TCDRs, this figure illustrates the same connections.

Figure 3 illustrates the difficulty of maintaining version control. To produce a single version of the Level 1 data, the source code and coefficients that convert the Level 0 data to Level 1 must be kept constant for the time span of the data in the version. To produce a single version of the Level 2 data, there needs to be a homogeneous version of Level 1 data, as well as a single version of the Level 1 to Level 2 source code and a single version of the Geophysical Conversion Coefficients. A homogeneous Level 3 TCDR version adds additional requirements on the homogeneity of the algorithms and coefficients for the time and space averaging process.

In practical terms, then, the production of CDRs must adhere to the highest standards of configuration management. To avoid introducing inhomogeneities into CDRs, the production system must ensure that the input data sources are as homogeneous as possible and that source code and input coefficient changes are as small as possible, but still support generation of homogeneous time series. As our field matures, it appears probable that interdisciplinary CDRs will become more common. Such data records are particularly challenging for climate research because they require maintenance of homogeneous input versions. Thus, attempts to produce CDRs with input data sets whose source

code changed in an irregular fashion are likely to be less valuable than CDRs whose input data has homogeneous versions.

A Model for Maturity of Data Sets

The CDR characteristics just mentioned seem daunting. Furthermore, it appears necessary to allow different degrees of stringency for different kinds of data – some applications are more demanding than others. It is useful to draw on the experience of the software development community, in which a Capability Maturity Model, Chrissis (2003) has proven useful. In this model, maturity is quantified according to the degree of reproducibility a team exhibits in estimating and producing a software product. At a low level of maturity, the results are scattered; the team does not have the ability to estimate and produce results on time and within budget. At a high level of maturity, the process becomes consistent.

Adopting this suggestion, the SDS program is pursuing an approach to quantifying the maturity of a data set based on three axes:

- Scientific Maturity – a set of measures of the scientific quality of a data product
- Preservation Maturity – a set of measures of the long-term sustainability of the information in a data set
- Societal Benefit and Impact – a set of metrics that assess the potential value of a data set

The intent of these axes of maturity is to provide a systematic measurement approach to ensuring that data sets are valuable and ready for long-term preservation.

In quantifying the level of maturity for each axis, we have found it helpful to develop a hierarchical breakdown of the appropriate attributes. At the most detailed level, each attribute can be ranked non-dimensionally, avoiding the troubles that arise with different units for different attributes. The lack of units makes it easier to quantify maturity and rank the attributes of various properties.

Scientific Maturity attributes break down into the following categories:

- Physical understanding of the measurement process, including
  - Measurement of spectral sensitivity
  - Measurement of Point Spread Function (PSF) (spatial sensitivity)
  - Pre-launch calibration
- Capability to detect important changes in calibration
  - Changes in spectral sensitivity
  - Changes in PSF
  - Changes in calibration parameters
- Public accessibility of data production processes
  - Documentation of data flow diagram
  - Algorithm Theoretical Basis Documents
  - Documentation of data editing algorithms
  - Availability of source code for modification

- Rigorous validation
  - Documentation of a validation plan
  - Documentation of validation data and results
  - Understandable uncertainty analysis

Preservation Maturity attributes break down into the following categories:

- Low Total Cost of Operation, including automated operations
- Highly Reliable Operations, including
  - Reliable and automated configuration management
  - Balanced approach to redundancy and dispersed site storage
  - Robust and graceful exception handling
- Evolvability
  - Documentation of designer intent and design evolution
  - Modularity of architecture while being traceable to user needs
  - Formalization of Operations and of operational procedure evolution
  - Outside participation in design, development, and evolution
- Integrity Maintenance
  - Intellectual Property Rights Considered in design
  - Permanent file and data naming registration
  - Ability to track provenance
  - Transactional Basis for system operation and auditing

Societal Benefit and Impact attributes are more difficult to quantify. This difficulty arises because of the diversity of the communities that produce and use CDRs. Each user group has a distinctive dialect, a distinctive set of data world views (that include data structures and formats that are "easiest" or "most natural" for a particular kind of data), and a distinctive set of user "customs". The latter include data search strategies and use of particular visualization tools.

A second aspect of the difficulty of quantifying societal benefit and CDR impact lies in the fact that there are at least two measures of data value.

One measure of impact or benefit arises from the ability of data to confirm or negate cause-and-effect hypotheses. In other words, data in a CDR may be sufficiently accurate to confirm that a particular cause-and-effect relationship holds – or it may not be that accurate. Thus, this approach to valuing data is perhaps best quantified in terms of the uncertainty in the data or in terms of its ability to answer important scientific questions.

A second measure arises from the flow of economic impact a CDR can produce. While our attention is often drawn to the ability of data to assist emergency first responders, we should also keep in mind that CDRs may help quantify the probability of long-term extreme conditions and can assist with planning for mitigating the effect of those conditions. With this kind of metric, data value may be measured by cost avoidance for particular conditions. Thus, to the extent that we could quantify

savings associated with better water or energy management, CDRs might be evaluated in terms of the net present value of future cost savings for expenditures.

The Intergovernmental Panel on Climate Change has noted that scientific value and monetary value are not measurable on the same scale – and therefore we need to develop methods that appropriately balance the valuations that these two different systems use. It seems useful to think of the monetary valuation as emphasizing near-term data uses, while the scientific valuation emphasizes the long-term use.

Conclusions

Recent progress within NOAA's NCDC has resulted in greatly improved archival, access, and assessment of NOAA's environmental satellite data. Work is underway for the expansion of these services and products to include support for a service oriented architecture and a suite of climate data record products.

References

Chrissis, M. B., Konrad, M., and Shrum, S., 2003: *CMMI: Guidelines for Process Integration and Product Improvement*, Addison-Wesley, Reading, MA.

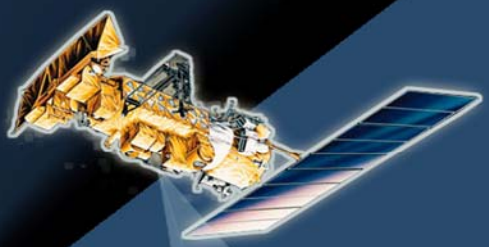
National Research Council, 2004: *Climate Data Records from Environmental Satellites*, National Research Council, Washington, DC.

INTERNATIONAL ATOVS WORKING GROUP

Sharing ideas, plans and techniques

to study the earth's weather

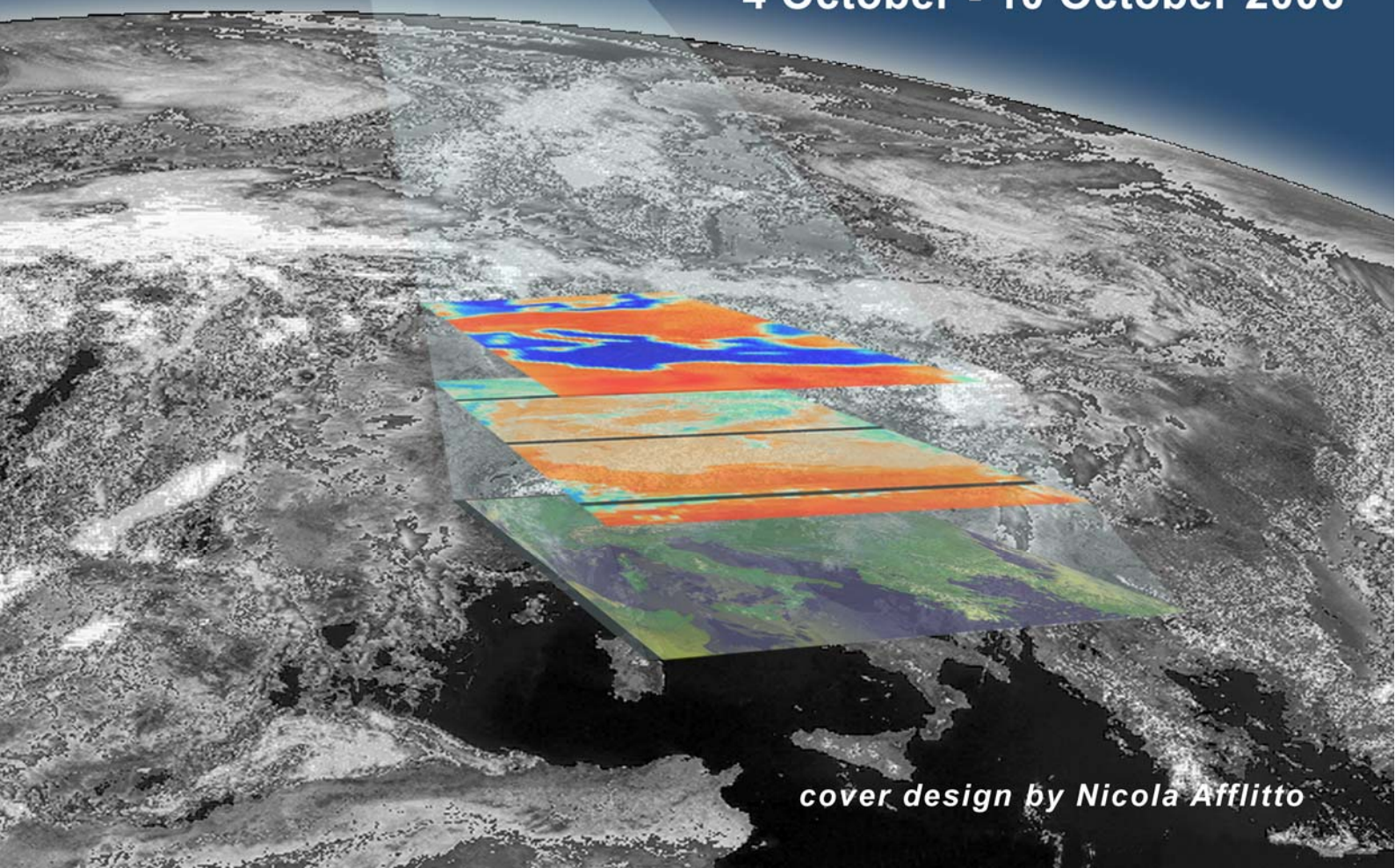
using space-based observations



Proceedings of the Fifteenth International TOVS Study Conference

Maratea, Italy

4 October - 10 October 2006



cover design by Nicola Afflitto