EOSDIS Core System: System Overview, Internet Access of Data and Information, Operator Tools, and System Benefits

Timothy Gubbels and Robert Plante
Raytheon Systems Company
Information Technology Systems Division
1616 McCormick Drive
Upper Marlboro, MD 20774
tgubbels@eos.hitc.com

ABSTRACT

EOSDIS Core System handles the ingest, archiving, production, and distribution of most of the data coming from NASA's Earth Observing System (EOS). The system is now operational, handling the archiving and distribution of Landsat-7 data, and soon the production and distribution of data from the Terra platform (projected launch 11/99). Although EOS data is primarily research-oriented and intended more for climatological research than for operational meteorology, key features of the data and information system offer the potential to enhance operational data management systems used for medium-range forecasting. The objective of this paper is to explain the EOSDIS Core System and its features, in order to stimulate ideas of those involved in medium-range forecasting, and those involved in developing or upgrading real-time operational meteorological systems.

INTRODUCTION

NASA's Earth Science Enterprise (ESE) program turns NASA's pioneering, space-based observing technology and scientific expertise to the study of our planet, and forms a key component of the US Global Change Research Program. This integrated, multi-disciplinary research program is driven by a set of science priorities that include land-cover and land-use change, seasonal-to-interannual climate variability and prediction, natural hazards research and applications, long-term climate variability and change, and atmospheric ozone research.

The Earth Observing System, a planned 15-year program of integrated earth observation, is the centerpiece program within NASA's Earth Science Enterprise. The EOS is composed of three distinct elements: a new constellation of earth observing satellites, a multidisciplinary science research program, and a data and information system known as the EOS Data and Information System (EOSDIS). The first spacecraft to be supported by EOSDIS Core System, Landsat-7, was launched in April of this year. The next, Terra (formerly EOS AM-1), will launch in November. EOSDIS will support all future EOS spacecraft including ACRIM, METEOR-3M, EOS PM-1, ADEOS-II, Jason-1, ICESat-1, and EOS CHEM-1 (see http://www.earth.nasa.gov).

EOSDIS is composed of three major parts: EOSDIS Data and Operations System (EDOS), EOSDIS Backbone Network (EBNET), and EOSDIS Core System (ECS). EDOS handles the reception of downlinked telemetry, production of the Level-0 data stream, and distribution of the data stream through EBNET, a high-performance backbone network, to distributed installations of ECS. ECS functions encompass spacecraft command and control; health and safety; and data ingest, archiving, processing, and distribution. ECS produces Level-1 and higher standard products from the Level-0 data stream. (Figure 1) Dedicated Science Investigator-led Processing Systems (SIPS) also participate in the production of standard products, which are distributed by ECS. The ECS installations are physically distributed throughout the US and are located within larger data centers known as Distributed Active Archive Centers or DAACs. The centers are:

- The EROS Data Center, specializing in land processes.
- The Goddard Space Flight Center, specializing in upper atmosphere, ocean color, and meteorology.
- Langley Research Center, specializing in radiation budget and upper atmosphere.
- National Snow and Ice Data Center, specializing in snow and ice.
- Alaska SAR Facility, specializing in synthetic aperture radar imagery.
- Oak Ridge National Laboratory (ORNL), specializing in biogeochemical dynamics.

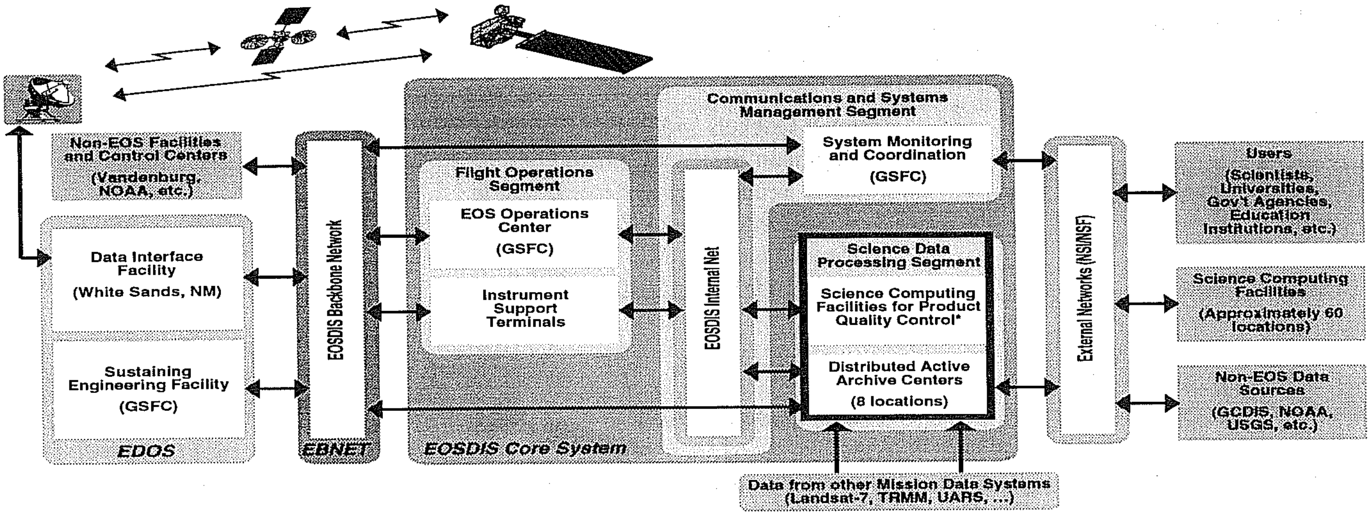

Figure 1. EOSDIS Architecture

ECS SYSTEM OVERVIEW

ECS is comprised of three major elements: ECS Mission Operations Segment (EMOS), Communications and Systems Management Segment (CSMS), and Science Data Processing Segment (SDPS). The SDPS resides at the DAACs and is the data management system at the heart of EOSDIS. These SDPS "nodes" vary dramatically in size and disciplinary focus, but the software infrastructure is nearly identical. The review of SDPS that follows may have value for those that develop meteorological operational systems and wish to replicate specific design features, operator tools, or client/distribution services. For those responsible for developing medium-range forecast products, the instrument and geophysical products supplied by the system may have specific value in the operational data stream.

SDPS handles data ingest and archiving, and science data production and distribution. It serves both as a "production line" to create standard products and a "digital library" to store and provide access to these products. SDPS consists of seven logical subsystems that intercommunicate. The subsystems work together to perform three basic functions: Ingest, Processing, and Distribution. (Figure 2)

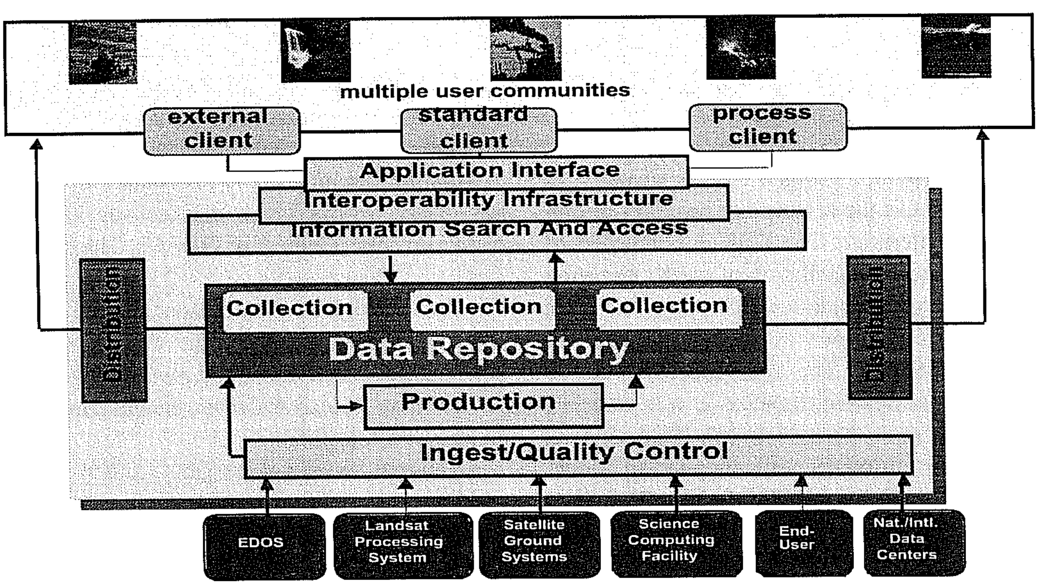

Figure 2. EOSDIS Core System Functional Architecture

• For Ingest, the data are received, verified, and stored in an archive that is managed by a data server. The searchable inventory database is updated. The event is registered by the system.

• For Processing, a pre-established processing plan with production rules is executed, often triggered by completion of an ingest event. The source data are retrieved from the data server and moved to the processing system, where a new standard product is created. The information products are then inserted into a data server and the searchable database is updated.

• For Distribution, an end user performing an inventory search interacts with a client that communicates with the archive to execute the search and return references to data. The user may use the references to retrieve the data item itself or share these references with colleagues. In addition to the search and order function, a user may place a subscription to receive notification of new data availability or to have new data automatically pushed to a specified location. Notification may be triggered by completion of an insert event.

SDPS today is capable of handling over 2 terabytes of data throughput per day using over 100 gigaflops of distributed processing power within 6 installations. Over the next two years, the archive will grow to exceed 2 petabytes with an intended persistence of 15 years or more. The system will handle more than 1,500 discrete data types and is scaled for more than 10,000 users.

The system is designed for a high degree of operational resiliency. Routine operations, including automated monitoring of system performance and usage, are highly automated. Robust failure recovery adds further resiliency to operational data management.

KEY ARCHITECTURAL FEATURES

In addition to handling a massive scale of data throughput and providing a large archive, the system has several key architectural features that distinguish it from other legacy systems and that offer potential for adding value to operational meteorology systems through design reuse. Each of these architectural features has evolved and matured through our implementation experience with the first round of EOS sensors, ETM+, MODIS, MISR, MOPITT, ASTER, and CERES.

The first key architectural feature, the ECS data model together with the HDF-EOS data standard, originated from a requirement to facilitate interdisciplinary collaboration. The ECS data model is rich and extensible, consisting of more than 280 defined core attributes and specific provision for large sets of product-specific attributes. The model is a result of an extensive earth science data modeling effort. Just as the ECS system is data quality focused, the data model includes extensive data quality documentation metadata and provisions for both calibration and validation documentation. The HDF-EOS standard data format is an extension of NCSA's self-describing Hierarchical Data Format (HDF). It allows the core and product-specific metadata to be conveyed, hides the structural metadata, and provides API-level access to a variety of services (e.g. read/write, inquiry, subset) on a variety of data structures (e.g. point, grid, swath). The advantage of HDF-EOS is that it accommodates substantial heterogeneity within earth science data and streamlines software development associated with the exploitation of that data. HDF-EOS can readily accommodate any form of meteorologic data. The key components of successful development of both the data model and the new data format included: 1) testing with a wide variety of data types; 2) fusing elements of emerging standards; 3) involving the community; 4) establishing rigid version control of both software and descriptive data.

Efficient algorithm integration is a second key architectural feature; it originated from the requirement to allow new research algorithms to be readily integrated and easily versioned. Satisfying this requirement necessitated the development of new tools and methods, as well as an integrated standard process for science software integration and test. Within EOSDIS, the developers of the instruments, known as the Instrument Teams, develop the algorithms associated with each standard data product. The teams also code the product generation software by encapsulating the algorithms within executable product generation software that is then integrated within ECS. ECS provides guidance to the teams in the form of coding standards and a software toolkit. Algorithm development is intended to occur at associated science computing facilities, with eventual insertion into ECS once the algorithms are stabilized. ECS includes a well-defined, standardized process that can be applied by the developers from the beginning of their algorithm development and which significantly reduces the difficulties of integration into the operational environment. The key components of smooth algorithm integration included: 1) precise coding standards and a mature toolkit; 2) a process, whereby integration of new algorithms and data types into the system coexists with normal system operations. To accommodate this, SDPS operates in multiple modes simultaneously; operational and test modes coexist.

Efficient data typing is a third key architectural feature; it originated both from the requirement to allow new data types to be readily defined to or removed from the system and to allow for the easy addition and dynamic provision of new services. Many algorithms are quite complex and involve large numbers of intermediate products. The total number of data types now exceeds 1,300 and is growing; these are needed in order to accommodate large numbers of standard products, intermediate products, and ancillary data needed for production. Efficient and rapid definition of data types required the development of a new tool called Metadataworks, which defines the metadata for a given datatype, which allows the metadata to be propagated around the system. (Figure 3) This tool serves to hide the underlying complexity of the data model from the user and allows for automated metadata checking and validation. Other tools allow the definition and easy modification of the services available upon a given datatype within the system; the information-rich data model allows for a rich array of potential services. New standard products and new services are advertised through the Global Change Master Directory (GCMD) as soon as they are available. Future plans for improving datatyping include a transition to XML.

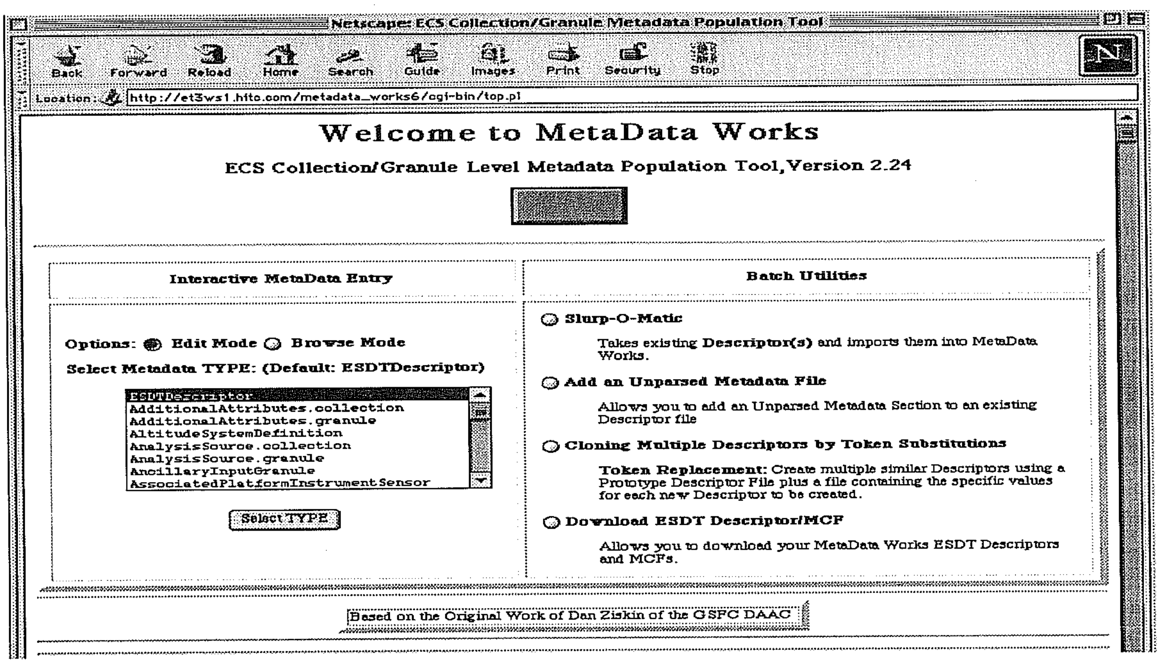

Figure 3. Metadataworks Graphical User Interface

TOOLS FOR CONTROL OF THE OPERATIONAL SYSTEM AND DATA FLOW

For each of the three major system functions, ingest, processing and distribution, an array of operator tools exists that facilitate the operational control of the function. A fourth function, archive management, supports all three of these system functions. Below we review the principal tools for the three major system functions in order to illustrate their value.

The ingest servers and Graphical User Interface (GUI) allow ECS to support many data providers using several different protocols for both automated ingest and manual ingest. For example, during automated ingest of data from EDOS, the Ingest GUI monitors a polling process on a secure FTP site where the data arrives by secure, fault-tolerant Kerberos FTP. The GUI allows the operator to monitor multiple simultaneous ingest events, including their degree of completion, their source, and unique identification. A history log is continually updated and can be examined in a variety of ways. For the media ingest, the GUI manages the process of multiple reads from a variety of media. For both the automated and manual cases, the Ingest Subsystem performs a quality assurance verification step transparently to the operator. In a scenario of an operational meteorological system, the tool could allow an operator to manage large numbers of discrete FTP processes on a highly heterogeneous group of inputs.

For processing, the workhorse of the operator tools is the Production Request Editor (PR Editor). The PR Editor is used to define processing jobs and to submit them to the planning system for scheduling. Up to 4,000 discrete jobs per day will have their execution managed in this manner. Specifically, the PR Editor can be used to select input data in several different ways, including orbit number, data collection time, or time of data insertion into the archive. The PR Editor associates the input data with the appropriate Product Generation Executive (PGE). The PGE contains a science algorithm and metadata extraction routines as integrated code elements. It also includes production rules that codify a multiinput request. The output of the PR Editor is a set of Data Production Requests, which are submitted for processing using a tool known as a Planning Workbench. In a scenario of an operational meteorological system, these tools could be used to manage the preprocessing of all of a model's inputs. Multiple production runs for multiple inputs could be submitted and managed easily, keeping the need for operator intervention to a minimum while maximizing operator productivity.

For distribution, a key tool is the Subscription Server Tool (SS Tool). The SS Tool is used by DAAC operators to place subscriptions into the system after receipt of a subscription request from an end user in the form of a telephone conversation or a formatted email. A subscription may specify simple notification or notification plus delivery. Notification allows users to be automatically alerted to the availability of newly produced data products. Subscription/delivery allows users to subscribe to a product and have it automatically sent to them once it has been produced. In order to service user requests, the operator can use the SS Tool to select the event to be subscribed to, commonly a production completion/archive insertion event for a newly produced granule of a desired standard product, and the destination of the notification (email address). The operator may also associate a notification with an automated delivery by FTP, by entering a destination FTP site, or delivery on media by a mail service.

Subscription services are innovative new system enhancements that have substantial value to users and have not been available at this scale within NASA data distribution systems of the past. In a scenario of an operational meteorological system and reuse of this feature, the tool could be used to manage the entire array of subscriptions placed by the user community for a data center. In some centers where the present system is manual and labor-intensive, significant efficiencies could be gained. In a scenario of utilizing EOS data within an operational medium-range model, once the value of a given standard product has been verified in the course of research and development, the placement of a subscription and association of automatic delivery would allow automated delivery on a routine basis. This could allow for reliable, automated delivery of EOS data into an operational medium-range forecast system every time a new product of interest became available.

INTERNET ACCESS OF DATA AND DATA DESCRIPTIONS

The ECS provides a variety of services for Internet access of data and online data descriptions. Most of the services are accessed via a general-purpose client known as the EOS Data Gateway (EDG). The EDG allows the user to discover new data collections within EOSDIS, review extensive on-line (guide) documentation that includes documentation of the algorithm used to generate a given standard product, the algorithm's theoretical basis, processing methodology, and sources of error. User registration through the EDG allows the creation of an account in which search results can be stored, allowing users to order, and that enables accounting and billing. Inventory search can be performed by graphically selecting a spatial extent using an interactive JAVA mapplet, then selecting a temporal extent. Other selectable attributes include a very large number of metadata elements such as desired parameter, data quality, sensor, and processing level. An on-line data dictionary can be reviewed to clarify the definitions of potentially searchable attributes. Once the search is completed and the search results have been received, visual summaries of the data, in the form of browse imagery, can be reviewed. All identified granules are uniquely associated with a persistent Universal Reference (UR), facilitating user understanding of the search results. Once a selection is made, an order can be placed. For some datasets within the system, additional services such as browse may be available. Users may choose to save customized settings of the client and share these with collaborators; search results and URs may be shared in the same manner. In the scenario of an operational meteorological system, the client would likely be used by researchers experimenting with new medium range models incorporating new inputs selected from the EOS standard product array. The online data documentation and the search and order functions would be used to locate, review, and retrieve data to be used in model improvement research.

ROLE AND VALUE OF THE ECS FOR OPERATIONAL MEDIUM-RANGE WEATHER APPLICATIONS

The EOSDIS Core System plays three roles of value to operational medium-range weather applications: 1) The production of data products that are potentially usable in medium-range operational weather forecasting by NOAA and other organizations; 2) Interface with the NASA Goddard Data Assimilation Office (DAO), providing the DAO with an array of lower level EOS standard products and distributing the DAO level 4 model outputs; 3) Technology leadership in the arena of operational satellite remote sensing data processing and management systems.

In the existing plans for the interface between ECS and NOAA, NOAA serves primarily as a data producer and provides a variety of meteorological products as ancillary inputs for a number of EOS standard products. NOAA is also working with the EOS Project Science Office to establish the capability of subscribing to the MODIS L1A data stream from ECS and processing this into higher-level products for operational meteorological applications. The current plans call for NOAA to receive the MODIS processing software for this higher-level processing.*

In the existing plans for the interface between ECS and the DAO, the DAO will serve as both a consumer of low-level EOS standard products and a producer of assimilated L4 data products to be distributed by ECS. The mission of the DAO is to produce the research-quality datasets for studies of the earth system and global change. The DAO will place an array of subscriptions into the ECS in order to obtain their EOS standard product model inputs. ECS will ingest the DAO L4 model products and provide archive and distribution services. These L4 products may have value for other producers of operational medium range forecasts for comparison purposes.

Independent of the specific content being managed by the system, key aspects of ECS offer potential for the enhancement of operational data management systems used for medium-range forecasting. In the preceding discussion, we divided these aspects into key architectural features, operator tools, and internet access of data and data descriptions. The benefits for those who may be involved in the design or evolution of new operational meteorological data management systems range from reuse of design or concept to specific reuse of software.

For the key architectural features discussed, value may be gained through reuse of concept or the software itself. All of these key architectural features have been thoroughly "road tested" in the course of the experience with the Landsat-7 and Terra instruments. The specific value propositions are:

• The ECS data model and the HDF-EOS data standard have value for the interoperability they promote. For those involved in operational medium-range forecasting who may want to experiment with the utilization of EOS products, the rich metadata will have value for precise documentation of data accuracy, precision, and quality. The HDF-EOS data standard will obviate the need to write a new interface based on an Interface Control Document; the data is self-describing and access is provided by the HDF-EOS API.

• For those involved in algorithm development or versioning, value may be gained by studying and/or reusing elements of the ECS software toolkit for algorithm developers, coding standards, and the science software integration and test process.

• For those involved in data typing, study of the principal data typing/metadata population tool "Metadataworks" and the associated methodologies may hold value.

* Questions on this NOAA project, should be directed to the EOS Project Scientist, Dr. Mike King.

For the operator tools discussed, value is gained from simplification of workflow, information hiding, and process automation. For the ingest, processing, and distribution functions, operators are able to manage and control large numbers of discrete processes with minimal intervention:

• The Ingest GUI, if reused in an operational meteorological system, could allow an operator to manage large numbers of discrete FTP processes; cases of contention could be resolved by adjusting process priority or other ingest parameters. It could also give the operator maximum flexibility for automatically reading data from a variety of media.

• The PR Editor could be used to manage the preprocessing of all of a medium range model's inputs.

• Subscription services and the associated ECS tool, the SS Tool, have a dual value for those involved in operational medium range forecasting: reuse of the tool/feature and utilization of the ECS service. In the scenario of reuse of the tool, it could be used to manage the entire array of subscriptions placed by the user community for a given data center. In the scenario of automated delivery of EOS data into an operational medium-range forecast, a subscription or group of subscriptions would make elaborate machine-to-machine interfaces, or other special arrangements unnecessary.

Two types of benefits are derived from EOSDIS client services. For researchers involved in improving operational medium range forecasts direct access to EOS online data and data descriptions holds value. In this instance, client services of greatest value include user registration, guide/document search, inventory search and order, and browse. The second type of benefit involves reuse of specific services as an enhancement to an existing operational medium range forecasting data management system. The reuse of the guide/document search function may allow users to better understand the nature and background of a new weather product. The reuse of browse could allow users to examine a product's content prior to ordering a dataset.

THE FUTURE

The operational EOSDIS Core System provides a powerful tool to support a complex and evolving NASA global observing capability. Although now operational, this is only the first release and additional capabilities will be added in successive future releases. Substantial possibility of reuse exists for many of the systems elements; this is both welcomed and encouraged.

The timely, relevant data that will be made available from NASA's EOS program will have broad applicability and benefit to a diverse group of national and international users, including the meteorological community. The tools that will be made available through EOSDIS will help facilitate easy access to a growing catalogue of data collections of importance to society. This ease of access, combined with the low cost of the data and its direct relevance to global environmental studies, will herald a new era in the practical use of satellite-based Earth observations.

ACKNOWLEDGMENTS

This work was undertaken as part of the Raytheon Systems Company's (RSC) contract to NASA for the EOSDIS Core System contract (contract NAS5-60000).

Raytheon Systems Company's Information Technology Systems (ITS) business unit, located in Landover, Maryland is the prime contractor for the EOSDIS Core System. The 10-year contract began in 1993 and extends through 2002. Subcontractors include Lockheed-Martin Corporation, Electronic Data Systems, Science Applications Corporation, NYMA Inc., Earth Observation Sciences Ltd., and Raytheon Training and Services Company. RSC ITS has developed substantial expertise in the development of vary large scale, high reliability, complex science data management systems in this project and in other work for defense, government, and commercial customers. For further information on these services, please contact Joseph T. Senftle, Director of Business Development, at (301) 925-2035 or jsenftle@eos.hitc.com.

Raytheon Systems Company is a subsidiary of Raytheon Company, based in Lexington, Mass. Raytheon is a global technology leader that provides products and services in the areas of commercial and defense electronics, engineering and construction, and business and special mission aircraft. Raytheon has operations throughout the United States and serves customers in more than 80 countries around the world.