A Pilot Environmental Data Grid Distribution System

Gary Walter[1*], Alice Gilliland[1*], Ellen Cooter[1*], Robert Gilliam[1*], Kevin Cavanaugh[2] , Lynne Petterson[3]
[1]AMD/NERL/ORD, U.S. EPA, Research Triangle Park, North Carolina *On assignment from ARL/NOAA
[2]ORMA/ORD , U.S. EPA, Research Triangle Park, North Carolina
[3]NERL /ORD, U.S. EPA, Research Triangle Park, North Carolina

Pilot Objectives

- Implement, operate and evaluate a workable, distributed access, format neutral "data-on-demand" prototype data distribution system inside EPA's intranet
- Assess potential leading edge technologies for distributed data access, data mining and data provisioning
- Plan and develop multiple mechanisms for distributed data access, analysis, visualization, product generalization, and sub-setting over EPA's Data Grid to external partners

Global Earth Observation System of Systems

"An extraordinary international effort is now underway to promote and plan the development of a comprehensive, coordinated, and sustained Earth observation system of systems among governments and the international community to understand and address global environmental and economic challenges." -- At the Earth Observation Summit in Washington D.C., July 2003, participants recognized the following needs related to Earth observation:

- Affirmed need for timely, quality, long-term, global information as a basis for sound decision making. Recognized need to support:
  - Comprehensive, coordinated, and sustained Earth observation system or systems;
  - Coordinated effort to address capacity-building needs related to Earth observations;
  - Exchange of observations in a full and open manner with minimum time delay and minimum cost; and
  - Preparation of a 10-year Implementation Plan, building on existing systems and initiatives by European ministerial in late 2004

EPA Model Archive Data Distribution System (E-MADDS)

To overcome a deficiency in model data access, scientists are actively engaged in a grass-roots effort to develop a framework to share data and research findings over the Internet. Some current examples of technologies to distribute data include: OPeNDAP, ADDE, and netCDF access via the HTTP protocol. Many agencies are actively pursuing strategies to improve model data access. E-MADDS, EPA Model Archive and Data Distribution System, is a distributed data services pilot for format independent access to environmental data, addressing the summit call for the exchange of observations in a full and open manner.

E-MADDS Why now?

Current data access speeds and other considerations have highlighted deficiencies in model data access, interoperability and distribution. A data distribution system is needed for the following reasons:

- There is a need for more robust ways to distribute data and models both inside and outside EPA
- We need to promote model evaluation, product development and product deployment
- We need to foster research within the geo-science communities to study multiple earth systems using collections of distributed data
- We need to develop institutional partnerships via distributed open technologies

Model Data Access….Why is OPeNDAP/DODS a solution?

The data users experience is often frustrating—

- What existing data is of interest to me?
- Is a particular dataset going to be useful?
- How can I obtain data in a format I can use?

Time and effort is frequently wasted on data access and format issues. As a result, data is often under-utilized and model inter-comparison is frequently problematic. As a result of this, EPA and NOAA are collaborating on a pilot climate and air quality distribution system based on OPeNDAP/DODS which is currently the building block of NOAA's data distribution system NOMADS. OPeNDAP/DODS is a framework that simplifies all aspects of scientific data networking, providing tools for making local data accessible to remote locations regardless of local storage format. OPeNDAP/DODS is also a binary-level protocol designed for the transport of scientific data subsets over the Internet.

E-MADDS -- What else is needed?

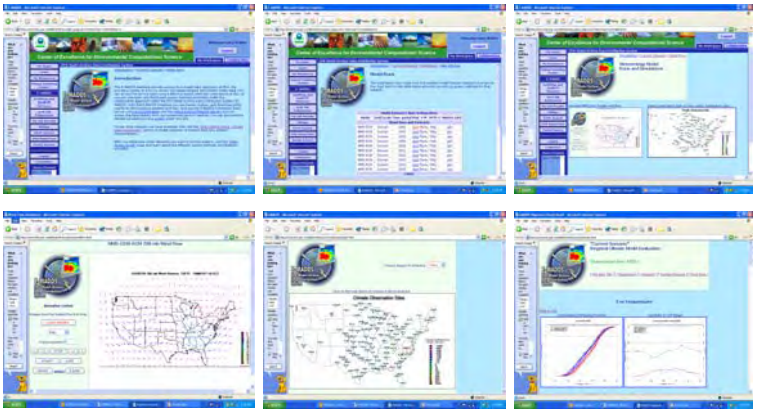
OPeNDAP/DODS is the initial building block of the E-MADDS system. However, the E-MADDS system needs a number of additional components to enable collaboration outside EPA's intranet including the following: EPA's Compute and Data Grid; EPA's Special Security Zone for Science; and EPA's Portal. Components that are being developed that are separate from the E-MADDS interface are described below:

Grid Services -- EPA is in the process of building Grid Services (Computational and Data) to provide Agency researchers and trusted partners with seamless access to enterprise-wide computational and data storage resources. The E-MADDS pilot uses one of EPA's Data Grid Access Servers (DGAS) to store datasets. Grid middleware distributes E-MADDS data to available storage resources, while maintaining the integrity of the data. Grid middleware also provides E-MADDS users transparent access to the data they need, without having to know where the data is located.

Special Security Zone for Science (SSZ) -- The Science FTP Server was one of the first capabilities created under the SSZ initiative. In May 2004, EPA institutionalized the Science FTP Server (SFS), located at: http://scienceftp.epa.gov. The SFS supports both inbound and outbound file transfers through EPA's Internet firewall. E-MADDS will use the Science FTP Server and other SSZ capabilities to move data to collaborators outside of EPA.

EPA's Portal – The current interface to the E-MADDS system is web html based. At some point EPA's portal will be the interface to this pilot data distribution system. Some sample portal screens are provided below.

E-MADDS Interface Screens



Summary

E-MADDS capabilities scheduled for completion by December 2005.

- Format Neutral Data Distribution – Data formats: cf, DIF, GRIB, GRIB2, BUFR, HDF, NetCDF, ascii, others to follow with libraries built as necessary
- Internal Client access – Direct client access is possible through GrADS (GDS), Ferret, MatLab, IDL, IDV, Web browsers or any OPeNDAP enabled client on EPA's intranet
- External Client access – The EPA Science Portal will be the access point to the data with data distribution through the Science ftp server.
- Data provisioning on the fly – A GDS sub-setting capability will allow users to retrieve a specified temporal and/or spatial sub domain from a large dataset, eliminating the need to download everything simply to access a small relevant portion of a dataset. (GDS is built on top of OPeNDAP)
- CIRAQ datasets – The Climate Change on Regional Air Quality (CIRAQ) datasets will be the first datasets to be distributed by E-MADDS. Regional climate runs and simulations for current climate will be available to authorized users.
- Search/Catalog – catalog / search of metadata will be implemented through the Portal