A Prototype for a Distributed Space Physics Data System

Charles Falkenberg & Chuck Goodrich
Advanced Visualization Laboratory
University of Maryland
College Park, MD 20742
{csfalk, ccg}@avl.umd.edu

James Gallagher & Peter Cornillon
Graduate School of Oceanography
University of Rhode Island
Narragansett, RI 02882-1197
{jgallagher, pcornillon}@gso.uri.edu

Glenn Flierl
Massachusetts Institute of Technology
Cambridge, MA
glenn@mead.mit.edu

December 23, 1996

Abstract

The collaborative analysis of data within the Space Physics community is hindered, in part, by the wide number of data formats and the wide distribution of data archives. In an attempt to address these two problems we have implemented a prototype which retrieves datasets, stored in different data formats at several remote locations. Our prototype uses the Key Parameter Visualization Tools (KPVT) and the Distributed Oceanographic Data System (DODS) to view data from the ISEE1, ISEE2, and ISTP programs. Our goal is to demonstrate the ability to access and use several types of remote data and existing analysis tools.

The work described demonstrates the power of an expressive data model, like the one in DODS, for converting and transmitting space physics data. Furthermore, since the DODS system architecture (and associated data model) was developed to meet oceanographic needs, the fact that it works well for use within the space physics community suggests that the DODS approach will also work well as a data distribution mechanism for the other earth science sub-disciplines. Given the growing interest in interdisciplinary work in the earth sciences the existence of a data model/system capable of spanning the various sub-disciplines is significant.

INDEX TERMS: Scientific data formats, scientific database management, distributed data systems, Space Physics Data System, Distributed Oceanographic Data System.

This work was supported by NASA grant NAGW-4580 to the University of Maryland and NASA grants NAGW 3784 and NAGW 3790 to the University of Rhode Island.

Introduction

Space Physics in general, and Magnetospheric Physics in particular are moving from the exploratory phase to one of detailed investigation. There is thus an increasing focus on global studies of phenomena and structures, such as the earth's magnetosphere. Effective understanding of these phenomena require the combined analysis of data from diverse instruments on various spacecraft along with ground based observations and computational simulations. An enormous amount of space physics data is available for these studies, with an order of magnitude more coming from current missions like the International Solar-Terrestrial Physics Project (ISTP).

However, the combined analysis of the different data sets needed by these studies is hindered by several factors. These impediments to interoperability include the various formats in which the data are stored, the physical locations of the archives, and the semantics of the data themselves. The number of data formats is almost as large as the number of past and present PIs involved in collecting and distributing data. The archives are spread across the country at the institutions where the original PIs were located and often human intervention is needed to make the data available. Finally, the variables can be stored in different units and coordinate systems. This can give a different semantic meaning to measurements of similar physical phenomena.

The challenges of using disparate and distributed datasets are shared by other scientific disciplines as well. As data collection techniques mature and the methods of storage evolve, many sciences are left with a legacy of difference too. The analyses of atmospheric, astrophysical, and oceanographic data are all hindered by the same set of problems. These problems, along with some interesting software solutions, were reported by many researchers at the recent Science Information Systems Interoperability Conference (SISIC).

One of the systems presented at the SISIC meeting was the Distributed Oceanographic Data System (DODS), designed to address some of these problems for the oceanographic community. DODS is a client/server data delivery system which allows disparate datasets to be retrieved from remote sites through an interface that mirrors one of the standard scientific file formats. We wanted to experiment with DODS as a possible solution to some of the dataset problems faced in space physics. Our goal was to utilize DODS, along with other existing software tools, to build a running prototype of remote interoperability between known software and space physics datasets in several different formats. Our intention was to find possible solutions in which the effort required by the data supplier was kept to a minimum and the flexibility provided to the data user was maximized.

DODS is designed to supply remote data using an interface which matches one of the common scientific data formats. Therefore we looked for a graphical analysis application which was familiar to the space physics community and retrieved data in one of these formats. We chose the ISTP Key Parameter Visualization Tools (KPVT) which are written in IDL to plot several project specific datasets, stored in Common Data Format (CDF). We modified the KPVT to retrieve data through DODS and built two data servers which read different datasets into DODS. One data server reads CDF files and the other server uses a product called FreeForm from the National Geophysical Data Center to access both ASCII and binary data.

The prototype integrates these three components; DODS, the KPVT and FreeForm, to allow disparate data from the ISEE1, ISEE3, and ISTP projects to be retrieved from remote sites and displayed together. Our results demonstrate the power of a canonical data model to simplify the conversion between data formats, and lay the groundwork for future analysis tools which can ignore format differences. Our result, however, highlight the larger problem of semantic differences between datasets.

The paper begins with an overview of the various software tools which were used, including the DODS system, FreeForm and a brief description of the KPVT. This is followed by the questions which lead to the prototype and some of the details of the prototyping exercise. The results of the the exercise include the answers to the prototyping questions and are followed with an outline of future work suggested by our experiences.

The KPVT, DODS, and FreeForm

The Key Parameter Visualization Tools are a project specific toolkit, written in IDL, to support the CDF datasets which are being produced by the ISTP program. The KPVT will plot multiple variables from several CDF files using the variable names and coordinates which are standard to ISTP. These tools are used by many of the scientists in ISTP and therefore offered an excellent opportunity to built a prototype with a familiar face.

Many scientists use applications like the KPVT to access data files which are stored in ASCII or binary files or a standard scientific format like CDF or netCDF. These scientific formats provide the application developer with an interface (API) for reading and writing large arrays of scientific data and they support various data types for both variables and metadata attributes.

The Distributed Oceanographic Data System was built as part of a collaborative effort between the University of Rhode Island and the Massachusetts Institute of Technology to allow interoperability for remote oceanographic data in different data formats through the World Wide Web (WWW). DODS is designed to support several data access interfaces which exactly match existing scientific data formating APIs. It uses its own model of a dataset as an intermediate representation of the data being transfered over the internet between a data server and data client. This means that DODS can be linked into existing applications and provide read access to remote files without modifying the application software. Several different remote servers can then read data from any semantically similar dataset into DODS. Detailed information related to DODS is available via the Web at *http://dods.gso.uri.edu/DODS/home/home.html*

Figure 1 compares a traditional scientific application and an application written with DODS. A traditional application accesses its data through the interface to a standard scientific format. Local file names are passed through the interface and the software opens the files and supplies the data. In addition to local file names, the DODS application can pass remote file names in the form of Uniform Resource Locators (URLs) through the interface. DODS examines the file name and if it is a URL the request is passed across the internet to the HTTP server and the appropriate data server program is executed.

The data servers are similar to the I/O portion of the traditional application and are part of the Common Gateway Interface (CGI) programs that are executable by an HTTP server. The server opens the file which is local to the server and the results are returned through the WWW and supplied back to the application program. DODS passes any data constraints along with the file name so that only the requested subset of data is returned over the net. The remote request is slower but it presents no change in processing for the application program.

DODS supports its own data model as an intermediate representation of a dataset. This allows for the translation between formats with similar data semantics. Figure 2 shows this use of the DODS canonical data model as an intermediate representation through which data can be converted. The data models for netCDF, CDF and HDF are all quite similar and the process of converting between them does not present major problems. Other formats, however, like JGOFS or relational database tables present a different data model which is more difficult to standardize.

One of the advantages of DODS is that the servers are quite simple and relatively easy to write. Most

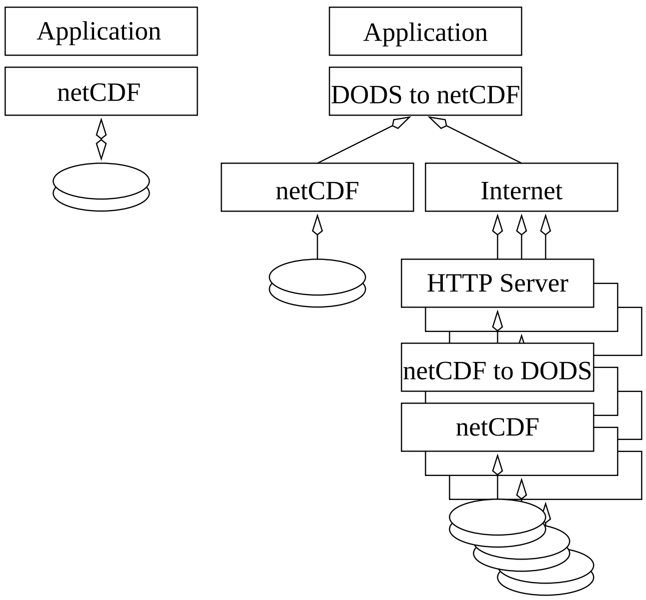

Figure 1: Traditional applications interface directly with a scientific format library like CDF. (left) The DODS layers provide a canonical representation of the datasets along with the remote access. (right)

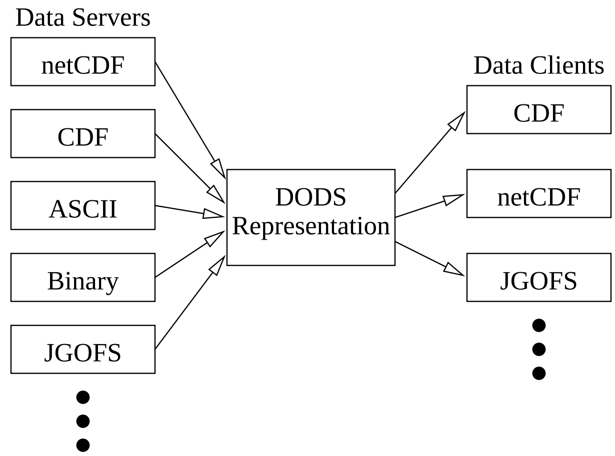

Figure 2: DODS canonical data representation provides flexible data format conversion.

of the complexity is found in the client software which mirrors the interface to a particular scientific data format. This meets one of our initial goals and is part of the reason we chose to experiment with DODS.

Currently two data clients have been written which mirror the interfaces to different scientific formatting standards. These are a netCDF client and a JGOFS client. Several data servers exist as well and as part of the prototype described here we created data servers for CDF as well as an ASCII and binary data server using FreeForm.

FreeForm is a tool for reading and writing files in ASCII, Binary or dBase formats. FreeForm uses a text based file description to define the variables and attributes in the data file. Once a text description has been created an application can use the FreeForm API to read and write the data file. FreeForm includes several utilities for printing out the data file and for converting between a wide range of units of measure. FreeForm supports data files with several levels of header information and data representation.

Using FreeForm allowed us to create one DODS server which provided access to two different types of datasets. A text description file had to be provided for each type of dataset but a single DODS server was able to access both types of data.

Prototype for Space Physics Data

Our prototype was built with the goal of experimenting with interoperability between datasets in the space physics community. It integrates the KPVT, DODS and FreeForm and demonstrates the access to remote files in three different formats. One of the goals was to better understand the challenges which would be faced during the development of a larger scale production version of a distributed space physics data system. The work includes adapting the KPVT to use the existing netCDF interface to DODS and building data servers for CDF and FreeForm. The FreeForm server used different text description files to supply ASCII and binary data.

Our prototyping exercise was tailored to answer the following questions:

1. What are the obstacles (if any) when DODS is re-linked to an existing application? DODS is designed to mirror existing interfaces to scientific data software and therefore should be easy to use in place of that software.

2. Can data servers be set up to supply data with a minimum of effort? One of our goals was to minimize the effort on the part of the data suppliers to provide data to the system.

3. Can FreeForm be used effectively as a data server for delivering ASCII and binary files to DODS? Can it successfully describe space physics data?

4. What is the performance of remote access to datasets through DODS?

5. What are some of the semantic obstacles to interoperability between different generations of space physics datasets.

6. How would the ability to retrieve data in several formats from many remote locations change the way tools like KPVT are written.

Many of the issues dealt with the implementation and performance problems of the target system. We therefore wanted to create a running prototype which would have a familiar interface and add support for a wide range of useful space physics data. We chose the KPVT because of its familiarity and because it is currently project specific. In addition to Wind, we picked two other sources of data which are in proprietary formats.

The architecture of our prototype is compared in figure 3 to the original architecture of the KPVT. On the left is original KPVT which are written in IDL and access the CDF files through IDLs CDF interface. The architecture of the prototype, on the right, shows a re-linked version of IDL which accesses DODS through its own netCDF interface. DODS handles the remote communications to the CDF and FreeForm data servers which read and return data from CDF, ASCII and Binary files.

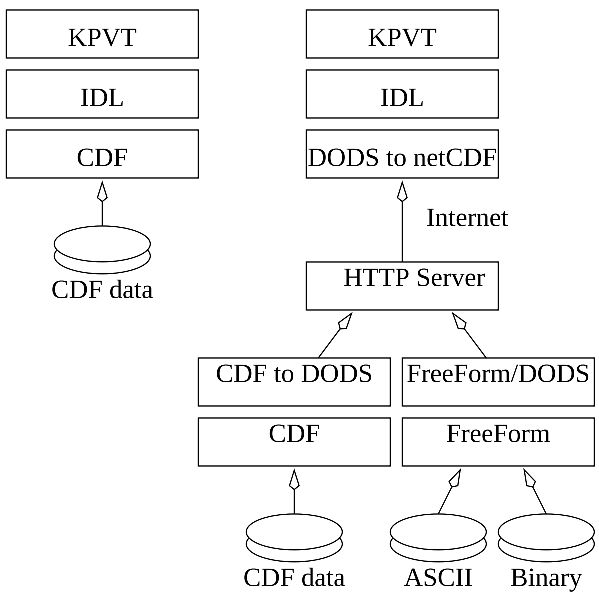

Figure 3: Left: Architecture of the Key Parameter Visualization Tools. Right: Architecture of the space physics prototype.

One of the key parameters that the KPVT were designed to plot is the magnetometer data from the Wind Satellite. These files are in CDF which is the ISTP file format of choice. We chose to add magnetometer data from ISEE3 and ISEE1 because they are the same type of measurement, taken more than 10 years ago, and they are stored in very different data formats. The ISEE3 data is in an ASCII format from Los Alamos and the ISEE1 data is in a binary flat file format and archived at UCLA.

The FreeForm data server reads a dataset definition for the ISEE1 and ISEE3 datasets and supplies the data to DODS. The CDF server supplies the original Wind file in CDF. Since writing a CDF client for DODS was not in the scope of this effort we modified the KPVT to use the existing DODS netCDF interface. We then re-linked IDL to the DODS netCDF interface to allow the KPVT to retrieve data through DODS. The result was a prototype which read remote CDF, ASCII, and binary files through IDLs netCDF interface. Figure 4 shows the Wind, ISEE1, and ISEE3 data plotted on the same page of output from the KPVT.

Results

The answers to most of our prototyping questions are encouraging and so our results are quite positive. Although FreeForm was not as expressive as we had hoped and the performance of remote access is a bit slow, the other aspects of the system are promising. The prototyping questions are addressed in sequence below.

1. Re-linking IDL with DODS took us a fair amount of time but it had little to do with DODS. Once this was done the netCDF version of the KPVT ran without any difficulty. Therefore, although some technical experience is needed for this step, the software logic of other typical scientific software tools should not need to be changed to accommodate DODS.

2. We were also pleased with how simple the servers were to write. The CDF server only took us a little over a week to complete, the FreeForm server required about the same amount of time. With a system like FreeForm in place, however, special servers might not even need to be written for many ASCII or binary data files. A data archive must also be running a Web server and have the data servers available as CGI scripts. Making these scripts available to the Web server was no problem.

3. The FreeForm software did not quite live up to our high hopes. The server was a little more complicated than necessary and the FreeForm data definition language did not capture the concept of large arrays as we had hoped. However, the concepts behind FreeForm are very good and we can foresee ways to expand upon its functionality in the future.

4. The performance of remote access is a bit slow, especially for large files. This can be mitigated by selecting hyperslabs of data through DODS. Unfortunately the KPVT subsets variables after they have been read and so they cannot take advantage of this feature of DODS. Remote access may prove very valuable for browsing a large number of datasets to find one of interest but then down loading the files to a local archive may be preferred.

5. Reconciling the semantic differences between datasets proved to be a time consuming part of the exercise. The different dataset use different conventions and different time coordinates. The Wind data uses a special CDF time type, the ISEE1 data has a count of milliseconds since 1966 and the ISEE3 data is broken down into a standard number of averages per day. In addition the difference in the coordinate systems of the data variables prevent them from being compared directly. Addressing these differences will be difficult but removing the format impediments is the first step.

6. Finally, if the functionality provided by this prototype was available to tool developers it would motivate several changes in the design of tools. This prototype reconciles some of the data format issues and the distributed data issues. Using this as a starting point tools can be designed to be much more general and accommodate datasets from may different origins.

Future work

In general this prototype demonstrates the power of an expressive data model, like the one in DODS, for converting and transmitting data. Several areas should be pursued in order to move the functionality of this prototype into production quality software for the space physics community.

First, a DODS client library with a CDF interface needs to be developed and tested with the CDF server. This a necessary before applications which uses CDF files can make significant use of DODS.

Second, FreeForm should be enhanced or a similar system developed to allow ASCII and binary datasets to be described and augmented with some meta-data. A system of this type might encompass more scientific and database formats as well, greatly reducing the effort required of data suppliers.

Several modifications to DODS should also be considered. The ability to display the directories of remote files should be included along with improved data translation capability. Enhanced data translation will be needed to convert between data models which are not as similar as CDF and netCDF.

In conclusion, our work shows a possible solution to some of the challenges of disparate and distributed datasets within space physics and suggests a similar solution for data access across the broad set of earth science sub-disciplines. We think that future software analysis tools can make use of the functionality provided by DODS to reduce the complexity of using remote datasets which are in multiple data formats. We look forward to feedback from the space physics and earth science communities about other possible uses.

Acknowledgments

The work described in this manuscript was supported by the National Aeronautics and Space Administration's Office of the Mission to Planet Earth and Office of Space Science Information Systems Program via grants # NAGW 3784 and # NAGW 3890, to the University of Rhode Island and grant(s) # NAGW 4580 to the University of Maryland.

The authors would also like to express their special appreciation to Joe Bredekamp of NASA for encouraging this collaboration and to Research Systems Inc. for their help in relinking IDL with the DODS core software. We further thank C. T Russell of UCLA and Los Alamos National Laboratory for providing the ISEE 1 and ISEE 3 magnetometer data.

References

[1] C. Falkenberg and J. Purtilo. "Parallel I/O using a distributed disk cluster: An exercise in tailored prototyping". Technical Report UMIACS-TR-95-18, Institute for Advanced Computer Studies at the University of Maryland, College Park, MD, February 1995.

[2] National Geophysical Data Center, Boulder, Co 80303. *FreeForm, A Flexible System of Format Descriptions for Data Access, Users Guide, Version 3.1.*

[3] Russ Rew, Glenn Davis, and Seve Emmerson. *NetCDF User's Guide.* Unidata Program Center, University Corporation for Atmospheric Research, April 1993.

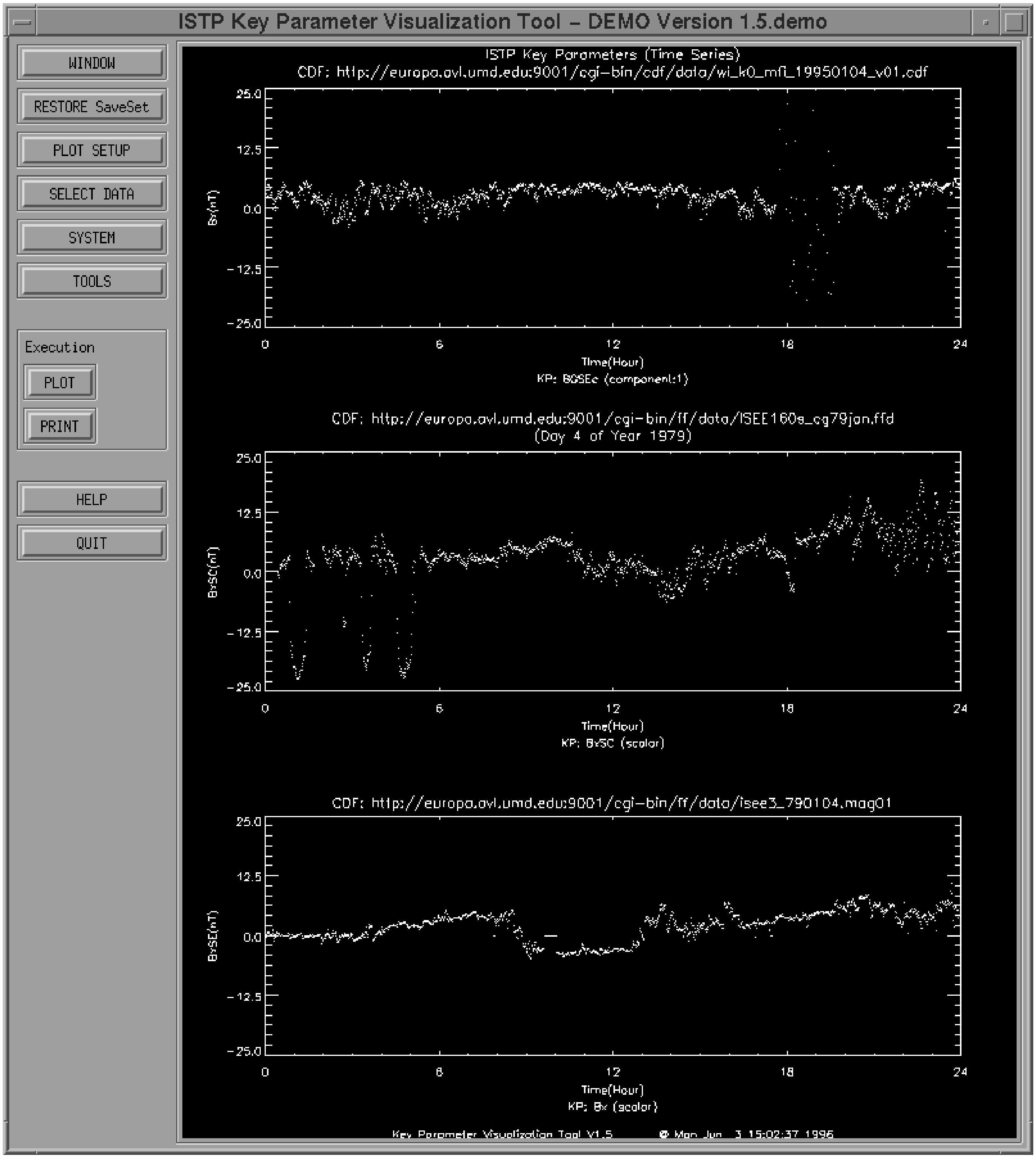

Figure 4: Magnetometer data from all three data formats plotted in the KPVT. File names are URLs.