

Earth Science Markup Language

A Solution to Address Data Format Heterogeneity Problems in Atmospheric Sciences

BY RAHUL RAMACHANDRAN, SUNDAR A. CHRISTOPHER, SUNIL MOVVA, XIANG LI, HELEN T. CONOVER, KEN R. KEISER, SARA J. GRAVES, AND RICHARD T. MCNIDER

Scientists are often confounded by various data types, formats, and systems used in Earth science. For example, regional atmospheric air pollution models have to deal with over 20 datasets in different formats, where each format has a large user manual describing the structure and format of the data. It is almost impossible for individual scientists or even scientific groups to have expertise at their disposal in each of these data formats. Often, important new data are not incorporated into models because they are in a format that is new to the research group. Dealing with heterogeneous formats can become a major bottleneck in the process of data analysis. Since most scientists are not programmers, much of their time and effort are required to understand and decode data in unfamiliar formats.¹ Scientists have to either write a data decoder or reader module for a new data format or translate the data into a format that their analysis tool can handle.

One solution to this problem would be to have all data-producing and data-consuming communities agree upon a single self-describing format that includes the metadata defined in a community-accepted convention. The metadata annotations in this “standard” data format would allow software developers to

write analysis tools that can decode any data in the scientific domain. Although a single, “standard” self-describing data format would greatly simplify the use, integration, or fusion of disparate data types, the science community has found this solution to be impractical for the following reasons: First, there is a considerable volume of legacy datasets. Converting all of these datasets to a standard format would be exceedingly time-consuming and expensive. Furthermore, it is difficult to design a standard format that can match the capabilities of specialized formats designed for specific uses. For example, some formats were designed for efficient storage and exchange across a network, whereas other formats were designed for efficient data retrieval.

The Earth Science Markup Language (ESML) provides a more flexible solution to this data usability problem, supporting multiple data formats with metadata-based interchange technologies. ESML is a specialized markup language for Earth science metadata based on XML (eXtensible Markup Language), allowing analysis tools to seamlessly utilize datasets in heterogeneous formats. ESML description files contain metadata with content and structural information for the corresponding data file format. These descriptions can be generated by either the data producer or the data consumer at any time to allow data/analysis tool interoperability. Because ESML description files are external files, they do not modify the analysis tool or the data file itself. The ESML schema defines the XML grammar for writing ESML description files. The schema has been designed such that a combination of a few elements can be used to describe numerous datasets. The ESML library is used by analysis tools to parse the relevant ESML description file for structural information about a data file, and to read the data from the file. Together, these three components comprise a system of machine-readable and interpretable markups that allow analy-

¹ Informal surveys and anecdotal evidence indicate time spent on such data issues to be 50% or more.

AFFILIATIONS: RAMACHANDRAN, MOVVA, LI, CONOVER, KEISER, AND GRAVES—Information Technology and Systems Center, University of Alabama in Huntsville, Huntsville, Alabama (UAH); CHRISTOPHER AND MCNIDER—Department of Atmospheric Science, UAH

CORRESPONDING AUTHOR: Rahul Ramachandran, S 333 Technology Hall, ITSC, University of Alabama in Huntsville, Huntsville, AL 38599

E-mail: rramachandran@itsc.uah.edu

DOI: 10.1175/BAMS-86-6-791

©2005 American Meteorological Society

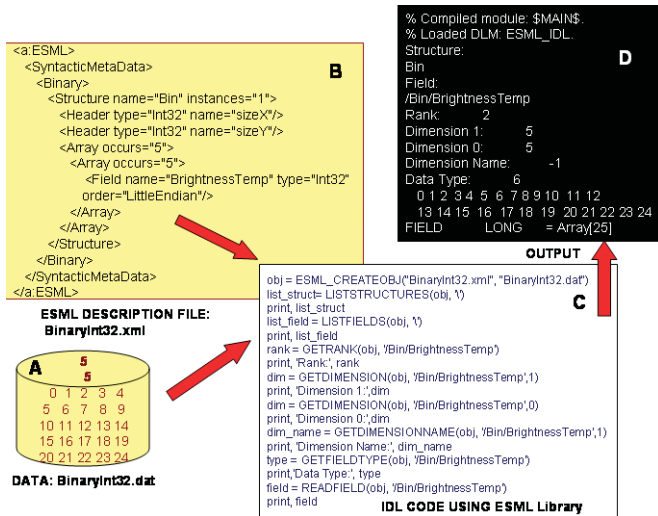


FIG. 1. (a) Example data file; (b) ESMIL description file for the example data file; (c) function calls made in IDL using the ESMIL Library API to read the data file; and (d) output generated by the IDL function calls.

sis tools to parse ESMIL descriptions to read the data files, eliminating the need for additional data conversion software.

ESML greatly simplifies a scientist’s job by removing the overhead for handling heterogeneous data formats, enabling them to use virtually any data format in their analysis tools. Since ESMIL description files are external to the data files, they can be easily created, modified, and viewed using any text editor. Scientists can view an ESMIL description file as a set of instructions to the analysis tool on how to read and understand a data file (a machine-readable README file). Furthermore, if the structure of the data format changes for any reason (e.g., a new version of the dataset) no software modifications are required; rather, a new ESMIL description file is created for the modified dataset.

ESML also provides several advantages to data archiving centers by allowing the flexibility to store data in their native formats, rather than converting them to some standard format. The archive center need not provide an ESMIL description file for every data file, but instead can write a single ESMIL description file for all data files belonging to the same data class, where a data class defines a set of data files that are structurally and semantically similar. This feature provides an inexpensive solution for dealing with legacy datasets. Data centers can easily create ESMIL description files for all of their legacy datasets with minimal effort in terms of time and labor. The exist-

ing legacy datasets then become a more valuable data resource for scientists, because they can be easily used more efficiently and effectively.

The ESMIL library is written in C++ for Windows and LINUX operating systems. The ESMIL library application programming interface (API) has been designed to be intuitive and easy to use. By using the ESMIL library, software developers can build “ESML-enabled” science analysis tools with a single reader component for all the various data formats, rather than separate reader components for each format. Currently, ESMIL supports ASCII and binary data formats; several complex, structured, self-describing data formats [namely Hierarchical Data Format-Earth Observing System (HDF-EOS); HDF-5; Gridded Binary (GRIB); and network Common Data Form (netCDF)]; and the format specific to WSR88D (Weather Surveillance Radar, 1988, Doppler) radar data called Next Generation Radar (NEXRAD) Level II. An easy-to-use graphical editor is available to write the XML markups for

the ESMIL description files along with a data-browser tool to browse the data and the metadata values within a file. Also available are Python and Interactive Data Language (IDL) interfaces for the ESMIL library. Python is a modern high-level programming language and is very portable, modular, and extensible. Python enjoys the support of a strong community of developers and users in the science community. IDL is a commercial product from Research Systems Inc., commonly used by scientists for data analysis and visualization. ESMIL plug-ins for Python and IDL extend capabilities of these tools to read heterogeneous data formats.

ESML developers are also addressing data transport issues: an extended ESMIL library supports remote file access via cURL, an open-software command-line tool for transferring files with URL syntax. In addition, the ESMIL team is working together with OPeNDAP developers to extend that data transport system to more types of data. OPeNDAP software (<http://opendap.org>) makes local data accessible to remote locations regardless of local storage format. OPeNDAP also provides tools for transforming existing analysis tools into OPeNDAP clients (i.e., enabling them to remotely access OPeNDAP served data). An OPeNDAP-ESML data server that allows a single OPeNDAP server to distribute datasets in multiple formats to an OPeNDAP client is being beta tested.

An ESMIL example for a simple data file in binary format is shown in Fig. 1. The data file contains two

header values followed by a two-dimensional data field (Fig. 1a). Fig. 1b is one possible ESML description of the data file. After declaring the format <Binary>, the subsequent sequence of declarations follows the data structure of the file. The entire data file is described within a single <Structure> element that contains two <Header> elements. These elements contain the type attribute defined to read integer data in base ten. The actual data are described in two nested <Array> elements with size specified for each dimension using the occurs attribute. Data values in the array are described by the <Field> element, with a name and a data type definition to read integer data in base ten. Figure 1c shows a sequence of calls made in IDL using the ESML Library API. Once an ESML object has been created within IDL for a given data file (BinaryInt32.dat) and the associated ESML description file (BinaryInt32.xml), all the structures and fields can be listed. To retrieve a particular data field or the metadata for a data field, the user provides a path to the field and the ESML object reference. The output generated by the IDL code is shown in Fig. 1d. For complex data formats such as HDF-EOS, minimal ESML declarations are required. The ESML library simply uses the format's native software library to query the data file to retrieve the relevant metadata required for reading the data. Thus, the user is able to access data in multiple complex formats through a single API and with only minimal additional ESML descriptions.

FIG. 3. Examples from two atmospheric science applications using ESML: (a) example of MODIS/CERES data collocation; (b) morning surface skin temperatures for 8 Aug 1999 from three different data sources: (left) GOES satellite-derived, (middle) MM5 model, and (right) NCEP-NCAR Reanalysis data. The time is 1200 UTC for all data types except the satellite skin temperatures, which are for 1400 UTC. Units are degrees Kelvin, and the color interval is 5 K.

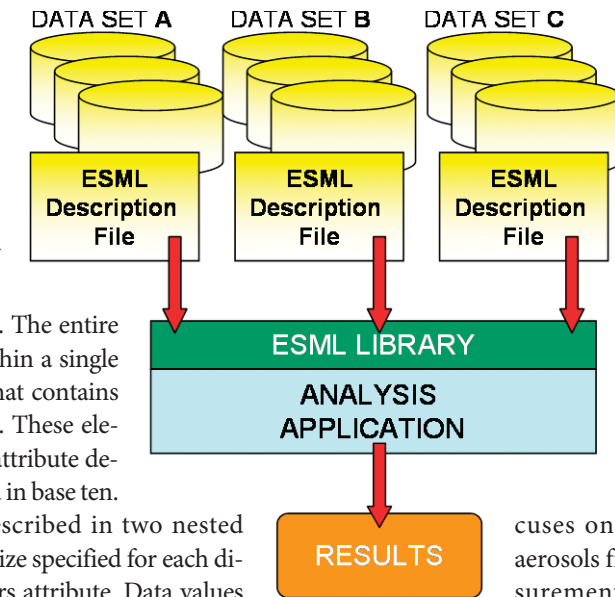
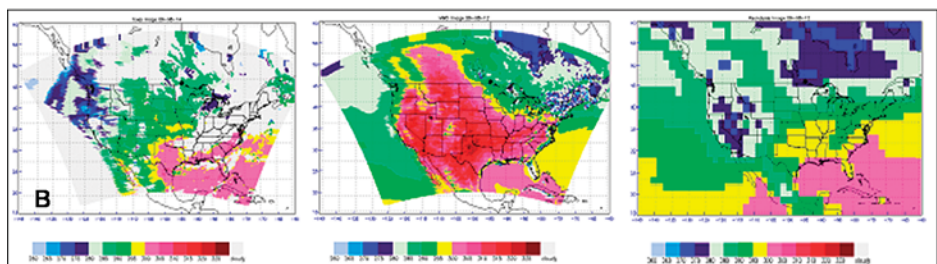
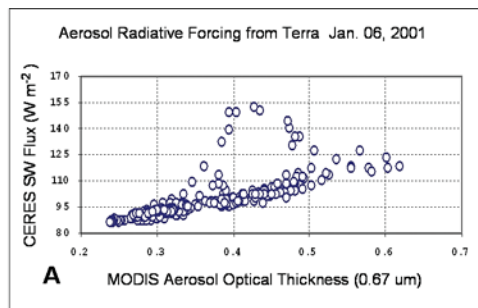


FIG. 2. Schematic representation of how ESML can be used by science applications.

A schematic representation of analysis tools using ESML is given in Fig. 2. Two ESML applications in satellite remote sensing and numerical modeling are described here, with their results shown in Fig. 3a and 3b, respectively. The first application focuses on analysis of atmospheric aerosols from different satellite measurements. One of the major improvements expected within the life-

time of NASA's Earth Observing System mission is to combine different satellite sensors with ground-based and in situ measurements to examine the role of aerosols in the Earth-atmosphere system. Compared to traditional numerical modeling simulations, the approach of using combined satellite measurement from sensors on the *Terra* and *Aqua* satellites provides an independent method for studying the impact of aerosols on climate. Both *Terra* and *Aqua* have multiple instruments that can be used to examine the effect of aerosols. The Moderate Resolution Imaging Spectroradiometer (MODIS) is a multichannel satellite imager used to detect aerosols and provide a global picture of aerosol distribution and thickness. However, there is a need to combine this information with radiation measurements



from the Clouds and the Earth's Radiant Energy System (CERES) instrument onboard *Terra*. Additional information can also be obtained from other instruments, such as the Multiangle Imaging SpectroRadiometer (MISR). Since these data must be collocated in space and time for the analysis, researchers developed a collocation tool which reads the data in different formats by making the same function calls to the ESML library. By writing different ESML descriptions for the different datasets, the scientists were able to incorporate multiple satellite datasets into their collocation analysis.

Atmospheric numerical modeling requires the use of a broad range of data types to initialize and verify the model output. In the second application, ESML capabilities were used to evaluate a specific atmospheric variable. Three measures of skin-temperature data produced in three different formats were chosen for evaluation: satellite-derived temperatures from Geostationary Operational Environmental Satellite (GOES) satellite data, model-predicted temperatures from the fifth-generation Pennsylvania State University–National Center for Atmospheric Research (PSU–NCAR) Mesoscale Model (MM5), and tem-

peratures from the National Centers for Environmental Prediction (NCEP)–NCAR reanalysis data. By coupling the analysis tool (in this case, IDL) with the ESML library and writing ESML description files for the three datasets, it was possible to quickly evaluate the skin temperature from the different sources. An example result is shown in Fig. 2c, where the morning skin temperatures for 8 August 1999 for the three different data sources were read using ESML. Again, the scientists were able to easily visualize and analyze data in different data formats using ESML combined with their visualization software.

These two examples demonstrate the benefit of ESML to the scientists in handling multiple data formats in their analysis work. Additional information about the concepts, tools, and products described in this article can be obtained at the ESML Web site (<http://esml.itsc.uah.edu>). The source code for the ESML schema and library is also available at the Web site.

ACKNOWLEDGEMENTS. This work was funded by the Earth Science Technology Office, Goddard Space Flight Center, NASA.

FOR FURTHER READING

- Christopher, S. A., and J. Zhang, 2004: Cloud-free short-wave aerosol radiative effect over oceans: Strategies for identifying anthropogenic forcing from Terra satellite measurements. *Geophys. Res. Lett.*, **31**, L18101, doi:10.1029/2004GL020510.
- Ramachandran, R., M. Alshayeb, B. Beaumont, H. Conover, S. Graves, X. Li, S. Movva, A. McDowell, and M. Smith, 2002: Interchange technology for applications to facilitate generic access to heterogeneous data formats. *2002 IEEE International Geoscience and Remote Sensing Symp.* and the *24th Canadian Symp. on Remote Sensing*, Toronto, Canada, IEEE.
- , H. Conover, S. J. Graves, and S. A. Christopher, 2003: Earth Science Markup Language: A solution to the Earth science data format heterogeneity problem. *19th Int. Conf. on Interactive Information Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, Long Beach, CA, AMS, CD-ROM, 15.10.
- , S. Graves, H. Conover, and K. Moe, 2004: Earth Science Markup Language (ESML): A solution for scientific data-application interoperability problem. *Comput. and Geosci.*, **30**, 117–124.