2016 Workshop: Data Integration in Geospace

Long title
Data Integration and Software Practices in Geospace Science
CEDAR-GEM
Conveners
Asti Bhatt
Russell Cosgrove
Ennio Sanchez
Michael Ruohoniemi
Philip Erickson
Description

The geospace community relies on data from disparate sets of instruments and various modeling frameworks involving different boundary conditions. Facilities like CEDAR and later Madrigal database and CCMC have come into existence to store and curate the data and models produced by the community. However, we are still lagging significantly in making effective use of all the data that is generated from a multitude of instruments with high cadence. We are making some progress towards creating data integration tools that would allow models like AMIE to imbibe the data with ease. But the individual researchers still face challenge in navigating the data/software landscape to both use and share with ease. While there are some who have embraced the open source ways of documenting and sharing their software, many tools used in our communities reside on researchers' own computers in proprietary languages. How do we navigate these diverse sets of resources and make them available for effective collaboration?

The data and software are going to only increase as we go forward. Both NSF and NASA have put significant resources towards data management that has also brought in computer scientists to collaborate effectively with geoscientists. For example, the NSF has put their energy behind 'EarthCube', a program that explicitly brings together geoscientists and computer scientists to solve the modern problems in data and software management in geosciences. We want to start the conversation on effective practices of data and software sharing that the field of Geospace sciences can employ.

This session will include a tutorial on 'Geoscience paper of the future (GPF)' concept by Yolanda Gil from University of Southern California, who is a computer scientist with a focus on geoscience. The GPF is an initiative to encourage geoscientists to publish papers together with the associated digital products of their research. The session will also have an invited talk from Tanu Malik, a computer scientist who works on data reproducibility practices in geosciences.

We invite short 2-slide presentations or demos on available and upcoming resources, tools, needs, challenges in data integration and software sharing practices in our community. We would like to have time for discussion on improving the data and software infrastructure for geospace science.

Agenda

This initiative is also part of the NSF EarthCube activities. At the session on Data Integration and Software Practices in Geospace Science on Monday afternoon, many current and proposed EarthCube projects will be discussed along with other initiatives in data integration space. The lineup is as follows:

1:30-2:15 Yolanda Gil: Geoscience Paper of the Future
2:15-2:30 Tanu Malik: Reproducibility in Geosciences/Introduction to GeoDataspace, and EarthCube building blocks project
2:30-2:40 Phil Erickson: Open source Python data integration effort
2:40-2:50 Russell Cosgrove/Todd Valentic: Integrated Geospace Observatory, and EarthCube integrative activity project
2:50-3:00 Xueling Shi: DavitPy, a Python library for SuperDARN data processing
3:00-3:10 Tomoko Matsuo/Liam Kilcommons: Distributable AMIE
3:10-3:20 Jesper Gjerlov: Magnetosphere-Ionosphere-Atmosphere-Coupling Project (MIAC)
3:20-3:30 Michael Hirsch: Curating, distributing and managing large datasets

We look forward to enthusiastic participation from the larger CEDAR community.

Justification

Characterization of geospace system relies on observations from various instruments, models with a variety of frameworks and inputs, effective analysis of data and assimilation of these data into the models to improve predictive capability of the models. The 2012 Decadal Survey for Solar and Space Physics recognized the

need to create an effective data environment to enhance the space physics research. This includes coordinated development of data systems infrastructure, community based data mining and assimilation tools, exploitation of emergeing technologies and community oversight of emerging, integrated data systems. The decadal survey also recognized that much of the data infrastructure and analysis tools development has happened in an uncoordinated fashion. This is largely true for the geospace communities including CEDAR and GEM.

Reference: 2012 Decadal Survey for Solar and Space Physics Appendix B

The 2011 CEDAR strategic plan thrust #6 calls out the need to Manage, Mine and Manipulate Geoscience Data and Models. The rationale for this approach is to discover correlations, understanding causalities, and contribute to determining the evolution of geospace using data spanning space and time. This includes developing and implementing standardized data formats, and developing data assimilation schemes to integrate data with models among other things.

NASA and NSF have put significant resources in the cyberinfrastructure and data science development in recent years. One of the latest NSF initiatives is called EarthCube that aims to bring geoscientists and computer scientists together to solve scientific challenges in geoscience. Geospace sciences is woefully underrepresented in the EarthCube community even though the concepts beind developed through it are equally applicable to geospace sciences. Many other geoscience communities have developed data and software standards and frameworks that aid the scientists in curating, managing and mining both the data and software tools. We would like to start the conversation on the topic of data and software standardization, mining, assimilation and curation in CEDAR and GEM fields. We anticipate this to be a recurring workshop at subsequent CEDAR and GEM meetings.

View PDF