Solar-Terrestrial Research

Executive Summary:

Science-Driven Cyberinfrastructure Needs in Solar-Terrestrial Research

Held at New Jersey Institute of Technology, 2014 Aug. 13-15

Steering Committee: ***Gelu M. Nita, Dale E. Gary, Andrew J. Gerrard, Gregory D. Fleishman, Alexander G. Kosovichev, Vincent Oria, Marek Rusinkiewicz***

Introduction

More than 80 domain scientists and students from three sub-disciplines of Geospace research (solar/heliospheric, magnetospheric, and upper-atmospheric research), as well as computer science, met at the *Center for Solar-Terrestrial Research* at *New Jersey Institute of Technology* for a 3-day workshop to examine the field's current state of cyberinfrastructure (CI) and its future needs. To prepare for the workshop, the steering committee identified 17 CI-knowledgeable leaders (listed at http://workspace.earthcube.org/solar-terrestrial-end-user-workshop/) who represent each of the NSF Geosciences programs SHINE, GEM and CEDAR, as well as computer science. This scientific organizing committee identified an additional 40+ scientists for invitation to the workshop, as well as NSF program managers Eva Zanzerkia (Earthcube), Ilia Roussev (SHINE), Anne-Marie Schmoltner (CEDAR), and Raymond Walker (GEM).

We endeavored to balance the demographics among the sub-disciplines and in relative experience of the participants. Approximately 25% of participants were early-career (8 students, 7 young scientists), 25% mid-career, and 50% in senior positions. The sub-discipline participation was nearly evenly split, with 34% SHINE, 23% GEM, 23% CEDAR, and 19% computer science. The preponderance of solar participation reflects mainly the concentration of solar research among the local NJIT participants. The organizers believe that the workshop successfully captured the expertise and experience of the Geospace research community, and that the findings herein represent the consensus view of leaders and practitioners in science-driven cyberinfrastructure among space-science researchers.

The Geospace disciplines are somewhat unique in the Geosciences for at least two reasons: (1) the disciplines are dominated by highly dynamic phenomena, and hence the data are organized mainly (though not entirely) on events and time rather than primarily spatially; and (2) the science drivers in these disciplines are studied in depth and decided upon as a broad-based community endeavor culminating in a decadal survey report every 10 years. The most recent report, *Solar and Space Physics: A Science for a Technological Society* (National Research Council, The National Academies Press) was released in 2013, and serves as the main guide for science drivers examined during the workshop. None of the findings below are meant to conflict in any way with the national science goals outlined in this decadal survey.

In addition to science goals, the NRC Decadal Survey also recommended, as a high priority, the implementation of an integrated initiative (DRIVE) to develop critical new technological capabilities in order to address the decadal survey's complex scientific topics. In particular the decadal survey encourages the development of a "data environment that draws together new and archived satellite and ground-based solar and space physics data sets and computational results from the research and operations communities." This included "community oversight of emerging, integrated data systems" and "exploitation of emerging information technologies" with "virtual observatories as a specific component of the solar and space physics research-supporting infrastructure."

Science Issues and Challenges

Important science drivers:

The latest NRC Decadal Survey in Solar and Space Physics outlines four overarching key science goals for solar-terrestrial studies in the coming years. Below are more-focused science goals, consistent with the Decadal Survey goals, that we anticipate will benefit most from investments in cyberinfrastructure during the next 5 - 15 years:

- **Understanding the couplings among physically different domains ranging from the solar interior to the Earth's atmosphere**: The advent of "Big Data" (the aggregation of large, complex, heterogeneous data sets) in observations and numerical modeling holds promise for rapid progress in solar-terrestrial research. Space- and ground-based observatories will provide important constraints for models in terms of boundary conditions and synthetic observables. New observational data and computational advances provide new opportunities to develop cutting edge, data-driven models for the evolution of the magnetic flux below and above the solar surface, its influence throughout the heliosphere, and its impact at Earth. New cyberinfrastructure is required to improve our knowledge of the transfer of physical drivers across different physical domains from observational data and numerical simulations.

– **The study of the fundamental processes through which magnetic energy is generated, stored, released, and propagated**: This is critically dependent on an advanced cyberinfrastructure that enhances our ability to assemble, analyze, and visualize multi-instrument, multi-wavelength datasets covering multiple temporal and spatial scales in combination with detailed physical models. The application of computer vision and machine learning techniques to identify features across different physical dimensions and to better mine large, distributed databases will be needed to enable event identification and statistically driven analysis. Of particular interest is understanding the process of magnetic reconnection, the primary mechanism for energy release in solar flares and coronal mass ejections, which controls the occurrence and severity of magnetic storms through transport of mass, energy and momentum both at the sunward side of the magnetosphere and in the magnetotail.

– **Predicting the solar wind and Interplanetary Magnetic Field in the near-Earth environment.** Understanding the origin of magnetic flux structure at the Sun, and how it evolves during magnetic eruption and propagation through the heliosphere to produce the relevant spatial scale of B_z variation near Earth that drives magnetic storms, will depend critically on in situ and remote sensing observations from the *Solar Dynamics Observatory*, *Magnetospheric Multiscale, Solar Probe Plus* and *Solar Orbiter* and other spacecraft, as well as ground-based facilities, combined with modeling techniques capable of simulating CME flux ropes from the Sun to the Earth. The many disparate types of data and the broad range of spatial and temporal scales involved in both observations and models present a substantial cyberinfrastructure challenge.

– **Understanding the acceleration of particles throughout the Sun-Earth system.** Acceleration of electrons and ions, often to extremely high energies, is ubiquitous throughout the solar atmosphere, heliosphere, magnetosphere, and ionosphere, and creates hazards for humans and technological systems (spacecraft, communication and navigation systems, and even aircraft) everywhere within Geospace. In every region, important tasks remain, such as: identifying the acceleration mechanisms that operate in the various regions of the Sun-Earth system; determining which mechanisms are most important at different times and locations; identifying common *vs*. distinct mechanisms in different regions; identifying the more important plasma instabilities that operate in the different regions and the role they play in particle acceleration under varying conditions; and following the propagation of accelerated particles within and across regions of the Sun-Earth system.

- **Understanding and forecasting the effects of forcing** on the coupled Ionosphere-Thermosphere-Mesosphere (ITM) system. The ITM system presents a unique challenge in that strong coupling between charged and neutral species dominates physical processes. The system is responsive to external forces, e.g. reconnection, which impose global electric fields and magnetic currents, but also to internal processes, e.g. tropospheric heating and upward transmission of tidal forces, ionospheric instabilities, ion-neutral collisions and frictional drag. The coupled system demands cross-disciplinary study involving data acquired over multiple time and distance scales from ground and space observatories. Our ability to facilitate telecommunication and navigation, prevent catastrophic failure of the power grid during magnetic storms, or protect space assets from collisions demands accurate forecasting of the ITM response to forcing. Unique to this effort, international collaborations often require the participation of poorer countries with desirable locations for observations, but without the means to install instrumentation or distribute data in optimal ways.

Current Challenges to High-Impact, Interdisciplinary Science:

The main challenges identified by workshop participants center around bridging the gaps among the Geospace sub-disciplines, to foster interdisciplinary research.

Challenges in finding / discovering data

- Users do not know how to search for data across multiple repositories, and in general what data sets/resources exist. Data are hard to find, and even harder to transform into the form needed for further analysis.
- Semantic techniques should be available to enable broad discovery and use of data. Tools/libraries that enable the generation of metadata (annotations) in an automated fashion would be preferred.
- Joint data discovery ideally makes use of of centralized data repositories or search facilities where all the metadata (and pointers to the data) are queried and made available through a common interface. Complementary to this would be the implementation of a system based on semantic web technologies, which would require that a widely accepted standard vocabulary/ontology (suitable for our community) be put in place that the community agrees to abide to.
- There is a need for encouraging adoption and consistent usage of metadata standards for the essential attributes of both observational and modeling data sets, as well as an agreement on vocabulary to use.
- Getting to a set of “widely accepted standards” is itself a challenge. Also needed are translation tools (“ontology alignment”) between different sets of standards, especially where there are already multiple sets of established practices.
- The Geospace disciplines increasingly need better tools for mining our spatiotemporal datasets for features, both known and unknown
- The tools need to be scalable, to work for both large and small datasets.
- Data query: enabling the easy and effective querying of very specific subsets of data in order to tailor the results according to a specific science objective, thus reducing the volume of the data transfer. Good metadata and strong quick-look tools play a big role in this.
- Data volumes are becoming prohibitively large. It is not feasible to co-locate all data sets, or even apply the “old model“ of requiring users to download all the datasets of interest onto their own computers to manipulate them locally. Analysis increasingly needs to be co-located with the data, but this is problematic for analysis of multiple datasets, located in different places. Processing and user-driven analysis carried out at these large data centers may provide a solution to this coming problem, but mechanisms need to be in place to allow these providers to develop and support these (potentially costly) capabilities.

Challenges in working with data

- Continuity of data sets (both space and ground-based) over time has an increasing value as our ability to mine and probe these large data collections grows. Ensuring continuity should be a factor in funding decisions. (For example, there are concerns about several older instruments with no successor at the moment.)

- There are similar issues of continuity in the development of data analysis tools as well as instruments.
- Getting the most out of existing or legacy data; ensuring things do not get lost over time as missions or groups end.
- Information about assumptions, sources of error, and methodologies should be included along with the data.
- Need methods to ensure scientific reproducibility by allowing citation of specific data products and processing steps used in a scientific study.
- Need a mechanism for ensuring proper attribution of data sources in publications. It is critical to record provenance of all data to improve future reuse.
- Need better benchmarking/validation of data catalogs for researchers in different disciplines: it is important to have clear quality metrics that allow users to determine which data points are "good" or "bad" for their purposes.
- It is important not to "re-invent the wheel." If someone has "solved" a problem, other communities need to be able to find out about this and make use of it.
- The wide variety of analysis tools and languages in current use inhibits the development of a common set of analysis tools. Clearer documentation and use of software development best practices would help mitigate this confusion.
- There is a need for a strong leadership structure: a project should be run by a single, strong entity with broad community buy-in to ensure coordination.

Challenges in cross-disciplinary science / working with data outside our sub-discipline.

- Data from outside a researcher's field is difficult to find and learn how to analyze.
- An impediment to cross-disciplinary research is that while the same problems might be studied in different sub-disciplines, the observables, scales, and parameter regimes may be quite different.
- It is difficult to find sources of funding for cross-disciplinary research.
- Researchers using data from outside their areas of expertise need trusted catalogs of events and categorizations
- Data integration is needed to enable interfacing and interoperability among diverse datasets.
- Need better support for 'sun-to-mud' efforts. Solutions may be to have more common workshops, and classes offered online by multiple institutions.

Modeling-specific challenges

- It is important to compare and address discrepancies between data and models. Tools are generally not readily available to directly compare model outputs and observations.
- If these tools were available, iteration between modeling and data comparison could take place, allowing ongoing improvement of both.
- While data are often open and analysis code is sometimes open source, the same is not generally true for models (although it should be).
- In terms of modeling: there is a need for better flexibility/modularity in large model design so various groups could "plug and play" their components.

Educational, societal, and public outreach challenges

- There is a dearth of data-science and cyberinfrastructure-related content in the domain-specific academic curricula, impairing the ability of students to incorporate existing tools and best practices into their research.
- Scientists often do not know how to scale up their cyberinfrastructure usage from the desktop to make use of high-performance computing (HPC).
- Students and practicing researchers need training on how to use GPUs and other advanced computing resources.
- Scientists want to share their data in the public domain, but may worry about potential misuse or misinterpretation of the data.

Technical Issues/Challenges

Many of the interdisciplinary science challenges noted above are rooted in technical issues that must be addressed in order to successfully overcome them. The breakout sessions devoted to technical challenges included moderators who are computer scientists, in order to encourage new thinking.

- There is a need to develop computationally efficient capabilities for searching and expressive querying of Large/Diverse/Distributed Data Sets including provenance and data quality. What is of interest to scientists can be very complex to define. With today's high-volume databases, it is increasingly important to locate and download only the portion of data of interest. Propagation delays from one regime to another within the Geospace system make event searches challenging—e.g. how to do correlations to find linked events among data sets with such delays, without downloading all of the data.
- There will be a continuing need to discover, search, and utilize historical datasets, which must be preserved and, if necessary, modernized through metadata indexing to bring them into discoverable form.
- Data providers, especially new and actively maintained services, need to include well-documented APIs (application programming interfaces) and service interfaces, to aid in development of flexible workflows for utilizing the data resources.
- Some metadata standards already exist, but translators/converters are needed for searches bridging solar-terrestrial environments (solar, heliosphere, magnetosphere, ionosphere/upper-atmosphere) to promote interdisciplinary science. Additional efforts to agree on a wider standard of keywords, vocabulary and ontologies would be useful, but difficult.
- A platform and standards for data and software citations need to be further developed and widely adopted. A scheme for searching ranked databases and software according to popularity, usage, and quality would be a useful addition.
- Workshops/tutorials and academic curricula are needed to teach standard tools and techniques for interdisciplinary research to the community (e.g., orbital discovery tool). Community-developed toolkits (e.g. those at SolarSoft, sunpy.org, itk.org) are important sources of cross-platform tools for general analysis. Community involvement in further open-source tool development (e.g. through Github) should be strengthened and encouraged.

- Tools are needed for generalized Event/Object recognition in space and time, and for visualizing multi-dimensional data in large data volumes

Community Next Steps

Since this Solar-Terrestrial Cyberinfrastructure workshop occurred rather late in the process of Earthcube governance, we have the advantage of knowing the context of the program within which we should coordinate our efforts. Many of the challenges identified during the workshop have also been identified by other domain workshops, and hence our community can form Earthcube working groups or join with others already forming within Earthcube. In addition, our community can undertake the following steps, and also encourage NSF to provide Earthcube funding opportunities to address these areas:

Tools and Standards

- Make/collect a list of useful tools and services (with user reviews)
- Provide additional tools for generating metadata from existing data and manipulating metadata in the form of plots, indexing
- Support development of community-led general analysis toolkits
- Provide translators between standard data formats
- Provide translators between metadata (e.g. keywords) standards
- Develop standard service interfaces (such as APIs)
- Develop “one-stop-shopping facility” to aggregate data, or facilitate ordering/delivery of data

Cross-disciplinary CS/domain scientist collaborations

- Assemble domain scientists and computer scientists to attack specific and realistically achievable high-value science goals as identified by the decadal survey
- Identify and list the most widely-used data-sets in the relevant disciplines and design data integration tools according to the above-mentioned science goals
- Create hyper-dimensional visualization tools
- Develop the capability for advanced semantic queries for nearest-neighbor matching of widely dissimilar data
- Develop the capability to construct queries of what is missing (identifying gaps and dealing with intermittency in data coverage)

Education (community and academic)

- Adding cyberinfrastructure and computer visualization components to solar-terrestrial curricula.
- Educating domain scientists on scaling up their applications from desktop to HPC
- Access to HPC resources for training in solar-terrestrial research
- Education on how to utilize GPU and other advanced computing resources
- Advanced data analysis techniques (e.g. inverse theory, forward fitting, data assimilation)

Data management

- Searching and querying long-term archived databases with access control and provenance
- Use of DOIs and alternatives for data and software citations
- Tools and standards for creation of metadata that tracks database use (who, for what purpose, popularity)
- Cloud storage and HPC processing
- Support for creation, population, and operation of new databases based on new instruments and modeling efforts
- Capability for creation of quick-look data products

Model input/output

- Develop techniques for data-assimilation, data-driven modeling, and cross-domain model coupling
- Metadata concepts for model output (descriptive of format)
- Develop standards and guidelines for making model output shareable and comparable
- Search tools for integrating observational and model output data

Quantifying data quality

- Include valid error estimates together with data
- Include information about data quality, completeness, and fitness for use
- Research methods and practices for quantifying errors (random, systematic)
- Biases introduced by data processing

Encouraging good practices

- Study feasibility of creating cloud-storage for data, whose use would enforce good practices as a prerequisite for use
- Create or join an Earthcube working group to identify and share information and tools for enforcing metadata standards
- Include software engineering and development techniques as part of academic training