EXECUTIVE SUMMARY: WORKSHOP RESULTS

(Liping Di, George Mason University)

Earth Cube Workshop Title: Engaging the Atmospheric Cloud/Aerosol/Composition Community

Introduction: Scientists working on the atmospheric cloud/aerosol/composition (ACAC) domain typically develop theories, models, and predictions on the state and dynamics of the atmosphere and its constituents by acquiring, processing, analyzing, integrating, assimilating, and modeling with data from diverse, multi-disciplinary sources, both in-situ and through remote sensing methods. The volumes of the data used in the research can be small or very large ("big data") and the data could be from live sensors, archived at the big data centers, or at the hand of individual scientists. Such diversity on the data poses great challenges to ACAC scientists on their research and education activities. Therefore, common cyber-based infrastructures, such as EarthCube, for handling and managing diverse data and facilitating information extraction and knowledge discovery from the data are urgently needed. The purpose of the workshop is to gather community inputs on current challenges and requirements to the EarthCube.

A total of 67 scientists from the ACAC community participated in the workshop. Among all participants, 60% of them are from universities and 40% from government agencies, industry, and non-profit research centers. All of the participants are affiliated with domestic organizations of the U.S. The main outcomes of the EarthCube workshop discussions are summarized below.

SCIENCE ISSUES AND CHALLENGES

1. **Important science drivers and challenges:** Participants identified several high-priority science questions that will be the focus of interdisciplinary efforts during the next 5-15 years.
    - What are the sources and the removal mechanisms of chemical species in the atmosphere?
    - A lot of work has focused on improving Ozone but other species, for example, Methane and NO_x, have been neglected in the process. The entire atmosphere system is sensitive to changes in NO_x and needs to be considered. What are the effects of industrial NO_x on atmosphere composition?
    - What are the exact roles of the clouds in the cloud systems and in the entire earth system? Several outstanding cloud-related deficiencies in the climate modeling are well documented by the research community and need to be addressed in the next 5-15 years, including the double ITCZ problem, poor MJO, too short and too regular ENSO periodicity, diurnal cycle and frequency of precipitation, inconsistent representation of radiation and clouds.
    - How do clouds affect the cloud feedback on climate sensitivity?
    - What is the role of clouds on biosphere or ecosystems or vice versa?
    - What is the spatial, temporal, size distribution and composition distribution of aerosol particles in the atmosphere and the aerosol particle emissions globally?
    - What are the exact roles of aerosols in the cloud and climate?
    - What is the impact of aerosol on severe marine storms?
    - What are the changes to Cloud Condensation Nuclei (CCN) with changes in aerosol loading?

2. **Current challenges to high-impact, interdisciplinary science:** The participants, through mentally exercising inter-disciplinary research procedures for solving the above mentioned science challenges, came out a set of consistent barriers and challenges that prevent the above-mentioned science challenges from being solved easily.

- The inter-disciplinary research requires the use of diverse data from diverse sources. Significant challenges still remain for scientists to discover, access, integrate, and use those data in their research.
- Long-term Earth observation through remote and in-situ sensors is one of the major data sources for the inter-disciplinary ACAC research. However, the continuity of satellite & sensor is an issue since the current fleet is aging and not sure what the future holds, including the transition from research-based satellites (NASA) to more operational (NOAA) platforms. Another issue is how to obtain and use data from satellites of other countries. Currently there are multinational efforts (e.g., Europe, US, and China) on satellite measures of ACAC. Enabling the interoperability of data between those efforts and the U.S. efforts is a challenge.
- More data products and/or higher data resolution mean a lot of more data needs to be transferred. This introduces the bandwidth issue both for satellite downlinks and at the user-end (internet). Increased onboard computing power (compression, data sampling) may address this. Combining operational and research platforms is a challenge. The research community is competing with an exponentially increasing data hungry society (Netflix, online streaming, telecommunications, remote working, etc.).Who pays the bills for the mounting cost of the network backbone support?
- Comparing to satellite remote sensing data, in-situ data are less accessible (but high value content) data. Most often it is only accessible by personal requests, making it very hard to develop a harmonized dataset for regional and global models. The model data are by far the least accessible (but also high value content) data. The un-accessibility of model data makes the model diagnoses and inter-comparison difficult.
- Inadequate metadata on data quality and provenance makes scientific use of data from other sources difficult since scientists have hard time to understand if the obtained data are useful. No standard is available yet as to when data is useful (some properties, e.g., ice crystal number are orders of magnitude off, so perhaps that may not be useful)
- The inter-disciplinary research often requires integration of data across sensors and platforms. However, not much has been accomplished in this area yet. The major issue is co-location of sensors with each other and models with sensors. Recent progress on sensor web technology may relieve this issue a bit.
- Many needed global datasets are not yet available or with bad quality. An example of such datasets is the cloud hydrometeors (crystal number, droplet number and cloud phase) for statistical evaluation of models. Aerosol cloud particle precursors climatologies are lacking (CCN, IN) but that is becoming better. Vertical velocities are critical but nonexistent. Cloud lifecycle datasets are nonexistent. The lack of datasets impedes our ability to evaluate conceptual models of cloud development and aerosol-cloud interactions.
- Significant insider knowledge on the data and IT skills are needed for full utilizations of these data resources. No adequate tools and services are available for readily integrating data from multiple sources and across disciplines.
- Modeling is the major method in the inter-disciplinary ACAC research. Data and products obtained from Earth observation sensors are extremely useful in the model initialization, verification, and validation, and as model constraints. However, it is a challenge to make sensor observation data easily consumed by models. Common methodology, framework, and tools for easy sensor-model coupling and integration are needed.
- Lack of community standards (or too many standards) on data format, file types and metadata make the interoperability and sharing of interdisciplinary datasets difficult.
- Many current researchers have not been trained to collect and report data in an interoperable way so that it can be reused by others. There is also lack of formal mechanism for rewarding scientists who share their data, algorithms, and services.

TECHNICAL INFORMATION/ISSUES/CHALLENGES

Data Access Challenges

- Different types of users need different types of support (some, for example in developing countries, just want to import data into Excel)
- Cross-community access is the biggest challenge (within an domain community, it is generally understood how to work with data formats and tools)
- Need to understand the data characteristics (quality, provenance)
- Enable scientists to find relevant, reliable data regardless where the data are archived and obtain the data in the form specified by the scientists
  1) Search by location, date, topics, etc.
  2) On-line services for providing automated data customization
  3) Globally available data; data in other country's agencies
  4) Able to integrate data from different platforms and repositories
  5) System interoperability (inter-agency and international sharing)
  6) Better metadata and standards for data understanding and usage
- Can EarthCube provide a better search engine tailored for communities?

Non-domain Understanding

- Challenge is in having users understand the uncertainty and errors associated with data.
- Documentation is critical. Needs to be understood by others outside of the immediate discipline that created the data.
- Data processes change over time
- Need education/training about data (perhaps a service EarthCube could provide?)
- Can EarthCube fund training for cross-disciplinary data management and informatics?

Supporting small datasets in EarthCube

- Many groups, (e.g., research labs) have small, individually maintained datasets and do not have a large infrastructure to support them in the publication and management of them
- EarthCube should support these small datasets
- One example, might be an EarthCube sponsored cloud data management and publication service to simplify the process for smaller groups

Supporting Data Management

- Challenge is in funding data management - for example, many research groups don't have the funding or time to do metadata creation.
- Interest in an EarthCube supported cloud service or tools
- Employ people within EarthCube who have library science and similar skills to help organize and provide access to data

Data Quality and Standards

- Need a set of EarthCube recommended standards and best practices to facilitate the interoperability and sharing.
- Bad data needs to be flagged
- Need a rating system to help determine and convey what is the quality and type of published data.
- User need to understand when they're using data at their own risk or when it is peer reviewed
- Need mechanisms to catch uncertainties and errors in data before and after they are published
- Unique dataset IDs are created to link datasets to publications and datasets to each other, Suggestion to only provide a dataset an ID if it is peer reviewed and determined to be acceptable.
- Should provide supporting documentation that describes how dataset was derived (algorithms, software used, etc.). And need to track data processes as they change over time.

Supporting modeling and integrated analysis

- On-line data integration and analysis services
- Tools and services to manage, archive, and disseminate model outputs for facilitating modeling comparison
- Sensor-model coupling for facilitating model verification and validation with observation data
- Sensor web and models as services

Merging Existing and New Infrastructures

- Need to transition existing systems to EarthCube, not requiring an overhaul of existing systems
- Need translators, converters, adaptors
- Strive for common standards and practices where most effective

Assessment of current tools and cyberinfrastructure capabilities/best practices

- Provide testbed of developing cyberinfrastructure components, tools, systems, and etc.
- Provide tools to help professors incorporate new datasets and tools into classes.
- Create an EarthCube working group to be involved in assessing value and usability of tools and improve teaching with actual data.
- Focus on the establishment and enforcement of metadata standards.
- Enforce the use of common accepted metadata language.
- Focus more on the development of extended metadata and less on data formats.
- Support the development of converts, translators and readers from one data format to metadata and back to data in a different format.
- Define the languages to read metadata.
- Support interoperability for hardware and software components.
- Improve understanding of users and their needs: students, scientists, public, government agents, university professors, etc.
- Maintain a standard digital object identifier (DOI) system. This is a character string (a "digital identifier") used to uniquely identify an object such as an electronic document.
- Improve access methods. Focus on the development of tools and schemes that help users to find the data and products that they need and show how to access them.
- Is it possible to build one step for all?
- Provide information and access to data properties and quality.
- Favor interdisciplinary approaches.
- Search for ways and methods for multi-sensor, data and products combination.
- Need for instrument simulation capabilities.
- Allow integration and retrieval of data and products.
- Favor long-term continuity of data and products (the issue of trust and lineage between different data sets).
- In case of data/products gap, offer ways to fill them.
- Provide tools for data/products discovery, analyses and integration.
- Enforce documentation standards.
- Provide visualization tools.

Making scientists easy to contribute their data and sources to EarthCube

- A set of easy-to-use tools for scientists to document and publish their data and cyber-resources (e.g., algorithms, models, and computing facilities) for sharing
- The academic community needs to change the way for valuing the academic achievement. Researchers should get credit for sharing their data and resources.