Geophysical Research Abstracts
Vol. 19, EGU2017-17396-1, 2017
EGU General Assembly 2017


# Enabling a new Paradigm to Address Big Data and Open Science Challenges

Mohan Ramamurthy and Ward Fisher
University Corporation for Atmospheric Research, Unidata, Boulder, United States (mohan@ucar.edu)

Data are not only the lifeblood of the geosciences but they have become the currency of the modern world in science and society. Rapid advances in computing, communi¬cations, and observational technologies — along with concomitant advances in high-resolution modeling, ensemble and coupled-systems predictions of the Earth system — are revolutionizing nearly every aspect of our field. Modern data volumes from high-resolution ensemble prediction/projection/simulation systems and next-generation remote-sensing systems like hyper-spectral satellite sensors and phased-array radars are staggering. For example, CMIP efforts alone will generate many petabytes of climate projection data for use in assessments of climate change. And NOAA's National Climatic Data Center projects that it will archive over 350 petabytes by 2030.

For researchers and educators, this deluge and the increasing complexity of data brings challenges along with the opportunities for discovery and scientific breakthroughs. The potential for big data to transform the geosciences is enormous, but realizing the next frontier depends on effectively managing, analyzing, and exploiting these heterogeneous data sources, extracting knowledge and useful information from heterogeneous data sources in ways that were previously impossible, to enable discoveries and gain new insights. At the same time, there is a growing focus on the area of "Reproducibility or Replicability in Science" that has implications for Open Science. The advent of cloud computing has opened new avenues for not only addressing both big data and Open Science challenges to accelerate scientific discoveries. However, to successfully leverage the enormous potential of cloud technologies, it will require the data providers and the scientific communities to develop new paradigms to enable next-generation workflows and transform the conduct of science. Making data readily available is a necessary but not a sufficient condition. Data providers also need to give scientists an ecosystem that includes data, tools, workflows and other services needed to perform analytics, integration, interpretation, and synthesis - all in the same environment or platform. Instead of moving data to processing systems near users, as is the tradition, the cloud permits one to bring processing, computing, analysis and visualization to data – so called data proximate workbench capabilities, also known as server-side processing.

In this talk, I will present the ongoing work at Unidata to facilitate a new paradigm for doing science by offering a suite of tools, resources, and platforms to leverage cloud technologies for addressing both big data and Open Science/reproducibility challenges. That work includes the development and deployment of new protocols for data access and server-side operations and Docker container images of key applications, JupyterHub Python notebook tools, and cloud-based analysis and visualization capability via the CloudIDV tool to enable reproducible workflows and effectively use the accessed data.