DATA-DRIVEN SCIENTIFIC WORKFLOWS

A Summary of New Technologies and Datasets Explored at the Unidata 2018 Workshop

Kevin H. Goebbert, John T. Allen, Victor A. Gensini, and Mohan Ramamurthy

Every three years, the Unidata Users Committee organizes a weeklong workshop that delves into important issues surrounding computing and the broader atmospheric sciences. The 2018 workshop theme, "Reducing Time to Science: Evolving Workflows for Geoscience Research and Education," focused on the challenge of accessing ever-growing datasets from across the atmospheric sciences in an efficient manner and using robust computing techniques to investigate those datasets. A primary goal of the workshop was to illustrate methods to reduce "time to science" for researchers by using cloud computing and the Python programming language to facilitate scientific workflows. This was accomplished through a combination of lectures and hands-on activities that demonstrated some of the topics discussed.

2018 UNIDATA USERS WORKSHOP

What: Over 70 participants from academia, research organizations, and industry interested in how they can evolve their workflows met to learn how to utilize the latest computing technology to access and visualize different datasets across the atmospheric sciences.
When: 25–28 June 2018
Where: Boulder, Colorado

DATA-PROXIMATE WORKFLOW. The increasing size of datasets and how to facilitate timely analysis of these data are an ongoing challenge for atmospheric science and, more broadly, the geosciences. Data-proximate approaches address this challenge by allowing users to query or analyze data without retrieval to their own devices or local storage. The growing size of climate model output is one example of this problem, with simulations from phase 6 of the Coupled Model Intercomparison Project (CMIP6) expected to exceed 20 PB in stored data. One approach to address this challenge was demonstrated using the European Network for Earth System Modelling (ENES) Climate Analytics Service, a methodology that focuses on providing simple access mechanisms for researchers to interface with remotely hosted multimodel climate simulations. This framework allows for batched or parallel data-proximate processing for interpretation of these data using a series of prebuilt operations that are capable of chain application. The main goal of this approach is to reduce the need to download independent copies of these datasets.

Extending this more broadly to commercially available cloud providers, an interesting presentation

AFFILIATIONS: Goebbert—Valparaiso University, Valparaiso, Indiana; Allen—Central Michigan University, Mount Pleasant, Michigan; Gensini—Northern Illinois University, De Kalb, Illinois; Ramamurthy—Unidata, UCAR, Boulder, Colorado
CORRESPONDING AUTHOR: Dr. Kevin H. Goebbert, kevin.goebbert@valpo.edu

DOI:10.1175/BAMS-D-18-0265.1

In final form 28 September 2018

from Amazon Web Services (AWS) demonstrated the feasibility of approaches extending from prebuilt machine learning tools in AWS Sagemaker. The speaker described different ways to approach parallel instances and sensitivity model simulations using prebuilt instances of the Weather Research and Forecasting (WRF) Model through dedicated clusters in the AWS cloud. Also highlighted were the substantial range of options to access existing datasets (via the Big Weather Web) and the potential to provide economical long-term redundant storage (Glacier) for research applications. Perhaps of greatest interest to attendees were the details of the emerging options for serverless computing, which allows researchers to "plug and play" their own code to perform analysis—without the requirement of creating a new virtual cluster.

The potential for relatively inexpensive or limited free time via the AWS Research Cloud Program (https://aws.amazon.com/rcp) was also discussed and provides an excellent opportunity to try out cloud computing or use these resources for education. An application of cloud implementation for data analysis was provided by the Pangeo project (https://github.com/pangeo-data/pangeo), focusing on providing efficient analysis tools in Python to analyze the large datasets found across the geosciences. Attendees worked through examples analyzing oceanic sea surface temperatures and multiterabyte climate data using only a few lines of code. Most impressively, all attendees were able to log onto a single server managed via Jupyterhub and perform an analysis of this large dataset in parallel, with a data-proximate workflow on the remote cloud server, highlighting the substantial progress toward more efficiently handling large data analysis for research applications and collaborative computing at institutions around the country.

MACHINE LEARNING APPLICATIONS. Machine learning is rapidly becoming a popular research methodology in atmospheric science, and the growing availability of large datasets in cloud environments [e.g., Next Generation Weather Radar (NEXRAD) via the Big Weather Web] is further driving this growth. While various interpretations and approaches exist, the potential to handle multidimensional data and apply statistical methodologies to extract signals from large datasets suggests considerable opportunities. A perspective from Google offered new methodologies that can allow mining of large datasets or incoming data streams, for example, using convolutional neural networks (CNNs) to classify coastlines or image data. Another example more germane to atmospheric science included the use of CNNs to provide insight into postprocessing of model data by identifying simulated mesocyclones, using more filters than just updraft helicity or reflectivity, and offering ways forward to handle ensemble data. In a novel approach, CNNs were also adapted to classify types of mesoscale convective systems from the National Oceanic and Atmospheric Administration (NOAA) NEXRAD imagery archive, showing considerable promise in identifying these features. Given the widespread use of maps, images, and graphics across the geosciences, there is no doubt that various aspects of machine learning will be used to examine large weather and climate datasets in the coming years.

ACCESS AND USE OF *GOES-16* DATA. In recent years, there has not been a more anticipated new dataset than that from the latest series of geostationary satellites [Geostationary Operational Environmental Satellite (GOES)-R series]. This new generation of satellites provides increased spatial, temporal, and wavelength coverage, vastly increasing opportunities for operational meteorology, teaching, and research. A series of presentations provided information about new aspects of *GOES-16*, how it is used in the weather service to create derived products, an introduction to the Global Lightning Mapper (GLM), and how these new datasets can be used to teach elements of atmospheric radiation (Fig. 1).

The critical piece for any dataset is its accessibility. Workshop participants were introduced to a number of different methods for accessing data emanating from *GOES-16*, including the Thematic Real-Time Environmental Distributed Data Services (THREDDS) server hosted by Unidata. Data are primarily distributed in netCDF format, which allows the files to contain full metadata. There are two primary ways in which the observations are stored within the files: 1) as radiance values and 2) as reflectance/brightness values.

The radiance data are the primary data coming across the GOES Rebroadcast (GRB) feed and can be used to teach elements of radiation including the Planck function and Stefan–Boltzmann law. Workshop attendees participated in a hands-on session that used the functionality of the Python programming language in the Jupyter Notebook environment to remotely access *GOES-16* data, calculate IR brightness temperatures using the inverse Planck function, and display a final image.

ENSEMBLE MODEL VISUALIZATION AND ANALYSIS. One important development in meteorology has been the use of various ensemble techniques

for addressing uncertainty in weather forecasts. There is no doubt that ensemble systems are the way of the future, and several presentations discussed opportunities for condensing large amounts of ensemble information into easy-to-use graphical interfaces. For example, a convective-allowing ensemble designed for short-term prediction of severe convective storms uses the Data-Driven Documents (D3) Javascript library to reduce data transfer and provide a seamless web browser experience for users. In addition, the widely used Python programming language is becoming a go-to option for researchers wishing to display a variety of ensemble postprocessed numerical weather prediction output (e.g., paintball plots, probability of exceedance plots, postage stamps, and spaghetti plots). Two hands-on sessions allowed attendees to try out some of these techniques and see how they might be able to use them for their own workflows.

Fig. 1. Workshop presenter Fred Mosher from Embry–Riddle Aeronautical University describes data available from *GOES-16*.

BENEFITS AND OUTCOMES. Workshop attendees were exposed to a wide range of ideas that take advantage of the latest technologies to conduct scientific inquiries. The demonstrations of data-proximate workflows challenged attendees to think differently about how we access and process increasingly large datasets, especially when it does not make sense to move the datasets around. Additionally, attendees learned about the increasing use and viability of cloud platforms that can economically assist in reducing the time to science by allowing for scaling of resources on demand to complete tasks in a more time efficient manner. The generation of new datasets, like those available from the new GOES-R series, and the ever-expanding suite of model output, including ensemble forecasts, mean we must rethink how we use, analyze, and visualize data within the field. Attendees gained a great appreciation for new and innovative methods that can be readily incorporated into their research and educational practices.

ACKNOWLEDGMENTS. The Unidata Users Committee and the Unidata Program Center would like to acknowledge NSF's support in funding this workshop through Award 1822272. In addition, the workshop would not have been possible without the generous contribution of time and expertise by the many workshop speakers and participants and the assistance of staff.