EOS
EOS, TRANSACTIONS, AMERICAN GEOPHYSICAL UNION

VOLUME 94 NUMBER 32
6 August 2013
PAGES 277–288

Computational Earth Science: Big Data Transformed Into Insight

PAGES 277–278

More than ever in the history of science, researchers have at their fingertips an unprecedented wealth of data from continuously orbiting satellites, weather monitoring instruments, ecological observatories, seismic stations, moored buoys, floats, and even model simulations and forecasts. With just an internet connection, scientists and engineers can access atmospheric and oceanic gridded data and time series observations, seismographs from around the world, minute-by-minute conditions of the near-Earth space environment, and other data streams that provide information on events across local, regional, and global scales. These data sets have become essential for monitoring and understanding the associated impacts of geological and environmental phenomena on society.

This increasing amount of data has led us into the era of "big data," or the "fourth paradigm," as described in essays based on Jim Gray's vision of data science in the book *The Fourth Paradigm: Data-Intensive Scientific Discovery* [*Hey et al.,* 2009]. Big data, however, brings an inherent problem: How can researchers extract usable information from such overwhelming quantities of numbers and variables?

To help better understand, describe, and model data, scientists need an effective means of analyzing massive amounts of it. Doing this efficiently involves, at its heart, the use of computer programs, machine learning, and statistical techniques that view and analyze Earth and environmental events as "objects."

Object-Oriented Analysis

An object can be thought of as an identified item, event, or instance with distinct attributes and statistics representing the existence of the entity in space and/or time—for example, a storm, an earthquake, an ecological region, or a sea surface temperature anomaly. The attributes and statistics associated with these objects can be analyzed using statistical modeling algorithms to identify structural relationships between different characteristics, as well as time periods corresponding to different physical systems and phenomena interactions, leading to enhanced knowledge of what trends the data hold.

Specifically, object-oriented data analysis can be thought of as the study of the statistics of populations of objects. The analysis can include defining objects contained within digital images or photographs, gridded data sets, and animations, which allow for the objects to be analyzed over time and space.

If such algorithms are run in a computer environment designed to home in on characteristics of objects or events of interest, then the data can be crunched even more efficiently, allowing insights from big data to be revealed at a quicker pace. Such machine learning evolved from artificial intelligence research and focuses on developing models that are based on the behaviors and characteristics of empirical data. Capturing the behaviors and characteristics from data and determining their underlying probability distributions can provide new knowledge regarding the object or characteristic of interest. Typically, the properties or "true" underlying probability distributions of the observed variable of interest are not explicitly known. However, by seeking to define or describe these underlying probability distributions, data mining can help scientists learn or discover unknown properties and patterns contained in the data. This is particularly useful with complex systems and data sets.

The object-oriented approach has been widely used in various Earth and environmental science fields, from researching geographical terrain morphology [*Mitasova et al.,* 2012] to determining better methods for verifying the forecasts of numerical weather prediction (NWP) models [*Davis et al.,* 2006a, 2006b; *Mittermaier and Bullock,* 2013] and tornado forecasting [*Clark et al.,* 2012]. The

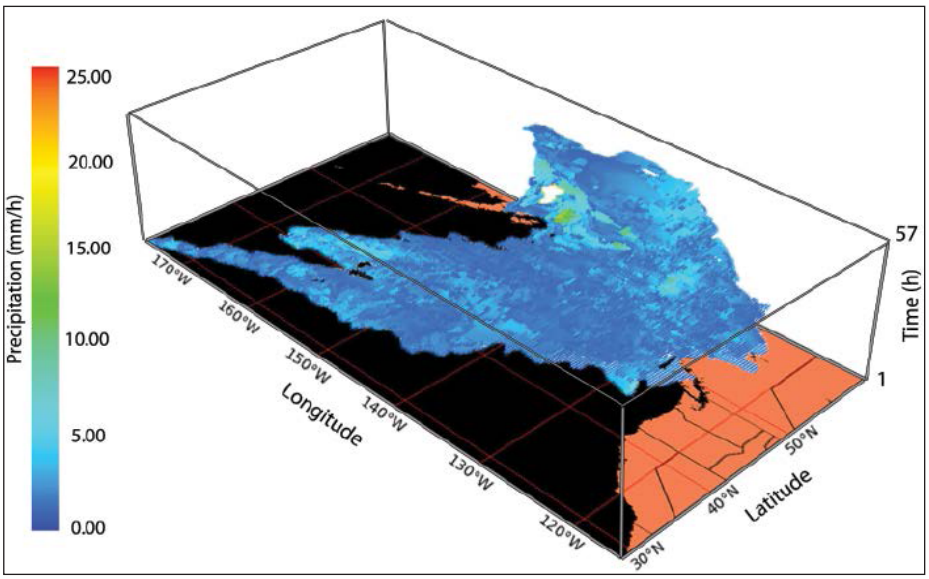

Fig. 1. A connected four-dimensional atmospheric river, or "precipitation object," extracted from the PostgreSQL database. The atmospheric river originated in the eastern Pacific and affected the western United States from 28 to 30 December 2005.

By S. Sellars, P. Nguyen, W. Chu, X. Gao, K. Hsu, and S. Sorooshian

recent work by *Mitasova et al.* [2012] uses object-based methods with lidar terrain data to scientifically visualize landscapes and landforms.

In the atmospheric sciences, *Davis et al.* [2006a, 2006b] describe the challenges encountered with traditional verification methods used with NWP models, which involve determining model performance scores (e.g., based on pixel-to-pixel comparisons, probabilities of accurately detecting weather that does occur, or the rate of false alarms), and they demonstrate how geometric and object-oriented attributes, such as shape, orientation, and size of meteorological fields such as precipitation, provide more informative results that help ease efforts to validate the performance of NWP models.

The focus of *Davis et al.* [2006a, 2006b] is verifying and describing the performance of model forecasts compared to observations using object-based methods. Machine-learning algorithms can then be used to learn from these populations of precipitation object attributes to better understand the underlying processes associated with the evolution of the population. As a result, state-of-the-art object-relational databases and machine-learning approaches may provide new and creative research insights and allow scientists to extract enhanced information from data.

Building an Object-Oriented Analysis Tool: A Case Study With Precipitation Data

Object-oriented approaches are becoming increasingly valuable for studying climate and weather. In particular, precipitation, or the lack thereof, with its direct connection to the hydrological cycle, is not only the most common atmospheric phenomena affecting society but also one of the most commonly monitored variables. In recent decades, it has become possible to observe precipitation processes in detail at near-global coverage, through advanced satellite sensor platforms, sophisticated retrieval algorithms, the use of ground-based radars, and integrated observational gauge networks.

However, even though there have been significant advances in observation, modeling, and prediction over the last few decades, there is still a significant amount of information that is unknown about the fundamental processes that govern shifts in climate, which cause widespread variability in drought and/or extreme flooding [*Bader et al.,* 2008]. To provide a new perspective on these governing processes, the researchers at the Center for Hydrometeorology and Remote Sensing (CHRS) at the University of California, Irvine designed an approach embracing the "fourth paradigm" as envisioned by Jim Gray.

CHRS's data-intensive object-oriented approach allows for the accomplishment of three objectives: (1) transformation or segmentation of the precipitation data into decipherable units—large-scale storm systems such as typhoons, hurricanes, atmospheric rivers, or even a sustained but gentle rainfall—with defined event characteristics; (2) organization of the data into an advanced database that contains the segmented precipitation objects and their associated characteristics; and (3) application of machine-learning algorithms for learning from the data.

How does this work? In computer science, a data point can be represented on a three-dimensional (3-D) grid (latitude, longitude, and time) as a volumetric pixel or "voxel," which has recently been increasingly used for graphical applications such as 3-D terrain features in computer games and medical imaging. In CHRS's method, instead of employing the traditional data analysis approaches, which look at information at each data point within a data set, scientists focus on using an object-based connectivity algorithm. This algorithm organizes data points into 4-D objects (latitude, longitude, time, and intensity) that better characterize the events and their structure.

The algorithm is designed to ensure that all voxels of precipitation estimates are connected in both space and time, allowing for the feature to be analyzed as a 4-D object (for example, the atmospheric river in Figure 1, identified by the algorithm as a precipitation object). In other words, at each time step, an object consists of connected voxels in direct neighborhood locations during that time step and also in the previous and future time steps. Organizing the data into a 4-D object helps one to visualize the dynamical changes to the precipitation object in time and space, enabling empirical characteristics to be calculated for each object.

This higher-dimensional approach is a departure from the traditional time series and grid-based methods of analyzing precipitation because it segments different objects into different cross-relational data sets. This object-relational database, called PostgreSQL, is searchable and can handle complex search queries—including geographical location, object characteristics, and temporal considerations—which allows for manipulation and subsetting of the data in a meaningful way.

The data that have been used to test this approach are the near-global precipitation estimates generated by an algorithm called the Precipitation Estimation from Remotely Sensed Information Using Artificial Neural Networks (PERSIANN), which processed observations from 60°N to 60°S globally from 1 March 2000 to 1 January 2011 every hour, at 0.25° resolution in both latitude and longitude. The estimates from PERSIANN are stored in a PostgreSQL database hosted at CHRS. PERSIANN uses artificial neural networks to learn on its own and estimate precipitation rates from infrared geostationary satellite data [*Sorooshian et al.,* 2000; *Hsu et al.,* 1997].

The connectivity algorithm is then applied to the PERSIANN data to segment out precipitation events. To remove a huge number of small precipitation events for this application, three criteria were chosen for an object to be considered an event. The algorithm applied a minimum threshold of precipitation intensity at 1 millimeter per hour (mm/h), applied a duration threshold of precipitation at 24 hours (i.e., the precipitation event must exist for at least 24 hours), and required connectivity to neighboring voxels along at least one face of the voxel's cube (which means that if two voxels have one face connection, they are grouped into the common object). The current version of the data set contains 55,173 precipitation objects worldwide for the nearly 10-year period. For access to the files containing the 55,173 objects and additional documentation, visit http://chrs.web.uci.edu/research/voxel/intro.html.

A descriptive statistics algorithm is then applied to all objects contained within the database. These descriptive statistics are stored in an $N \times d$ "design" matrix, where N is the number of objects and d is the number of dimensions (or features/characteristics). In statistics the design matrix contains the explanatory variables to be used in statistical models. Characteristics in the $N \times d$ matrix include: volume (in cubic meters), maximum precipitation intensity (mm/h), average intensity (mm/h), duration (h), average speed (kilometers per hour), and center of mass, or "centroid" (latitude and longitude coordinates).

More characteristics of the precipitation objects will be added to the database in the future, such as their axis angles, aspect ratios, curvature, track, connectivity index, area index, shape index [*AghaKouchak et al.,* 2011], time of year (season), atmospheric wave number that describes the large-scale energetic properties of the atmosphere surrounding the feature, El Niño–Southern Oscillation phase, meteorological classification (e.g., tropical cyclone or extratropic cyclone mesoscale convective system), etc. The $N \times d$ matrix is in optimal form for machine learning, experimental designs, and statistical modeling in that there is no limit to the number of variables that can be added. This optimal form would allow time periods and statistics to be compared with various climate indices or indicators to diagnose and/or determine possible physical processes governing the precipitation object characteristics.

From Big Data to Big Discoveries

Ever-increasing amounts of data are out there, waiting for interpretation. Learning from these data in a manner that is organized and interpretable requires enhanced computation abilities to detect patterns within data. This will enable scientists and decision makers to gain greater confidence in newly discovered knowledge of the fundamental causes of change to the environment. The advancement of such fundamental knowledge will assist in quantifying how these

changes affect society and may help identify future sustainability options that can be used for the development of adaptation and mitigation strategies to protect society.

In the case of CHRS's efforts, the information differentiated and cross-correlated within the PostgreSQL database of remotely sensed precipitation objects will support other potential applications in many science and engineering fields, including climate, oceanography, and weather studies; the development of new hypotheses around global precipitation; and further development of methods, models, and knowledge within different disciplines. The approach can be seen as an attempt to bridge computational sciences and Earth sciences and to provide new data sets to be explored by the various communities as they work to solve world problems related to environmental change. This, and similar endeavors in other fields, is where big data can take scientific discovery as a whole into a new paradigm of integrated data-intensive investigation and research.

Acknowledgments

We thank the University of California, Irvine Graduate Division's Public Impact Fellowship program, the Cooperative Institute for Climate Studies (CICS) (NOAA award NA09NES4400006), the U.S. Army Research Office (Award W911NF-11-1-0422), and NASA (award NNS09AO67G) for their contribution and financial support. We also sincerely thank Dan Braithwaite for IT support.

References

AghaKouchak, A., N. Nasrollahi, J. Li, B. Imam, and S. Sorooshian (2011), Geometrical characterization of precipitation patterns, *J. Hydrometeorol., 12*(2), 274–285, doi:10.1175/2010JHM1298.1.

Bader, D., C. Covey, W. Gutowski, I. Held, K. Kenneth, R. Miller, R. Tokmakian, and M. Zhang (2008), Climate models: An assessment of strengths and limitations, report, U.S. Dep. of Energy, Washington, D. C. [Available at http://digitalcommons.unl.edu/usdoepub/8.]

Clark, A., J. Kain, P. Marsh, J. Correia, M. Xue, and F. Kong (2012), Forecasting tornado pathlengths using a three-dimensional object identification algorithm applied to convection-allowing forecasts, *Weather Forecast., 27,* 1090–1113, doi:10.1175/WAF-D-11-00147.1.

Davis, C., B. Brown, and R. Bullock (2006a), Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas, *Mon. Weather Rev., 134,* 1772–1784, doi:10.1175/MWR3145.1.

Davis, C., B. Brown, and R. Bullock (2006b), Object-based verification of precipitation forecasts. Part II: Application to convective rain systems, *Mon. Weather Rev., 134,* 1785–1795, doi:10.1175/MWR3146.1.

Hey, T., S. Tansley, and K. Tolle (Eds.) (2009), *The Fourth Paradigm: Data-Intensive Scientific Discovery,* Microsoft, Redmond, Wash.

Hsu, K., X. Gao, and S. Sorooshian (1997), Precipitation estimation from remotely sensed information using artificial neural networks, *J. Appl. Meteorol., 36,* 1176–1190, doi:10.1175/1520-0450(1997)036<1176:PEFRSI>2.0.CO;2.

Mitasova, H., R. S. Harmon, K. W. Weaver, N. J. Lyons, and M. F. Overton (2012), Scientific visualization of landscapes and landforms, *Geomorphology, 137,* 122–137, doi:10.1016/j.geomorph.2010.09.033.

Mittermaier, M. P., and R. Bullock (2013), Using MODE to explore the spatial and temporal characteristics of cloud cover forecasts from high-resolution NWP models, *Meteorol. Appl., 20,* 187–196, doi:10.1002/met.1393.

Sorooshian, S., K. Hsu, X. Gao, H. Gupta, B. Imam, and D. Braithwaite (2000), Evaluation of PERSIANN system satellite–based estimates of tropical rainfall, *Bull. Am. Meteorol. Soc., 81,* 2035–2046, doi:10.1175/1520-0477(2000)081<2035:EOPSSE>2.3.CO;2.

Author Information

Scott Sellars, Phu Nguyen, Wei Chu, Xiaogang Gao, Kuo-lin Hsu, and Soroosh Sorooshian, Department of Civil and Environmental Engineering, University of California, Irvine; E-mail: scott.sellars@uci.edu