

Introduction to Big Earth Data Applications

Christopher Lynnes¹ and Tiffany Vance²

¹NASA/Goddard Space Flight Center (U.S. Government Civil Servant)

²NOAA/Integrated Ocean Observing System (U.S. Government Civil Servant).

Corresponding author: Christopher Lynnes (christopher.s.lynnes@nasa.gov)

Key Points:

- The enormous volumes from climate and weather modeling drive new ways to enable iterative analysis, with an emphasis on identify and highlight “interesting” data for the scientist to focus on.
- Sensor and satellite technology are driving ever larger data volumes, while spawning new applications.
- Operational applications often scale up from small, local studies to larger spatial scales with more analysis targets.

Abstract

Climate and weather modeling generate enormous volumes that make iterative analysis challenging, spurring the development of new ways to work with the data. At the same time in the Earth Observation area, technology advances are enabling new sensors and satellites that will increase data volume, velocity and application variety. Scaling up can also be seen when operational applications expand from small, local studies to larger spatial scales with more analysis targets.

The aim of most analytics activities is to wrest some kind of insight from the data being analyzed. Insight derivation is often iterative, rather than a one-time process: the answer to one question posed to the data spurs further questions. As Earth Observation data grow, the ability to iterate, to pose one question of the data, succeeded by a follow-up question, is eroded by two factors: the simple length of time that each analysis run consumes and the time and effort spent on subsetting, filtering, or otherwise reducing the data in order to accelerate analysis.

It is thus no surprise that many Big Earth Data Analytics problems are being encountered and addressed in the area of climate modeling and its close cousin weather modeling. These disciplines deal with global datasets over long time periods, with spatial resolution limited only by our knowledge of the physics at small scale. Data volumes are further multiplied by the incorporation of the vertical dimension. The success of ensemble modeling adds yet another multiplier to the final data volume. Furthermore, climate and weather modeling have been one of the prime applications of supercomputers, resulting in computer models that can efficiently produce high-volume datasets for analysis. As a result, it is difficult for iterative analysis to keep up with the volumes that can be generated by the models. One approach is to enlist extensive computing power via cloud computing, in which the elastic nature of resource allocation and costing matches well with the bursty nature of the iterative analysis problem. An example of this approach is given in the chapter “Giving Scientists back their flow: Analysing Big Geoscience Datasets in the Cloud”. An alternate approach to analyzing climate model output is to reduce the effective amount of data being analyzed by using pattern recognition algorithms to detect interesting phenomena in the data. The chapter “Topological Methods for Pattern Detection in Climate Data” proposes a topological algorithm approach to finding extreme events in climate data, allowing researchers to zero in on specific areas of the data fields for further analysis.

However, even surface Earth Observations can generate data volumes that are resistant to iterative analysis, especially if researchers are interested in relatively small scale features close to the limit of the spatial resolution. Similar to the above climate modeling case, focussing on interesting values in the dataset can in many cases significantly reduce the amount of data that needs to be analyzed (and thus wrangled and managed). This is particularly the case during the data exploration phase of studies. An example approach to extracting extreme values from a land surface temperature dataset is demonstrated in the chapter “Demonstrating Condensed Massive Satellite Data Sets for Rapid Data Exploration: The MODIS Land Surface Temperatures of Antarctica”. Similarly, the chapter “Exploring Large Scale Data Analysis and Visualization for Atmospheric Radiation Measurement Data Discovery Using NoSQL Technologies” also explores the process of selecting data for retrieval by their measurement values.

Advances in Earth Observation instrumentation can also drive tremendous increases in data volume. Some of the increase comes from advances in lasers that enable huge increases in the number of shots from satellite-borne Lidar instruments such as ICESat-2 and GEDI. Advances in the field of spaceborne radar will exceed even those: the Surface Water Ocean Topography (SWOT) interferometric altimetry radar will generate about 20 TB of data per day. Its wide swath and high vertical accuracy and spatial resolution will open up new areas of study to radar altimetry, such as water resource management and flood monitoring, posing both challenge (volume) and opportunity.

Even more voluminous than SWOT is the NASA-Indian Space Research Organization Synthetic Aperture Radar (NISAR), which will have a duty cycle averaging at 30%, providing much higher coverage of land and ice surfaces than present day SAR satellites, producing more than 80 TB of data per day. As with SWOT, the increased coverage and resolution have the

potential to open up new applications for SAR data. These include aiding responses to a variety of disasters: oil spills, earthquakes, volcanic eruptions, landslides, floods (Kumar et al., 2016). These often levy stiff requirements for low latency, bringing into play the Velocity aspect of Big Data. In addition, the SAR community has engaged in significant capacity building both in terms of easy-to-use SAR processing tools such as the SentiNel Application Platform (e.g., Fournelis et al., 2018) and workshops and mini-courses (e.g., Rosen et al., 2020). These data volumes are high enough that conventional analysis method that rely on downloading data to a local computer to analyze will no longer be sustainable for many users. As a result, NASA is preparing to host SWOT and NISAR data in commercial cloud (Behnke et al., 2019) in order to allow analysis-in-place. In addition, the European Space Agency is teaming with NASA to develop a cloud-based analysis platform named the Multi-Mission Algorithm and Analysis Platform for biomass study using high volume Lidar and SAR data (Albinet et al., 2019). The MAAP is intended to support community development of biomass estimation algorithms but will also include robust analysis and visualization capabilities in order to validate algorithms on the data at scale. The combination of cloud-based analysis-in-place and mechanisms for sharing research software and outputs also enables the scientists to collaborate more efficiently.

Another advance in hardware that drives Big Data is the miniaturization of satellites and the instruments that fly on them. This has spawned a surge of commercial satellites, often flown in large constellations. The data may be offered directly for sale to the user or be purchased by governments for research and applications use. Again, the high resolution of these satellites drives new applications. For example, Brandt et al. (2020) applied a machine learning approach to high-resolution commercial imagery to count the number of individual trees in and near the Sahel area of Africa.

The Veracity aspect of Big Data puts a premium on Earth Observation satellite data that is coincident with any kind of data collected in the same vicinity (both spatial and temporal) nearer the ground, such as in situ or airborne. This is one of the most common methods of validating satellite data retrievals of geophysical parameters. Ironically, the combination of spatial and temporal dimensions makes this match-up problem difficult computationally. The match-up problem is further complicated by the tendency of different researchers or science problems to require different spatial and temporal tolerances. The chapter on the “Distributed Oceanographic Matchup Service” presents an approach for on-demand determination of matchups.

While the above research applications typically evolve toward ever larger volumes via improvements in spatial or temporal resolution, operational applications often scale differently. Typically, applications begin as case studies in a particular area, and then scale up to regional or global scales. An example of this is the use of Automated Information System (AIS) tracking of ship positions, which has been used for a number of oceanographic applications, including fishery investigations and search and rescue. While the computational needs for a small set of ship data in a particular region are modest, scaling up to ocean basins or global problems requires a significant infrastructure of compute power and storage, such as that described in “Developing Big-Data Infrastructure for Analyzing AIS Vessel Tracking Data on a Global Scale”.

References

Albinet, C., Whitehurst, A.S., Jewell, L.A., Bugbee, K., Laur, H., Murphy, K.J., Frommknecht, B., Scipal, K., Costa, G., Jai, B. and Ramachandran, R., 2019. A joint ESA-NASA multi-mission

algorithm and analysis platform (MAAP) for biomass, NISAR, and GEDI. *Surveys in Geophysics*, 40(4), pp.1017-1027.

Behnke, J., Mitchell, A. and Ramapriyan, H., 2019. NASA's Earth Observing Data and Information System—Near-Term Challenges. *Data Science Journal*, 18(1).

Brandt, M., Tucker, C.J., Kariryaa, A., Rasmussen, K., Abel, C., Small, J., Chave, J., Rasmussen, L.V., Hiernaux, P., Diouf, A.A. and Kergoat, L., 2020. An unexpectedly large count of trees in the West African Sahara and Sahel. *Nature*, 587(7832), pp.78-82.

Kumar, R., Rosen, P. and Misra, T., 2016, May. NASA-ISRO synthetic aperture radar: Science and applications. In *Earth Observing Missions and Sensors: Development, Implementation, and Characterization IV* (Vol. 9881, p. 988103). International Society for Optics and Photonics.

Rosen, P.A., Meyer, F.J., Hensley, S., Donnellan, A., Davis, H.B., Bekaert, D.P., Fattahi, H. and Funning, G., 2020, December. A Modular SAR/InSAR Imaging Geodesy Training Course for Capacity Building. In *AGU Fall Meeting Abstracts* (Vol. 2020, pp. SY002-0004).