Introduction to Big Earth Data Applications

Christopher Lynnes[1] and Tiffany Vance[2]

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

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

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

Key Points:

- Climate and weather modeling generate enormous volumes that make iterative analysis challenging, spurring the development of new ways to work with the data.
- A theme going across applications is the need to identify and highlight "interesting" data for the scientist to focus on.
- Operational applications often scale up 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. However, insight derivation is intrinsically iterative in nature, not a batch process. 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 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".

Another characteristic that makes specific Earth Observation data interesting, i.e., promising for further analysis, is coincidence of satellite data with any kind of data collected in the same vicinity (both spatial and temporal) nearer the ground, such as in situ or airborne. Ironically, the combination of spatial and temporal dimensions makes this match up problem computationally difficult, which is further complicated by the tendency of different researchers to prefer different spatial and temporal tolerances, or to have different tolerance requirements for different problems. 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 the scale of complete 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”.