LEVERAGING DATA INTENSIVE COMPUTING TO SUPPORT AUTOMATED EVENT SERVICES

Thomas L. Clune, Shawn M. Freeman, Kwo-Sen Kuo

NASA GSFC
Earth Science Division
Greenbelt, MD 20771

1. INTRODUCTION

A large portion of Earth Science investigations is phenomenon- or event-based, such as the studies of Rossby waves, mesoscale convective systems, and tropical cyclones. However, except for a few high-impact phenomena, e.g. tropical cyclones, comprehensive records are absent for the occurrences or events of these phenomena. Phenomenon-based studies therefore often focus on a few prominent cases while the lesser ones are overlooked. Without an automated means to gather the events, comprehensive investigation of a phenomenon is at least time-consuming if not impossible.

An Earth Science event (ES event) is defined here as an episode of an Earth Science phenomenon. A cumulus cloud, a thunderstorm shower, a rogue wave, a tornado, an earthquake, a tsunami, a hurricane, or an El Niño, is each an episode of a named ES phenomenon, and, from the small and insignificant to the large and potent, all are examples of ES events. An ES event has a finite duration and an associated geolocation as a function of time; its therefore an entity in four-dimensional (4D) spatiotemporal space. The interests of Earth scientists typically rivet on Earth Science phenomena with potential to cause massive economic disruption or loss of life, but broader scientific curiosity also drives the study of phenomena that pose no immediate danger. We generally gain understanding of a given phenomenon by observing and studying individual events – usually beginning by identifying the occurrences of these events. Once representative events are identified or found, we must locate associated observed or simulated data prior to commencing analysis and concerted studies of the phenomenon. Knowledge concerning the phenomenon can accumulate only after analysis has started. However, except for a few high-impact phenomena, such as tropical cyclones and tornadoes, finding events and locating associated data currently may take a prohibitive amount of time and effort on the part of an individual investigator. And even for these high-impact phenomena, the availability of comprehensive records is still only a recent development.

A major reason for the lack of comprehensive records for the majority of the ES phenomena is the perception that they do not pose immediate and/or severe threat to life and property and are thus not consistently tracked, monitored, and catalogued. Many phenomena even lack commonly accepted criteria for definitions. However, the lack of comprehensive records is also due to the increasingly prohibitive volume of observations and model data that must be examined. NASA Earth Observing System Data Information System (EOSDIS) alone archives several petabytes (PB) of satellite remote sensing data and steadily increases. All of these factors contribute to the difficulty of methodically identifying events corresponding to a given phenomenon and significantly impede systematic investigations.

In the following we present a couple motivating scenarios, demonstrating the issues faced by Earth scientists studying ES phenomena.

This work is supported by NASA's Earth Science Technology Office

Heat Wave *Heat kills by taxing the human body beyond its abilities. In a normal year, about 175 Americans succumb to the demands of summer heat. Among the large continental family of natural hazards, only the cold of winternot lightning, hurricanes, tornadoes, floods, or earthquakestakes a greater toll.* –National Weather Service[1]

Heat waves pose a serious public health threat, yet a standard definition of heat wave does not exist. Many researchers argue for a changing threshold in heat wave definition ([2, 3, 4]). A researcher interested in understanding heat waves will have to first devote a great deal of time in research before associated, concurrent observations (ground-based or remote sensing) can be obtained and analyzed. A literature search for heat wave definitions will yield multiple conflicting definitions. Some decision on what qualifies as a heat wave would need to be made, followed by a search for the incidences that satisfy the definition. Each investigator, who is not in collaboration or association with other researchers of like interest, must repeat a similar process. Intense and prominent episodes of heat waves are easier to find and not likely to raise questions on definition. Thus, a few intense cases are likely studied repeatedly with great scrutiny while the less intense cases are largely ignored or never identified.

An intriguing possibility with undesirable consequences is that emphasis on intense cases (or cases that impact populated regions) has been known to induce biases in past investigations of some phenomena. A systematic treatment of events could thus be an important mechanism for eliminating such bias in certain types of investigation.

Blizzard According to the National Weather Service glossary: A blizzard means that the following conditions are expected to prevail for a period of 3 hours or longer: 1) sustained wind or frequent gusts to 35 miles an hour or greater; and 2) considerable falling and/or blowing snow (i.e., reducing visibility frequently to less than mile). Consequently, Blizzard is also an ES phenomenon that does not have an unambiguous definition, because both "considerable" and "frequently" are vague and not quantified. There exists an opportunity to explore for a more definitive specification.

2. THE AUTOMATED EVENT SERVICE

We are in the process of Automated developing the Event Service (AES) system that methodically mines custom-defined events in the reanalysis data sets of atmospheric general circulation models. Our AES will enable researchers to specify their custom, numeric event criteria using a user-friendly web interface to search the reanalysis data sets. Searches can also be performed using our Event Specification Language (ESL) to afford more flexibility and versatility. Investigators will be able to subscribe to event searches and get notified of new results when data sets are updated with the latest additions. Moreover, we will implement a social component to enable dynamic formation of collaboration groups for researchers to cooperate on event definitions of common interest. AES will also interface with the ESDIS ECHO service[5], making it possible to convert the spatiotemporal coordinates returned by AES queries into complimentary searches for EOSDIS remote sensing observations.

The AES System is comprised of several components:

- **Event Specification Language** a simplified programming language that allows researchers to specify event identification functions that are more complex than simple parameter searches.
- **Event Portal** provides a browser based interface to allow users to define events, save their definitions and search results, share their definitions with individuals, groups or the entire community, and discover other shared events. The Event portal will serve not only as an entry point but also as a collaboration platform.
- **Event Web Services Interface** provides an application programming interface (API) for accessing the search and querying functionality of the system. This allows for researchers to incorporate the functionality of the Event Services system into their own applications and workflows.
- **Distributed Event Database** used to store and recall event specifications as well as the results of searches using those specifications. It will use a distributed database for performance and reliability.

- **Event Search Engine** processes event specification queries and searches the data archive for events that match. The search engine will utilize a Map-Reduce framework to parallelize the search and associated computations. Adjacent regions in space and/or time that satisfy the criteria are merged and treated as part of the same event.
- **The Distributed Data Archive**contains multiple data collections stored on a distributed file system. Each data collection may itself be stored in multiple layouts and/or with with multiple precomputed indexing schemes[6] to further optimize queries of such high-dimensional sets of data.

3. CONNECTION TO MOVING OBJECTS DATABASE

The Moving Objects Database (MODB) technology, which has been developed in part with funding from ESTO, combines a series of separate snapshots of a moving event into an integral entity and catalogs it in a relational database. When applied to an Earth Science (ES) event (e.g. tropical cyclone, TC), the MODB enables sophisticated spatiotemporal queries concerning the events in the database, such as "What are all the names of the hurricanes that made landfall in North Carolina between 12-6 PM local time?".

Although the MODB has allowed novel exploration and analysis of moving ES phenomenal, further extensions and complementary tools are necessary to realize the full potential of the approach. At the very least, techniques must be developed which automate the identification of events to ingest required data into the MODB for cataloging. Currently the recognition that observations from different times are of the same event still requires human effort, even when definitive, quantitative criteria exist for the event. In the case of TCs, National Hurricane Center documents the time and location of their eyes, which are then entered into the MODB through a process which is largely manual. Additionally, although the MODB currently treats objects as point trajectories and thus does not directly support events with spatially extended and discontiguous components such as convective cells in a mesoscale convective system (MCS), automated tracking of such events can and should collect data which characterizes their spatial distribution.

4. THE MCS TRACKER PROTOTYPE

With initial seed funding from NASA ESTO, we have developed a flexible automatic tracking system and successfully applied it to identify and characterize tornado-producing mesoscale convecting systems in a manner suitable for ingest to the MODB. MCS events are identified by a simple thresholding scheme along with an algorithm to associate active elements that are adjacent in space or overlap in time. In this manner, a single MCS event may consist of multiple cells that merge and diverge over time. For each MCS event a variety of characteristics are computed including the centroid, area, perimeter, max/min intensity, etc. The pluggable architecure permits users to provide custom search criteria and derived characteristics through lighweight scripts and thereby encourages extensions by the user community.

As a demonstration of the capabilities of our prototype system, we have applied it to identify tornado-producing MCS's in North America over the past 3 years. This effort synergistically combined a number of data sources including the tornado occurrence and trajectory data from the Tornado History Project (THP), historical NWS warnings, 1-km 5-min quantitative precipitation data , 4-km 1/2-hourly GOES-East Rapid Scan Operation imagery, and 12-km 3-hourly North America Mesoscale Model analysis (NAM-ANL). The THP and NWS warnings were used to automatically prune MCS events which did not spawn tornados. A rendering of representative output is shown in figure 1.

5. CONCLUSION

Our Automated Event Service is an ideal example of a new generation of scientific analysis tools that are empowered by the rapid growth of facilities tailored for data-intensive computing. AES will greatly reduce the effort on the part of investigator to

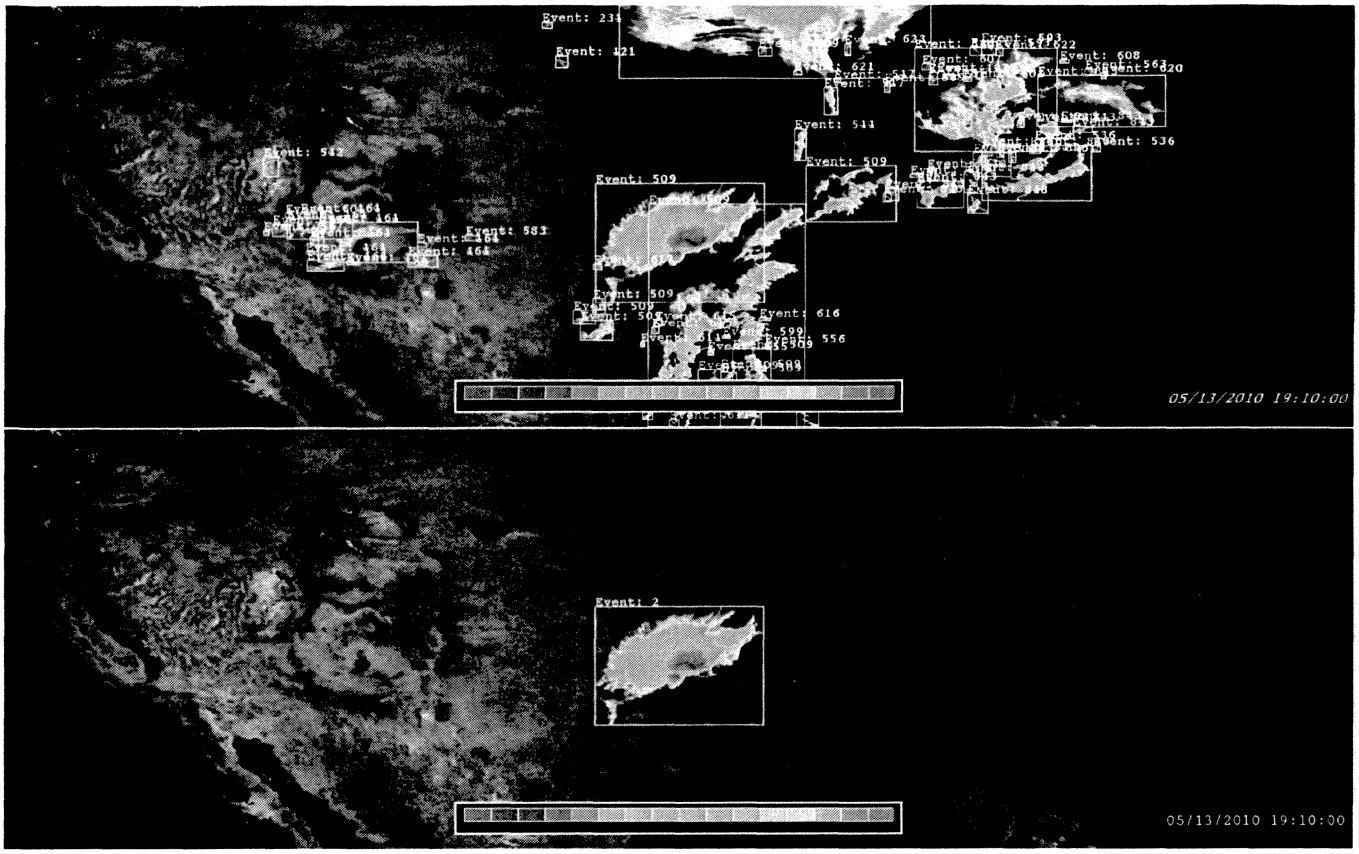

Fig. 1. Visualization of output from automated tracking of MCS events based upon GOES imagery. The event specificiation is for cloud top temperatures between 200K and 240K and a minimimum size of 300 km^2. The upper image includes all MCS events while the lower includes only those MCS events associated with tornado activity.

systematically search for interesting correlations and test hypotheses while also freeing researchers from the burden of managing the exploding volume of data. By supporting the ability to exchange event specifications and query results, AES greatly aids collaboration among investigators. We anticipate that AES will ultimately lead to entirely novel lines of investigation.

6. REFERENCES

[1] "Heat wave: A major summer killer," `http://www.nws.noaa.gov/om//brochures/heat_wave.shtml`.

[2] P. J. Robinson, "On the definition of a heat wave," *J. Appl. Meteorol.*, vol. 40, pp. 762–775, 2001.

[3] J. Abaurrea, J. Asin, A. C. Cebrian, and A. Centelles, "On heat wave definition," *Geophys. Res. Abs.*, vol. 7, pp. 10014, 2005.

[4] J. Abaurrea, J. Asin, A. C. Cebrian, and A. Centelles, "On the need of a changing threshold in heat wave definition," *Geophys. Res. Abs.*, vol. 8, pp. 09142, 2006.

[5] "The nasa-developed earth observing system (eos) clearinghouse (echo)," `http://www.echo.nasa.gov`.

[6] B. Nam and A. Sussman, "Analyzing design choices for distributed multidimensional indexing," *Journal of Supercomputing (to appear)*, 2011.