Earth Cube in cross-disciplinary Earth system science -- and opportunities for a comprehensive data infrastructure to facilitate integrated Earth system science.

Chris Hill, M.I.T.

Many of the key Earth system challenges (and opportunities) of the 21st century involve (at some level) measuring, monitoring and attributing changes in the Earth system that transcend regional and national boundaries and that are reflected in multiple different measurable quantities associated with different Earth system science "disciplines". A goal for Earth cube should be to develop, demonstrate and deploy enabling technological frameworks that can accelerate cross Earth system science discipline discovery and knowledge generation in both theoretical and applied domains.

As one example consider sea-level change globally and regionally. Global sea-level rise is highly likely to impact human and natural environments in the coming 100 years. The causes of any rise, together with the impacts of any mitigation strategies are many and varied. Sea-level involves hydrological cycles with time scales of decades to tens of thousands of years, coupled to lithospheric processes with even longer time scales. Impacts are non-uniform and indeed global mean sea-level rise is not inconsistent with regional sea-level drops. As a consequence coordinated mitigation and preparedness is challenging, as is the quantitative attribution of causes as well as effective mitigation.

To simply understand sea-level rise in the Earth requires detailed information on mass-balance, heat redistribution and geodetic effects that crosses many traditional Earth science disciplines (glaciology, geodesy, oceanography, meteorology and so forth). Anticipating sea-level change impacts further involves data around storm surge likelihood and vulnerability. Current computational infrastructure is not ideally construed to support science that synthesizes knowledge and measurement from such different communities.

Earth Cube could make bold investments in a range of technology activities that emphasize and enable integrated Earth system science scenarios including for example around understanding, anticipating and managing sea-level rise. For example, observations associated with sea-level change include the gravitational potential over the Earth, ocean density, liquid and frozen water precipitation rates, isostatic adjustments and even large scale changes in atmospheric winds and pressure.

Currently de-convolution of the forcings and signals associated with sea-level changes requires synthesis of some or all of the data sources highlighted above. There is no single source for information for this, nor are there collaborative and interactive forums for quality control and validation. Instead researchers must grapple with a wide array of different sources of information, some very high quality and some less so.

In this note we speculate that a different approach could be beneficial. In particular we hypothesize that **all** the digitized information of interest for a sea-level study (and for studies of biodiversity, ocean acidification, carbon cycle dynamics, permafrost change and so forth) could (and should through Earth cube) ultimately co-exist -- in some form of virtual spatio-temporal database. Such a virtual database (which could in fact be a federation of distributed resources) would provide detailed information, for example, on every temperature measurement over the Earth, together with diverse other measurements such as geopotential, spectral irradiance and metagenomic sequences– all through one unified spatio-temporal "interface". Projects such as the NSF OOI and Earthscope initiatives would provide their data to such a system in geo-spatio located form and that information would be available for both direct analysis and for automated ingestion into modeling and analysis tools. Data from remote sensing satellite programs and from autonomous sensor networks would all be connected into this database and therefore much more readily available to researchers (and interested citizen scientists) across disciplines.

Implemented well (with careful attention to technical aspects of provenance, quality tracking, curation and to social aspects of engagement and attribution) such a database abstraction could be truly transformative for Earth science. Whether the question is what are drivers for present and future melting of coastal glaciers in Greenland or Antarctica or what causes the saw tooth cycles in atmospheric CO2 over the last million years – a single database abstraction that could be queried across space and time for relevant information would be of immeasurable value today and of increased value as richer streams of digital information (for example the NSF Pioneer Array) about the Earth system come online.

Technologically, the traditional database community has already begun thinking hard about petabyte, exabyte (and even zettabyte) data stores. This thinking has been demonstrated in action in projects such as the Large Hadron Collider (LHC) and is further evolving in projects such as the Large Synoptic Survey Telescope (LSST). These projects are demonstrating integrated (although physically distributed) data infrastructures at unprecedented scales for the fields of particle physics and astrophysics. An Earth science thrust that would allow similar levels of integrated analysis by a community spanning GPS monitoring to marine microbial measurement would be a bold further step, but one that could yield significant benefits to existing science projects as well as enabling new science that was previously impractical.

Clearly this white paper leaves many questions unanswered (what would be the temporal scope and resolution of the proposed system, what spatial resolution(s), which disciplines would be included) however with strategic investments under the Earth cube umbrella these questions could be answered and real progress achieved in this area. Alternately, without investments in this area, Earth system science will continue to span a hodge-podge of impressive, but largely disconnected data islands.

Such balkanization may be inevitable, given both the disparate breadth of the field and also the individual nature of much science – but such a view probably does not reflect where much of the field on Earth science would like to be headed in the future. Indeed one of the most powerful sites for climate information is the Ingrid server at Lahmont-Doherty (http://ingrid.ldeo.colubia.edu). This site, though far from a truly integrated repository is nevertheless an invaluable resource for questions around certain recent climate related data.

A full-blown solution should go well beyond the sorts of facilities currently available, including supporting query concepts such as those embedded in research around the SciDB system, and enabling seamless interaction from compute agents (models, synthesis tools etc…) as well as less structured querying. In the same vein a true solution would ultimately transcend (though maybe not initially) the traditional realms of the NSF GEO directorate to focus on all aspects of the planet past, present and future. Such an activity would require many steps from planning workshops, to competing prototype activities, to full-blown implementations. This would provide one strong and useful focus for Earth cube.

References

LSST – Large Synopti Survey Telescope, http://www.lsst.org/lsst/
SciDB – http://www.scidb.org/
LHC – Building a Database for the LHC--the Exabyte Challenge, Shiers, J., Proceedings of the 15 th IEEE Symposium on Mass Storage Systems.
Ingrid – http://ingrid.ldeo.columbia.edu