ANALYTICS CENTER FRAMEWORK FOR ESTIMATING THE CIRCULATION AND CLIMATE OF THE OCEAN

Thomas Huang[1], Ian Fenty[1], Patrick Heimbach[2]

[1] NASA Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, USA
[2] University of Texas, Austin, USA

ABSTRACT

As Alvin Toffler had eloquently put it "You've got to think about big things while you're doing small things, so that all the small things go in the right direction." [10] We have a long history of building many innovative solutions. A quick search on the web we can find various of tools that offer similar capabilities such as search, visualization, subsetting, analysis, etc. The community is very good on building domain-specific solution for specific applications. The lack of cohesiveness among these tools introduces technology gaps, which lead to even more stovepipe solutions. An Analytics center framework is an architectural concept to establish an extensible, reusable software framework for specific research community. This paper discusses the application of an open source data analytics framework NASA is developing through its Advanced Information Systems Technology (AIST) program that is the platform to enable Sea Level research and estimating ocean circulation modeling.

***Index Terms*—** Big Data, Analytics Centers, Cloud Computing, Open Source, Data Discovery, sea level, ocean circulation

1. INTRODUCTION

The Estimating the Circulation and Climate of the Ocean (ECCO) global ocean state estimation system is the premier tool for synthesizing NASA's diverse Earth system observations into a complete physical description of Earth's time-evolving full-depth ocean and sea ice system. ECCO state estimates are of particular significance to NASA because on their own, all satellite observations, although global in coverage, remain sparse in space and time relative to the inherent scales of ocean variability, and are blind to the ocean's interior. With increased streams of data better spatial resolution the scientific utility of the product is increasing limited by 1) the inability to automate observing network ingestion and update in a rapid, robust manner, and 2) the lack of capabilities to search, and perform efficient online data analysis. ECCO-Cloud is a NASA Advancing Collaborative Connections for Earth System Science (ACCESS) project. Its goals are to tackles the two limitations by leveraging serverless cloud-based solution to automate product generation and to establish a cloud-based analytics center for ECCO products for fast access and interactive analysis using prior NASA AIST-funded technology, the Apache Science Data Analytics Platform (https://sdap.apache.org).

2. SERVERLESS WORKFLOW

Satellite observations, although global in coverage, are still sparse in both time and space relative to the inherent scales of ocean variability, and are blind to much of the ocean's vertical structure which can only be probed by in-situ observations. Even with the more than 10^6 in-situ vertical profiles of temperature and salinity measured by the global array of Argo floats [8] only sample a small fraction of the ocean's volume. Simultaneously analyzing these diverse and heterogeneous data streams to comprehend our changing ocean exceeds the capacity of any one individual.

The ECCO ocean and sea-ice state estimation system (ECCO-SES) is a mathematically rigorous, mature, state-of-the-art data tool for synthesizing NASA's diverse Earth System observations, especially remote sensing data, into a complete and physically-consistent description of the three-dimensional time-varying global ocean and sea-ice state (e.g. [12]). By combining more than 10^8 measurements with an ocean and sea-ice general circulation model, the ECCO-SES is able to reproduce more than two decades of observed ocean and sea-ice variability [1] (See Figure 1). Distributed as a monthly-averaged product of nearly 100 physical variables across 2+ TB, it is a clear example of a large and complex NASA Earth Science data set. ECCO-SES is running under the Pleiades supercomputer at the NASA Ames Research Center (ARC).

An ECCO state estimate (ECCO-StE), produced by ECCO-SES, requires dozens of distinct ocean and sea ice datasets constructed by identifying a set of ocean model initial conditions, parameters, and atmospheric boundary conditions such that free-running simulation of the ocean model reproduces the observations as a result of the governing equations of motion. It reproduces observations of ocean variability and the explanation for the underlying physical causes and mechanisms responsible for bringing them into existence (e.g. [9]).

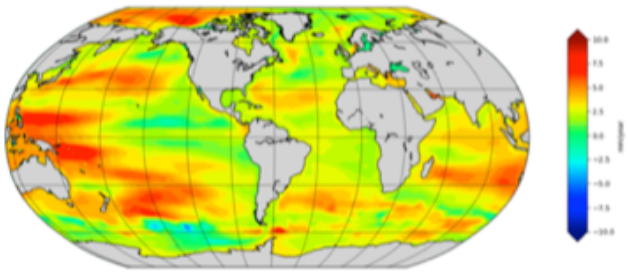

Figure 1: Linear trends in Global Mean Sea Level Rise between 1/1993 - 12/2015 from the latest ECCO product. Compare to the altimetry map on NASA's Sea Level Change Portal

An ECCO state estimate (ECCO-StE), produced by ECCO-SES, requires dozens of distinct ocean and sea ice datasets constructed by identifying a set of ocean model initial conditions, parameters, and atmospheric boundary conditions such that free-running simulation of the ocean model reproduces the observations as a result of the governing equations of motion. It reproduces observations of ocean variability and the explanation for the underlying physical causes and mechanisms responsible for bringing them into existence (e.g. [9]).

Historically the generation of ECCO-StE was done in an ad-hoc manner by multiple, distributed organizations. The lack process oversight between teams lead to the organic expansion of the ECCO community. One of the goals of ECCO-Cloud is to streamline the processing of ECCO-StE using modern cloud computing solutions to provide the consortium visibility and traceability to the ECCO-StE. The cloud-based process (Figure 2) provides tracking and notification for each preprocessing and transformation step. It integrates with ECCO-SES running on the Pleiades supercomputer and streamline the packing the ECCO-StE reanalysis products according CF-convention and deliver to the NASA Physical Oceanography Distributed Active Archive Center (PO.DAAC).

The new cloud-based workflow will be designed and deployed under the Amazon Web Services (AWS) environment where it can take advantage of AWS' severless, messaging, and object storage solutions. Each processing and transformation will be tracked and managed for traceability and lineage. It notifies the ECCO coordinator on new ECCO-StE jobs for the Pleiades-based ECCO-SES.

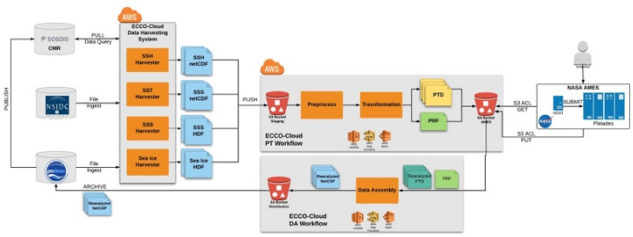

Figure 2: ECCO-Cloud Serverless Workflow System

3. ECCO ANALYTICS PLATFORM

An Analytics center framework is an architectural concept to encapsulate the scalable computational and data infrastructures and to harmonize data, tools and computation resources to enable scientific investigations. The goal is to create a webservice platform for researchers and tools developers to discover, interact and analysis massive amount of related data without having to move data between systems over the internet. The Apache Science Data Analytics Platform (https://sdap.apache.org) is a NASA-funded Analytics center framework (Figure 3) [3]. It is currently under the Apache Incubator.

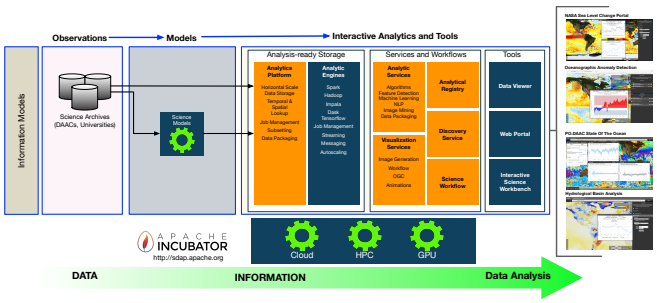

Figure 3: The Apache Science Data Analytics Platform (SDAP), an Analytics center framework

SDAP was originated from a collection of AIST-funded projects. They include OceanXtremes for cloud-based anomaly detection, Distributed Oceanographic Matchup Service (DOMS) for distributed in-situ to satellite matchup on the cloud, and MUDROD for data discovery using machine learning. These big data technologies were latest integrated to established an Analytics center framework for ocean science, called OceanWorks. Recognizing OceanWorks can be generalized to support various Earth science disciplines, the project collaborated with the Apache Software Foundation to establish SDAP.

The SDAP's analytics engine is called NEXUS [6][7]. It takes on a different approach for storing and analyze large collection of geospatial, array-based data by breaking the netCDF/HDF file data into data tiles and store them into a cloud-scale data management system. With each data tile has its own geospatial index, a regional subset operation only requires to retrieve the relevant tiles into the analytic engine. Our recent benchmark shows NEXUS can compute an area-averaged time series hundreds time faster than traditional file-based approach [5].

SDAP enables advanced analytics that can easily scale to the available computation hardware along the full spectrum from an ordinary laptop or desktop computer, to a multi-node server class cluster computer, to a private or public cloud computer. The architectural drivers are:

- Both REST and Python API interfaces to the analytics
- In-memory map-reduce style of computation
- Horizontal scaling so computational resources can be added or removed on demand
- Rapid access to data tiles that form natural spatiotemporal partition boundaries for parallelization
- Computation performed close to the data store to minimize network traffic
- Container-based deployment

The REST and Python API enables SDAP to be easily plugged into a variety of web-based user interfaces, each tuned to particular domains. Calls to SDAP from a Jupyter notebook enables interactive cloud-scale, science-grade analytics.

Built-in analytics are provided for the following algorithms:

- Area-averaged time series to compute statistics (e.g., mean, minimum, maximum, standard deviation) of a single variable or two variables being compared. Optionally apply seasonal or low-pass filter to the result.
- Time-averaged map to produce a geospatial map that averages gridded measurements over time at each grid coordinate within a user-defined spatiotemporal bounding box.
- Correlation map to compute the correlation coefficient at each grid coordinate within a user-specified spatiotemporal bounding box for two identically gridded datasets.
- Climatological map to compute a monthly climatology for a user-specified month and year range.
- Daily difference average to subtract a dataset from its climatology, then, for each timestamp, average the pixel-by-pixel differences within a user-specified spatiotemporal bounding box.
- In-situ match to discover in situ measurements that correspond to a gridded satellite measurement.

In addition, authenticated or trusted users may inject their own custom algorithm code for execution within SDAP. An API is provided to pass the custom code as either a single or multi-line string or a python file or module

SDAP is currently the analytics platform for the NASA Sea Level Change Portal (https://sealevel.nasa.gov) [4] to provide interactive analysis of observation and model parameters. Figure 4 is a screenshot of the Sea Level Data Analysis Tool that coordinates sea level time series between ECCO and the NASA Sea Level Anomaly product. The plot shows ECCO's accuracy in estimating sea level. For Sea Level science, SDAP also provide removal of seasonal signal, Hovmöller and profile plots, and in-situ analysis. SDAP is also the cloud-based analytics platform for the GRACE science portal (https://grace.jpl.nasa.gov). Like the Sea Level Data Analysis Tool, the GRACE Data Analysis Tool provides interactive analysis of GRACE parameters, including hydrological basins. And finally, SDAP is being integrated into the NASA PO.DAAC as its cloud-based analytics platform for all relevant physical ocean parameters. In addition to on-the-fly analytics, this platform could also serve as its fast-geospatial data subsetting solution.

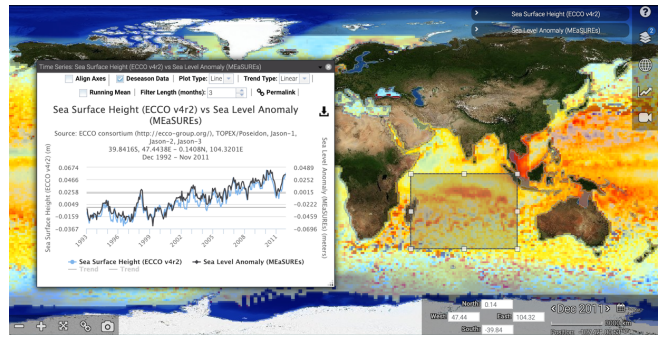

Figure 4: NASA Sea Level Data Analysis Tool comparing ECCO output with Sea Level Anomaly from MEaSUREs

The second goal of ECCO-Cloud is to establish an official analytics center for ECCO products. It provides the one-stop search and access to all ECCO products. In addition to the already out-of-the-box analytics, the ECCO-Cloud will tackle the dynamic data regridding and image generation. Currently approaches to data regridding and image generation create massive data silo, approaches that are unsustainable over time as our data volume from observation, model, and value-added processing continue to expand. SDAP has the architecture to fully leverage the elastic cloud to support these capabilities. The goal is to use SDAP as the manager of analysis ready data and with a collection of cloud-based, value-added services. The ECCO-Cloud team also plan to introduce complex GIS-based coordination and ability for users to run simulations based on ECCO's Global Circulation Model (GCM) on the AWS environment.

4. CONCLUSION

From a software architecture point of view an Analytics center framework is an architectural approach to establish reusable and extensible. With ECCO-Cloud, this platform must tackle both storage and software architecture together in order to fully leverage of its operating environment, such as the elastic cloud, without tying the users of the platform to a specific cloud provider and/or a specific underlying technology. SDAP was invented to established a community-supported big data framework for Earth science. This framework can also be extended to support Space science. As our data volume continue to expand, we are looking to integrate machine learning and deep learning solutions to reduce our time for search, retrieval, and analysis.

5. ACKNOWLEDGEMENT

The research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

© 2019 California Institute of Technology. Government sponsorship acknowledged.

6. REFERENCES

[1] Forget, G., J. M. Campin, P. Heimbach, C. N. Hill, R. M. Pointe, and C. Wunsch, "*ECCO version 4: an integrated framework for non-linear inverse modeling and global ocean state estimation*," Geosci. Model Dev., 8(10), 3071-3104, doi:10.5194/gmd-8-3071-2015.

[2] Huang, T. *Architecture for Distributed Earth Science Data Analysis*. In proceedings of the 2018 CEOS SIT Technical Workshop, Darmstadt, Germany, 2018.

[3] Huang, T., E.M. Armstrong, F.R. Greguska, J. Jacob, N. Quach, L. McGibbney, V. Tsontos, B. Wilson, S. Smith, M.A. Bourassa, J. Elya, S.J. Worley, T. Cram, and Z. Ji. *High Performance Open-Source Big Ocean Science Platform*. In proceedings of the 2018 Ocean Science Meeting, Portland, OR, 2018.

[4] Huang, T. *NASA Sea Level Change Portal – It is not just another portal site*. In proceedings of the 2017 American Geophysical Union Fall Meeting, New Orleans, LA, 2017.

[5] Huang, T., E.M. Armstrong, and N. Quach. *Metadata-Centric Discovery Service*. In proceedings of the 2012 Federation of Earth Science Information Partners (ESIP) Summer Meeting, Madison, WI, 2012.

[6] Jacob, J.C., F. R. Greguska, T. Huang, N. Quach, and B. D. Wilson. *Design Patterns to Achieve 300x Speedup for Oceanographic Analytics in the Cloud*. In proceedings of the 2017 American Geophysical Union Fall Meeting, New Orleans, LA, 2017.

[7] Lynnes, C., M. M. Little, T. Huang, J. C. Jacob, C. Yang, and K. Kuo. *Benchmark Comparison of Cloud Analytics Methods Applied to Earth Observations*. In proceedings of the 2016 American Geophysical Union Fall Meeting, San Francisco, CA., 2016.

[8] Riser, S.C., H.J. Freeland, D. Roemmich, S. Wijffels, A. Troisi, M. Belbéoch, D. Gilbert, J. Xu, Jianping, S. Pouliquen, A. Thresher, P. Traon, G. Maze, B. Klein, M. Ravichandran, F. Grant, P. Poulain, T. Suga, B. Lim, A. Sterl, Andreas, and S.R. Jayne, Steven, "*Fifteen years of ocean observations with the global Argo array*," Nature Clim. Change. 6. 145-153. 10.1038/nclimate2872.

[9] Stammer, D., M., Balmaseda, P. Heimbach, A. Kohl, and A. Weaver, "*Ocean Data Assimilation in Support of Climate Applications: Status and Perspective*," Annual Review of Marin Science, 8(1), 491-518. http://doi.org/10.1146/annurev-marine-122414-034113

[10] Toffler, A., *Future Shock*, Random House, New York, 1970.

[11] Wilson, B., E.M. Armstrong, T. Chin, K. Gill, F. Greguska, T. Huang, J. Jacob, and N. Quach, "OceanXtremes: Scalable Anomaly Detection in Oceanographic Time-Series," in proceedings of the 2106 American Geophysical Union Fall Meeting, San Francisco, CA, 2016.

[12] Wunsch, C., and P. Heimbach, "*Practical global oceanic state estimation*," Physica D, 230(1-2), 197-208. http://doi.org/10.1016/j.physd.2006.09.040