Geophysical Research Abstracts
Vol. 20, EGU2018-12146, 2018
EGU General Assembly 2018

Pangeo: A Big-data Ecosystem for Scalable Earth System Science

Joseph Hamman (1), Matthew Rocklin (2), and Ryan Abernathy (3)

(1) National Center for Atmospheric Research, Research Applications Laboratory, Boulder, Colorado, United States (jhamman@ucar.edu), (2) Continuum Analytics, Austin, Texas, United States, (3) Columbia University, Lamont Doherty Earth Observatory, Palisades, New York, United States

Data intensive scientific workflows are at a pivotal time in which traditional local computing resources are no longer capable of meeting the storage or computing demands of scientists. In the Earth System Sciences (ESS) community, we are facing an explosion of data volumes where new datasets, sourced from models, in-situ observations, and remote sensing platforms, are being made available at prohibitively large volumes to store and work with using conventional computational approaches. Furthermore, there is a growing recognition that the fragmentation of software tools and environments renders most geoscientific research effectively unreproducible and prone to failure. While the private data-science community has been actively developing big-data and open-source solutions to these problems, the academic community has begun to lag, leading to a growing technology gap between the two sectors.

Pangeo is a community driven effort for open-source big-data in the Earth System Sciences. Pangeo's mission is to cultivate an ecosystem in which the next generation of open-source analysis tools for the geosciences can be developed, distributed, and sustained. In this presentation, we will introduce the core concepts driving the development throughout the Pangeo community and will describe our recent efforts towards 1) closing the gaps in scalability and technology and 2) building an interdisciplinary community to develop and sustain the technology needed to leverage big data in ESS. We will discuss the core software libraries of the Pangeo ecosystem, Python, Xarray, Dask, and Jupyter and will provide examples of how these tools, when deployed on both high-performance computing (HPC) and cloud computing systems are allowing scientists to scale their workflows on large datasets.