# THE PANGEO BIG DATA ECOSYSTEM AND ITS USE AT CNES

*Guillaume Eynard-Bontemps, Joseph Hamman, Ryan Abernathey, Matthew Rocklin, Aurlien Ponte*

CNES, NCAR, Columbia, Anaconda, Ifremer

**ABSTRACT**

Pangeo[1] is a community driven effort for open-source big-data intially focused on the Earth System Sciences. It represents at the same time a collaboration of people and a platform composed of open source scientific python packages like Jupyter, Dask and Xarray. One of its goal is to improve scalability of these tools to handle petabyte-scale datasets on HPC or public cloud infrastructure. In this paper, we will first describe Pangeo: its motivation, community, the underlying technology stack and associated deployments, different applications and the on going work. On a second part, we will present its use in CNES: HPC deployment, some simple and more complicated use cases, and what we are planning to do.

***Index Terms*—** Pangeo, Dask, Jupyter, HPC, Cloud, Big Data, Analysis

## 1. PANGEO

### 1.1. Motivations, mission and goals

There are several building crises facing the geoscience community:

- Big Data: datasets are growing too rapidly and legacy software tools for scientific analysis cant handle them. This is a major obstacle to scientific progress.
- Technology Gap: a growing gap between the technological sophistication of industry solutions (high) and scientific software (low).
- Reproducibility: a fragmentation of software tools and environments renders most geoscience research effectively unreproducible and prone to failure.

Pangeo aims to address these challenges through a unified, collaborative effort.

Our mission is to cultivate an ecosystem in which the next generation of open-source analysis tools for ocean, atmosphere and climate science can be developed, distributed, and sustained. These tools must be scalable in order to meet the current and future challenges of big data, and these solutions should leverage the existing expertise outside of the geoscience community.

To accomplish this mission, we have identified three specific goals.

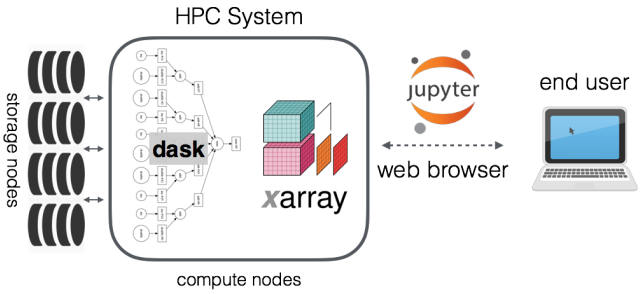



**Fig. 1**. Pangeo platform main components.

- Foster collaboration around the open source scientific python ecosystem for ocean / atmosphere / land / climate science.
- Support the development with domain-specific geoscience packages.
- Improve scalability of these tools to handle petabyte-scale datasets on HPC and cloud platforms.

### 1.2. Community

One crutial attribute of Pangeo is to be community driven. The goal is of course to have the most wider and open comunity as possible. All the effort are made in the open on github, any one can join or get involved in the community.

The community is already quite diverse, from academic research, going through government agency up to open source developers. A lot of different nationalities are represented too: from USA of course, but also UK, France or Australia to name a few.

### 1.3. Technology stack

Jupyter/Dask/Xarray scheme **??** and explanation.

Present all three, the cloud and HPC also.

### 1.4. Applications

Go through one Panego exmple and also the last blog post from Scott Henderson.

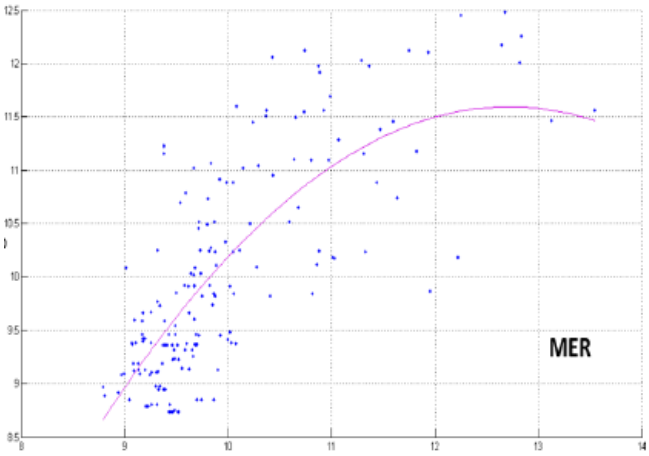



**Fig. 2**. Result 1.

### 1.5. On going work

Binder, Community governance, subdomains and Continuous deployment, Cluster Manager rationalization...

## 2. CNES DEPLOYMENT AND USE CASES

### 2.1. Context and HPC System

CNES various science, heavy simulations through launchers or flight dynamics, but majority of data processing : remote sensing, astronomy, climate.

Main processing platform is our cluster HAL: modestly sized High Performance Computer: 8000 cores, 6PB storage. PBS Pro.

### 2.2. From embarrassingly parallel to more complex workflow with Dask

Embarrassingly parallel simulation

Before: complex and unreadable batch scripts, launching PBS arrays, writting millions of small results on shared storage.

With Dask: elegant python code, scaling easily, in memory data exchange...

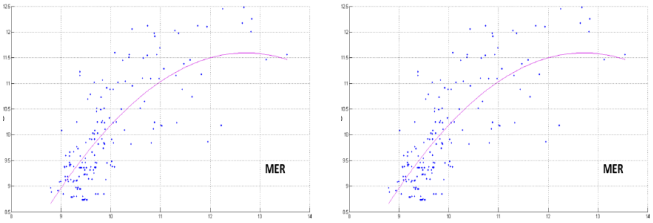
Add some post processing

Add some CSV file generation, build of a dask dataframe, reduction of results, launch of a new simulation...

### 2.3. Simulating remote sensing data through dask array

Generating a lot of Dask array that needs to be sum. Memory problem, rechunking...

1 and Fig. 2.




(a) Results 2 (b) Result 3

**Fig. 3**. Example of placing two figures with experimental results.

## REFERENCES

[1] Abernathey, Ryan; paul, kevin; hamman, joe; rocklin, matthew; lepore, chiara; tippett, michael; et al. (2017): Pangeo NSF Earthcube Proposal. figshare. Figshare paper.