PANGEO
THE STATE OF THE ART
AT NCAR & BEYOND

KEVIN PAUL, ANDERSON BANIHIRWE, HAIYING XU,
JULIA KENT, JOSEPH HAMMAN, MATTHEW LONG

NATIONAL CENTER FOR ATMOSPHERIC RESEARCH,
BOULDER, CO, USA

JUNE 12, 2019
What drives progress in climate & weather research?

Observations

Simulations

Ideas

\[
\begin{align*}
\frac{D u}{D t} - f v &= - \frac{\partial \phi}{\partial x} \\
\frac{D v}{D t} + f u &= - \frac{\partial \phi}{\partial y}
\end{align*}
\]
Big Geoscience Data is here!

From Remote Sensing Platforms:
- New sensors / platforms
- Continuous observations
- Multiple versions of derived datasets

From Earth System Models:
- Higher resolution
- More process representation
- Larger ensembles
- On track for exabytes by CMIP7

Projected NASA Cloud Storage

<table>
<thead>
<tr>
<th>Year</th>
<th>Cumulative Archive Size (PB)</th>
<th>Archive Growth Rate (PB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>15.0</td>
<td>2.6</td>
</tr>
<tr>
<td>2016</td>
<td>17.7</td>
<td>2.9</td>
</tr>
<tr>
<td>2017</td>
<td>21.6</td>
<td>3.0</td>
</tr>
<tr>
<td>2018</td>
<td>26.8</td>
<td>5.2</td>
</tr>
<tr>
<td>2019</td>
<td>33.0</td>
<td>5.2</td>
</tr>
<tr>
<td>2020</td>
<td>37.7</td>
<td>5.2</td>
</tr>
<tr>
<td>2021</td>
<td>57.6</td>
<td>18.4</td>
</tr>
<tr>
<td>2022</td>
<td>103.4</td>
<td>47.7</td>
</tr>
<tr>
<td>2023</td>
<td>151.1</td>
<td>47.7</td>
</tr>
<tr>
<td>2024</td>
<td>208.9</td>
<td>47.7</td>
</tr>
<tr>
<td>2025</td>
<td>246.6</td>
<td>47.7</td>
</tr>
</tbody>
</table>

Size of CMIP Archives

- CMIP: 1.00E+09
- CMIP2: 5.00E+11
- CMIP3: 3.60E+13
- CMIP5: 3.30E+15
- CMIP6: 1.50E+17

300 PB

150 PB
What is Pangeo?
Pangeo is a community that promotes open, reproducible, and scalable science. …just to name a few.
Pangeo is a funded *collaboration* dedicated to the advancement of scalable, interactive, easy-to-use data analysis for the climate and weather community.
Pangeo is a platform that is deployable on HPC or cloud.
Pangeo is about integration.
Pangeo is open source.

GitHub
https://github.com/pangeo-data/pangeo

GITTER
NCAR is an Institutional Partner with Pangeo.
NCAR has invested in two full-time software engineers on Pangeo.
NCAR’s contributions to Pangeo

- JupyterHub on Cheyenne
- Data Cataloging with Intake
- Domain-specific tooling on top of Xarray
- Alternative I/O Studies
- Benchmarking and Scaling studies
JupyterHub on Cheyenne

Spawner Options

- **Job Name (-N)**: Jupyter
- **Enter Queue or Reservation (-q)**: share
- **Specify your project account (-A)**
- **Specify N node(s) (-I select=N)**: 1
- **Specify N CPUs per node (-I ncpus=N)**: 1
- **Specify N MPI tasks per node (-I mpiprocs=N)**: 1
- **Specify N threads per process (-I ompthreads=N)**: 1
- **Specify wall time (-I walltime=HH:MM:SS) (12 Hr Maximum)**: 02:00:00

[Spawn button]
Dask-jobqueue & NCAR-jobqueue
Intake & Intake - ESM
ESMLab
Training & Education Activities

Helping Port Trusted Analysis Code to Pangeo
  • From IDL
  • From NCL

Education
  • Giving Tutorials
  • Planning Hackathons
  • Micro IT Support
Benchmarking & Scaling Studies
Questions?