The 800 lb Gorilla: Really Big Data

UCAR SEA Conference 2012
Gary Strand
strandwg@ucar.edu
supercomputer

A machine that takes one problem (computation) and turns it into more - I/O, storage, and management.
A little order-of-magnitude exercise

**mega (M, \(10^6\))**  
*Cray 1A (1970s-1980s; 160 MF)*  
small single netCDF file (today, 100s of MB)

**giga (G, \(10^9\))**  
*blackforest (2000; 906 GF)*  
large single netCDF file (today, 10s of GB)

**tera (T, \(10^{12}\))**  
*bluesky (2003-2007, 8 TF)*  
*bluefire (2008-; 77 TF)*  
CMIP3 archive (2004, 10 TB from NCAR, 35 TB total)  
local disk array (silver, 125 TB disk)

**peta (P, \(10^{15}\))**  
*yellowstone (2012-? 1.5 PF)*  
CMIP5 archive (2013-, 400 TB from NCAR, 2+ PB total)  
All GCM data (3.5 PB)  
NCAR HPSS (15 PB), GLADE (1.8 PB)

**exa (E, \(10^{18}\))**  
*next NWSC machine? (2017-2018; 1-2 EF)*  
SKA (2014-2015, 1 EB/day)  
CMIP6? (circa 2020, total 1-20 EB?)

Friday, February 24, 2012
“The volume of worldwide climate data is expanding rapidly, creating challenges for both physical archiving and sharing, as well as for ease of access and finding what’s needed, particularly if you are not a climate scientist.”
Typical CESM output arrangement

Useful arrangement

\[ \begin{array}{c|c|c|c|c|c} \hline t_0 & t_1 & t_2 & \ldots & t_m \\ \hline f_1 & f_1 & f_1 & \ldots & f_1 \\ \hline f_2 & f_2 & f_2 & \ldots & f_2 \\ \hline \ldots & \ldots & \ldots & \ldots & \ldots \\ \hline f_n & f_n & f_n & \ldots & f_n \\ \hline \end{array} \]
CESM “1°” output volumes (TB)

- red: other
- orange: CMIP5 format
- green: post-processed
- blue: model output
CESM Data Management

- What counts as CESM data?
- Who is responsible and what are their obligations?
- What gets released and when?
- For how long are the data stored?
- Standards and more standards - conventions too.
- Future challenges
Major Categories of CESM Data

**Development**

- Evaluation
- Testing

Typically short duration, local use

**Production**

- “Control”
- Experiment

Typically long duration, external use

*Dictates many aspects of the CESM DMP*
Ownership Rights & Responsibilities

Ownership

• Principal Investigator (including SSC)
• Working group
  • First right-of-use

Responsibilities

• Adherence to policy
  • Guidelines on release timeline
Data Release Timeline
Development and Production

• All are originally “Protected”
• Six months sole use - with caveats
• 6 months to 12 months, WG access
• 12+ months, public access

Caveats

• Discretion of SSC
• Additional per-instance restrictions
• Strongly advisory - not strictly mandatory

Friday, February 24, 2012
Retention of CESM Data

Development

• All data for no less than 3 years
• Removal unless exceptions made
• SSC has final authority

Production

• All data for no less than 4 years
• Stepwise reduction for next 3 years
• Latitude for maintenance costs
Metadata Standards/Requirements

- Climate and Forecast (CF) conventions
  - Adhered to as closely as possible
- Model Intercomparison projects (MIPs)
  - Including CMIP5, CFMIP, ACCMIP, etc.
- Others...
CMIP5 metadata requirements

Standard model output for specific variable

```c
float TS(time, lat, lon) ;
TS:units = "K" ;
TS:long_name = "Surface temperature (radiative)" ;
TS:cell_method = "time: mean" ;
```

As required by CMIP5

```c
float ts(time, lat, lon) ;
ts:standard_name = "surface_temperature" ;
ts:long_name = "Surface Temperature" ;
ts:comment = ""skin"" temperature (i.e., SST for open ocean)" ;
ts:units = "K" ;
ts:original_name = "TS" ;
ts:cell_methods = "time: mean (interval: 30 days)" ;
ts:cell_measures = "area: areacella" ;
ts:history = "2011-07-22T00:05:32Z altered by CMOR: replaced missing value flag (-1e+32) with standard missing value (1e+20)." ;
ts:missing_value = 1.e+20f ;
ts:_FillValue = 1.e+20f ;
ts:associated_files = "baseURL: http://cmip-pcmdi.llnl.gov/CMIP5/dataLocation gridspecFile: gridspec_atmos_fx_CCSM4_historical_r0i0p0.nc areacella: areacella_fx_CCSM4_historical_r0i0p0.nc" ;
```
CMIP5 metadata requirements

Standard model global attributes

As required by CMIP5
Challenges

Increasing resolution
Challenges
More components

Courtesy Caitlin Alexander, ClimateSight
Friday, February 24, 2012
Challenges
“Exotic” grids

The Parallel Ocean Program (POP) “tripolar” grid

## CESM CMIP5 simulations

<table>
<thead>
<tr>
<th>CMIP5 type</th>
<th>Description</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>piControl</td>
<td>pre-industrial control</td>
<td>11</td>
</tr>
<tr>
<td>1% CO2 increase</td>
<td>1 percent per year CO2</td>
<td>5</td>
</tr>
<tr>
<td>historical</td>
<td>Simulate 20th century climate</td>
<td>24</td>
</tr>
<tr>
<td>historical variations</td>
<td>Single forcing runs, etc.</td>
<td>28</td>
</tr>
<tr>
<td>paleoclimate</td>
<td>Past climate (LGM, mid-Holocene, past 1000 years)</td>
<td>3</td>
</tr>
<tr>
<td>RCPs</td>
<td>RCPs 2.6, 4.5, 6.0, 8.5</td>
<td>32</td>
</tr>
<tr>
<td>Decadal predictions</td>
<td>Predictions (hindcast and forecast)</td>
<td>400</td>
</tr>
<tr>
<td>ESM</td>
<td>Earth System Model (BGC, carbon cycle, &amp;c)</td>
<td>10</td>
</tr>
<tr>
<td>Other</td>
<td>Sensitivity and “idealized” Earths</td>
<td>25</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td></td>
<td>538</td>
</tr>
<tr>
<td></td>
<td>subdaily</td>
<td>daily</td>
</tr>
<tr>
<td>----------------</td>
<td>----------</td>
<td>-------</td>
</tr>
<tr>
<td>atmosphere</td>
<td>197</td>
<td>74</td>
</tr>
<tr>
<td>land surface</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>ocean</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>sea ice</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td><strong>totals</strong></td>
<td><strong>201</strong></td>
<td><strong>86</strong></td>
</tr>
</tbody>
</table>
Conclusions

- Output data volume is maintaining a roughly 3:1 ratio to FLOPs
- Many data tools are still embarrassingly serial, but work is being done (re: Rob’s talk)
- Ever-higher resolutions in space and time
- Data management requirements from funding agencies
- Data distribution and access (clouds?)
- Demands of intercomparison projects
- Citability (DOIs?) of data
Websites

CESM website
http://cesm.ucar.edu

CESM Data Management Plan

CMIP5 website
http://cmip.llnl.gov/cmip5