Data Management for Climate Science
Lessons Learned and Future Needs

Ilana Stern, NCAR
ilana@ucar.edu
Outline

○ NCAR HPC and big data
○ CMIP3 and CMIP5
○ Evolution of CESM workflow
○ Future plans and needs
<table>
<thead>
<tr>
<th>SI prefix</th>
<th>flops</th>
<th>bytes</th>
</tr>
</thead>
<tbody>
<tr>
<td>giga-</td>
<td>1988</td>
<td>long ago</td>
</tr>
<tr>
<td>tera-</td>
<td>2000</td>
<td>&lt; 1986</td>
</tr>
<tr>
<td>peta-</td>
<td>2012</td>
<td>2002</td>
</tr>
<tr>
<td>exa-</td>
<td>2024?</td>
<td>2025</td>
</tr>
<tr>
<td>zetta-</td>
<td>2036?</td>
<td>?</td>
</tr>
</tbody>
</table>
CMIP3/AR4 data volumes by group

total: 35 TB
CMIP5/AR5 data volumes by group

CMIP3 total: 35 TB
CMIP5 total: 2,200 TB
NCAR ESG-CET Portal Downloads (TB)

Monthly download volume (GB), last 24 months


Ilana Stern  ilana@ucar.edu

Joint Facilities User Forum on Data-Intensive Computing
Workflow prior to late 1990s

model

time 1
header
field 1
field 2
...  
field n

time 2
header
field 1
field 2
...  
field n

time m
header
field 1
field 2
...  
field n

small disk

tape archive

later

post-processing/analysis
Workflow 2000-2012

model

<table>
<thead>
<tr>
<th>time 1</th>
<th>time 2</th>
<th>time m</th>
</tr>
</thead>
<tbody>
<tr>
<td>header</td>
<td>header</td>
<td>header</td>
</tr>
<tr>
<td>field 1</td>
<td>field 1</td>
<td>field 1</td>
</tr>
<tr>
<td>field 2</td>
<td>field 2</td>
<td>field 2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>field n</td>
<td>field n</td>
<td>field n</td>
</tr>
</tbody>
</table>

...}

TB scale disk

post-processing/analysis

<table>
<thead>
<tr>
<th>field 1</th>
<th>header</th>
<th>time 1</th>
<th>time 2</th>
<th>...</th>
<th>time m</th>
</tr>
</thead>
<tbody>
<tr>
<td>field 2</td>
<td>header</td>
<td>time 1</td>
<td>time 2</td>
<td>...</td>
<td>time m</td>
</tr>
<tr>
<td>field n</td>
<td>header</td>
<td>time 1</td>
<td>time 2</td>
<td>...</td>
<td>time m</td>
</tr>
</tbody>
</table>

TB scale disk

tape archive

publish

data portal

publish

tape archive
Workflow 2000-2012

Typical workflow for standard runs, CMIPs

- Model writes time-slice (aka “history”) files from disk to tape
- Files read back from tape to disk for postprocessing
- Post-processed files (aka “time-series”) written to tape
- If for MIP, convert time-series files with CMOR
  - If necessary for MIP, much easier to work with time-series data

Issues:

- Many writes/reads to/from tape
- Multiple copies of identical data
- Constrained by “small” disk space
- Hugely inefficient
Current workflow

**Model**
- time 1: header, field 1, field 2, ..., field n
- time 2: header, field 1, field 2, ..., field n
- time m: header, field 1, field 2, ..., field n

**Post-processing/analysis**
- field 1: header, time 1, time 2, ..., time m
- field 2: header, time 1, time 2, ..., time m
- field n: header, time 1, time 2, ..., time m

**PB Scale Disk**

**Tape Archive**

**Data Portal**
Current workflow

- Model writes time-slice files to disk only
- Files postprocessed in situ
- Only time-series files written to HPSS
- Currently uses separate serial (nc3) postprocessing scripts
  - Within 3 months: parallel (nc4) postprocessing scripts
  - Eventually: PIO (Parallel I/O) package to be rewritten to write single-field format directly from model

Advantages

- Many fewer writes to HPSS - no reads
- Possible because of sufficient disk space now available
- Single copies of data
- Much more efficient for analysis purposes
Near-term future Workflow

model

<table>
<thead>
<tr>
<th>field 1</th>
<th>header</th>
<th>time 1</th>
<th>time 2</th>
<th>...</th>
<th>time m</th>
</tr>
</thead>
<tbody>
<tr>
<td>field 2</td>
<td>header</td>
<td>time 1</td>
<td>time 2</td>
<td>...</td>
<td>time m</td>
</tr>
<tr>
<td>field n</td>
<td>header</td>
<td>time 1</td>
<td>time 2</td>
<td>...</td>
<td>time m</td>
</tr>
</tbody>
</table>

analysis

<table>
<thead>
<tr>
<th>field 1</th>
<th>header</th>
<th>time 1</th>
<th>time 2</th>
<th>...</th>
<th>time m</th>
</tr>
</thead>
<tbody>
<tr>
<td>field 2</td>
<td>header</td>
<td>time 1</td>
<td>time 2</td>
<td>...</td>
<td>time m</td>
</tr>
<tr>
<td>field n</td>
<td>header</td>
<td>time 1</td>
<td>time 2</td>
<td>...</td>
<td>time m</td>
</tr>
</tbody>
</table>

10s PB disk

tape archive

data portal

publish
Near-term future workflow

I/O package writes single-field format directly from model

Advantages

○ Many fewer writes to HPSS - no reads
○ Possible because of sufficient disk space now available
○ Single copies of data
○ Much more efficient
○ Potential MIPs are potentially written directly from model, no additional steps required

Issues

○ Potential difficulties with reruns or branches
○ Possible to create unwieldy very large (1 TB+) files
○ Granularity of file sizes very dependent on exact configuration
Exascale workflow

modeling

100s PB disk

tape archive (?)

data portal (analyzed/derived data only)

analysis

analyzed/derived data

publish (?)

publish

publish (?)
Current CESM big data projects

Current

CESM1-CAM5-BGC ensemble
• 40 runs, total ~7,500 model years, ~400 TB

Last millennium ensemble
• 26 runs, total ~26,000 model years, ~600 TB
(Both using newest workflow)

Longer-term big data

CMIP6 (2016-2017)
Potential additional -MIPs
Higher resolution (1/8° SE atm/Ind, 1/10° ocn/ice)
# Current CESM big data projects

<table>
<thead>
<tr>
<th>Category</th>
<th>CMIP5</th>
<th>Large Ensemble</th>
<th>Last Millennium Ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model(s)</strong></td>
<td>CCSM4, CESM1-CAM5, CESM1-BGC, CESM1-WACCM, CESM1-FASTCHEM</td>
<td>CESM1-CAM5-BGC</td>
<td>CESM1-CAM5</td>
</tr>
<tr>
<td><strong>Volume measure 1</strong></td>
<td>~1,600 TB</td>
<td>~750 TB</td>
<td>~1,000 TB</td>
</tr>
<tr>
<td><strong>Volume measure 2</strong></td>
<td>~180 TB</td>
<td>~300 TB</td>
<td>~420 TB</td>
</tr>
<tr>
<td><strong>Total simulated years</strong></td>
<td>~30,000</td>
<td>~7,500</td>
<td>~26,000</td>
</tr>
<tr>
<td><strong>Number of runs</strong></td>
<td>555</td>
<td>62 (+ at least 10 more)</td>
<td>26</td>
</tr>
<tr>
<td><strong>Output categories</strong></td>
<td>19</td>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td><strong>Number of fields</strong></td>
<td>951</td>
<td>1127</td>
<td>820</td>
</tr>
</tbody>
</table>
DOIs for data

- UCAR-wide and institutionally-supported effort to add Digital Object Identifiers (DOIs) to UCAR data holdings
  - Effort led by the NCAR Library with participation by the Research Data Archive (RDA), EOL, CESM, HAO, etc.
  - NCL and Yellowstone have been assigned DOIs, among a number of RDA-held datasets and EOL-held datasets
- Intent is to make datasets citable for “forever”, enabling reproducibility and traceability for UCAR science
- At this time, does not imply data are “peer-reviewed”
- A “permanent” pointer to a landing page for data
- Give credit to data creators and data managers and scientists for their work
- UCAR uses the EZID service to generate random DOIs with the assigned-to-UCAR prefix [http://dx.doi.org/10.5065](http://dx.doi.org/10.5065)
- DOIs are stored by DataCite and the responsible unit maintains them
CESM and DOIs

- CESM-CAM5-BGC Large Ensemble and Last Millennium Ensemble to serve as test cases
- Issues include “granularity” - at what level (experiment, component, component stream, individual file) are DOIs assigned?
- What about simulations done elsewhere (DOE, etc.)?
- Permanence (really?) and responsibility
- Coordination of DOIs with MIP projects - how do we link CMIP5 data to the original CESM output, once the latter has a DOI?
CESM plans and future needs

Issues

○ User community needs/wants drives all!
○ Modeling and analysis ~concurrently to avoid memory -> disk latency and all the other issues
○ Ongoing updates of workflow
○ Updates to CESM data management policy to reflect workflow and other changes
○ Longer-term viability of ESG/ESGF model - downloading PB isn’t sustainable (?)
○ Must have serious server-side analysis
○ Possibility of reruns instead of save everything
Thank you!

Questions?