Working with Climate Model Data (and more) as a Data Scientist

Gary Strand
strandwg@ucar.edu
How it began...

**STUDENT ASSISTANT II - #0591**

AAP - Global Climate Modeling Group  
Flat Rate: $6.90/hour  
DUTIES: Modifies existing atmospheric and ocean model computer programs to perform new functions and/or to improve their flexibility and ease of operation. Writes new utility computer programs for scientists which satisfy their requirements. Operates, as needed, computer programs for the storage and processing of data.

REQUIRES:
-- Basic FORTRAN programming skills
-- Skill in handling large data sets
-- Skill in transferring data between the computers, mass storage, and physical tape
-- Skill in understanding scientists' requirements and translating these into workable programs
-- Successful completion of one year of college level physics, mathematics, or atmospheric science

Staff Notes 21 (35), 29 Aug 1986
The Larval Years (1986-1989)

The “Interactive Data Subset Processor” (IDSP)

- Originally designed for global subsets
- DEC Fortran on VAX 11/750-780 series
The Larval Years (1986-1989)

The “Interactive Data Subset Processor” (IDSP)
- Ported to allow larger global datasets; large is relative (3 MB or so binary file)
- One job could take 8 hours+ to run, I made it more efficient and got it down to 1 hour wallclock
- Generated many plots and microfiche for later analysis

Data Management and Best Practices
Negligible
The Pupal Years (1990-1995)

Parallel Ocean Climate Model (POCM)
Generated ocean bathymetry
- very fine scale (1/12th°) global elevation dataset interpolated to 1°x1° land/sea mask
- 38 MB was too large for the Cray-1 to hold in memory

Data Management and Best Practices
Multiple versions of datasets created
Revision control of source code, input and output data
ACKNOWLEDGMENTS

We thank Gary Strand for his assistance in archiving the satellite data sets. His programming skills were invaluable to the project.
The Pupal Years (1990-1995)

Monthly Mean Global Satellite Data Sets Available in CCM History Tape Format

Data Management and Best Practices

Multiple versions of datasets created
Revision control of source code and input and output data
All data permanently on NCAR tape archive:

/CTSAT

Source code backed up and archived:
The Pupal Years (1990-1995)

Analysis of the mean annual cycle of the dissolved oxygen anomaly in the World Ocean


ACKNOWLEDGMENTS

[...] We are thankful for the superb programming support of Gary Strand and Chuck Reich. We are grateful to Steve Worley for his assistance in the unpacking and initial processing of the NODC data. [...]
The Pupal Years (1990-1995)

Analysis of the mean annual cycle of the dissolved oxygen anomaly in the World Ocean

Data Management and Best Practices

Multiple versions of datasets created

Revision control of source code and data

All data permanently on NCAR tape:

/home/strandwg/NODC

Source code backed up and archived:

1993/05/20: /home/strandwg/NODC/CODE/nodc_code.tar
1993/05/28: /home/strandwg/NODC/CODE/nodc_code2.0.tar.Z
1993/12/06: /home/strandwg/NODC/CODE/nodc_code3.0.tar.Z
The Juvenile Years (1996-1997)
Mostly some diversions from data stuff...

• Working on an early version of the CSM Sea Ice Model
What is a climate model?

![Diagram showing interactions between atmosphere, land, ocean, and sea ice models]

- Atmosphere (CAM) interacts with:
  - Solar radiation
  - Natural emissions
  - Anthropogenic emissions

- Coupler (CPL) connects:
  - Land ice (G-CISM)
  - Sea ice (CICE)
  - Land (CLM)
  - Ocean (POP)

Land use changes influence the couple.

Courtesy Caitlin Alexander, ClimateSight

DCERC Data Curation Workshop, June 3-4 2014
What is a climate model?

Processes

Grid
What is a climate model?
The climate model era (1998 to now)

Parallel Climate Model (PCM)

Department of Energy-funded climate model

- Designed to be run on DOE supercomputers for long-term high-resolution high-impact climate studies
- Data made available to the climate science community and anyone else via access to the NERSC HPSS system
- Community simulations and data distribution
- Use of standardized data formats (netCDF)
- Use of then-emerging data standards
- Data still (occasionally) in use today
The climate model era (1998 to now)
Parallel Climate Model (PCM)
The climate model era (1998 to now)

Data Management and Best Practices
- Sometimes multiple versions of datasets created
- Standardized post-processing
- Conversion/post-processing part of PCM configuration
- All data archived permanently on tape:
  NCAR: /PCM1
  NERSC: /home/$user
  ORNL: /f2|home/$user
- Data originally distributed via tape archive access at PCMDI
The climate model era (1998 to now)
Community Earth System Model (CESM)
DOE/NSF funded climate model

- Designed to be run on supercomputers for long-term high-resolution high-impact climate studies
- Source code readily available and supported
- Many possible configurations (SCAM to ESM)
- Data made available to the climate science community and anyone else
- Community simulations and data distribution
- NCAR’s flagship earth system model
The climate model era (1998 to now)

**Data Management and Best Practices**

- Sometimes multiple versions of datasets created
- Standardized post-processing
- Use of standardized data formats (netCDF) and data standards (CF)
- All data archived permanently on tape:
  
  - NCAR: /CCSM/csm
  - NERSC: /home/c/ccsm/csm
  - ORNL: /f2/ccsm/csm
- Data distributed via ESG
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.
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
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
General timeline summary

- % data kept
- % restricted

- end of run
- + 6 months
- +12 months
- +4 years
- +5 years
- +6 years
- +7 years
- later
The Earth System Grid (2001-today)

DOE-funded data distribution system

- Designed to address the issues with “manual” data distribution from PCMDI
- Built upon the early work on efficient, robust large-scale data movement between DOE centers
- Data to be made available to the climate science community via the tape archive at each center
- Uses web browsing, discovery metadata, access control, metrics, and several means of data download (http, ‘wget’, ‘DML’)
- Many projects (NARCCAP, NCL, TRaCE, etc.)
- Very heavy use, including to the present
The Earth System Grid (2001-today)

ESG and CMIP3/IPCC AR4

- Ported by PCMDI for the IPCC 4th Assessment report (IPCC AR4)
- Data distributed from the central repository at PCMDI
- Modeling groups transferred data by disk
- Total archive about 35 TB
- Hundreds of papers and thousands of users
- Data still heavily used and in active use
The Earth System Grid (2001-today)

NCAR ESG-CET portal cumulative download volume (TB)
The Earth System Grid Federation
ESG-F and CMIP5/IPCC AR5

- Led by PCMDI for the IPCC 5th Assessment report (IPCC AR5)
- International collaboration - UK, Australia, Germany, and more
- Federation of distributed data centers
- Modeling groups keep their data local
- Total archive about 2,500 TB
- Datasets assigned DOIs
- Hundreds of papers and thousands of users
- Data very heavily used and in active use
## The climate model era - MIPs

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Models used</td>
<td>2 (CCSM3 and PCM)</td>
<td>5 (CCSM4, CESM1-BGC, CESM1-WACCM, CESM1-FASTCHEM, CESM1-CAM5)</td>
</tr>
<tr>
<td>Total volume submitted</td>
<td>~9.2 TB (10 months)</td>
<td>~140 TB (one year), + ~30 TB since</td>
</tr>
<tr>
<td>Total volume generated</td>
<td>~120 TB</td>
<td>~1500 TB</td>
</tr>
<tr>
<td>Total simulated years</td>
<td>~14,900</td>
<td>~30,000</td>
</tr>
<tr>
<td>Number of model runs</td>
<td>107</td>
<td>555</td>
</tr>
<tr>
<td></td>
<td>73 (CCSM3)</td>
<td>91 (CCSM4 long-term)</td>
</tr>
<tr>
<td></td>
<td>34 (PCM1)</td>
<td>400 (CCSM4 decadal prediction)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>64 (other configurations)</td>
</tr>
<tr>
<td>Experiments requested</td>
<td>12</td>
<td>37</td>
</tr>
<tr>
<td>Output categories</td>
<td>6</td>
<td>19</td>
</tr>
<tr>
<td>Number of requested fields</td>
<td>137</td>
<td>951</td>
</tr>
</tbody>
</table>
CMIP5 distribution

- Modeling centers (24)
- Gateways (9)
- Nodes (14)
- Users (1000s)
IPCC AR5 metadata standard

Standard CESM output for a specific variable

float TS(time, lat, lon)
TS:units = "K"
TS:long_name = "Surface temperature (radiative)"
TS:cell_method = "time: mean"
IPCC AR5 metadata standard

Same variable as required by IPCC AR5

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"
IPCC AR5 metadata standard

Standard CESM global attributes

:Conventions = "CF-1.0"
:source = "CAM"
:case = "b40.20th.track1.1deg.006"
:title = "UNSET"
:logname = "mai"
:host = "be0809en.ucar.ed"
:Version = "$Name$"
:revision_Id = "$Id$"
:initial_file = "b40.1850.track1.1deg.006.cam2.i.0893-01-01-00000.nc"
:topography_file = "'/fis/cgd/cseg/csm/inputdata/atm/cam/topo/USGS-gtopo30_0.9x1.25_remap_c051027.nc"
:nco_openmp_thread_number = 1
Global attributes as required by IPCC AR5

:project_id = "CMIP5"
:product = "output"
:frequency = "mon"
:modeling_realm = "atmos"

:institution = "NCAR (National Center for Atmospheric Research), Boulder, CO, USA"
:institute_id = "NCAR"
:model_id = "CCSM4"
:source = "CCSM4 (tag: ccsm4_0_beta43 compset: B20TRCN)"
:resolution = "f09_g16 (0.9x1.25_gx1v6)"
:title = "CCSM4 model output prepared for CMIP5 historical"
:contact = "cesm_data@ucar.edu"

:acknowledgements = "The CESM project is supported by the National Science Foundation and the Office of Science (BER) of the U.S. Department of Energy. NCAR is sponsored by the National Science Foundation. Computing resources were provided by the Climate Simulation Laboratory at the NCAR Computational and Information Systems Laboratory (CISL), sponsored by the National Science Foundation and other agencies."

DCERC Data Curation Workshop, June 3-4 2014
Global attributes as required by IPCC AR5

:experiment = "historical"
:experiment_id = "historical"
:forcing = "Sl GHG Vl SS Ds SD BC MD OC Oz AA LU"

:realization = 1
:initialization_method = 1
:physics_version = 1

:parent_experiment = "pre-industrial control"
:parent_experiment_id = "piControl"
:parent_experiment_rip = "r1i1p1"
:branch_time = 937.

:tracking_id = "d33ccf77-a73c-4f55-8f02-3a0734d51151"

:creation_date = "2011-07-22T00:05:32Z"
:history = "2011-07-22T00:05:32Z CMOR rewrote data to comply with CF standards and CMIP5 requirements."
:Conventions = "CF-1.4"
:table_id = "Table Amon (27 April 2011) a5a1c518f52ae340313ba0aada03f862"
:cmor_version = "2.7.1"
Additional global attributes

:cesm_casename = "b40.20th.aero.1deg.008"
:cesm_repotag = "ccsm4_0_beta56"
:cesm_compset = "B1850CN"

:resolution = "f09_g16 (0.9x1.25_gx1v6)"

:forcing_note = "Additional information on the external forcings used in this experiment can be found at http://www.cesm.ucar.edu/CMIP5/forcing_information"

:processed_by = "strandwg on silver.cgd.ucar.edu at 20120205 -073243.338"

:processing_code_information
   = "Last Changed Rev: 525
    Last Changed Date: 2012-02-04 13:11:55 -0700
    Repository UUID: d2181dbe-5796-6825-dc7f-cbd98591f93d"
CMIP5 Data Quality Control

- Data Published > 10 PB
  - Data and Metadata QC L1
    - On globally distributed data nodes

- Metadata QC L2 passed?
  - YES
    - Data QC L2 passed?
      - YES
        - To be replicated among ESGF?
          - YES
            - Data QC L3 passed?
              - YES
                - DOI Assigned:
                  - Data formally citable
                  - Data can appear in IPCC-DDC
                  - Automatic access granted after filling in ESGF registration page
              - NO
                - Data NOT formally citable
                  - Data formally citable
                  - Modelling group control access manually.

            - NO
              - Discard data

          - NO
            - Data NOT formally citable
              - Data NOT formally citable
              - Automatic access granted after filling in ESGF registration page.

  - NO
    - Discard data

(Informal citation still requested where formal citation not available)
# CMIP5 Data Quality Control

<table>
<thead>
<tr>
<th>QC Level 1</th>
<th>QC Level 2</th>
<th>QC Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
<td>CMOR2 and ESG publisher conformance checks</td>
<td>Data consistency checks</td>
</tr>
<tr>
<td><strong>Data</strong></td>
<td>preliminary; no user notification about changes; performed for all data; metadata may not be complete</td>
<td>no user notification about changes; performed for CMIP5 requested metadata and data</td>
</tr>
<tr>
<td><strong>Access</strong></td>
<td>constrained to CMIP5 modeling centers</td>
<td>constrained to non-commercial research and educational purposes</td>
</tr>
<tr>
<td><strong>Access Control</strong></td>
<td>PCMDI on behalf of WMO/WGCM</td>
<td>PCMDI, BADC, WDCC/DKRZ core data archives on behalf of WMO/WGCM</td>
</tr>
<tr>
<td><strong>Citation</strong></td>
<td>no citation reference</td>
<td>informal citation reference</td>
</tr>
<tr>
<td><strong>Quality Flag</strong></td>
<td>&quot;automated conformance checks passed&quot;</td>
<td>&quot;subjective quality control passed&quot;</td>
</tr>
</tbody>
</table>
NCAR flops and bytes, 2000-2030

Terabytes

Projected TB

Projected TF

(Teraflops)

(Peta)


Terabytes

Projected TB

Projected TF

(Teraflops)

(Peta)
Current CESM 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**

**publish**

**data portal**

- netCDF-3
- netCDF-4

DCERC Data Curation Workshop, June 3-4 2014
Workflow ~2015

**model**

- **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

**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

10s PB disk

**tape archive**

**publish**

**data portal**
Exascale workflow

modeling

analysis

analyzed/derived data

100s PB disk

tape archive (??)

Data portal (analyzed/derived data only)

publish

publish (??)
Current big data projects

Near-term

CESM1-CAM5-BGC ensemble
• 40 runs, total ~7,500 model years, ~400 TB
“Last millenium” ensemble
• 26 runs, total ~26,000 model years, ~600 TB
(Both using newest workflow)

Longer-term

CMIP6 (2016-2017?)
Potential additional *MIPs
Higher resolution (1/8° SE atm/Ind, 1/10° ocn/ice)
CESM and the exascale

Issues

- Meeting user community needs/wants drives all!
- Modeling and analysis ~concurrently
- Ongoing updates of workflow
- Updating 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 rerunning simulations instead of saving output in bulk
References

CESM website
http://cesm.ucar.edu

CESM Data Management Plan

ESG website
http://www.earthsystemgrid.org

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