Climate Modeling as a Data Intensive Science

Robert Jacob
Mathematics and Computer Science Division, Argonne National Laboratory
Computation Institute, Argonne/University of Chicago

July 30, 2012
An old saying....
“Climate is what you expect, weather is what you get”

- Climate is the *average* of weather.

- The (predicted) high temperature today, Jul 30\(^{th}\), is 71F

- The average high temperature is 70F. This is calculated by taking the average of several (usually 30) Jul 30\(^{th}\) highs.

\[
\frac{(T \text{ Jul 30th, 1981}) + (T \text{ Jul 30th, 1982}) + \ldots + (T \text{ Jul 30th, 2010})}{30}
\]

To model the climate system, must model years of global weather.
Climate is the average of weather.

What makes weather? The Sun and the Earth’s rotation.
Need to simulate weather-scale phenomena over the entire globe.

Weather is embedded in the *general circulation* of the atmosphere.
Over may days, months, atmosphere circulation is dominated by interaction with surface.
Atmospheric General Circulation Model

- Algorithms to solve the primitive equations called “the dynamics”; “dynamical core” “dycore”
- Forcing terms: \( F(t,u,v,\phi) \)
  - Change in temperature due to radiative transfer
  - Effect of clouds on radiative transfer
  - Change in moisture due to cloud, rain formation
  - Change in temperature due to sensible heat transport through the boundary layer
  - Change in temperature due to release of latent heat
  - Change in momentum due to friction with surface.
- Algorithms for the above called “the physics” or “column physics”.
- Major groupings: longwave radiation, shortwave radiation, boundary layer, deep convection, cloud fraction, gravity wave drag.
- Can take as much or more computer time as the dynamics.
Ocean General Circulation Model

- Very Similar to AGCM except:
  - Presence of side boundaries. Nearly all OGCM’s are FD with z-coordinates.
  - Not as much “physics”
  - Motions are slower. Length scales are shorter.
  - Much higher heat capacity. The memory of the climate system is in the ocean.
Sea Ice Models

- Thermodynamics: formation, growth, melting, albedo, melt ponds.

- Dynamics: transport, internal stress, ridging
Land Surface Models

- Nearly all “physics”:
  - Vegetation composition, structure
  - Vertical heat transfer in soil.
  - Heat, radiation transfer between ground, canopy and free atmosphere
  - Hydrology of canopy, snow, soil moisture
  - River runoff

- Historically, was part of column physics in the atmosphere model.

Figure: Bonan, NCAR
Primitive Equations must be solved numerically
Other grids
NCAR/DOE Coupled Climate Model CCSM4/CESM1

- **Ocn (POP 2.0)**: 320x384x40
- **Atm (CAM 4.0)**: 128x256x26
- **Ice (CSIM 4)**: 320x384x5
- **Lnd (CLM 4)**: 128x256x10

Coupler (cpl7)
Take your coupled global climate model and calculate evolution of global weather for 100 years, 20 minutes at a time.

- CCSM3 (150km): 1 quadrillion operations/simulated year.
- After 100 quadrillion operations, what do you know about the climate?

NOTHING!
The data intensive part:

- Climate is revealed by calculating statistics on “climate” model output.
  - Averages over time and space.
  - Other moments
  - More sophisticated analysis: CCA, PCA, etc.
Climate model output

- Since running a model is very expensive AND
- Since the science comes from analyzing the output.

- Output everything!
  - Prognostic state variables
  - Derived quantities
  - Approximately 100 different variables. 25% 3D, rest 2D or 1D.

- ....But don’t save everything for all times
  - Monthly output of all variables.
  - Daily or 4-hourly output of some of the same variables.
<table>
<thead>
<tr>
<th>CMIP3</th>
<th>subdaily</th>
<th>daily</th>
<th>monthly</th>
<th>annual</th>
<th>totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>atmosphere</td>
<td>9</td>
<td>18</td>
<td>47</td>
<td>0</td>
<td>74</td>
</tr>
<tr>
<td>land surface</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>ocean</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>sea ice</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>totals</td>
<td>9</td>
<td>18</td>
<td>72</td>
<td>0</td>
<td>99</td>
</tr>
</tbody>
</table>

Gary Strand, NCAR
## CMIP3 vs. CMIP5

<table>
<thead>
<tr>
<th>CMIP5</th>
<th>subdaily</th>
<th>daily</th>
<th>monthly</th>
<th>annual</th>
<th>totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>atmosphere</td>
<td>100</td>
<td>75</td>
<td>223</td>
<td>8</td>
<td>406</td>
</tr>
<tr>
<td>land surface</td>
<td>3</td>
<td>5</td>
<td>82</td>
<td>0</td>
<td>90</td>
</tr>
<tr>
<td>ocean</td>
<td>1</td>
<td>3</td>
<td>127</td>
<td>79</td>
<td>210</td>
</tr>
<tr>
<td>sea ice</td>
<td>0</td>
<td>4</td>
<td>40</td>
<td>0</td>
<td>44</td>
</tr>
<tr>
<td>totals</td>
<td>104</td>
<td>87</td>
<td>472</td>
<td>87</td>
<td>750</td>
</tr>
</tbody>
</table>
Typical climate model data sizes

- Atmosphere Model (single output file of all variables, one time step)
  - 1 degree: 233MB
  - 0.5 degree: 821MB

- Ocean Model
  - 3 degree: 20 MB
  - 1 degree: 1.1 GB

- Sea Ice Model
  - 1 degree: 69 MB

- Land Model
  - 1 degree: 86 MB
# CMIP3 vs. CMIP5

<table>
<thead>
<tr>
<th>Modeling group</th>
<th>CMIP3 volume (GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCAR</td>
<td>USA</td>
</tr>
<tr>
<td>MIROC3</td>
<td>Japan</td>
</tr>
<tr>
<td>GFDL</td>
<td>USA</td>
</tr>
<tr>
<td>IAP</td>
<td>China</td>
</tr>
<tr>
<td>MPI</td>
<td>Germany</td>
</tr>
<tr>
<td>CSIRO</td>
<td>Australia</td>
</tr>
<tr>
<td>CCCMA</td>
<td>Canada</td>
</tr>
<tr>
<td>INGV</td>
<td>Italy</td>
</tr>
<tr>
<td>GISS</td>
<td>USA</td>
</tr>
<tr>
<td>MRI</td>
<td>Japan</td>
</tr>
<tr>
<td>CNRM</td>
<td>France</td>
</tr>
<tr>
<td>IPSL</td>
<td>France</td>
</tr>
<tr>
<td>UKMO</td>
<td>UK</td>
</tr>
<tr>
<td>BCCR</td>
<td>Norway</td>
</tr>
<tr>
<td>MIUB</td>
<td>Germany/Korea</td>
</tr>
<tr>
<td>INMCM3</td>
<td>Russia</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Modeling group</th>
<th>CMIP5 volume (GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPI</td>
<td>Germany</td>
</tr>
<tr>
<td>NCAR</td>
<td>USA</td>
</tr>
<tr>
<td>MRI</td>
<td>Japan</td>
</tr>
<tr>
<td>GFDL</td>
<td>USA</td>
</tr>
<tr>
<td>MIROC3</td>
<td>Japan</td>
</tr>
<tr>
<td>UKMO</td>
<td>UK</td>
</tr>
<tr>
<td>CNRM</td>
<td>France</td>
</tr>
<tr>
<td>IAP</td>
<td>China</td>
</tr>
<tr>
<td>U Reading</td>
<td>UK</td>
</tr>
<tr>
<td>EC</td>
<td>Europe</td>
</tr>
<tr>
<td>GISS</td>
<td>USA</td>
</tr>
<tr>
<td>INGV</td>
<td>Italy</td>
</tr>
<tr>
<td>IPSL</td>
<td>France</td>
</tr>
<tr>
<td>INMCM3</td>
<td>Russia</td>
</tr>
<tr>
<td>NorClim</td>
<td>Norway</td>
</tr>
<tr>
<td>CCCMA</td>
<td>Canada</td>
</tr>
<tr>
<td>CAWCR</td>
<td>Australia</td>
</tr>
<tr>
<td>CSIRO</td>
<td>Australia</td>
</tr>
<tr>
<td>METRI</td>
<td>Korea</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td></td>
</tr>
</tbody>
</table>
Future climate model data sizes

CAM-SE 0.125 degrees

- Single 3D variable: 616MB (real*8)
- Single 2D variable: 25MB (real*8)
- Total grid points per 3D variable: 3110402 x 26 (80M points)
- Single history file: 24GB

1 year: 392 GB
100 years: 39.2 TB

POP 0.1 degrees

- Single 3D variable: 1.45 GB (4 byte reals)
- Single history file: 18.94 GB
- Single restart file: 24.19 GB

1 year: 227 GB
100 years: 22.7 TB
The GCRM Tsunami

4 km, 100 levels, hourly data
~1 TB / simulated hour
~24 TB / simulated day
~9 PB / simulated year

Karen Schuchardt, PNNL
Commonly used tools for visualizing climate data

- **NCO - NetCDF common operators**: Command line tools to perform simple arithmetic (averages in space or time) on NetCDF files. Output is another NetCDF file.

- **NCL (NCAR), Ferret (PMEL), CDAT (LLNL)** -
  - Tools developed by climate community which understand climate specific viz needs such as spherical projections, continent outlines, specialized vertical coordinates (pressure, density).
  - Free! (but not all open source)
  - Mostly 2D and 1D plots. Very little 3D capability.
  - No animation capability.
  - Enter commands at interpreter prompt or write scripts (in custom language. NCL, CDAT have python interface.)

- Also IDL, Matlab, Mathematica.
NCAR Command Language (NCL)

A scripting language tailored for the analysis and visualization of geoscientific data

1. Simple, robust file input and output
2. Hundreds of analysis (computational) functions
3. Visualizations (2D) are publication quality and highly customizable

- Community-based tool
- Widely used by CESM developers/users
- UNIX binaries & source available, free
- Extensive website, regular workshops

http://www.ncl.ucar.edu/
**Argonne Leadership Computing Facility Hardware Layout**

**Intrepid**
- 40 racks / 160k cores
- 556 TF

**Eureka (Viz)**
- 100 nodes / 800 cores
- 200 GPUs
- 100 TF

**Networks**
- (ESnet, internet2, UltraScienceNet,...)

**I/O Switch Complex**
- 640 @ 10 Gig
- 100 @ 10 Gig

**(16) DDN 9900 – 128 file servers**
- /intrepid-fs0 (GPFS) 4.5PB
- /intrepid-fs1 (PVFS) 0.5PB
- Rate: 60+ GB/s

**(4) DDN 9550 – 16 file servers**
- Rate: 60+ GB/s

**(1) DDN 9900 - 8 file servers**
- /gpfs/home 100TB
- Rate: 8+ GB/s
- Tape Library 5PB
- 6500 LT04 @ 800GB each
- 24 drives @ 120 MB/s each
BER Program to address climate data issues:

- DOE LAB10-05: *Earth System Modeling: Advanced Visualization of Ultra-Large Climate Data Sets*. $5M/year for 3 years. Ending next year
  - Ultra-scale Visualization Climate Data Analysis Tools (UV-CDAT) PI: Williams
  - Visual Data Exploration and Analysis of Ultra-large Climate Data PI: Bethel
  - Parallel Analysis Tools and New Visualization Techniques for Ultra-large Climate Data Sets (ParVis) PI: Jacob
More Issues
In climate, the filesystem is the database.

ESM_SFASTv2_DMS_1850.cam2.h0.0046-06.nc  ESM_SFASTv2_DMS_1850.cam2.h0.0046-07.nc
ESM_SFASTv2_DMS_1850.cam2.h0.0046-08.nc  ESM_SFASTv2_DMS_1850.cam2.h0.0046-09.nc
ESM_SFASTv2_DMS_1850.cam2.h0.0046-10.nc  ESM_SFASTv2_DMS_1850.cam2.h0.0046-11.nc
ESM_SFASTv2_DMS_1850.cam2.h0.0046-12.nc  ESM_SFASTv2_DMS_1850.cam2.h0.0047-01.nc
ESM_SFASTv2_DMS_1850.cam2.h0.0047-02.nc  ESM_SFASTv2_DMS_1850.cam2.h0.0047-03.nc
ESM_SFASTv2_DMS_1850.cam2.h0.0047-04.nc  ESM_SFASTv2_DMS_1850.cam2.h0.0047-05.nc
ESM_SFASTv2_DMS_1850.cam2.h0.0047-06.nc  ESM_SFASTv2_DMS_1850.cam2.h0.0047-07.nc
ESM_SFASTv2_DMS_1850.cam2.h0.0047-08.nc  ESM_SFASTv2_DMS_1850.cam2.h0.0047-09.nc
ESM_SFASTv2_DMS_1850.cam2.h0.0047-10.nc  ESM_SFASTv2_DMS_1850.cam2.h0.0047-11.nc
ESM_SFASTv2_DMS_1850.cam2.h0.0047-12.nc  ESM_SFASTv2_DMS_1850.cam2.h0.0048-01.nc
ESM_SFASTv2_DMS_1850.cam2.h0.0048-02.nc  ESM_SFASTv2_DMS_1850.cam2.h0.0048-03.nc
ESM_SFASTv2_DMS_1850.cam2.h0.0048-04.nc  ESM_SFASTv2_DMS_1850.cam2.h0.0048-05.nc
ESM_SFASTv2_DMS_1850.cam2.h0.0048-06.nc  ESM_SFASTv2_DMS_1850.cam2.h0.0048-07.nc
ESM_SFASTv2_DMS_1850.cam2.h0.0048-08.nc  ESM_SFASTv2_DMS_1850.cam2.h0.0048-09.nc
ESM_SFASTv2_DMS_1850.cam2.h0.0048-10.nc  ESM_SFASTv2_DMS_1850.cam2.h0.0048-11.nc
ESM_SFASTv2_DMS_1850.cam2.h0.0048-12.nc  ESM_SFASTv2_DMS_1850.cam2.h0.0049-01.nc
ESM_SFASTv2_DMS_1850.cam2.h0.0049-02.nc  ESM_SFASTv2_DMS_1850.cam2.h0.0049-03.nc
ESM_SFASTv2_DMS_1850.cam2.h0.0049-04.nc  ESM_SFASTv2_DMS_1850.cam2.h0.0049-05.nc
ESM_SFASTv2_DMS_1850.cam2.h0.0049-06.nc  ESM_SFASTv2_DMS_1850.cam2.h0.0049-07.nc
ESM_SFASTv2_DMS_1850.cam2.h0.0049-08.nc  ESM_SFASTv2_DMS_1850.cam2.h0.0049-09.nc
ESM_SFASTv2_DMS_1850.cam2.h0.0049-10.nc  ESM_SFASTv2_DMS_1850.cam2.h0.0049-11.nc
ESM_SFASTv2_DMS_1850.cam2.h0.0049-12.nc  ESM_SFASTv2_DMS_1850.cam2.h0.0050-01.nc
ESM_SFASTv2_DMS_1850.cam2.h0.0050-02.nc  ESM_SFASTv2_DMS_1850.cam2.h0.0050-03.nc

Becomes really unworkable with ensembles.
CESM output data arrangement

CMIPn arrangement

<table>
<thead>
<tr>
<th>f₁:</th>
<th>t₀</th>
<th>t₁</th>
<th>t₂</th>
<th>...</th>
<th>tₘ</th>
</tr>
</thead>
<tbody>
<tr>
<td>f₂:</td>
<td>t₀</td>
<td>t₁</td>
<td>t₂</td>
<td>...</td>
<td>tₘ</td>
</tr>
<tr>
<td></td>
<td>t₀</td>
<td>t₁</td>
<td>t₂</td>
<td>...</td>
<td>tₘ</td>
</tr>
<tr>
<td>fn:</td>
<td>t₀</td>
<td>t₁</td>
<td>t₂</td>
<td>...</td>
<td>tₘ</td>
</tr>
</tbody>
</table>
Visualization issues:
Climate is a 2.5 dimensional system.

- Aspect ratio encourages 2D view:
  - Horizontal scale: 10,000km. Vertical scale: 10km
CCSM3 results: Sea Surface Temperature (1990 Control run)
CCSM3 results: Zonal average U-wind
Coupled ocean/atmosphere 3D view (rarely used).
Summary

- Data postprocessing is an essential part of climate modeling. It determines the climate from the weather-scale output of the model.

- Climate models are heading towards higher resolution
  - 20-80km for century-scale prediction.
  - 1-5km (GCRM) for inter-annual simulation.
  - Non-hydrostatic (full 3D vector velocity fields).
  - New non-cartesian, unstructured grids
  - Climate/weather model distinction goes away (aspect ratio gets better)

- Climate models are adding more degrees of freedom
  - Interactive carbon cycle (more tracers in all components)
  - Atmospheric chemistry (10s - 100s additional 3D tracers)

- Our current custom tools (NCL, Ferret) are breaking on multi-GB datasets.
Possible strategies

- Still save everything but:
  - Save it compressed
  - One file per variable, append in time up to X years. Better for deep storage.

- In-situ analysis:
  - Calculating averages costs more memory
  - Always need to compare with other climate simulations/observations.