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"

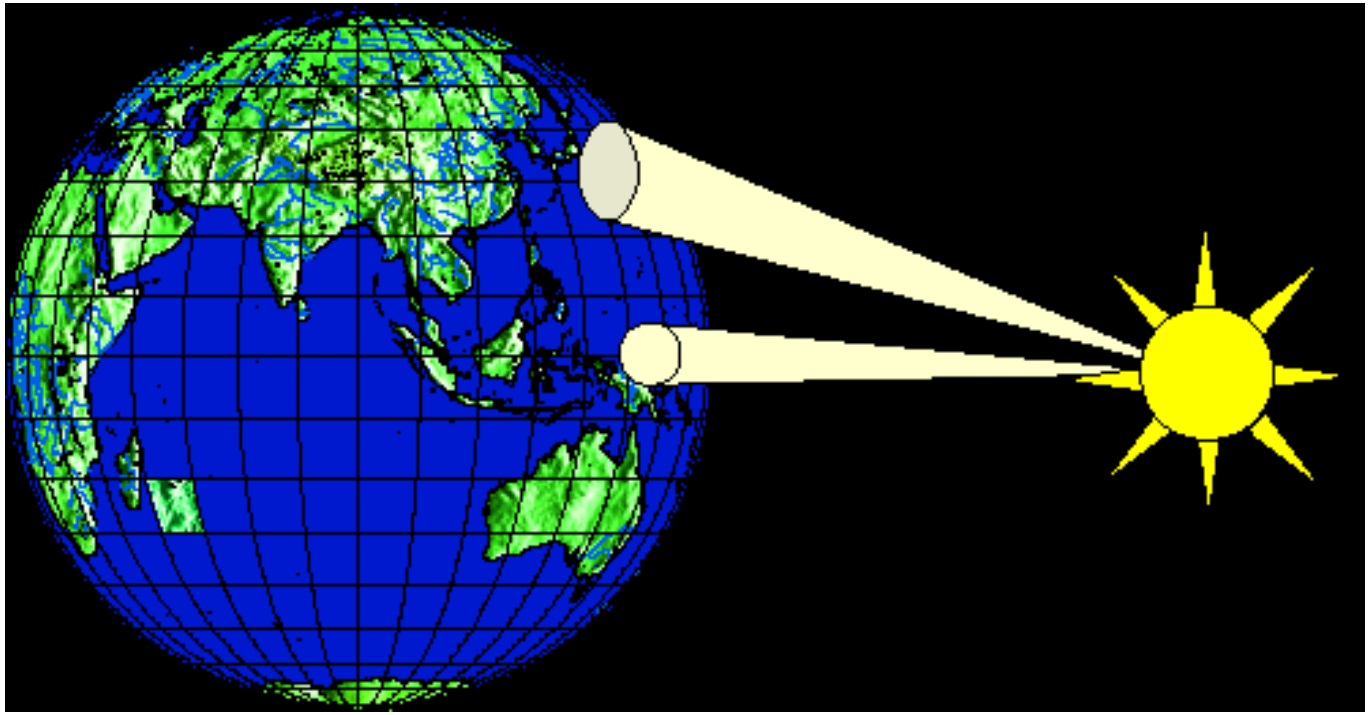
- Climate is the *average* of weather.
- The (predicted) high temperature today, Jul 30th, is 71F
- The average high temperature is 70F. This is calculated by taking the average of several (usually 30) Jul 30th highs.

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

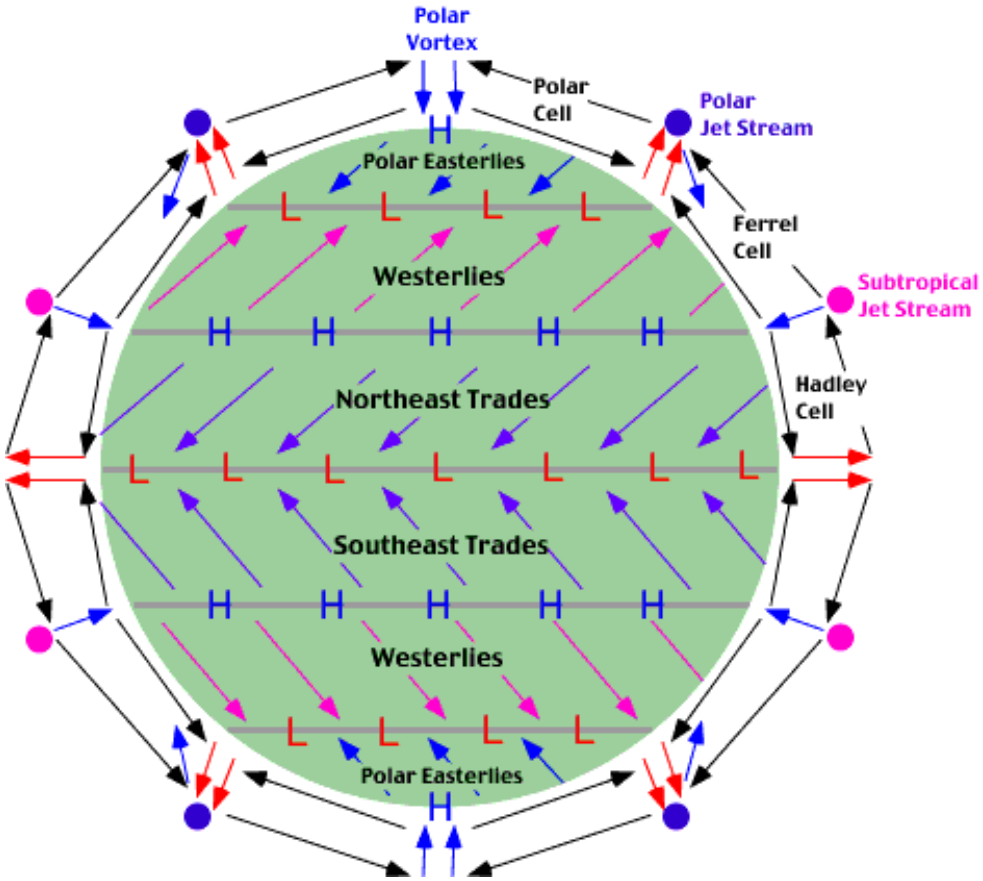
To model the climate system, must model years of global weather.

Climate is the average of weather.

What makes weather? The Sun and the Earth's rotation.

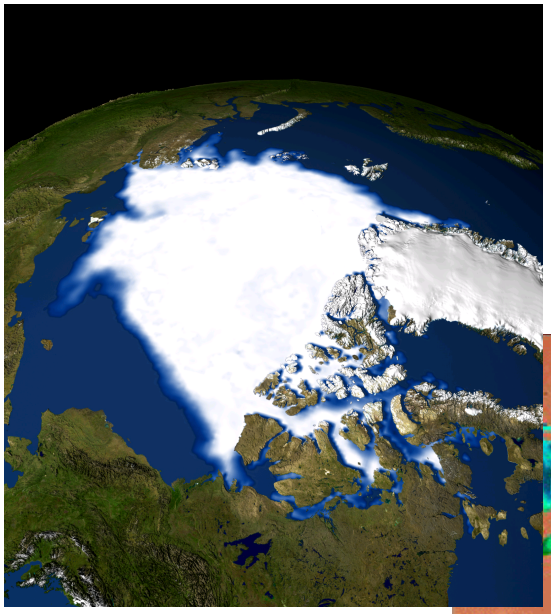


Need to simulate weather-scale phenomena over the entire globe.

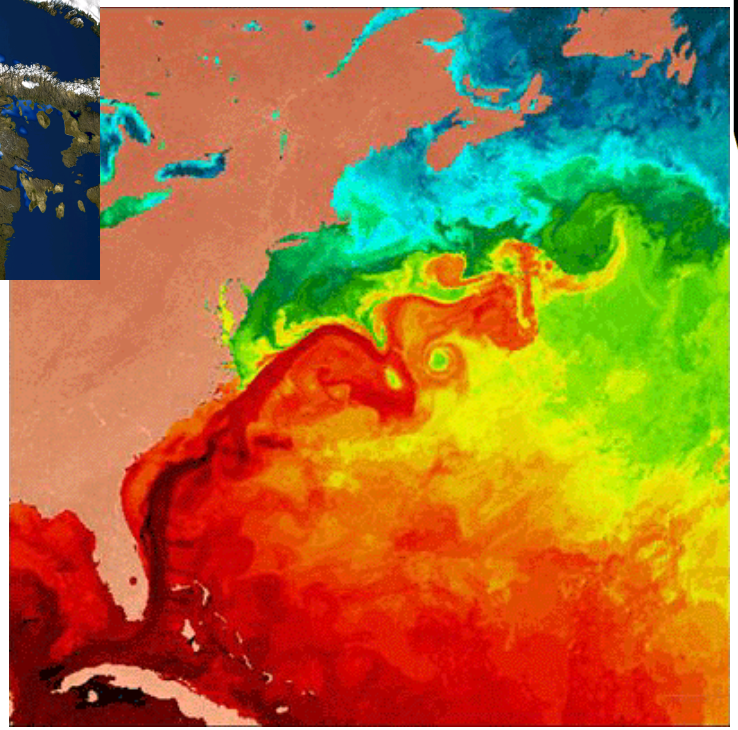


Weather is embedded in the *general circulation* of the atmosphere

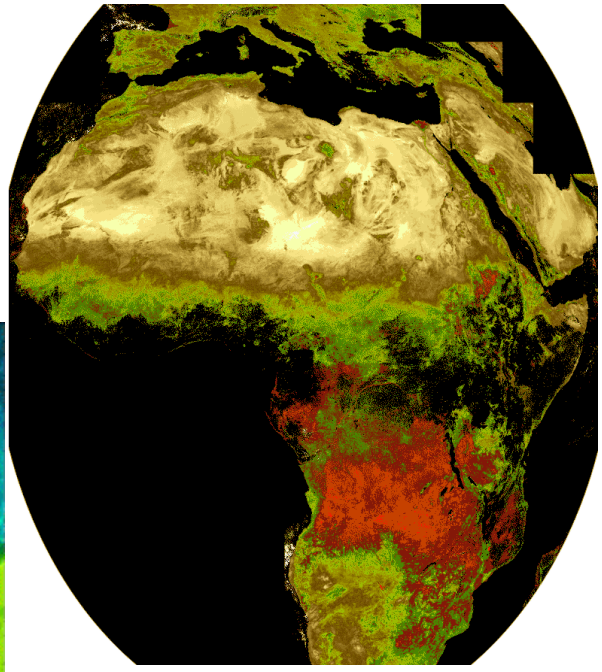
Over may days, months, atmosphere circulation is dominated by interaction with surface.



Sea Ice



Ocean

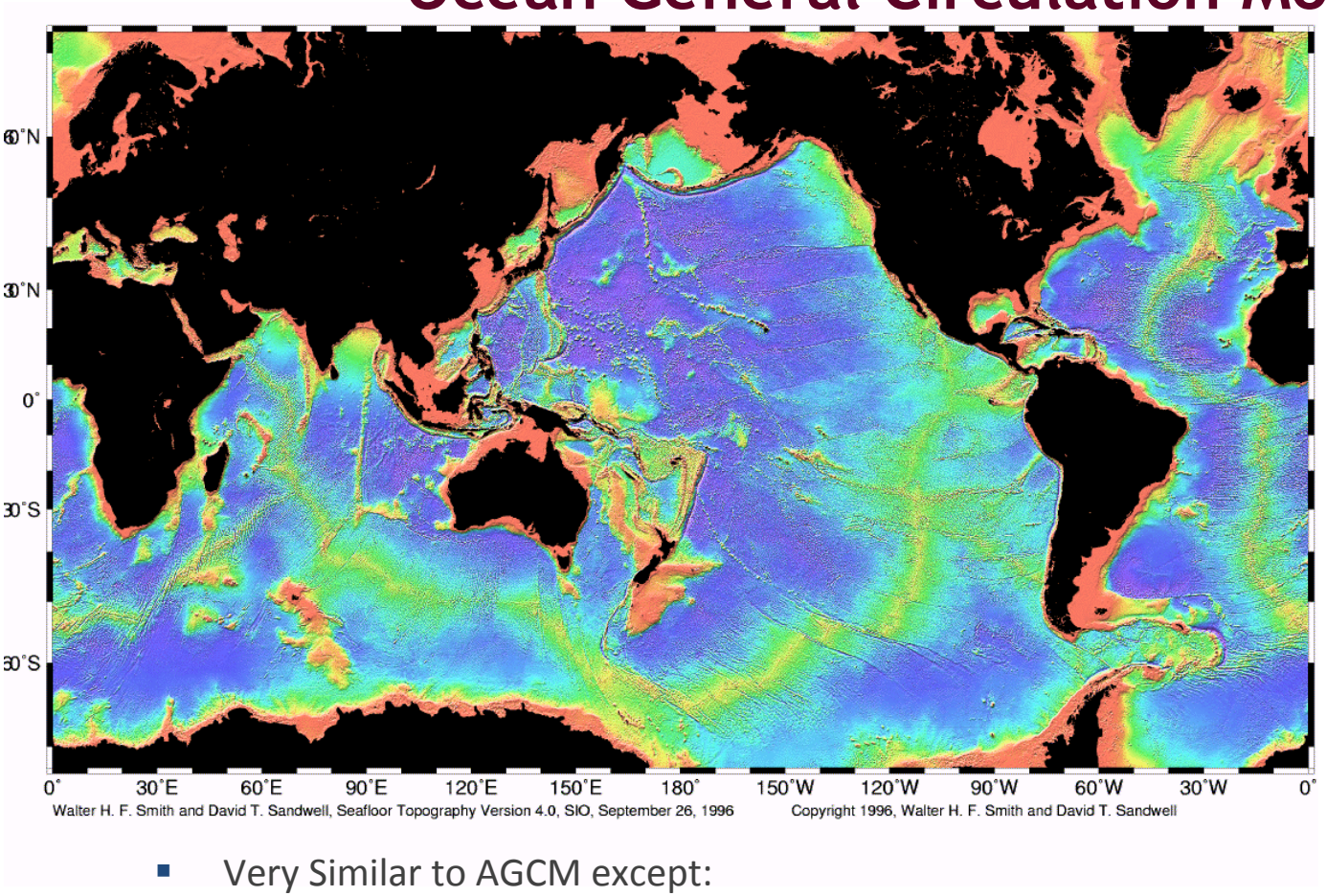


Land

Atmospheric General Circulation Model

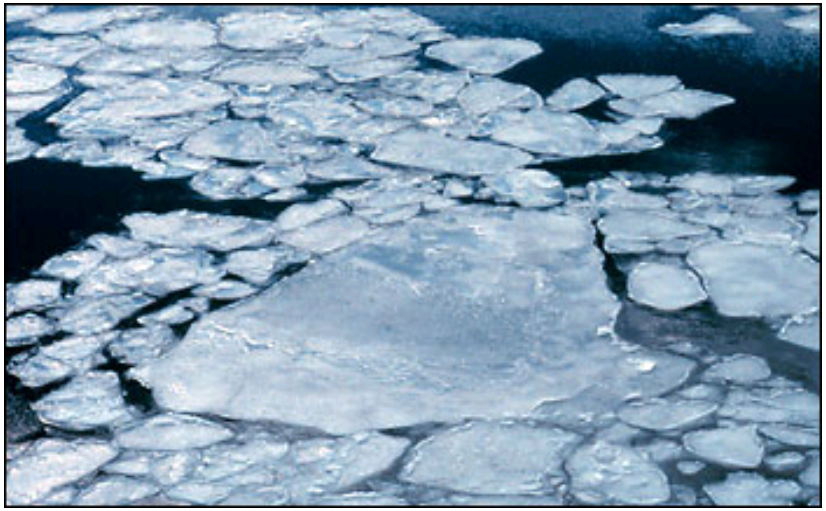
- Algorithms to solve the primitive equations called "the dynamics"; "dynamical core" "dycore"
- Forcing terms: **F(t,u,v,phi)**
  - *Change in temperature due to radiative transfer*
  - *Effect of clouds on radiative transfer*
  - *Change in moisture due to cloud, rain formation*
  - *Change in temperature 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



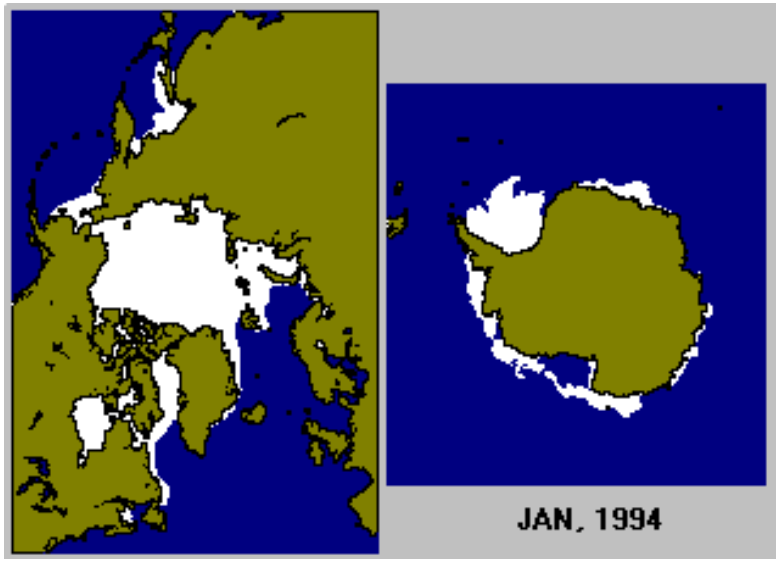
- Very Similar to AGCM except:
  - Presence of side boundaries. Nearly all OGCM's are FD with z-coordinates.
  - Not as much "physics"
  - Motions are slower. Length scales are shorter.
  - Much higher heat capacity. The memory of the climate system is in the ocean.

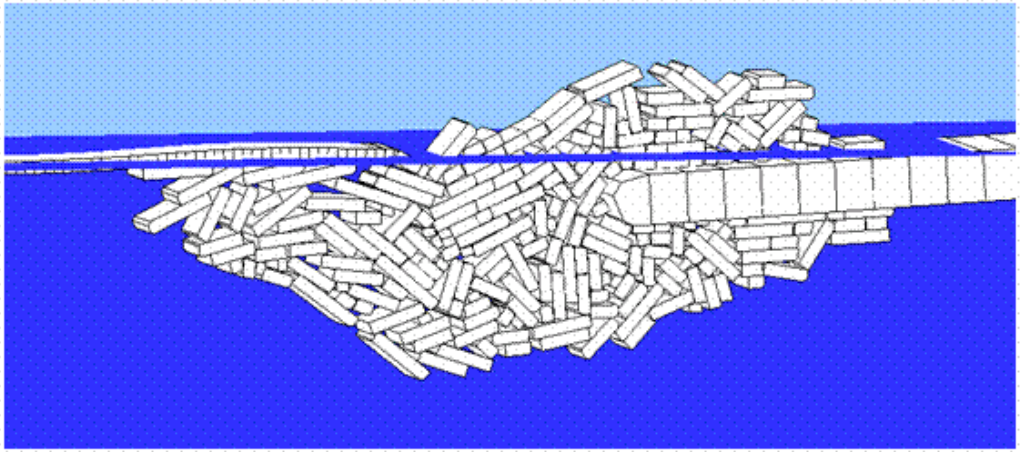
Sea Ice Models



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

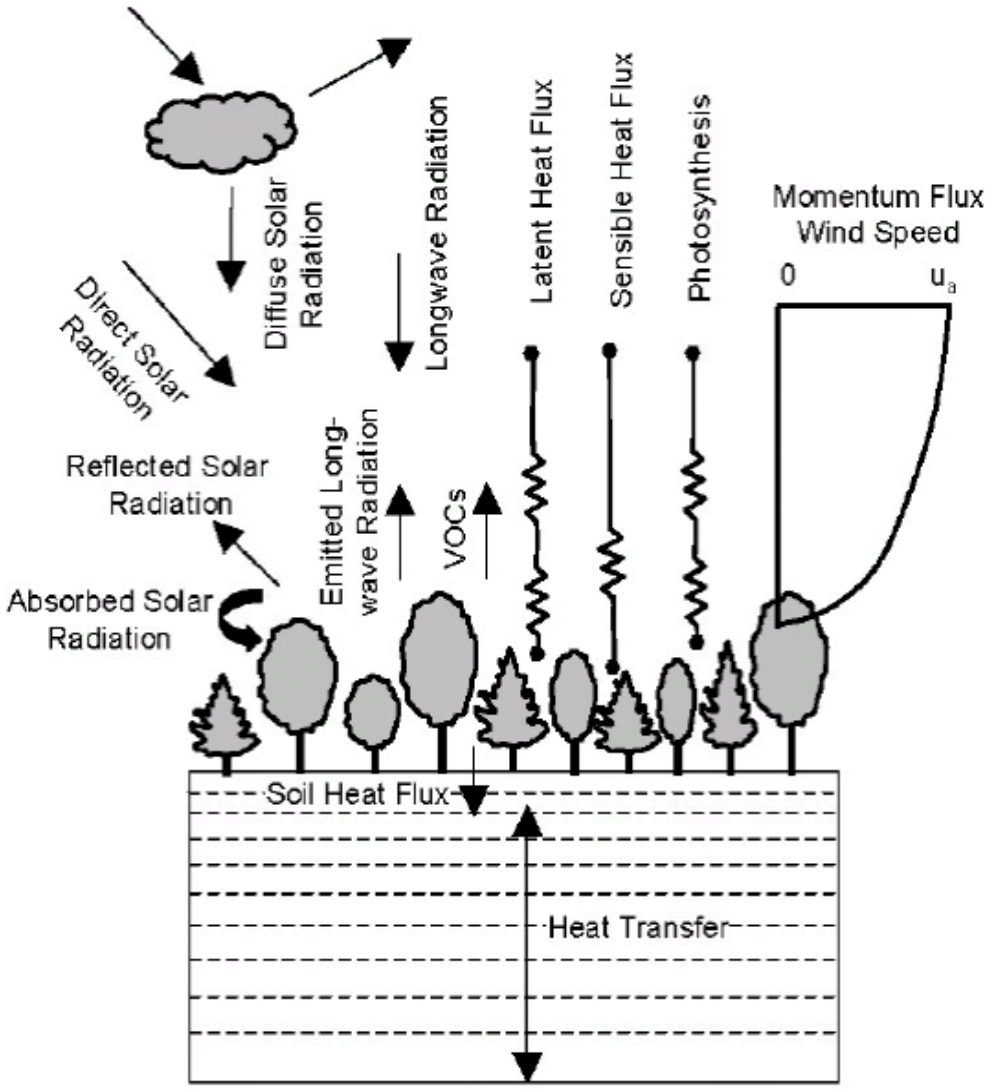
- Dynamics: transport, internal stress, ridging





Showing a scene from a pressure ridge simulation. The thin ice is 0.5 m thick and the thick floe is 2 m thick.

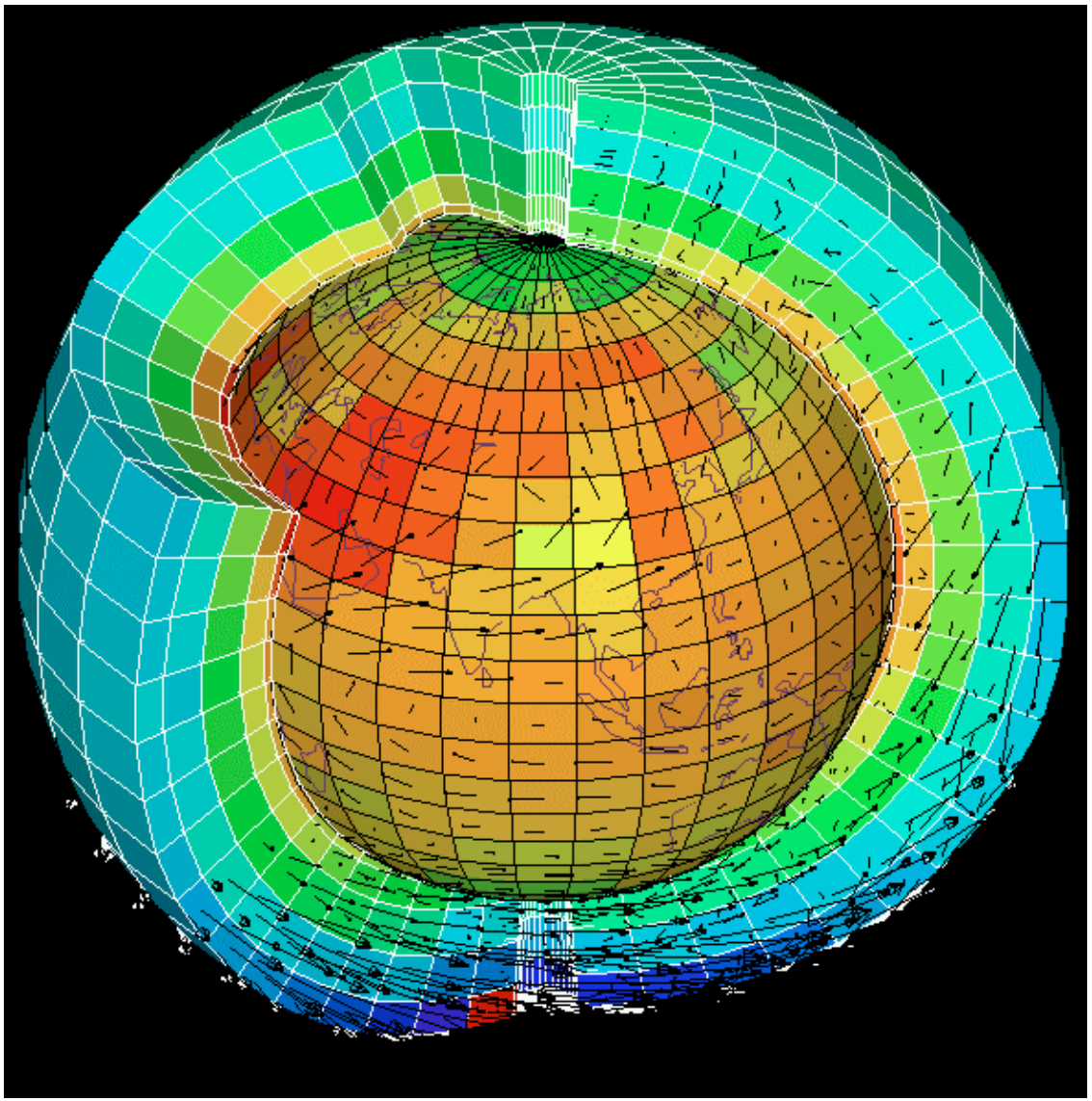
Land Surface Models



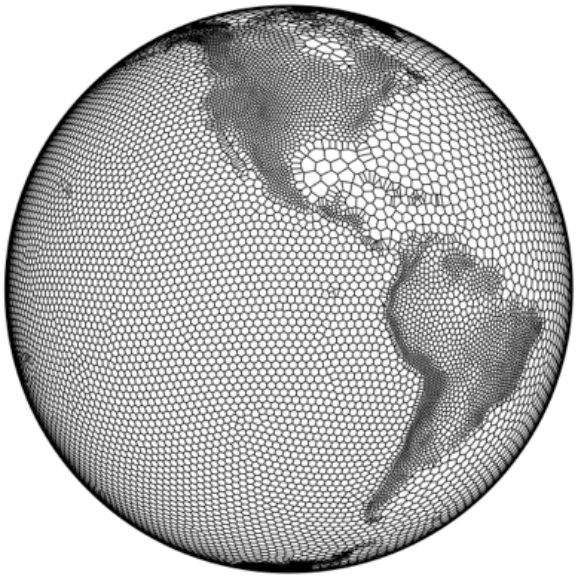
- Nearly all "physics":
  - Vegetation composition, structure
  - Vertical heat transfer in soil.
  - Heat, radiation transfer between ground, canopy and free atmosphere
  - Hydrology of canopy, snow, soil moisture
  - River runoff
- Historically, was part of column physics in the atmosphere model.

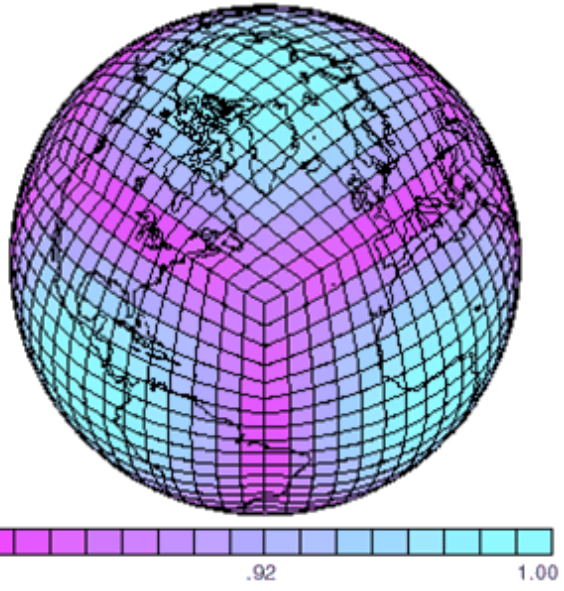
Figure: Bonan, NCAR

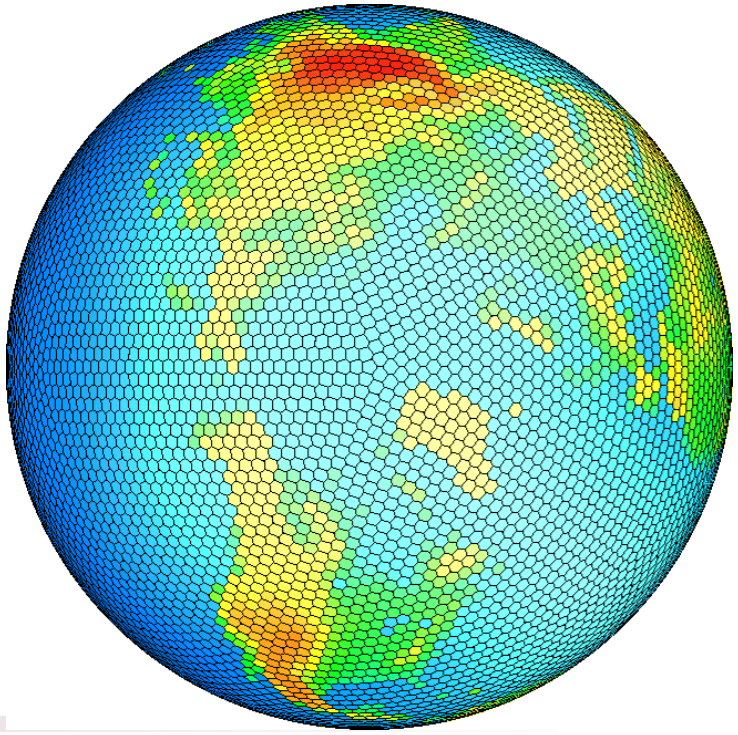
Primitive Equations must be solved numerically



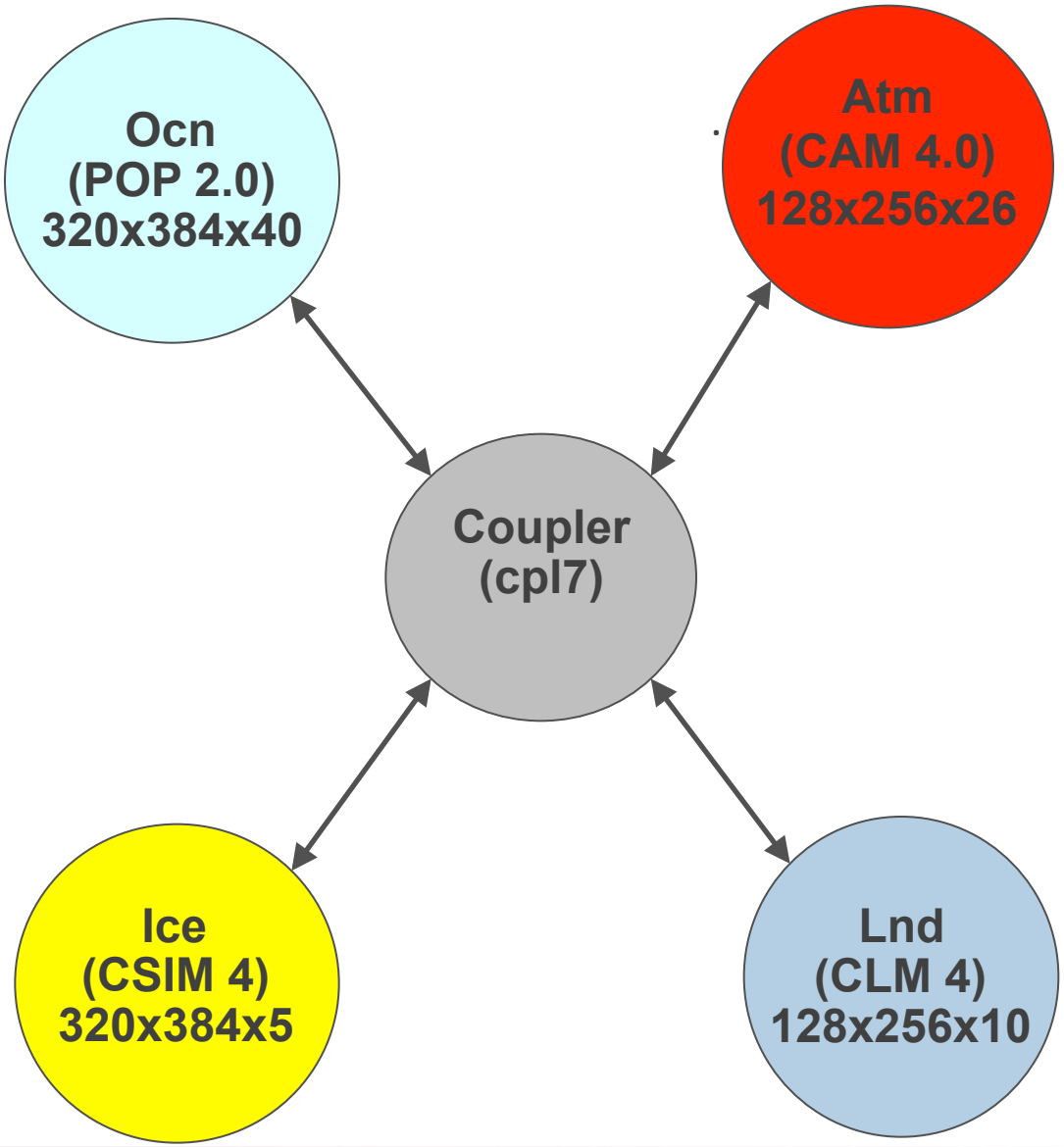
Other grids







NCAR/DOE Coupled Climate Model CCSM4/CESM1



Take your coupled global climate model and calculate evolution of global weather for 100 years, 20 minutes at 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.

CMIP3 vs. CMIP5

CMIP3	subdaily	daily	monthly	annual	totals
atmosphere	9	18	47	0	**74**
land surface	0	0	9	0	**9**
ocean	0	0	12	0	**12**
sea ice	0	0	4	0	**4**
totals	**9**	**18**	**72**	**0**	**99**

CMIP3 vs. CMIP5

CMIP5	subdaily	daily	monthly	annual	totals
atmosphere	100	75	223	8	**406**
land surface	3	5	82	0	**90**
ocean	1	3	127	79	**210**
sea ice	0	4	40	0	**44**
totals	**104**	**87**	**472**	**87**	**750**

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

Modeling group		CMIP3 volume (GB)
NCAR	USA	9,172.8
MIROC3	Japan	3,974.9
GFDL	USA	3,842.5
IAP	China	2,867.7
MPI	Germany	2,699.5
CSIRO	Australia	2,088.2
CCCMA	Canada	2,070.6
INGV	Italy	1,472.2
GISS	USA	1,096.8
MRI	Japan	1,024.5
CNRM	France	999.1
IPSL	France	997.7
UKMO	UK	972.8
BCCR	Norway	861.9
MIUB	Germany/Korea	477.2
INMCM3	Russia	368.2
Totals		**34,986.6**

Modeling group		CMIP5 volume (GB)
MPI	Germany	710,000
NCAR	USA	410,000
MRI	Japan	312,000
GFDL	USA	151,000
MIROC3	Japan	115,000
UKMO	UK	89,000
CNRM	France	64,000
IAP	China	63,000
U Reading	UK	63,000
EC	Europe	50,000
GISS	USA	50,000
INGV	Italy	50,000
IPSL	France	45,000
INMCM3	Russia	32,000
NorClim	Norway	30,000
CCCMA	Canada	29,000
CAWCR	Australia	21,000
CSIRO	Australia	20,000
METRI	Korea	13,000
Totals		**2,317,000**

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

2 km, 100 levels, hourly data

~4 TB / simulated hour

~100 TB / simulated day

~35 PB / simulated year

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.)

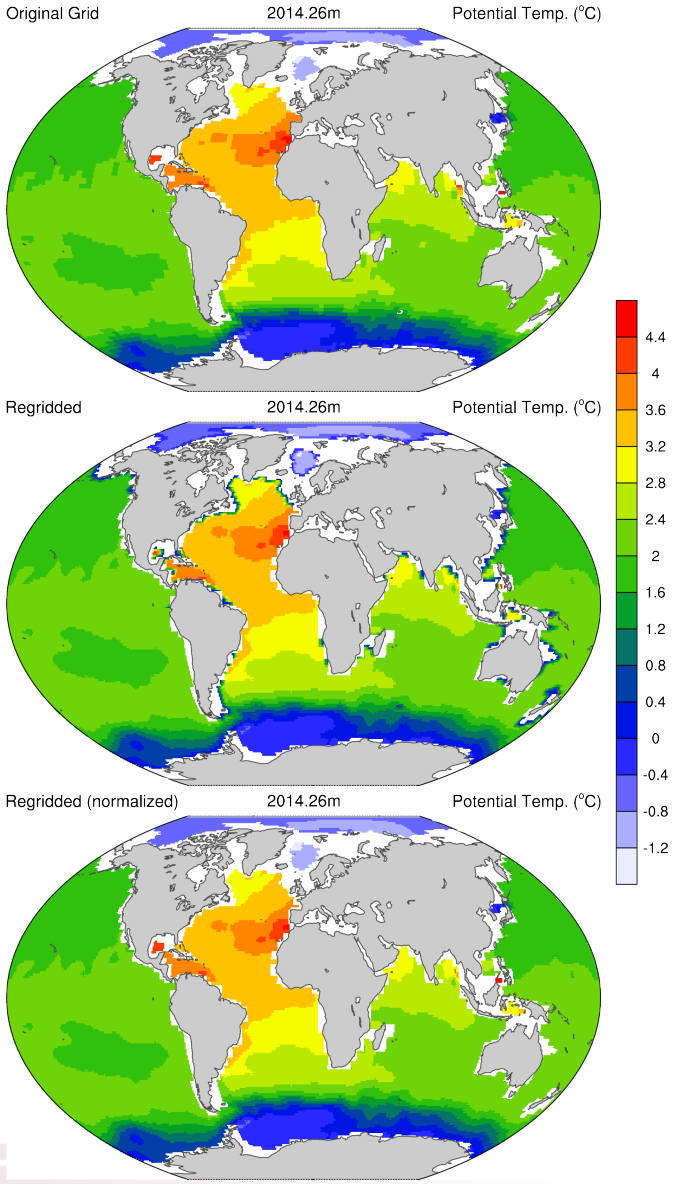
- Also IDL, Matlab, Mathematica.

NCAR Command Language (NCL)

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

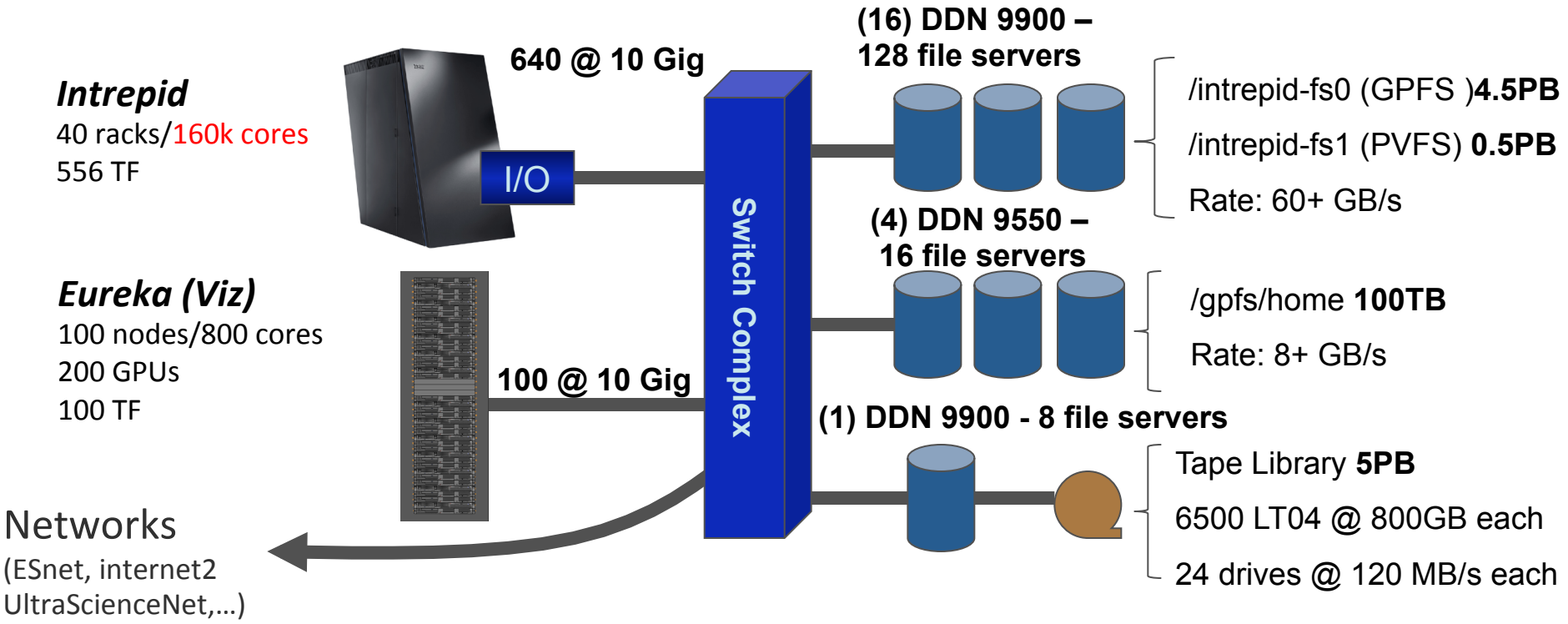
1. Simple, robust file input and output
2. Hundreds of analysis (computational) functions
3. Visualizations (2D) are publication quality and highly customizable

- **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



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

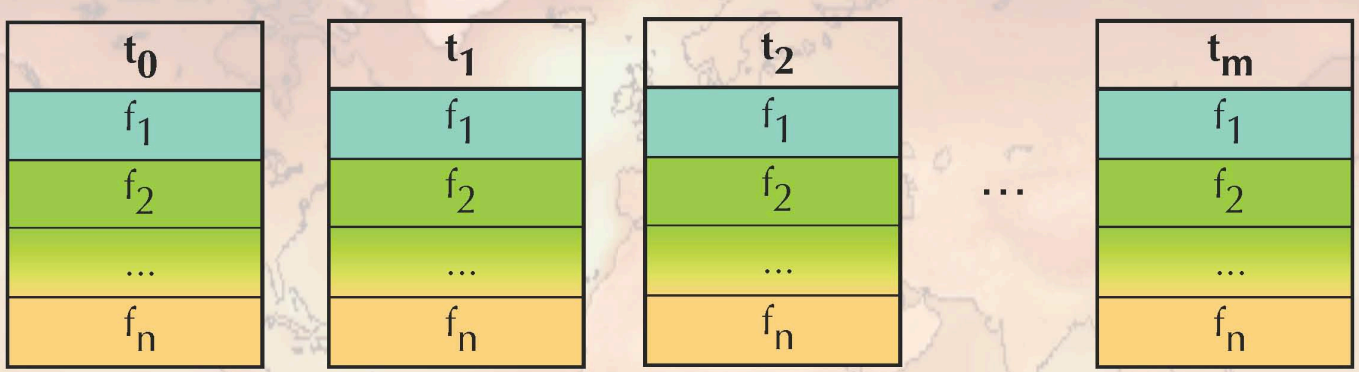
More Issues

In climate, the filesystem is the database.

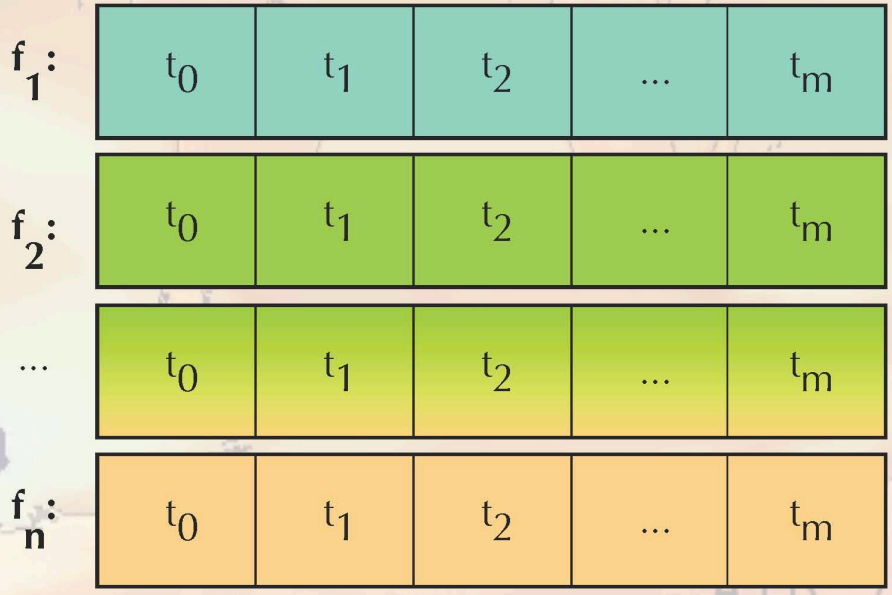
```
ESM_SFASTv2_DMS_1850.cam2.h0.0046-06.nc  ESM_SFASTv2_DMS_1850.cam2.h0.0046-07.nc
ESM_SFASTv2_DMS_1850.cam2.h0.0046-08.nc  ESM_SFASTv2_DMS_1850.cam2.h0.0046-09.nc
ESM_SFASTv2_DMS_1850.cam2.h0.0046-10.nc  ESM_SFASTv2_DMS_1850.cam2.h0.0046-11.nc
ESM_SFASTv2_DMS_1850.cam2.h0.0046-12.nc  ESM_SFASTv2_DMS_1850.cam2.h0.0047-01.nc
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

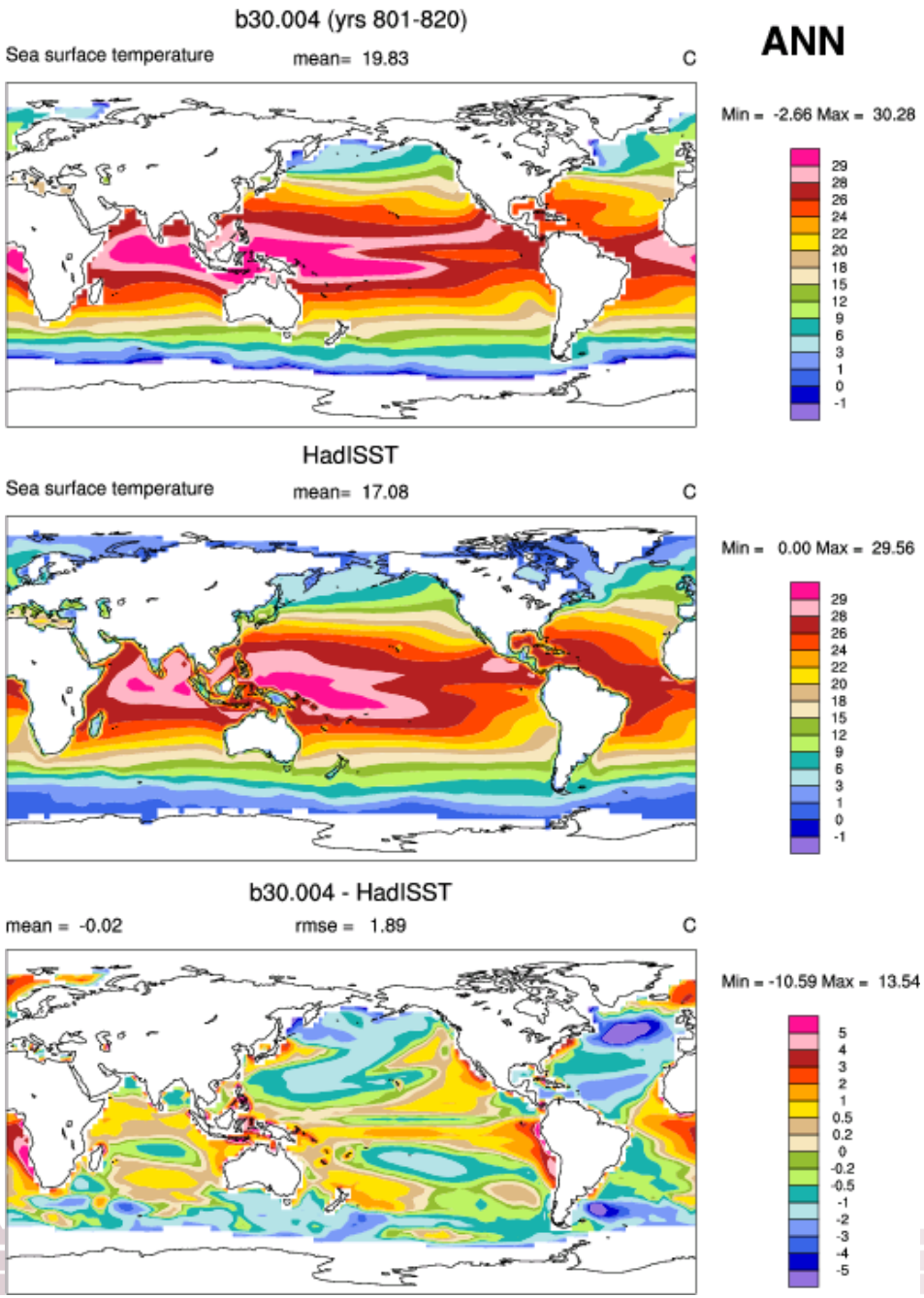


CMIPn arrangement

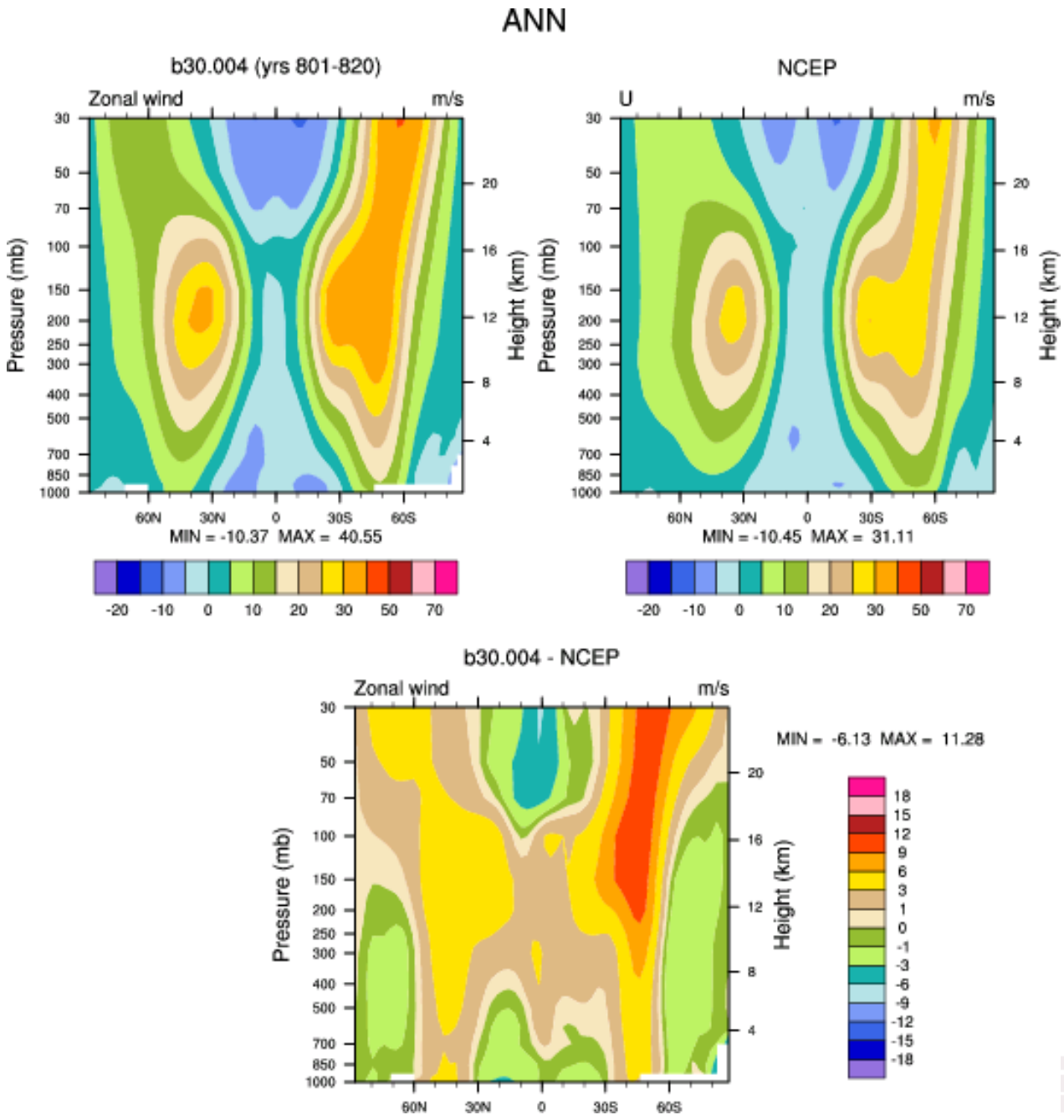


Visualization issues:
Climate is a 2.5 dimensional system.

- Aspect ratio encourages 2D view:
  - Horizontal scale: 10,000km. Vertical scale: 10km

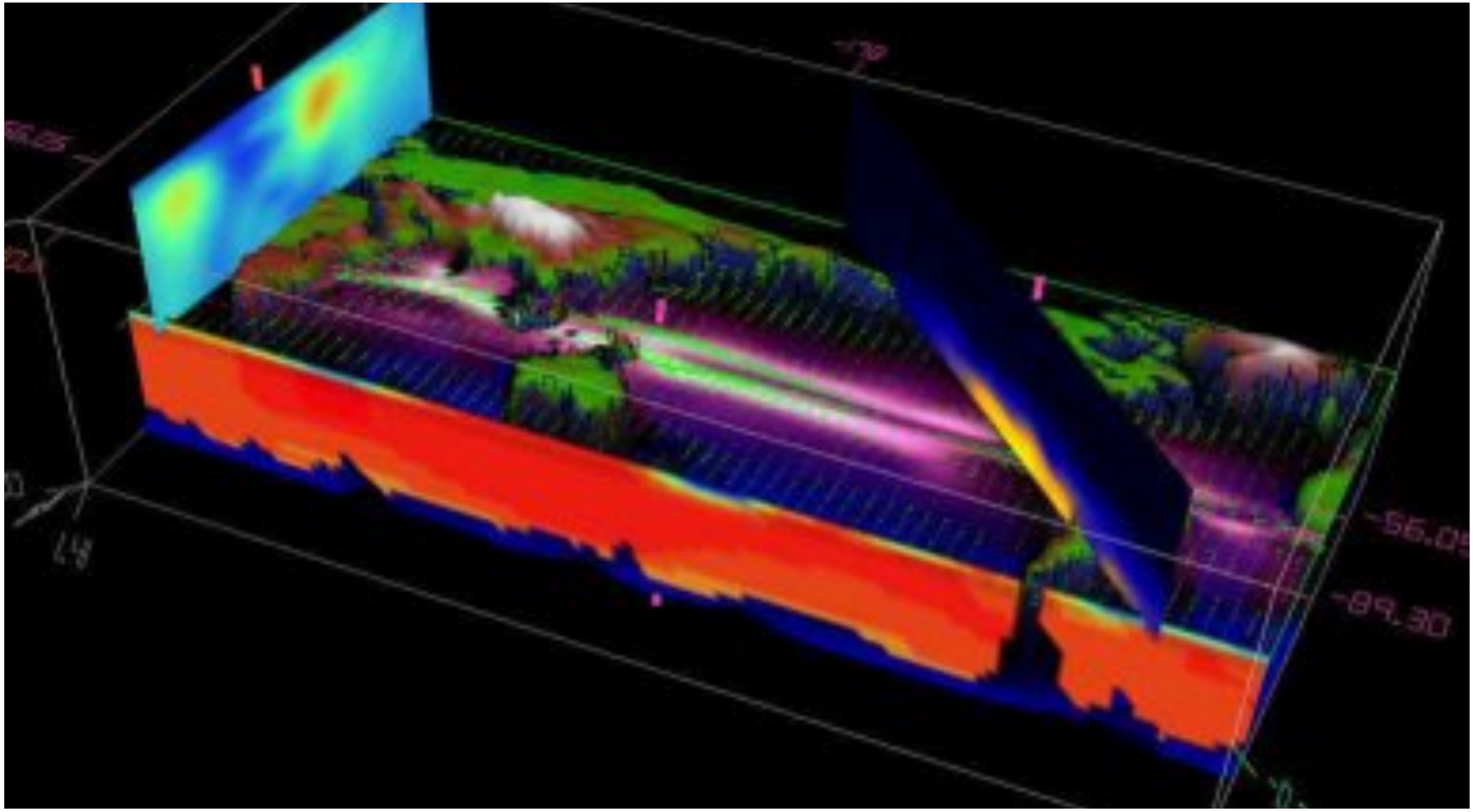


CCSM3 results: Sea Surface Temperature (1990 Control run)



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.