

The background of the slide is a stylized globe. It features a grid of latitude and longitude lines. Overlaid on the globe is a circular pattern of concentric rings, resembling a ripple in water or a target, centered near the top. The colors of the globe are muted, with greens, yellows, and oranges visible. The text is overlaid on this background.

The 800 lb Gorilla: Really Big Data

UCAR SEA Conference 2012

Gary Strand

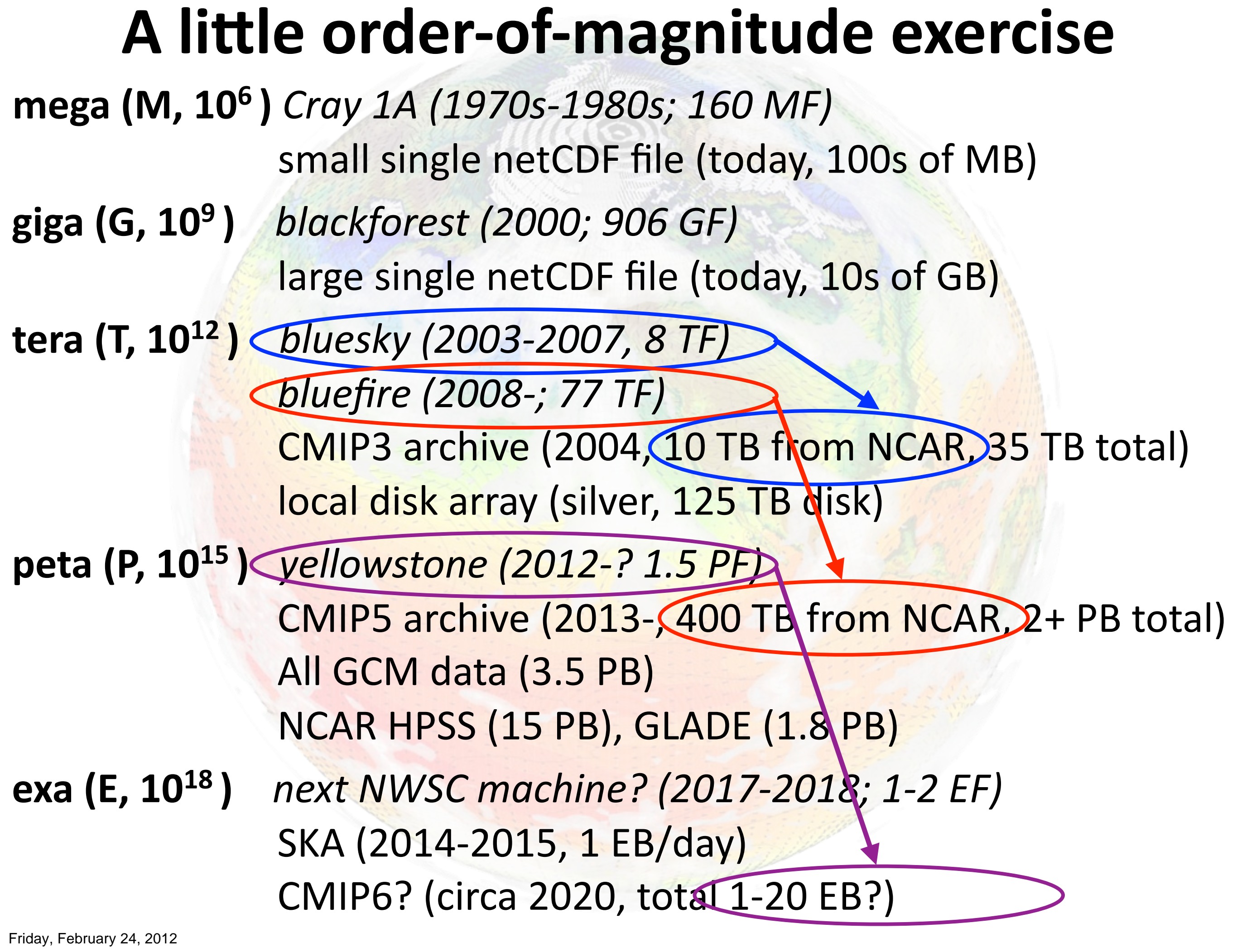
strandwg@ucar.edu

supercomputer

| 'so̯pər kəm , pyo̯tər |

**A machine that takes one problem
(computation) and turns it into more
- I/O, storage, and management.**

A little order-of-magnitude exercise

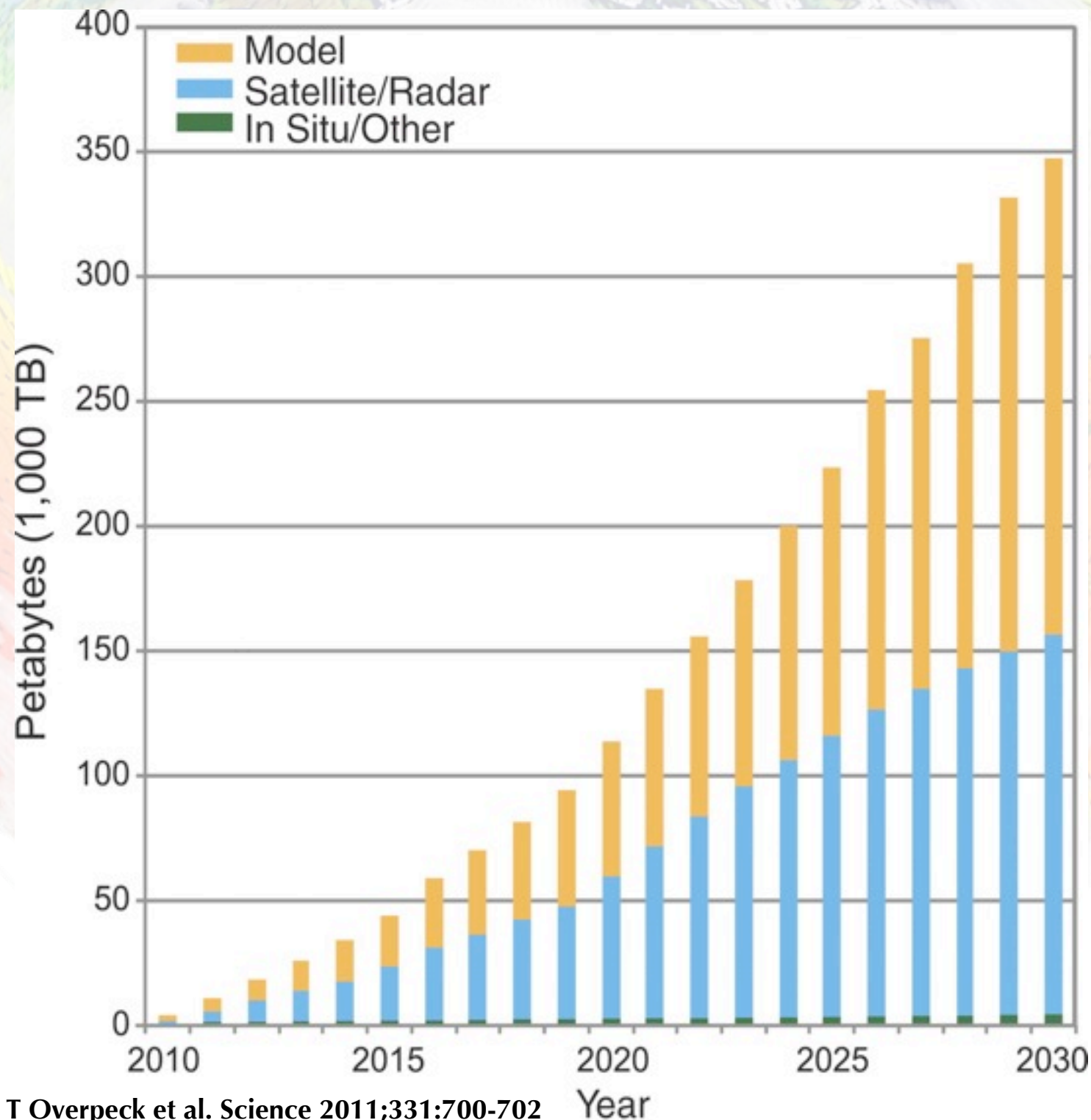


mega (M, 10^6)	<i>Cray 1A (1970s-1980s; 160 MF)</i> small single netCDF file (today, 100s of MB)
giga (G, 10^9)	<i>blackforest (2000; 906 GF)</i> large single netCDF file (today, 10s of GB)
tera (T, 10^{12})	<i>bluesky (2003-2007, 8 TF)</i> <i>bluefire (2008-; 77 TF)</i> CMIP3 archive (2004, 10 TB from NCAR, 35 TB total) local disk array (silver, 125 TB disk)
peta (P, 10^{15})	<i>yellowstone (2012-? 1.5 PF)</i> 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})	<i>next NWSC machine? (2017-2018; 1-2 EF)</i> SKA (2014-2015, 1 EB/day) CMIP6? (circa 2020, total 1-20 EB?)

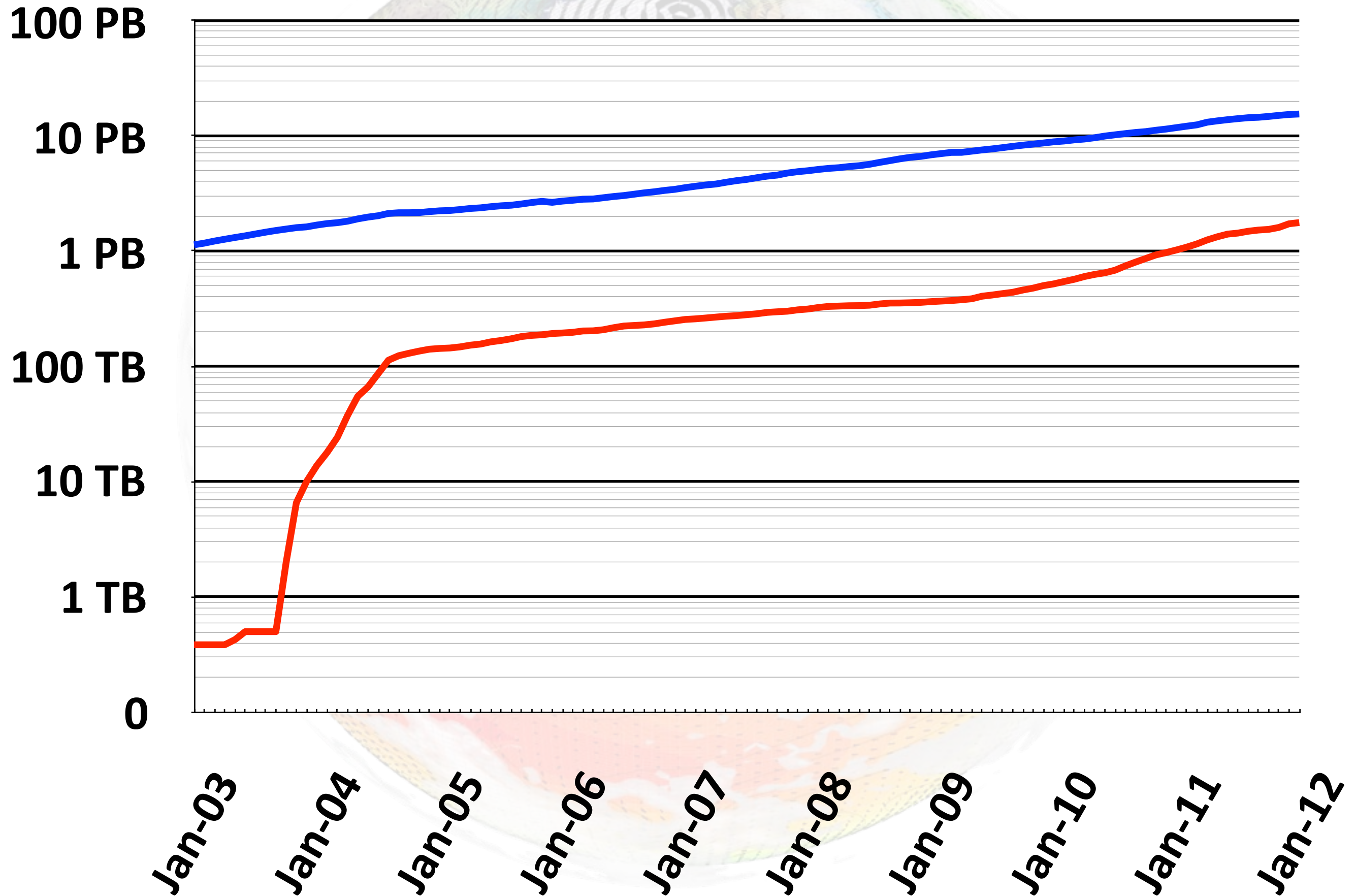
Diagram showing data flow and storage requirements:

- A blue arrow points from *bluesky* to the CMIP3 archive.
- A red arrow points from *bluefire* to the CMIP5 archive.
- A purple arrow points from *yellowstone* to CMIP6?

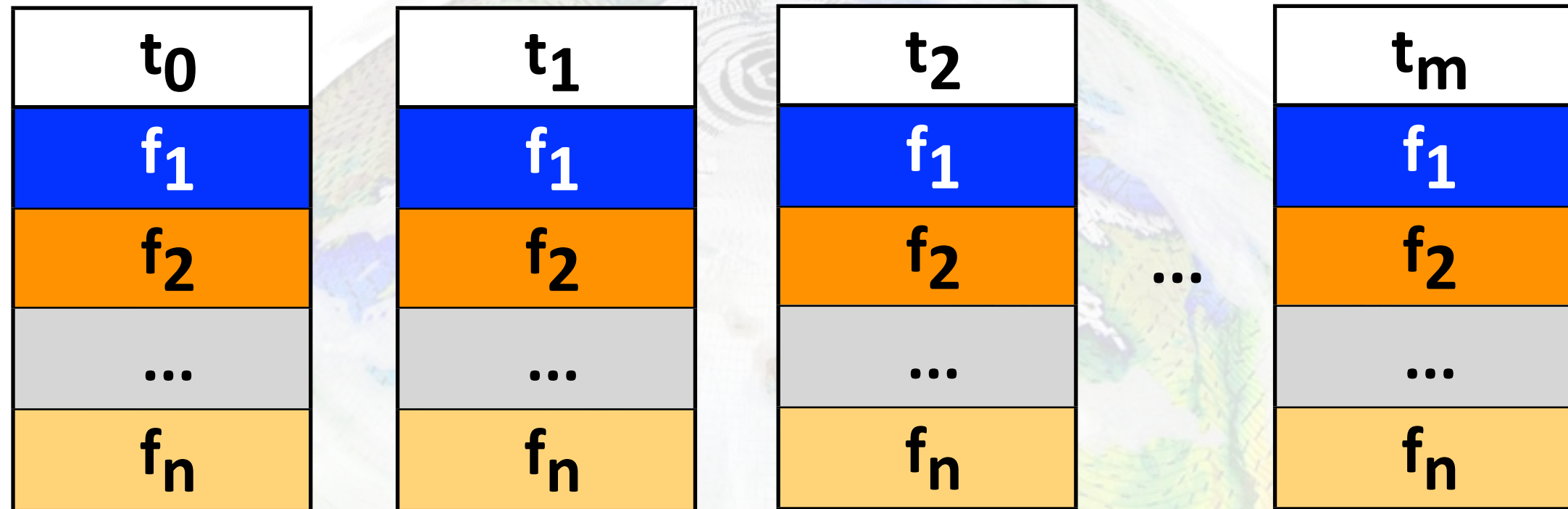
“The volume of worldwide climate data is expanding rapidly, creating challenges for both physical archiving and sharing, as well as for ease of access and finding what’s needed, particularly if you are not a climate scientist.”



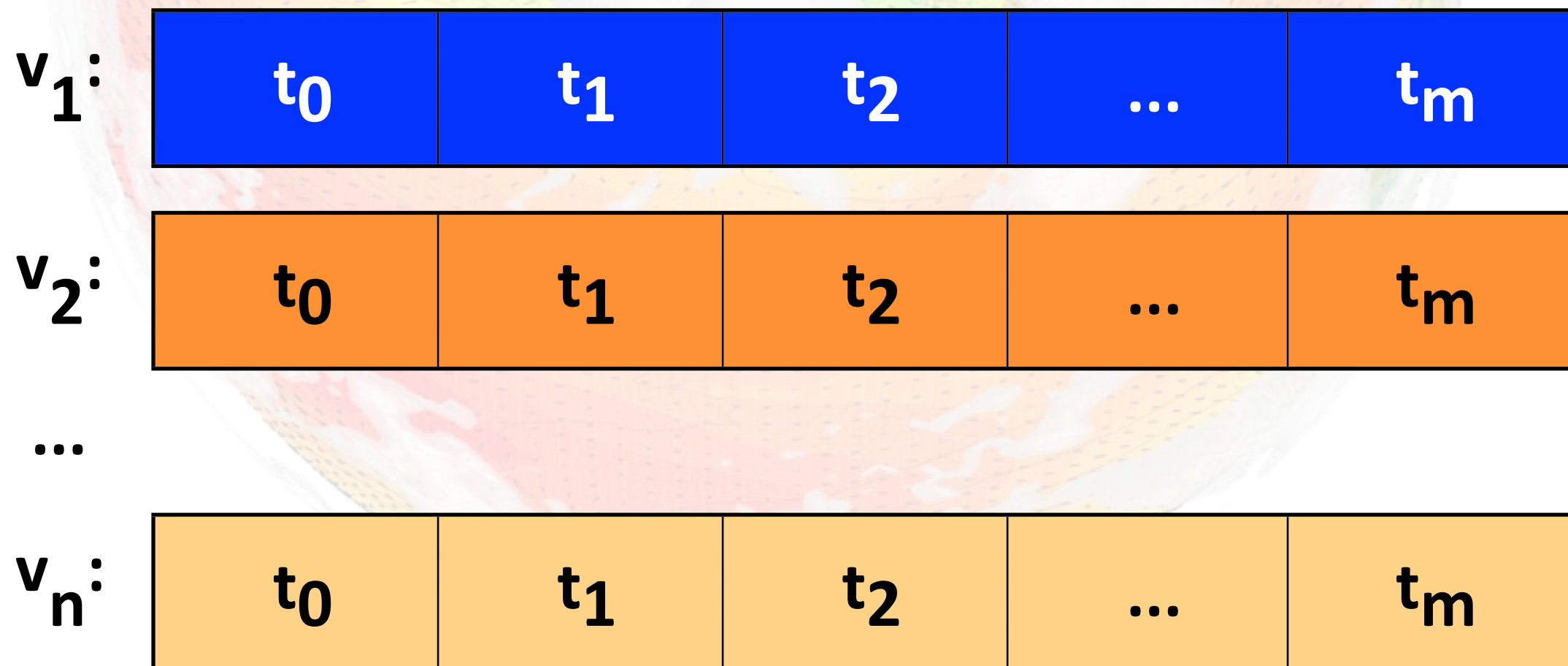
Total NCAR HPSS archive and /CCSM volume



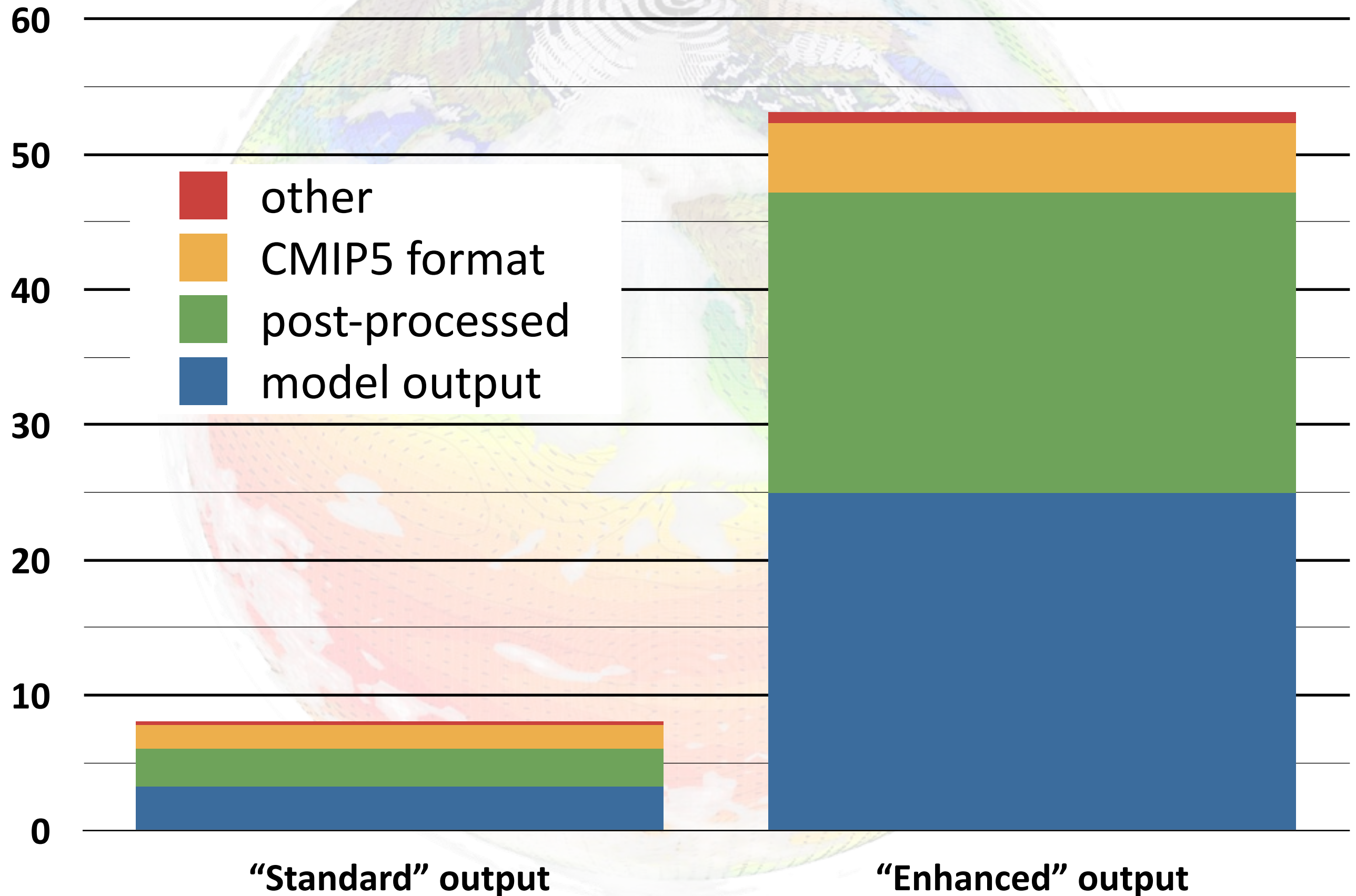
Typical CESM output arrangement



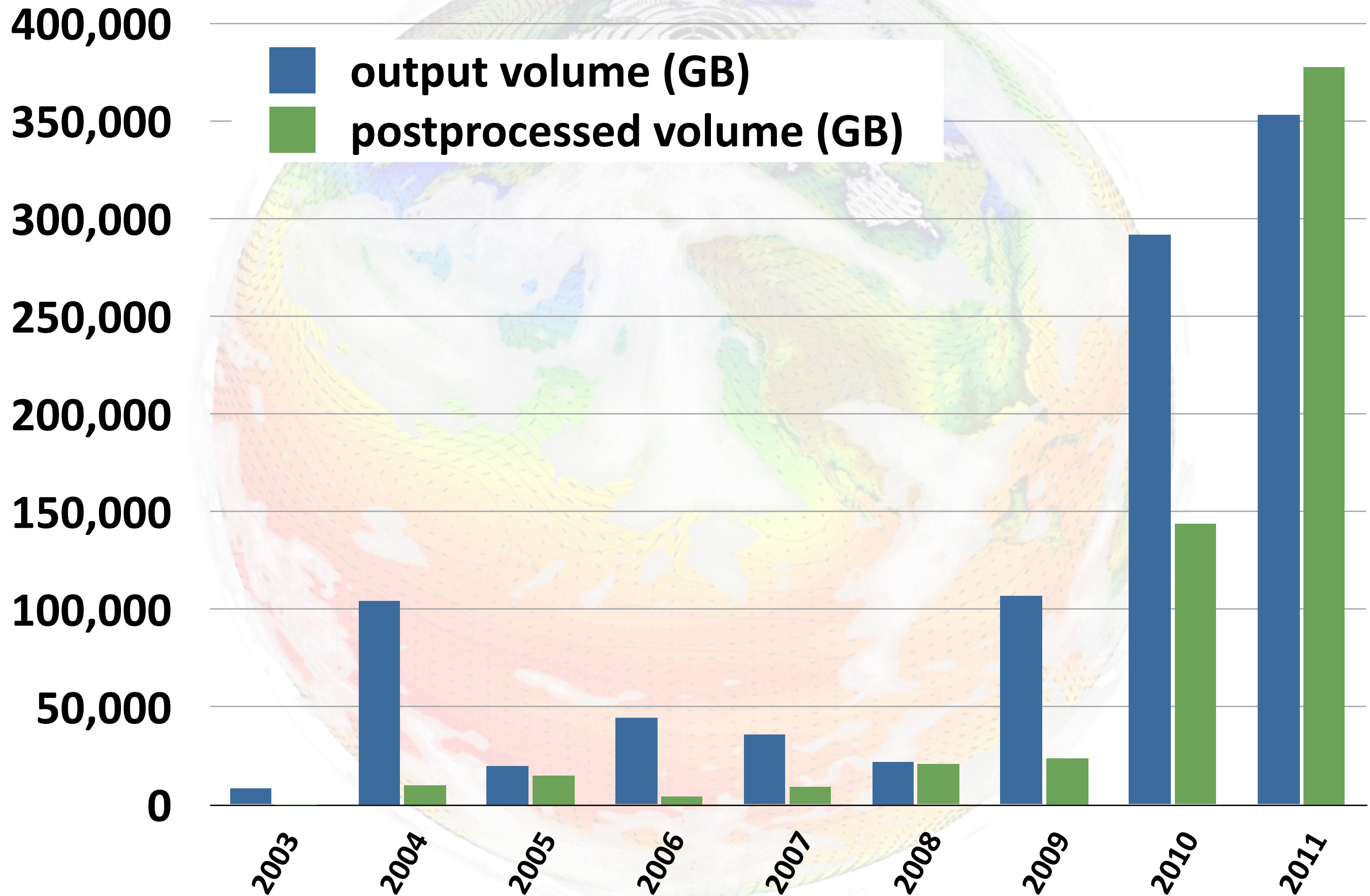
Useful arrangement



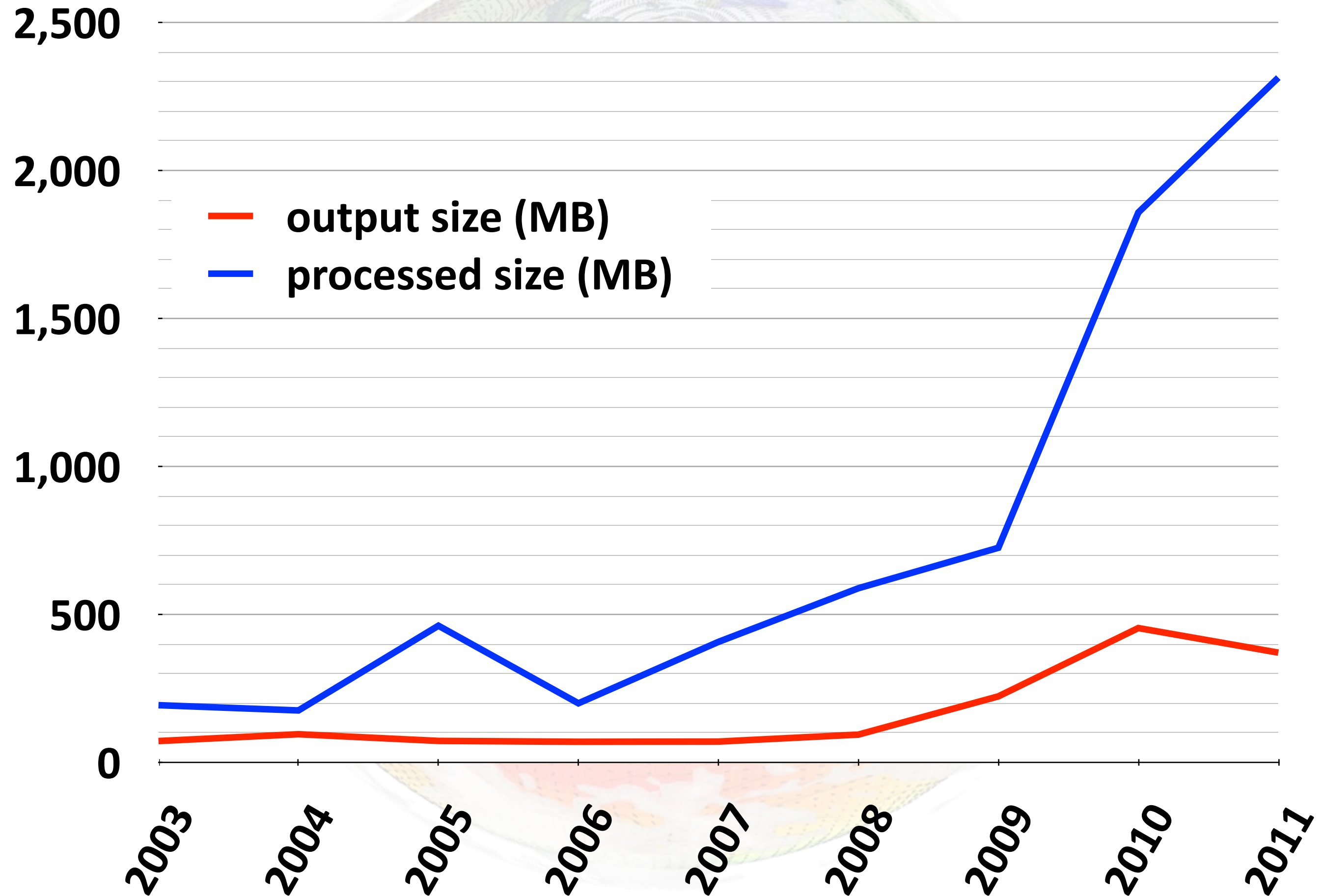
CESM “1°” output volumes (TB)



Output vs. postprocessed volume (GB)



Output vs. postprocessed file sizes



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

Retention of CESM Data

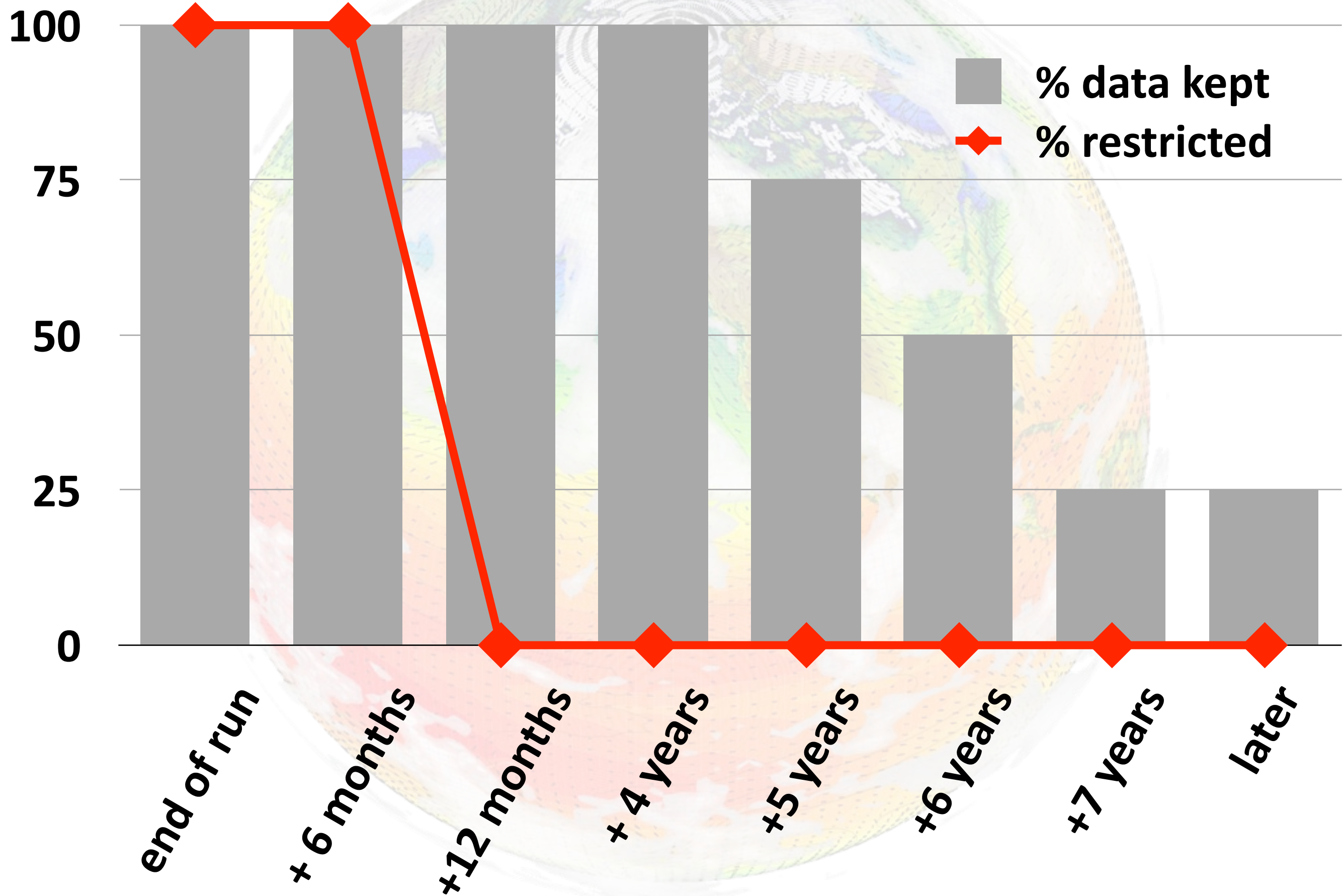
Development

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

Production

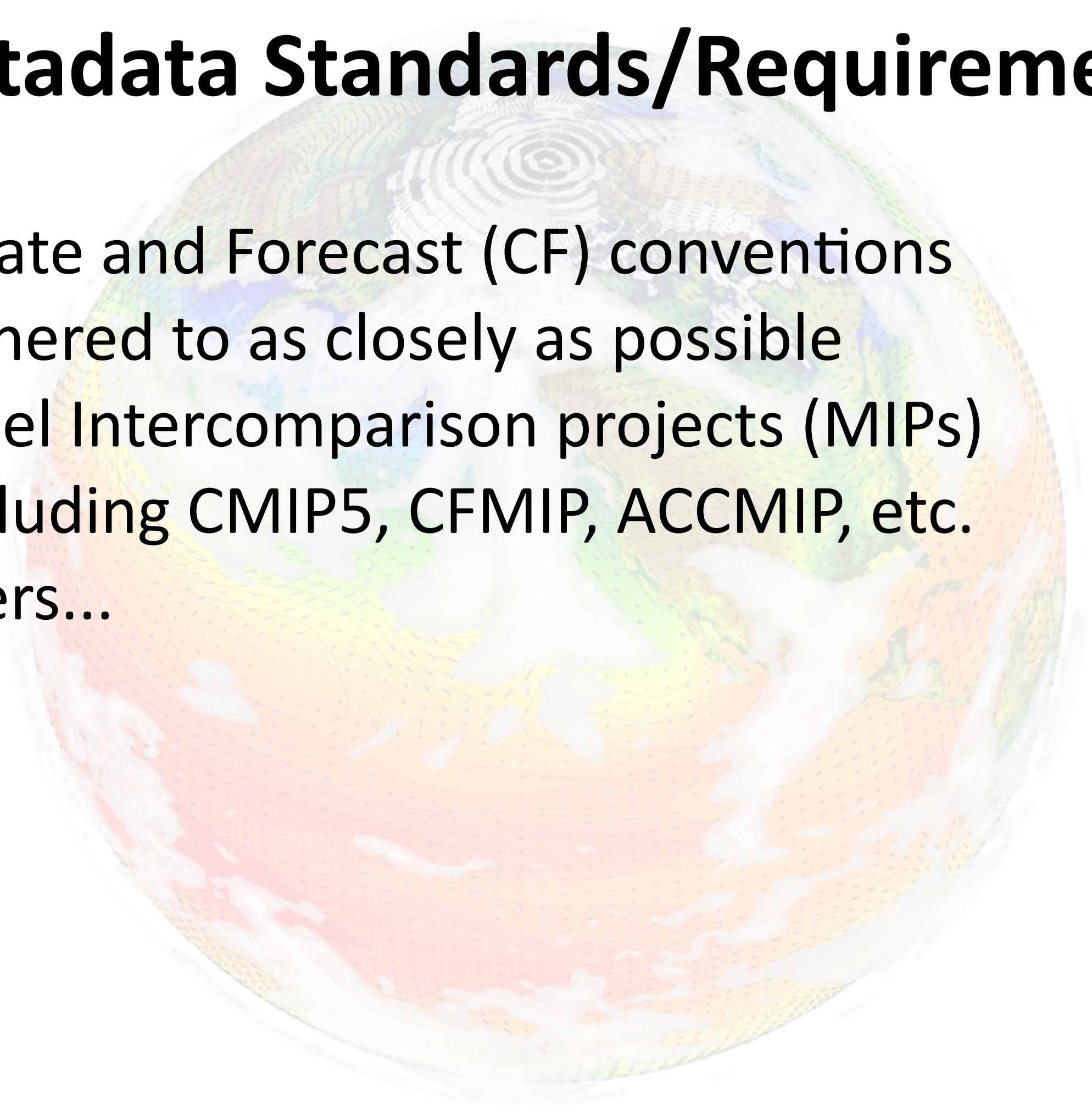
- All data for no less than 4 years
- Stepwise reduction for next 3 years
- Latititude for maintenance costs

General timeline summary



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

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

As required by CMIP5

```
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

```
: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_omp_thread_number = 1 ;
```

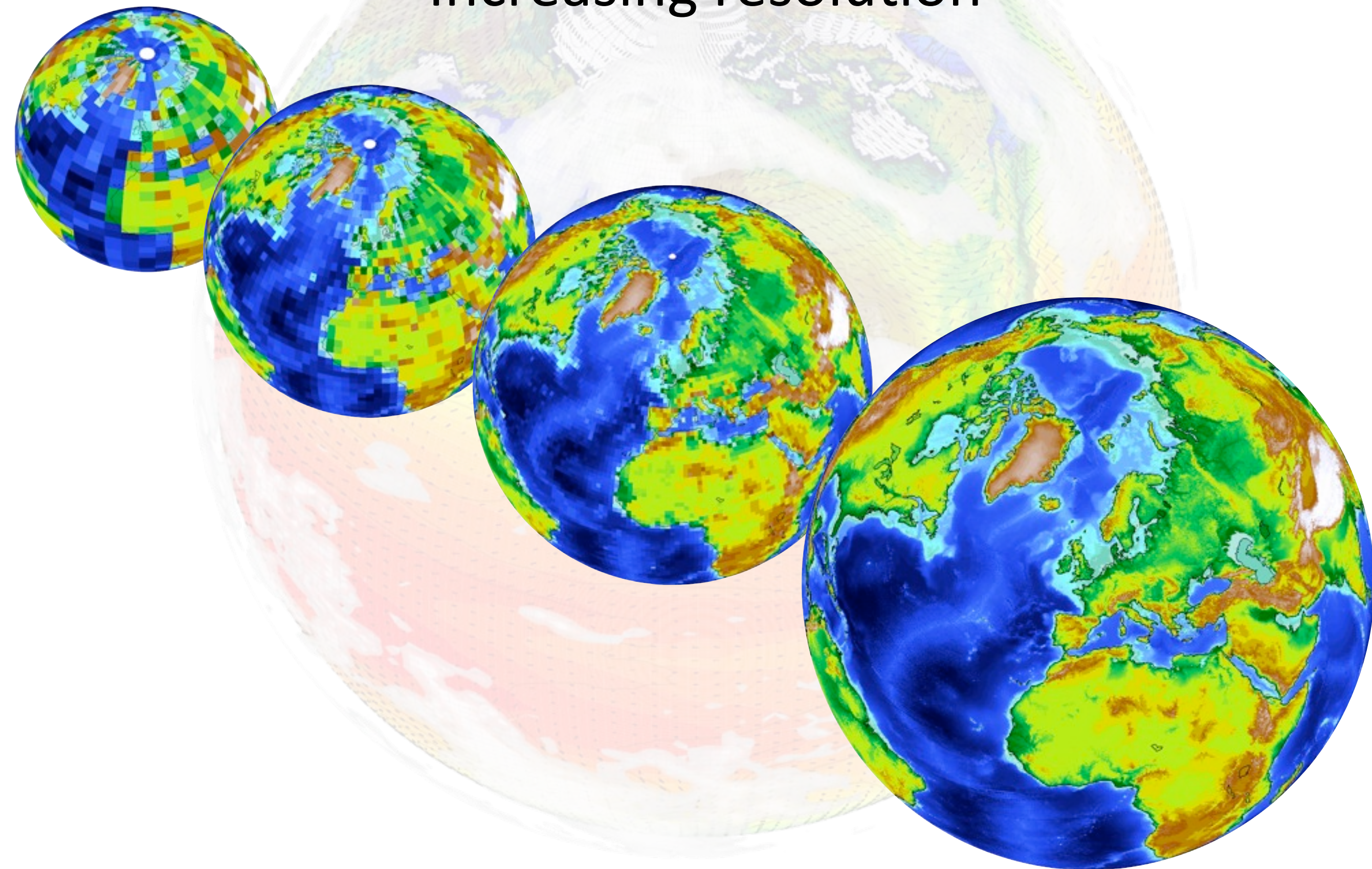
As required by CMIP5

```
:institution = "NCAR (National Center for Atmospheric Research) Boulder, CO, USA" ;
:institute_id = "NCAR" ;
:experiment_id = "lgm" ;
:source = "CCSM4" ;
:model_id = "CCSM4" ;
:forcing = "S1 GHG V1 SS Ds SD BC MD OC Oz AA LU (all fixed at or cycled over 1850 values)" ;
:parent_experiment_id = "N/A" ;
:parent_experiment_rip = "N/A" ;
:branch_time = 0. ;
:contact = "cesm_data@ucar.edu" ;
:references = "Gent P. R., et.al. 2011: The Community Climate System Model version 4. J. Climate, doi: 10.1175/2011JCLI4083.1" ;
:initialization_method = 1 ;
:physics_version = 1 ;
:tracking_id = "44ba7a25-6c75-4b07-9a7d-529bd07c70c8" ;
: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." ;
:cesm_casename = "b40.lgm21ka.1deg.003M" ;
:cesm_repotag = "cesml_0_beta05" ;
: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 mirage0 at 20120222 -160048.966" ;
:processing_code_information = "Last Changed Rev: 574 Last Changed Date: 2012-02-22 15:52:34 -0700 (Wed, 22 Feb 2012)
  Repository UUID: d2181dbe-5796-6825-dc7f-cbd98591f93d" ;
:product = "output" ;
:experiment = "last glacial maximum" ;
:frequency = "day" ;
:creation_date = "2012-02-22T23:00:56Z" ;
:history = "2012-02-22T23:00:56Z CMOR rewrote data to comply with CF standards and CMIP5 requirements." ;
:Conventions = "CF-1.4" ;
:project_id = "CMIP5" ;
:table_id = "Table day (12 January 2012) 7757d80c56ae0b9009f150afa4850c4e" ;
:title = "CCSM4 model output prepared for CMIP5 last glacial maximum" ;
:parent_experiment = "N/A" ;
:modeling_realm = "atmos" ;
:realization = 2 ;
:cmor_version = "2.8.1" ;
```

Friday, February 24, 2012

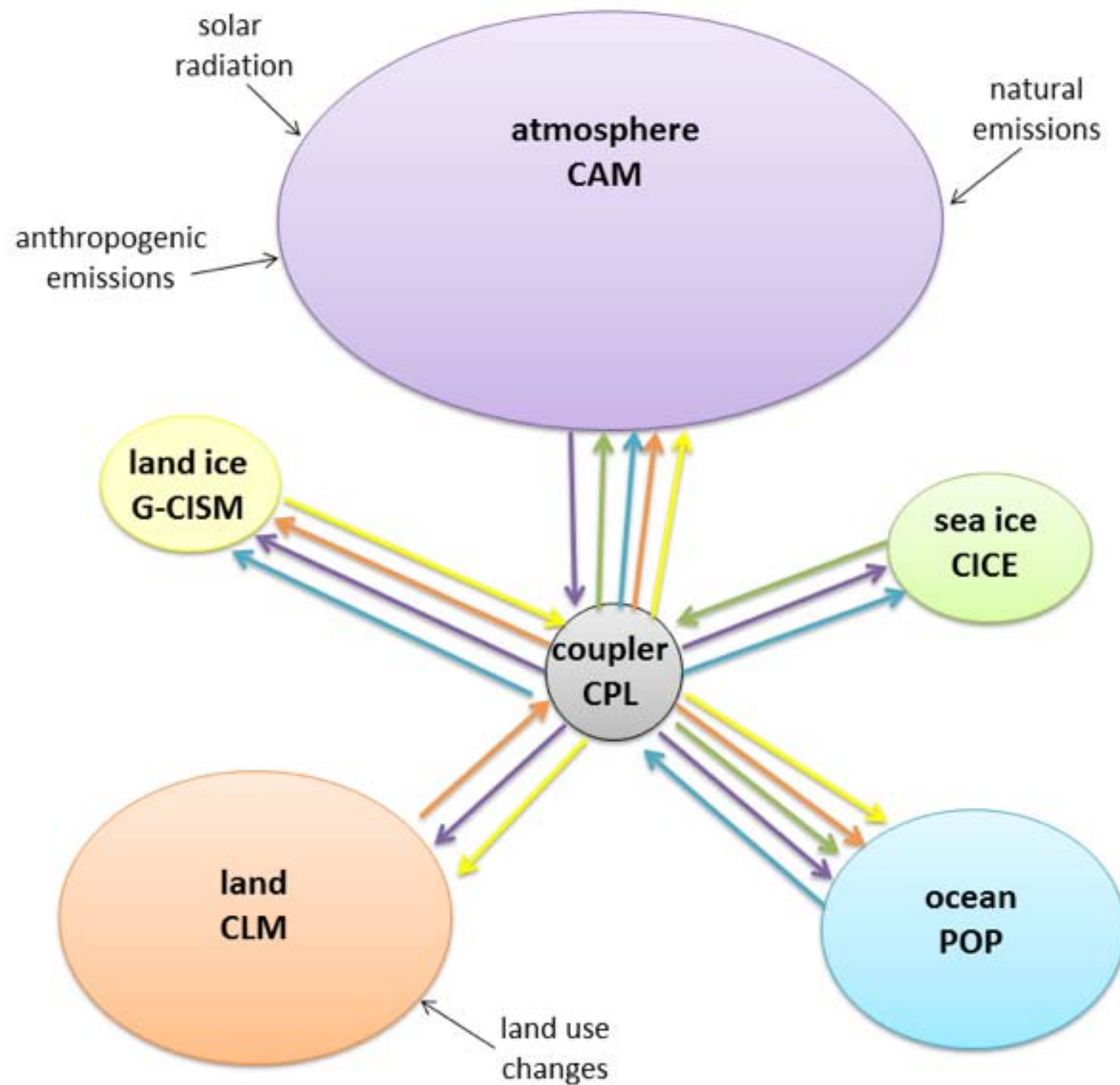
Challenges

Increasing resolution



Challenges

More components



Courtesy Caitlin Alexander, ClimateSight

Friday, February 24, 2012

Challenges

“Exotic” grids



Source: Randall, D. A. et. al., *Climate modeling with spherical geodesic grids*, Computing in Science and Engineering, 4, 5, 32-41.

The Parallel Ocean Program (POP) “tripolar” grid



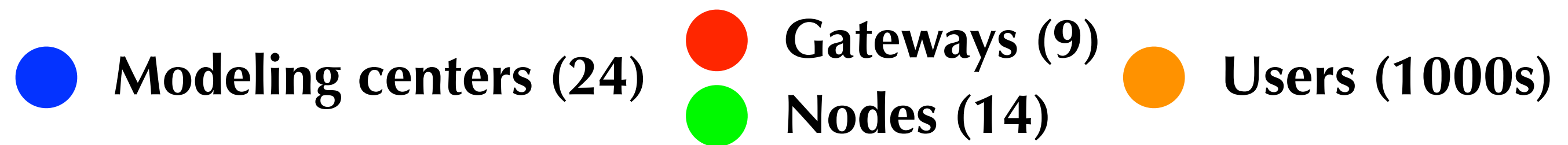
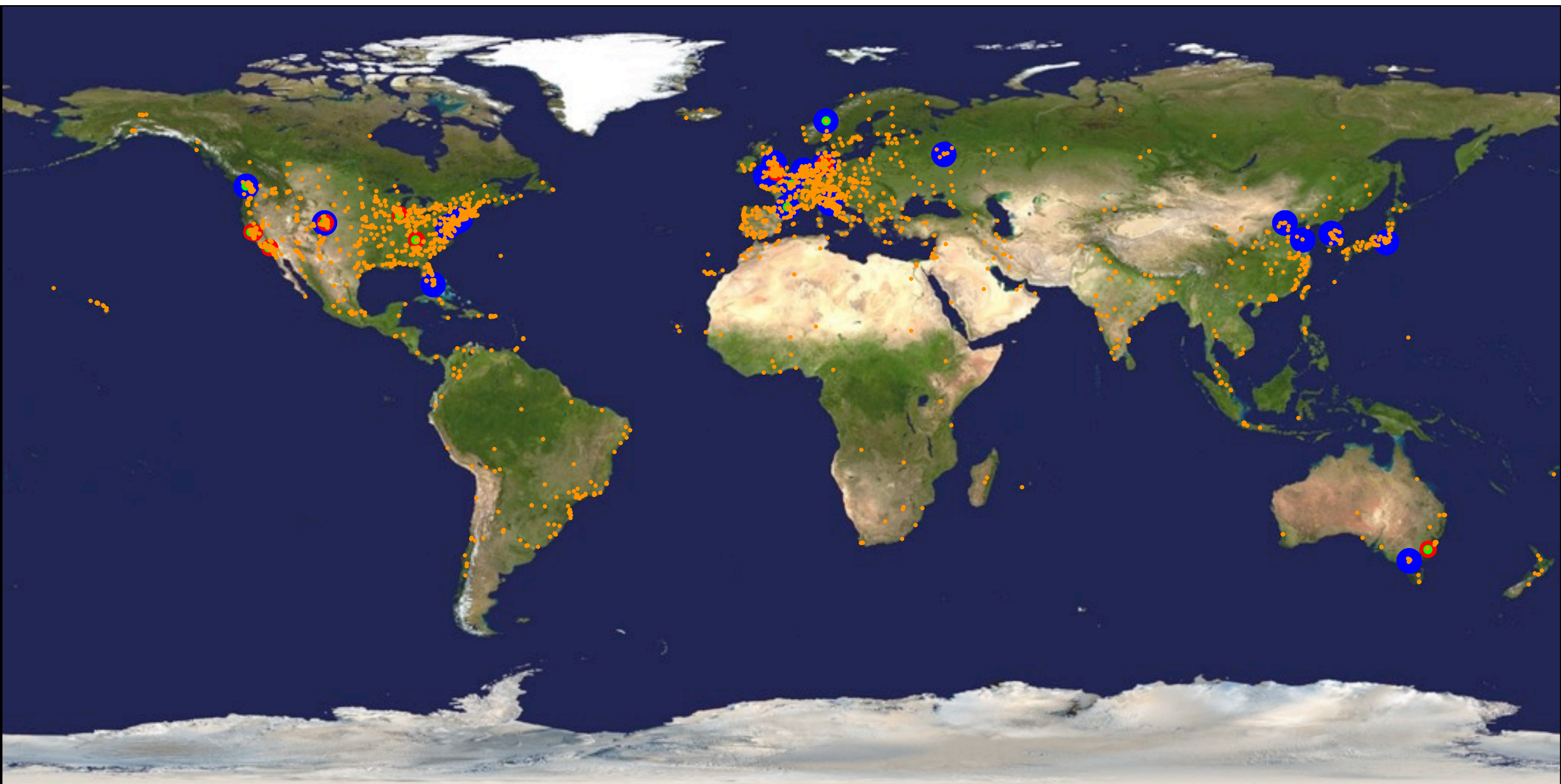
CESM CMIP5 simulations

CMIP5 type	Description	#
piControl	pre-industrial control	11
1% CO2 increase	1 percent per year CO2	5
historical	Simulate 20th century climate	24
historical variations	Single forcing runs, etc.	28
paleoclimate	Past climate (LGM, mid-Holocene, past 1000 years)	3
RCPs	RCPs 2.6, 4.5, 6.0, 8.5	32
Decadal predictions	Predictions (hindcast and forecast)	400
ESM	Earth System Model (BGC, carbon cycle, &c)	10
Other	Sensitivity and “idealized” Earths	25
Totals		538

CMIP5 variable counts

	subdaily	daily	monthly	annual	totals
atmosphere	197	74	222	0	493
land surface	3	5	73	0	81
ocean	1	3	220	71	295
sea ice	0	4	47	0	51
totals	201	86	562	71	920

CMIP5 distribution



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

<http://www.cesm.ucar.edu/management/docs/data.mgt.plan.2011.pdf>

CMIP5 website

<http://cmip.llnl.gov/cmip5>