Where Are We With CMIP6?

Sheri Mickelson
CISL/TDD/IOWA

Thanks to Gary Strand for CMIP5 information

June 19, 2019

This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977.

It’s been a group effort to get to this point

Software Developers
Sheri Mickelson, Kevin Paul, Alice Bertini

Data Management and Advice
Gary Strand

Advice and Guidance
John Dennis, Mariana Vertenstein

Running of the Experiments
Cecile Hannay, Peter Lawrence, Michael Mills, Simone Tilmes,
Robert Tomas, Brian Medeiros, Lantao Sun, Keith Lindsay,
Kate Thayer-Calder, John Fasullo

Publication and Documentation
Eric Nienhouse, Adam Phillips

And a big thank you to Mick Coady, Dave Hart, and others in CISL

So where are we?

- We're about 2/3^{rds} of the way through the tier 1 experiments
- To get this far, we've ran over 900 different CESM cases
  - We have about 200 left to run
- We've generated about 1 PB of compressed time series files
  - This would have been ~2.5 PB if uncompressed
- We've published to ESGF about 231 TB of compressed CMIP6 files from 728 CESM cases
  - This would have been ~575 TB if uncompressed
  - We have about 147 TB waiting to be published

Where are we with publication?

We have data published to ESGF under these MIPS …

- AerChemMIP
- CDRMIP
- CFMIP
- CMIP (all FV1 DECK simulations)
- LS3MIP
- LUMIP
- PAMIP

We have data ready to be published under these MIPS …

- C4MIP
- DAMIP
- GMMIP
- RFMIP
- Scenario
- And more from the column on the left

(These are just waiting for approval)

We have over 80,000 datasets that have been published with another large chunk coming within the next month or two.

This includes over 155,000 netCDF files that are ready to be published.

How does this compare with CMIP5?

CMIP5

- Timeline: 3 years from start to finish

- Generated about 2 PB of timeseries

- Published about 175 TB uncompressed

CMIP6

- Started first simulation the end of July 2018, postprocessing started near the end of September

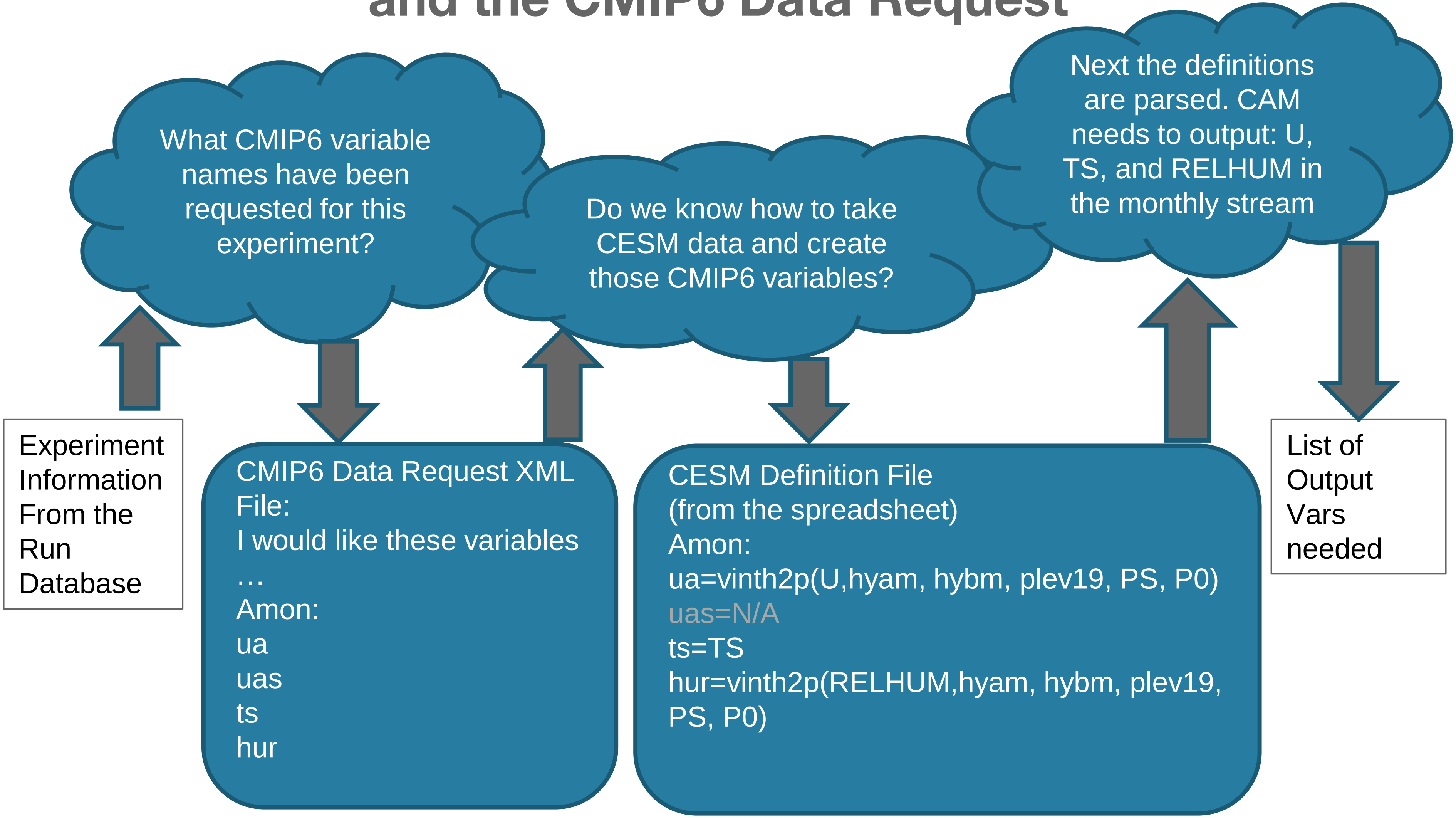
- We've generated about 1 PB of timeseries (2.5 PB uncompressed)

- Published about 231 TB compressed, with about 147 TB waiting to be published

For comparison, we've created 4x the amount of data for publication in 1/4th of the time
We've generated about the same amount of timeseries files in 1/4th the time

So what’s making the difference?

Tight Integration Between the Setup, Run Database, and the CMIP6 Data Request



Experiment Information From the Run Database

What CMIP6 variable names have been requested for this experiment?

CMIP6 Data Request XML File:
I would like these variables
…
Amon:
ua
uas
ts
hur

Do we know how to take CESM data and create those CMIP6 variables?

CESM Definition File
(from the spreadsheet)
Amon:
ua=vinth2p(U,hyam, hybm, plev19, PS, P0)
uas=N/A
ts=TS
hur=vinth2p(RELHUM,hyam, hybm, plev19, PS, P0)

Next the definitions are parsed. CAM needs to output: U, TS, and RELHUM in the monthly stream

List of Output Vars needed

This Integration Provides Us With These Benefits

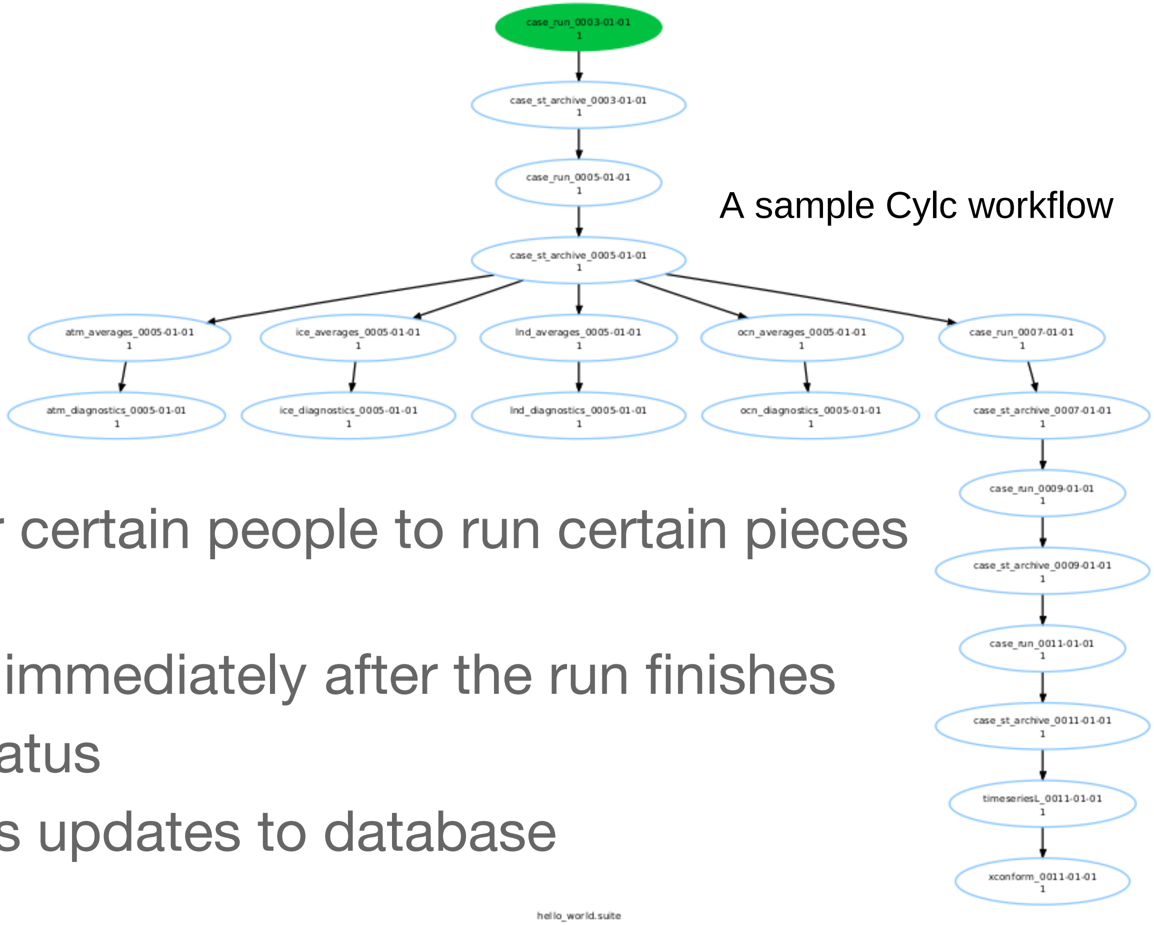
- We know exactly which variables need to be outputted by CESM to fulfill as much of the request as we can.

- The integration will automatically output exactly what CMIP6 is requesting for a particular experiment without anyone having to look up the experiment in the request up by hand and turn a crank on anything.

- The datarequest is queried in order to retrieve most of the variable and file attributes that are required for publication.
  - This integration allows for seamless updates between datarequest versions

- This allowed us to keep the data volume more under control

The Workflow Automation With Cylc

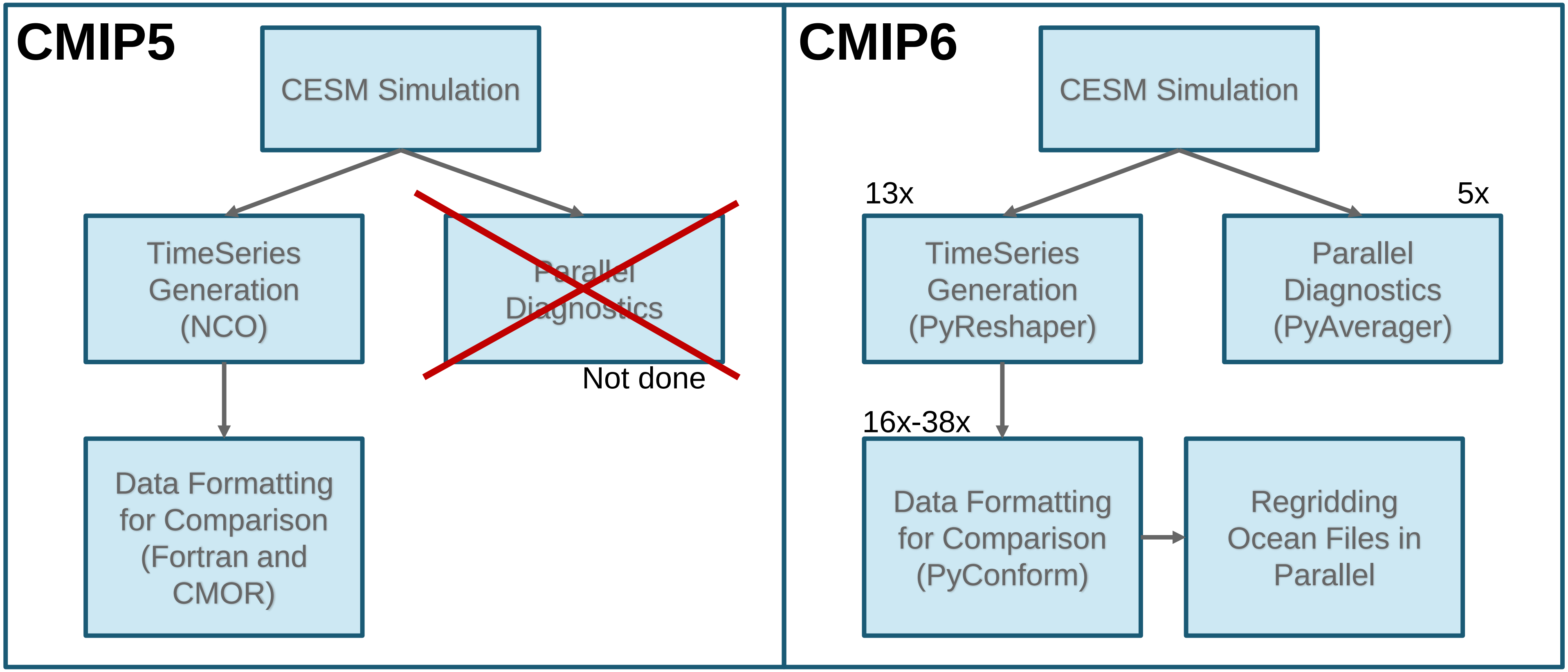


A sample Cylc workflow

- Eliminated the need for certain people to run certain pieces
  - No expertise is needed
- Data is postprocessed immediately after the run finishes
- Email notification on status
- Cylc providing progress updates to database

https://cylc.github.io/

The Parallel Post-Processing Tools


CMIP5
CESM Simulation
TimeSeries Generation (NCO)
Parallel Diagnostics
Not done
Data Formatting for Comparison (Fortran and CMOR)
CMIP6
CESM Simulation
13x
5x
TimeSeries Generation (PyReshaper)
Parallel Diagnostics (PyAverager)
16x-38x
Data Formatting for Comparison (PyConform)
Regridding Ocean Files in Parallel

Example Timings From a CAM/SSP3-7.0 Experiment

Task	Time	Number of Files Generated	Amount of Data
CESM Simulation	6 days	223,613	22 TB
Timeseries Generation (compressed)	3 hr, 25 min	5,079	9.5 TB
CMIP6 File Generation (compressed)	3 hr, 10 min	2,389	4.5 TB
Regridding Ocean Files (compressed)	11 min	399	289 GB

CISL Resources

- We have been running all experiments under one service account
  - This has prevented the need for asking for permissions to be changed in order to operate on data
  - Has allowed us to update runs as they progress
  - Made data management easier

- CISL has been very generous with the disk space they've given us
  - This includes extra space on scratch and collections
  - During CMIP5, it was required to put data on tape and pull it off for postprocessing and this was very time consuming
  - We have not had to move data around and we've been able to operate under CISL data policies with a carefully considered data plan

Other Things That Have Made the Process Smoother

- We've had more time to prepare
- Having regular development meetings
- The CMIP6 data request was better organized than it was for CMIP5
  - The request exists as an XML database that we query to get all information
- More people running the simulations
  - Taking out a lot of the expertise CMIP6 knowledge helped make this easier
- ESGF improved the publication process
  - Faster, all files are tracked better, and conventions are enforced

New Bottlenecks That Need To Be Considered

- The scheduling of the experiments was never enforced
  - This is the new bottleneck, we’re waiting for new experiments to start
  - We could have been running more experiments at a given time and this would have allowed us to meet the July deadline
- CESM must be able to write these variables out by default
  - This causes us to carry three copies of the data
  - As we go to the next round of CMIP, the data request will be larger and this will become a challenge under the current workflow
  - This would make it easier to run more non-cmip6 model inter-comparison projects

Questions?

The CESM project is supported primarily by the National Science Foundation (NSF). This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the NSF under Cooperative Agreement No. 1852977. Computing and data storage resources, including the Cheyenne supercomputer (doi:10.5065/D6RX99HX), were provided by the Computational and Information Systems Laboratory (CISL) at NCAR