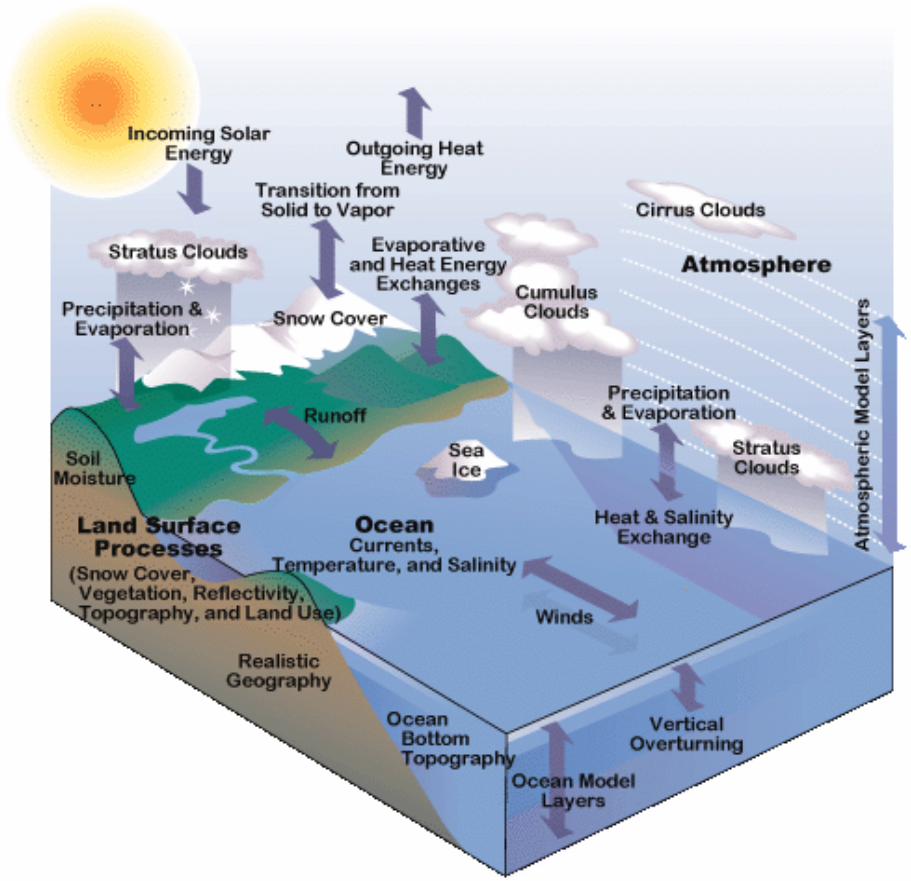
# CESM's New Data Workflow for CMIP6

***Sheri Mickelson,***
*NCAR/CISL/TDD*

**May 5, 2020**

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.

# Community Earth System Model (CESM)




CESM is a coupled climate model that contains separate models that simulate the atmosphere, ocean, land, sea ice, and land ice.

Each model component outputs its own datasets.

Data per component varies in the amount of variables per file, the grids used, attributes on the files, and how they interpret CF conventions.

Image credit: https://eo.ucar.edu/staff/rrussell/climate/modeling/climate_model_components_evolution.html

# What is CMIP6?

- CMIP6 is a large international project that consists of many centers around the world running the same simulations, in order to seek a better understanding of Earth processes.
- Requires all centers provide model output in a standard/specified format and the data must be made available through the Earth System Grid Federation (ESGF).
- This is done so researchers can more easily compare model output from different centers.

Image: https://cdn.pixabay.com/photo/2013/07/13/12/48/earth-160383_640.png

# CMIP6 Design

## CMIP6 Involves:

- A core set of experiments (DECK - Diagnostic, Evaluation, and Characterization of Klima)
  - Control
  - AMIPs
  - Historical
  - 1pctCO2 increase
  - Abrupt-4xCO2
- Other experiments branch off of the DECK experiments

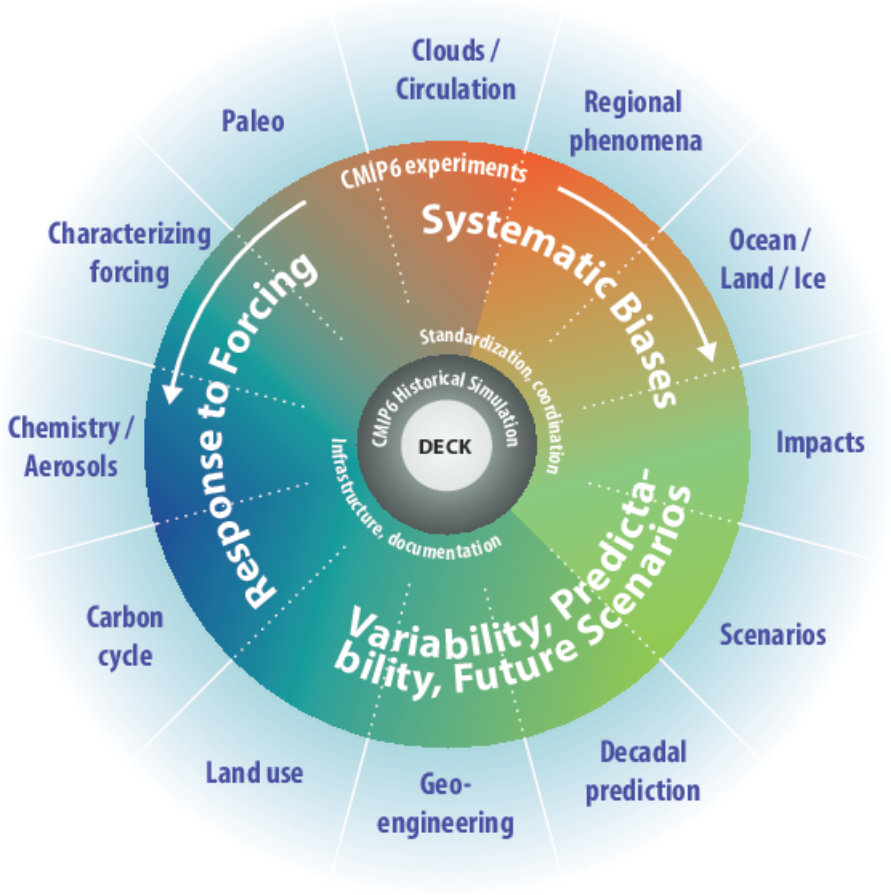



Image Credit: Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geosci. Model Dev., 9, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, 2016.
https://www.geosci-model-dev.net/9/1937/2016/gmd-9-1937-2016-f02.png

# Data Amounts
## CMIP5 vs. CMIP6

### CMIP5 Volume Statistics

- 54,632 datasets were contributed
- The volume is approximately 2 PB
- Data is from 39 different experiments, from 59 different models

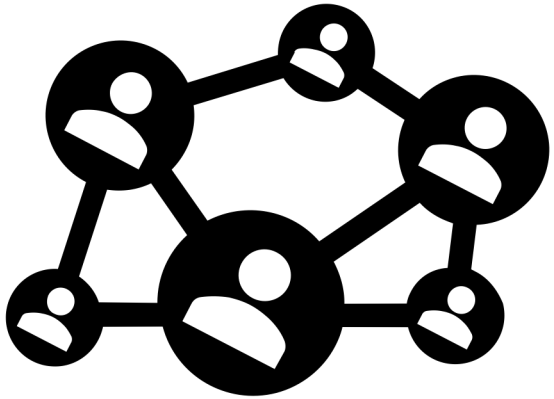
### CMIP6 Volume Statistics

- So far, over 3 million datasets have been contributed
- The volume is approximately 20 PB
- Data is from 312 different experiments, from 129 different models

# CMIP6 Contributions
# Total vs. NCAR

### Total Contributions

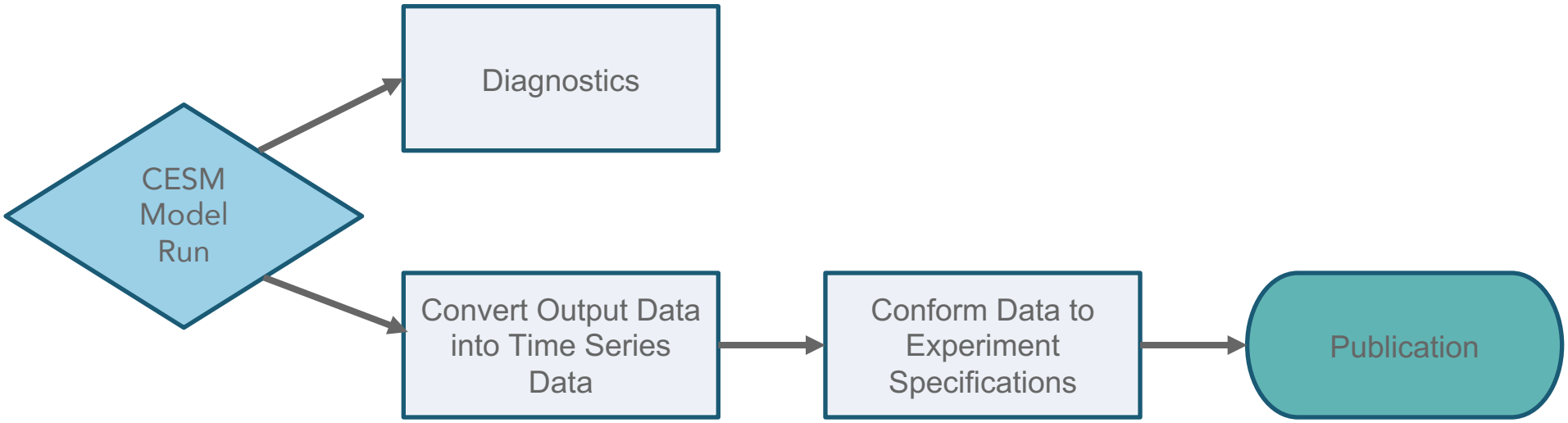
- Dataset Count: 3 million
- Volume: 20 PB
- 312 Total Experiments



### NCAR Contributions

- Dataset Count: 551,000
- Volume: 500 TB (all compressed)
- About 130 Experiments
- Consumed 190 million CPU hours

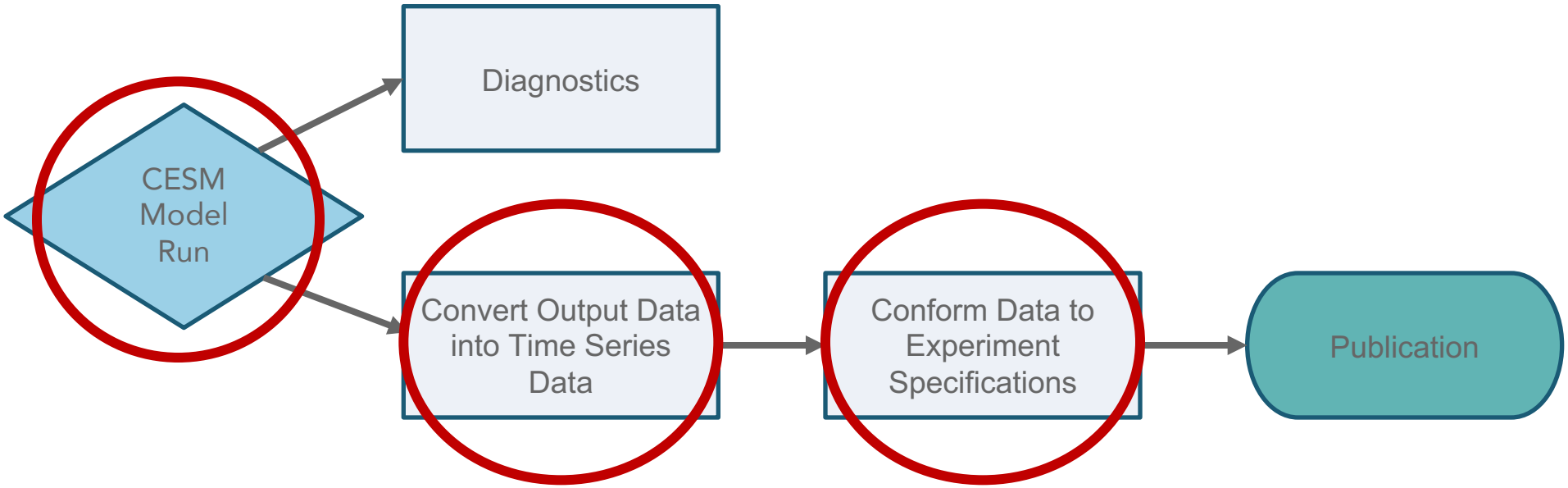
# CESM Data Flow




For CMIP5 we were the first model to finish their simulations, but the last to finish publishing our files.

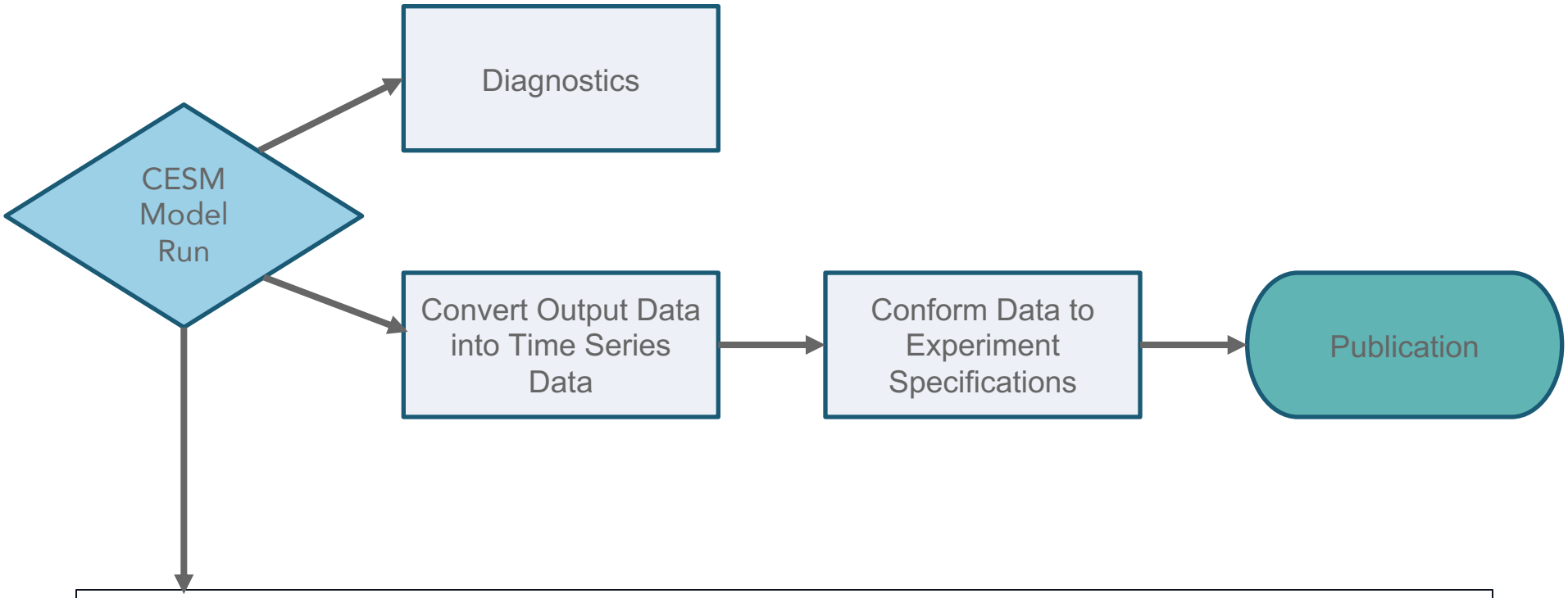
CMIP6 would stress our workflow even more.

# How We Improved the CESM Data Flow for CMIP6




We invested resources to improve these processes in order to handle the volume of data that needed to be post-processed within the allotted timeline of one year.

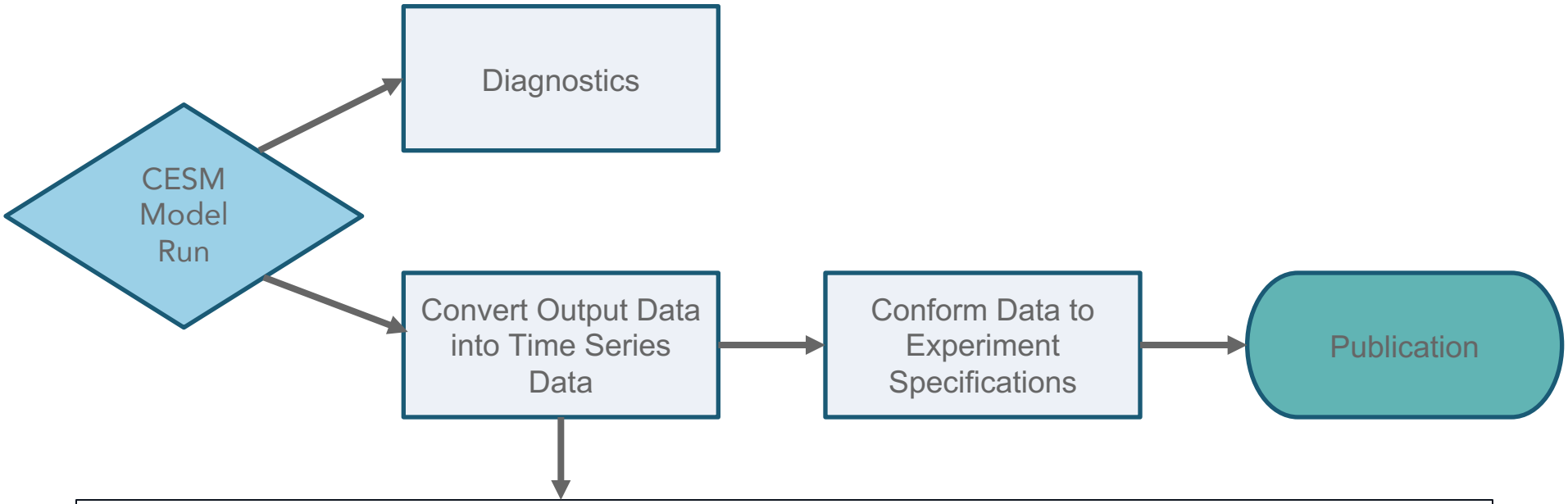
# How We Improved the CESM Data Flow for CMIP6




In order to reduce our data footprint we developed scripts that told our users exactly what needed to be outputted by the model in order to fulfill the data requested for that experiment.

**This data is not saved after it is post-processed**

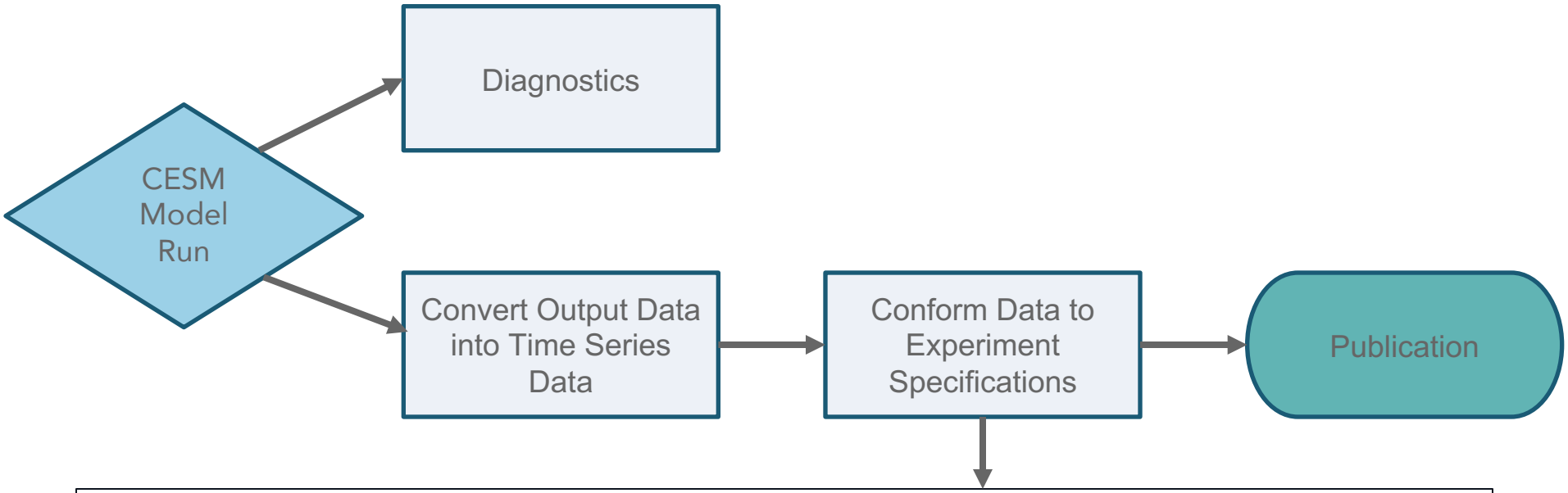
# How We Improved the CESM Data Flow for CMIP6




- The model outputs files where each file contains many variables and few time steps. We needed to convert these files into a format were they contained only one time series variable and many time slices.
- This was an expensive process during CMIP5.
- We developed a Python tool (PyReshaper) to create these file in parallel and output in compressed NetCDF format. This saved us a couple of PB worth of space.

**This data is saved**

# How We Improved the CESM Data Flow for CMIP6




- This step requires centers to format output to meet specifications.
- Requires us to combine, convert units, and rename our model output variables.
- This also required the use of controlled vocabulary for global attributes and correct file naming conventions.
- We developed a Python tool (PyConform) to create these file in parallel and output in compressed NetCDF format.

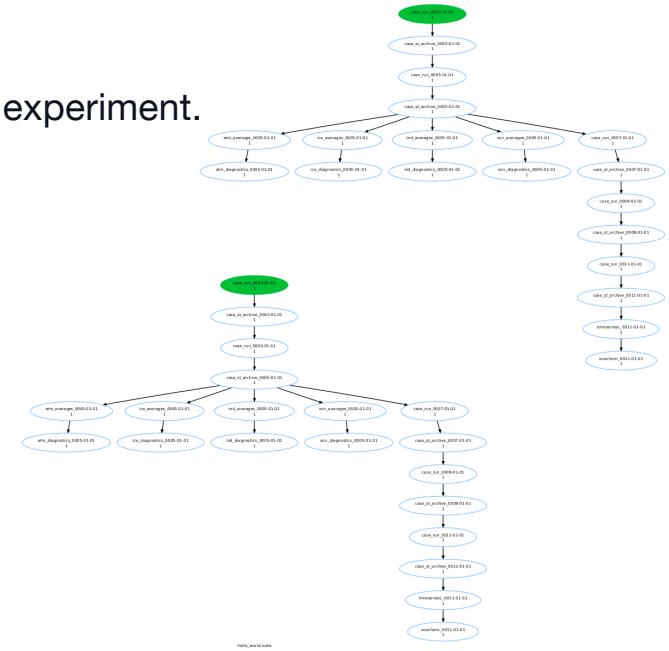
**This data is saved and published to ESGF**

# Automating Our Workflows

- We used Cylc to automate our workflow.
  - Because Cylc uses a script to determine when and how to run tasks, our workflow is preserved.

- After every model iteration, the experiment configuration files where archived in order to determine if values were changed.
  - Changes were recorded into our experiment database.
  - We can review these changes in order to reproduce the experiment.

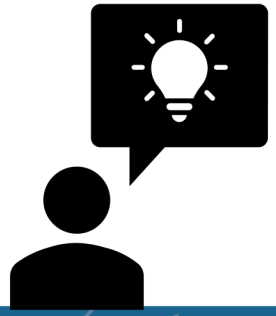
- While we cannot achieve BFB results if compilers or machines have changed, through our research we found that if the compiler passes acceptance testing (with pyECT*), we can accept these results as falling within the acceptable spread of an ensemble.



* doi:10.5194/gmd-8-2829-2015

# CMIP6 data formatting, standards, documentation

- All file names follow a specific naming convention
  ta_Amon_CESM2_piControl_r1i1p1f1_gn_000101-009912.nc
- Specific global attributes are added to each file and each follows a controlled vocabulary
  - These values are verified before the data can be published
- All of the climate models, experiments, and mips needed to be thoroughly documented before any data could be published
- All of the controlled vocabulary can be found here:
  https://github.com/WCRP-CMIP/CMIP6_CVs
- All documentation on the experiments, models, and mips can be found here:
  https://search.es-doc.org/
- Errors in data are also documented and can be found here:
  https://errata.es-doc.org/static/index.html
- Data citations are created for each collection and can be found here:
  https://cera-www.dkrz.de/ords/f?p=127:2



# 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.