Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2020-133-RC2, 2020

Geoscientific Model Development Discussions

Interactive comment on "A New End-to-End Workflow for the Community Earth System Model (version 2.0) for CMIP6" *by* Sheri Mickelson et al.

Anonymous Referee #2

Received and published: 29 July 2020

General comments :

This paper presents the work carried out to completely modify the CESM's post-processing workflow. It's interesting and useful to get an overview of such a process, but I think some information are missing for the paper to serve as an example for other communities.

During my reading I would have liked to know more information on the Cheyenne super-computer. For example, do you have some restrictions on the storage (volume quota, inodes quota), is this supercomputer dedicated only for CMIP6 experiments ? Did you have some restricitions on you cpu allocation for post-treatment ?

For each part, I think it can be useful to have an information on the human time and

FTE necessaries to realize the tool from scratch to the production.

It's really a great job to have created this workflow that can be used by a "normal" user, and that avoids the problem of knowing CMIP data that only relies on a few people.

Specific comments :

Introduction

- lines 24 & 25 : Can you add a graph in order to visualize calcul and post-treatment performances for NCAR and other climate models

Data Workflow

- line 41 : "it was time consuming" : can you precise if you are talking about "human time" (find the script, launch it, check it etc.) or cpu time ?

- Line 63 : can you explicite "FTE" before to use it for the first time ? How did you make the FTE estimation for the implementation of XIOS and for the development of your own new tools ?

Time Series Generation :

- line 96 to 104 : Can you precise in the text how many Time-series (493) are created by your evaluation. why did you stop the test to 144 MPI ranks and don't test with more MPI ranks ? Did you try with 493 MPI ranks ? Can you explain how finally you make your choice for the MPI ranks repartition you will use, I imagine there is a reflexion between the human time (5 $\frac{1}{2}$ hours with you previous workflow and now 4 $\frac{1}{2}$ minutes), the total CPU time (4 $\frac{1}{2}$ minutes * 144 = 10,8 hours), and your cpu allocation on Cheyenne. (this specific comment is done also for the other parts of your workflow)

- Line 102 : did you try to improve the way you done the variables distribution on MPI ranks ?

- Figure 3 : can you add the "ideal speedup" line on it ?

Diagnostics

- line 117 to 122 : can you add information on how the choice of subcommunicators's number was done, and of the MPI rank distribution on each subcommunicator.

- Line 128 to 130 : can you explain on which criterion was done the climatologies distribution on MPI ranks ?

- Line 135 : can you re-run the experiment on 32 MPI ranks, to fixed the distribution problem.

- Figure 5 : can you add the "ideal speedup" line on it ?

Conforming Data to Meet Specifications

- line 147 : can you explain what you mean by "flexible interface" ?

- Line 148 : can you describe the "task-parallel approach" you choose to implement ?

- Lines 152 a 153 : how users that are not experts on CMIP6 (as it's tell several times in the paper for example lines 218 & 219) can know which functionalities need to be create ?

Data Publication

- As far as I know PrePARE will check the correpondance between output metadata and what is wait by CMIP6. But it will not check outputs quality (for example : no missing time step on a time-series). Can you present how you manage the quality control of your cmip6 outputs files ?

- What happen if PrePARE return problems on outputs cmip6 files ?

Process workflow

- can you explain if learning how to use Cylc was easy or not ? Can you estimate time and FTE necessaries for this implementation ?

- Did you hesitate with another software ?

- Maybe it can be useful to add a graphic showing how cylc is incorporated to your workflow, with the call tree of all your tools.

- Line 213 & 215 : I don't understand the difference between "the users set the default values" and "users only needed to set experiment specific information". And if it's "default values" why users need to modified them ?

- Is Cylc workflow can solve all errors ? Or is there a need for human intervention from time to time?

Experiment Documentation

- Line 229 : "The experiments that . . . no provenance was obtained" : can you precise if it's only for NCAR simulations or for all groups's simulations ?

- Line 251 : can you precise how are managed "simulations that ran into problems" ?

Technical corrections

- Line 54 : it's finish by a "," instead of a "."

- Line 55 : "steps including;" need to be modified by "steps including:"

- Line 77 : "Instead the data", I'm not sure that you want to tell "instead", maybe "by consequences" or something like this.

- Line 91 : "this task base parallelism" need to be modified by "this task based parallelism"

- Line 187 : "CMIP6", I think you want to write "CMIP5"

- Line 200 : "in order keep track of the statues of all of the running tasks. In order to track the status of all of the tasks ...", maybe you can avoid to write two time "in order . . . tasks"

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-133, 2020.