Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC2
https://doi.org/10.5194/gmd-2021-290-RC2, 2021

Comment on gmd-2021-290

Patrick Bartlein (Referee)

Referee comment on "Analysing the PMIP4-CMIP6 collection: a workflow and tool (pmip_p2fvar_analyzer v1)" by Anni Zhao et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2021-290-RC2, 2021

General comments:

This manuscript describes a collection of datasets (the "curated ESGF replica"), a modified analysis package (a "paleo version" of CVDP), and a set of scripts written in NCL and Python for additional data analyses and visualization. Although not a model or stand-alone software package, it certainly codifies and advances the analysis of "CMIP-like" model-output data sets. The current approach involves downloading hundreds of files, a lot of fixing up and intermediate analysis, and finally summarization and visualization, with each investigator and group applying their own preferred set of tools, often involving other fix-ups and ad hoc analysis steps. The workflow described here goes a long way toward establishing a "best practices" approach to the analysis of CMIP-type data sets in general, and to PMIP paleoclimatic simulations in particular.

One thing that is absent from the beginning of the paper is a discussion of the "why do we need this?" question. A naïve reader might wonder if given a bunch of standardized files (from the ESGF), how hard could it be to produce some figures? Although the files may indeed be standardized, the models have, for example, differing resolutions, the simulations are of differing lengths, and there is considerable data-reduction and transformation that can not be transparently implemented using simple scripts, plus there are many, many variables—hence the utility of the present paper. The place for this discussion might be a new intro paragraph to Section 2 (or at the end of the Section 1) that describes the problem/motivation as in the previous sentence, with the solution being the workflow and tools described here. The sentence beginning on line 293 is a good overall summary, but it's at the end of the paper.

In fact, the workflow and tools could be better illustrated—the components are scattered across several GitHub repositories, and on first reading of the manuscript, I didn't immediately see how one would use the "curated ESGF replica" (or where it was). It took plowing around in repositories to figure that out. I think a figure that illustrates the data assembly, curation, development of second-order data sets (i.e. the CVDP output), and

production of summaries and illustrations would be good substitute for Fig. 2, which focuses on a different subject.

There are lot of technical terms, both climate- and IT-related. Readers might benefit from a few short "in-line" definitions on first use, or URLs to appropriate web pages.

The term "ensemble" is used in several distinct ways here, a) to represent the whole collection of simulations, as in the title; b) to describe all of the individual simulations for a particular experiment (i.e. "*lig127k* ensembles", line 87); and c) to represent multi-model means, e.g. lines 94, 163). I suggest that the whole collection (i.e. PMIP4-CMIP6, PMIP3-CMIP5, etc.) be referred to as the "collection", all of the simulations for a particular experiment as the "ensemble", and the multi-model means as "multi-model means". There's also a little fuzzy usage of "ensembles" (line 89), where the term could mean "ensemble average" or "ensemble members" ("the subset of models").

The paper does need some mechanical work and smoothing out for readability.

Specific comments:

line 9: "… to test the out-of-sample response to…" It's not the boundary conditions that are being tested, but instead the response of the models (to boundary conditions different from present).

line 15: I think the non-MIP-enabled reader might benefit from a short introduction, maybe a sentence, to the overall notion of using multiple models to simulate the climate while adhering to an identical experimental design, and evaluating those simulations using obervations.

line 16: "… improved forcings and boundary conditions by the new generation of climate models…". There's really two things there, updated experimental designs, and new models.

line 23: Not just topography, but also ice-sheet size.

line 32: Non-PMIP readers might wonder about the MIP-numbering scheme (and in particular what PMIP3-CMIP5/6 means, as well as "past2future").

line 41: Reverse the order of "Curating and Collating" to reflect the order that the work is done in. Also, I'm not sure "collating" is the right word; it's main definition includes the ideas of critically comparing or arranging items in order, while I think the work that was done here was perhaps more like "collecting" ESGF data, other data, and the results of the CVDP output.

line 45: How many source files are there? (i.e. just the PMIP-CMIP ones, not the whole repository).

line 50: You might do an in-line explanation of what "r1i1p1f1" means. Same for "areacello" a few lines down.

Line 58: "curated replica". I would reverse the order of discussion to state that you sought to build a replica of part of the ESGF data base (and why), and then how it was populated.

line 60: Explain what ncrcat is.

line 60: "varying years" Does this mean varying length of simulation? Because individual models use fake years for paleo simulations, there's really little overlap in years. Did you use the last, say, 100 years of each simulation as the common period?

line 74: "Calculation of regional mean temperature… used the adjusted monthly temperatures."

line 77: "midHolocene experiment."

line 79: I would substitute "while" for "despite".

line 80: "Interactive Atlas" This is the first mention of this, and probably should have URL.

line 85: "This has the disadvantage…" It's not really a disadvantage, it's just the way it is. Calculating annual mean temperature as the simple average of monthly mean value, without weighting for month length will always yield a different result than calculating the annual average from, say, daily data (unless the months have equal number of days).

line 89: “ensembles” or “ensemble averages”

line 101: “time spatial” What’s that?

line 102: “climate variabilities” Variables? Also “user-specified set”

line 106: “the variability (of something) in the CMIP6 simulations”?

line 106: “The resulting data” What resulting data?

line 109: Replace “look throughout the year” with “other seasons”?

line 100: “… although these were not used in the publication.”? Otherwise, I don’t understand.

line 115: “each individual monsoon season” Does this mean each year? Or was it intented to say “each individual monsoon region”?

line 116: Reword. (Too many “regional monsoons”.)

line 130-133: Wouldn’t you also want to define a fixed number of years for calculating climatologies? The standard error of the mean is strongly dependent on the number of observations.

line 136: What modification, and which manuscripts?

line 142: ambiguous “them”. The conflation of mean-state changes and variability, or the AMOC variations themselves?

line 149: Finish sentence. More computationally efficient that what?

line 150: Which associated web pages?

line 156: I don't understand "greater explanation of example cases".

line 164: I don't understand "retrospectively".

line 165: "The plotting routines allow 'resources'…". This is a generic feature of NCL and is not specific to the code described here. For the benefit of non-users of NCL, it might be good to include a couple of sentences that describe the basic architecture of the language.

line 174: Break up sentence.

line 187: Define "Binder" (or provide a URL).

line 189: Define "Docker" (or provide a URL), and "bespoke NCL analyses".

line 197: "These" Ambiguous. "summary data" or "CVDP output"?

line 198: "… collect together a single piece of information…" Single pieces?

line 204: "the global-mean surface temperature (something) also includes"

line 205: "equilibrium climate sensitivity"?

line 205: "temperature changes" Does this mean "long-term mean differences"?

line 209: "… climate models as reported in the Technical Summary…"

line 216: "… and monsoon (something)…"

line 222: I think "temperature change" is ambiguous. These are long-term mean differences, paleo minus present (the usual convention, but not everybody knows that).

Figure 4: No panel (e) in the figure.

line 251: Define "Iris cube" (or provide a URL).

line 251: "…all models are regridded onto…" Is this the same kind of regridding as in step 4 on page 10? If so, it probably should have been introduced then; if not, why is a different approach used here (and what was used for step 4)?

line 253: Why this particular scheme, and not iris.analysis.AreaWeighted()? Interpolation of AMOC data would seem to be an instance where conserving mass or volume would be good.

Figure 6: Should it be "grey line"?

line 266: "the standard CMIP6 model colour scheme" The previous citation to the color schemes only defines the colors, not their assignment to individual models.

Figure 8: I think this figure would "read" better if it were broken into two panels: the top one showing the latitude of the monsoon extent, and the bottom the experiment minus control difference in northward extent. The horizontal line at 10 sort of accomplishes that visually, but what's significant about 10?

line 274: "land monsoon" What does this mean?

line 275: "changes in monsoon characteristics relative to present"?

line 279: "… area-averaged monthly time series…" Replace with "area-weighted average monthly time series."

line 283: "NAF expansion" It's not the region (NAF) that's expanding, but the areal extent of the monsoon in that region.

line 295: "…this manuscript obviates…. The manuscript surely helps explain things, but it's the modification of the CVDP and the NCL and Python scripts that will reduce the need for reverse engineering.

Technical comments:

line 3: "there is overhead"?

line 21: Delete semicolon.

line 24: "large-scale"

line 29: "as well as to figures in that report"

line 36: "the following section" Specifically give the section number (for parallelism).

line 56: "This was"?

line 65: Delete "up"

line 67: Delete "a"

line 68: "stored in the general"

line 68: Reword: "The approach taken by the PaleoCalAdjust software is to interpolate from non-adjusted monthly averages to pseudo-daily values, and then to aggregate those values back up to 'monthly' resolution for each 30° segment of Earth's orbit…"

line 87: Begin new paragraph at "Subsetted files…". The theme really changes here…

line 92: Reword: “because Kageyama et al. (2021) included non-CMIP models.”

line 99: “interannual” See also line 104.

line 105: Reword: “For later use, output is saved in netCDF files that contain the data fields that are plotted in each .png image.”

line 108: Throughout, CVDP is sometimes used with an article (i.e. “the CVDP”); and sometimes not (e.g. “CVDP”, as here). It would be good to standardize.

line 112: Replace semi-colon with a comma.

line 113: Delete “as”.

line 121: Delete “that is”

line 122: Replace “are different to” with “differs from”.

line 123: Replace “is defined as where” with “is defined as the region where”.

line 153: Delete “started”.

line 154: Delete “then”.

line 155: The Google suggests “Python” should be capitalized.

line 157: Replace “are given the name” with “named”.

line 160: Delete “the loading of”

line 162: These functions are intended to operate on a directory containing the CMIP summary files…”

line 172: Delete “respectively”.

line 173: “documentation”

line 182: “accessed”

line 185: “Python notebooks allow documentation and output to be stored…”

line 187: “cloud-computing”

line 188: Replace “would want to” with “should”.

line 189: “a containerized”

line 190: “platform-agnostic”

line 191: Replace “uses” with “user”.

line 191: “and to act as a Jupyter notebook server that allows..” (for parallelism)

line 191: “Docker”

line 197: “comma-separated”

line 201: “area-averaged”

line 202: “newly created”

line 213: “is the spatial”

line 219: “output”

line 219: “These examples have generally already been featured…”

line 224: “the function”

line 226: “the function”

line 228: “The process used to create (these images? these directories?)…”

line 229: This list isn’t parallel. Change (4) to “regrid the difference”. (“Anomaly” in climatology generally refers to a particular observation minus its long-term mean.)

line 243: “built”

line 243: “We provide example code…”

line 249: delete “an”; hyphenate “ensemble-mean”

line 264: “The CVDP can also be used to analyse…”

line 268: Replace “at” with “is”?

Figure 8: “NAF monsoon”

line 271: “or in a format”

line 285: delete “to”