RC1: ['Comment on egusphere-2023-2720'](), Anonymous Referee #1, 04 Dec 2023

In their manuscript, Lee et al. present the PCDMI Metrics Package (PMP) version 3, which is an open-source Python software package that provides tools for comparisons of ESMs with each other as well as for comparisons of ESMs with observations. A large range of atmospheric processes can be assessed with this software package. Since this package has been developed within the CMIP comparison projects the PMP results can be produced for all model simulations contributing to CMIP6 and earlier CMIP phases.

The manuscript is generally well written and the package useful for the scientific community, but I have several comments that should be taken into account before publication in GMD. Generally, the authors should take care that their message comes better through. At the moment, I have the feeling that there is a lot written, but what is the take home message? Is this a unique package or are there other, similar packages available. What is new or unique for your specific software package?

We appreciate the reviewer's time for providing thoughtful and constructive comments. Please find our point-by-point response below (text colored in blue). Also, to better emphasize some of the key messages, we have reduced and/or simplified some content throughout the manuscript.

General comments:

- This is a quite extensive overview and I was wondering how a new user should get started with PMP. I saw from the links you provided that everything is quite well documented, however, this is so much information at once that I think that a new user will have really trouble getting started. Is there e.g. a documentation that could be downloaded as a single pdf file or does one really spend hours reading everything on the computer screen?

  Thank you for pointing this out. We have an online documentation website (http://pcmdi.github.io/pcmdi_metrics/) that includes instructions for installation and demo for quick start. We have added the link to it in the "Code and Data Availability" section of the revised manuscript. In addition, we also have added a brief description for the installation and also link to the instructions in Section 2 to respond to this comment. We also note that a typical user of the PMP may only make use of a few of the performance metrics available, and thus only need to read the documentation relevant for their purposes.

- How can a user install PMP on a computer? What requirements are needed? You provide a link to your github repository, where this information is provided, but I think this information should also be provided in the manuscript itself.

  As responded above, we have added the installation information in the manuscript, and also clarified the link to the instruction. We agree with the reviewer that it is helpful to point to this explicitly in the manuscript, but we would rather not add too many details.

● A general questions I was wondering about is, if you are offering workshops for training of new users?

We appreciate the reviewers' interest. We have provided demo notebooks that users can follow for self-guided tutorials, which are available from the online instruction and we have clarified this in the revised manuscript. In parallel, we are in the process of preparing such training sessions but dates and platforms are yet to be determined. Once decided, we will post it via the PMP website and other available channels to reach out to the community. We plan to provide recorded video of the tutorial via the PMP's online documentation website in the future for those who missed participating.

● In the abstract you state what you will discuss the history up to date, recent updates and future directions. The future directions are discussed to some part in Section 6, but I could not find any information or discussion on the history of PMP up to date or what the updates between former and this version were.

This information has been described at the beginning of Section 3 in the original submission, but to make this more prominent we have revised the paragraph to further clarify which updates have been made to the latest version of the PMP, and more generally to the origins of the PMP. We also have reorganized the Introduction section to more clearly describe some of the PMP's history.

● Generally, all sections seem to be a bit too lengthy and to my opinion not really coming to the point. Best example is the summary and future directions section. Although you have a discussion section, you provide a quite long summary and future directions section without really summarizing what you have presented.

We acknowledge the reviewer's point. To respond to this constructive feedback, we have reorganized the discussion and summary sections. In the revised manuscript, we have renamed the last two sessions as "Discussion" and "Summary and Conclusion", and some discussion pertaining to the future directions were moved to the "Discussion" section to make the "Summary and Conclusion" more clear and concise.

● Is this the only tool for analyzing CMIP data or have there also other tools been developed?

In the original submission we discussed other tools that are available in the community in the first paragraph of the Discussion section. To improve clarity and readability and to respond to the comment, we have moved the description of other fellow tools to the

Introduction section. We also further discussed the diversity of the tools in the following paragraph of the Discussion section:

"Current progress towards systematic model evaluation is exemplified by the diversity of tools being developed (e.g., the PMP, ESMValTool, MDTF, ILAMB, IOMB, and other packages). Each of these tools has its own scientific priorities and technical approaches. We believe that this diversity has made, and will continue to make, the model evaluation process even more comprehensive and successful. The fact that there is some overlap in a few cases is advantageous because it enables the cross-verification of results, which is particularly useful in more complex analyses. Despite possible advantages, having no single best or widely accepted approach for the community to follow, does introduce complexity to the coordination of model evaluation."

Specific comments:

P2, L47-48: When was version 1 developed (published)? What are the major changes/new developments you are presenting here?

The first version of the PMP was released in 2015 (https://doi.org/10.5281/zenodo.13673), and we have included this information as a part of the Introduction section in the revised manuscript.

To respond to the reviewer's concern for the lengthy paper from another comment, we have decided to focus more on current and future parts than the history part. Therefore, we have revised the text as follows. "In this paper, we provide an overview of the PMP including its latest capabilities, and discuss its future direction."

P3, L82: Also here you should clearly state when the first version was developed. Has this version been somewhere documented/published? Or has this version just been provided to the CMIP community?

This information has been described in the Code and Data Availability section. To further clarify, we have made the following revisions.

The original text, "To respond to the need, PCMDI has developed the PCMDI Metrics Package (PMP), to quantitatively synthesize results from the archive of CMIP simulations via performance metrics that help characterize the overall agreement between models and observations (Gleckler et al., 2016)" was citing Gleckler et al., 2016 for the first version of the PMP development.

To further clarify, we revised the sentence as follows: "To respond to the need, PCMDI developed the PCMDI Metrics Package (PMP) **and released its first version in 2015 (see Code and Data Availability section for all versions)**. A centralizing goal of the PMP then and now is to quantitatively synthesize results from the archive of CMIP simulations via performance metrics that help characterize the overall agreement between models and observations (Gleckler et al., 2016)."

We also added the following description in the Code and Data Availability section.

"PMP is available as an open-source Python package with **all released** versions archived on Zenodo DOI: https://doi.org/10.5281/zenodo.592790"

P4, L86: Here, some examples should be given. Which performance metrics or statistical measures are used?

To respond to the reviewer's comment, we have added the following description. "Common examples include a domain average bias, a root-mean-square error (RMSE), a spatial pattern correlation, or others, typically selected depending on the application."

P4, L101: Also here, add some examples.

Statistical measures that can be used as metrics are diverse depending on the climate characteristics being evaluated, which we therefore prefer to explain in each subsection of Section 3. Since we have now provided some examples in response to the comment above, we have revised the description as "via well-established statistics **those discussed in Section 3**".

P4, L102: Why only a subset? Specify.

The collection of CMIP experiments includes simulations for pre-industrial and historical periods, future projections with different scenarios, and many others. There are only a few CMIP experiments that are particularly well suited for comparison to recent observations. The most important examples of model evaluation via comparison with observations include the recent historical period as simulated in the CMIP Historical and AMIP experiments. To help clarify the above, we have rewritten the description as follows: "A subset of CMIP experiments, those conducted using the observation forcings such as "Historical" and "AMIP" (Eyring et al., 2016), is particularly well-suited for comparing models with observations." This addition is followed by detailed description for the Historical and AMIP experiments of CMIP.

P4, L102: What do you mean with CMIP class models? Models that participate in CMIP or the kind of models participating in CMIP?

To improve the clarity, we have revised "CMIP-class models" to "CMIP-participating models".

P4, L109: References? Are there any publications?

Yes, and these publications are referenced in each subsection of section 3. However, we also have now also added a few representative publications here to respond to the reviewer's comment.

P6, L170: The following subsections are rather the "processes" that can be assessed with PMP than metrics. The statistics you are using are the metrics for each of this processes.

We understand and appreciate what you have pointed out here, however, after considerable thought we prefer to retain our description. The reason for this is that it is not uncommon to use the word "process" for "process-oriented metrics" in this area of research, and thus to avoid any confusion with it, we would prefer to keep the description as is.

P6, L174: "well-established statistics" -> You should more clearly write here which statistics are used (at least some examples should be given).

We have added the following description: "such as RMSE, mean absolute error (MAE), and pattern correlation"

P6, L178: Provide here a typical example. What is a typical default model and what are the alternatives.

We have added "(e.g., see Table 1)" to clarify the "default" reference datasets.

P6, L189-191: Provide a short explanation how this is done or where this is described.

We have added the following description in the revised manuscript. "Detailed instructions can be found on the PMP's online documentation (http://pcmdi.github.io/pcmdi_metrics)."

P8, L237: Since "Performance", "Processes" and "Teleconnections" are not really metrics, I would suggest to rewrite the sentence as follows: "The ENSO metrics used to assess/evaluate the models are divided into three categories: ………..".

Thank you. Revised accordingly.

P16-18: Summary section is too lengthy and not really summarizing what has been presented in the manuscript. Some part of this should rather be part of the discussion section. Further, there should be a conclusion section e.g. stating clearly what is the gain for the community of this software package. Has it already successful applied for CMIP etc.

To clarify the take home messages, we have reorganized this section. Some discussion of the future directions were moved to the "Discussion" section, and the "Summary" section was renamed to the "Summary and Conclusion" section. We have revised text in both sections to clarify the message.

Figure 1 caption: Add in the figure caption what is shown in the boxes (thus, add that the separation of RMSE by season is shown there)

In the Fig.1 caption, we have added the following description. "The RMSE is calculated for each season (shown as triangles in each box)"

Technical corrections:

P4, L108: Add "assessment" or "model" after CMIP.

Revised accordingly.

P8, L247: Abbreviation "ITCZ" has not been introduced.

The full description, "intertropical convergence zone" is added. We also added appendix A that lists the acronyms used in this paper.

P9, L288: Abbreviation "DOE" has not been introduced.

"U.S. Department of Energy (DOE)" is added.

P11, L344: Abbreviation GoG has not been introduced. It is given in the figure caption of figure 7, but not in the text at L338: Should be done there, too.

"GoG" is replaced by "Gulf of Guinea".

P12, L401: Although the abbreviations "WGNE" and "WGCM" have been introduced, I would suggest to repeat it here.

We understand that it would be helpful to re-introduce WGNE and WGCM. However, they have been introduced earlier and the sentence includes multiple other abbreviations, so we decided to not re-introduce those abbreviations. Instead, to address your point we newly added a table of acronyms as an appendix A to help readers.

P12, L 466: Abbreviation "ESGF" has not been introduced.

Earth System Grid Federation (ESGF) was introduced in Section 2 on page 4, but again to improve the readability we have introduced the Table of acronyms.

P15, L493: …….Section 3.3 are respectively -> Section 3.3., respectively, are

Thank you for the suggestion. This sentence was removed in the revised manuscript during the process of condensing, consolidating, and simplifying the text in response to the reviewer's other comment.

P16, L542: use parenthesis instead of brackets

Revised accordingly.

P18, L605: check sentence.

The sentence was rewritten to be more concise and clear.

P41, Figure 6 caption, L1293: units should be given in an upright font.

The metric in Figure 6 is derived as a ratio of powers for eastward and westward wave propagation, thus the metric itself is unitless. To clarify this point, in the caption we indicated that the metric is a ratio and unitless.