Response to reviewer #1

In the manuscript titled “Evaluation of atmospheric rivers in reanalyses and climate models in a new metrics framework ”, the authors develop a new model evaluation package to systematically diagnose atmospheric river (AR) biases. The characteristic of this tool is the robust response to structural differences of AR detectors. So this tool is designed to intercompare the ARs as simulated by multiple climate models. There are still certain changes and clarifications that the authors should address prior to publication. For these reasons, I believe that the manuscript can be accepted for publication by the GMD after minor revision. Below, I have some general suggestions to the authors.

General comments:

The paper appears to be rushed, with inadequate detail and poor organization, and it seems that it was not carefully reviewed after completion. For instance, the introduction lacks systematic coverage and fails to logically present the structure of the package, which is contrary to the reader’s expectations. The crucial code section is that "full environment and python packages include AR metrics," it would be hard to find which section of PMP works for AR evaluation. The writing is not standardized, requiring readers to search backwards/proof/numbers for clarification. For instance, in Section 3, although the topic is metrics, specific numbers (metrics results) are not mentioned frequently, leading to a general lack of data support for qualitative analysis. Readers have to compare data themselves. Therefore, it is recommended to revise the writing based on similar highly-cited papers in AR evaluation.

Thank you for your review, comments and suggestions! We are taking the time to reorganize the manuscript with your feedback, including rewriting targeted parts of the manuscript. A new section describing the metrics workflow and code structure is also added, and some technical discussion is being moved to an appendix. Relevant figures are also being revised as suggested. Please see our responses below.

Specific comments:

Line #93, the introduction of the paper should not include information just for the sake of writing an introduction. This statement is not an actual argument; it is unnecessary and does not require citation of references to support it. This entire section covers very basic common knowledge and should be removed. The metrics are extremely common and there's no need to list them.

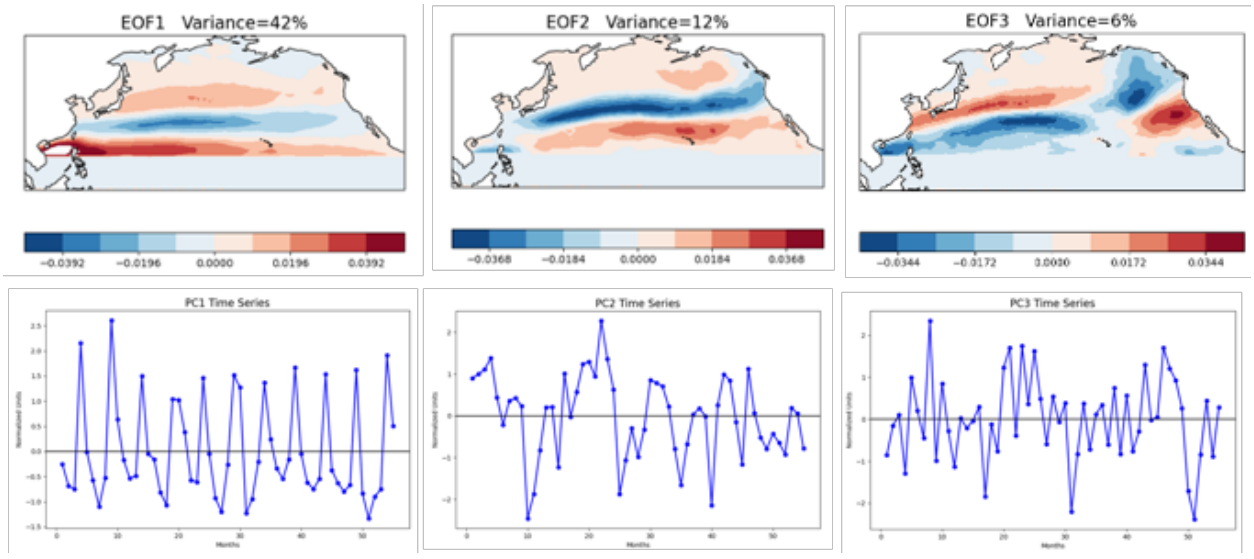
Thank you for your comments. Following your suggestion, we have deleted this section.

Line #144-145, the selected model (E3SM-HR, E3SM-LR) should include some basic tabulated information about its parameters, including grid resolution.
Thank you for your suggestion. We have added a table in the supplementary material detailing the E3SM-HR and -LR grid parameters.

Line #178, considering the target audience of the paper and the need for conciseness in scientific writing, the section introducing the computational methods should not be overly detailed. Lines 178 to 196 should be summarized in a few sentences. The same issue also appears in the pattern correlation section (lines #205-#214).
We have simplified this part by moving the statistical formula into the Appendix, leaving a concise description of the methods in the main text.

Line #235, could the author list the percentage next to this range? The proportion of the pattern variance that the principal components can explain should be quite low for 95%, such as first or second PC, etc. Additionally, within what range was this data calculated?

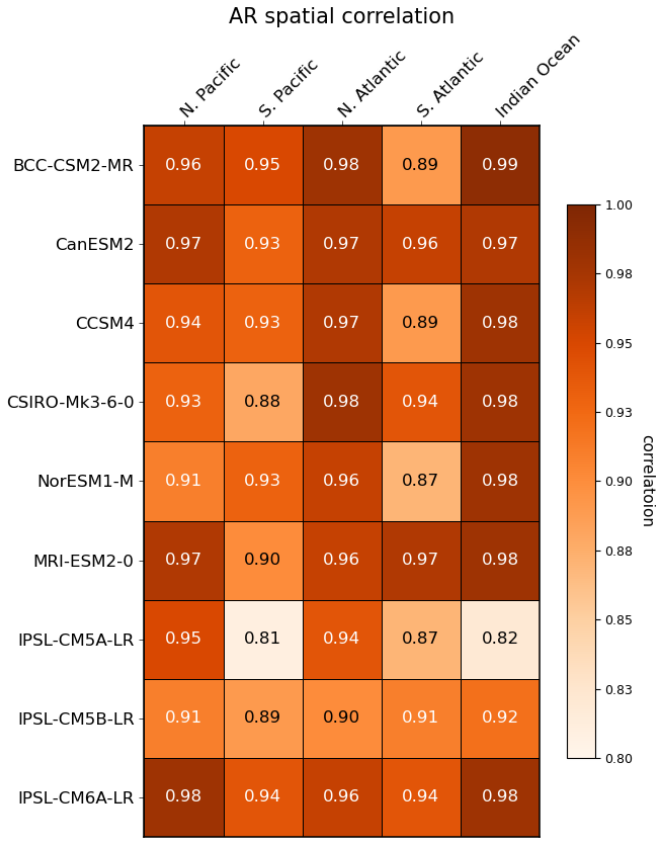
Thank you for your comments. As in our effective sample size (*Ne*) definition equation in section 2.3.2, *Ne* is the number of PCs (counting from the 1st PC to the Nth PC) such that the total variance sums to > 95% of the total pattern variance, and thus the range within which the data is calculated is from 0-95%. For example, Ne is 16 for North Pacific, which means the first 16 PCs account for more than 95% of the total variance. Therefore, the percentage (95%) is indeed implicitly linked with these numbers. Although the spatiotemporal variability of ARs is out of the scope of the current paper, in response to the reviewer's comment we analyzed the PCA pattern variance breakdown for North Pacific.



The 1st PC is the seasonal shift pattern (43% of the total variance), the 2nd PC is a year-to-year variability pattern (12% total variance), and the 3rd PC is a multi-year variability pattern (6% of total variance). The rest of the PCs account for weaker modes of spatial variability with each explaining a small amount (< 5%) of the total variance. It takes until the 16th PC for the accumulated variance to exceed 95%.

Line #272, what is the frequency of the data on which this calculation is based? Monthly?
Thank you for pointing this out! The data are all 6-hourly, as now clarified in the revised manuscript's data section.

Figure 2, the color scale intensity and the magnitude of the numbers are inversely related. I suggest adjusting the scale to consistently increase. Unless all values pass the significance test and are noted in the figure caption, additional markers, such as an asterisk next to the numbers, should be added to indicate significance test results.
Thank you for your comment. We have confirmed that all values in Fig. 2 are statistically significant. To clarify this, we added "the correlations are statistically significant for all models and regions" in the revised manuscript text.
Regarding the plot, we have made the changes per your suggestion as darker block showing stronger correlation.



Line #276, Fig.2 needs to be described, such as what percentage of models in a specific ocean area have correlations that pass the significance test, using numbers to support the qualitative description.

Thank you for your comments. Indeed, all the numbers in Fig. 2 have passed the significance test. We have included “the correlations are statistically significant for all models and regions” in the text.

Figure 3, this figure is missing subplot labels like (a), (b), etc. The scale for the difference should be placed at the bottom to avoid confusion.

Thank you for your comments. Panel labels are now included in the figures. We have moved the scale to the bottom of the plot.

Line #280, the spatial pattern correlation in S. Pacific is 0.88, and N. Atlantic is 0.98, these numbers in Fig.2 and the spatial gradient in Fig.3 should be discussed to support this statement. Additionally, just a suggestion—why not choose graphs with larger differences for comparison, such as BCC (0.99) and IPSL-CM5A (0.82) in the Indian Ocean? The comparison would be more intuitive given the same study area. Line #290 could be updated into “be better interpreted together with AR frequency maps with spatial gradient”.

Thank you for your comments and suggestions. We updated the text as you suggested, and added numbers and relevant discussions in the text as “The high spatial correlation (e.g., in Fig. 3, r = 0.88 in S. Pacific and r = 0.98 in N. Atlantic) is mainly a result of the similar spatial gradient (as in Fig 3a-b, and Fig 3d-e)”

Regarding the choice of models with larger differences for comparison, we agree this is a very good suggestion. Consequently, we compare BCC (0.99) and IPSL-CM5A (0.82) in the Indian Ocean. New figure panels have been added to Fig. 3.

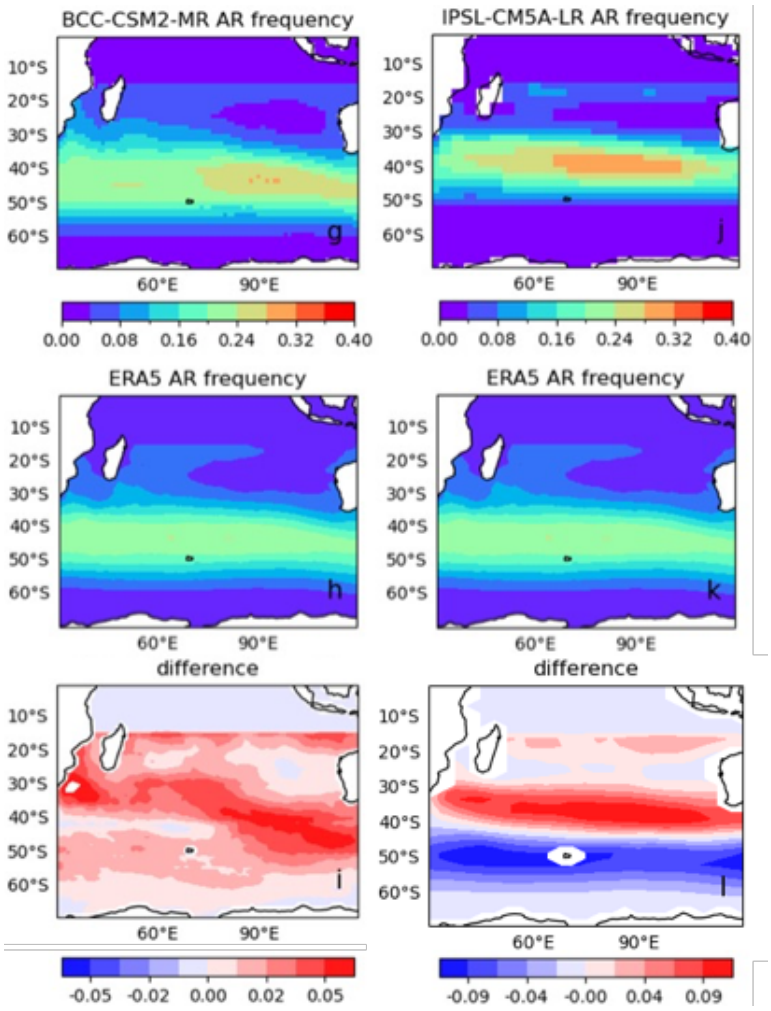
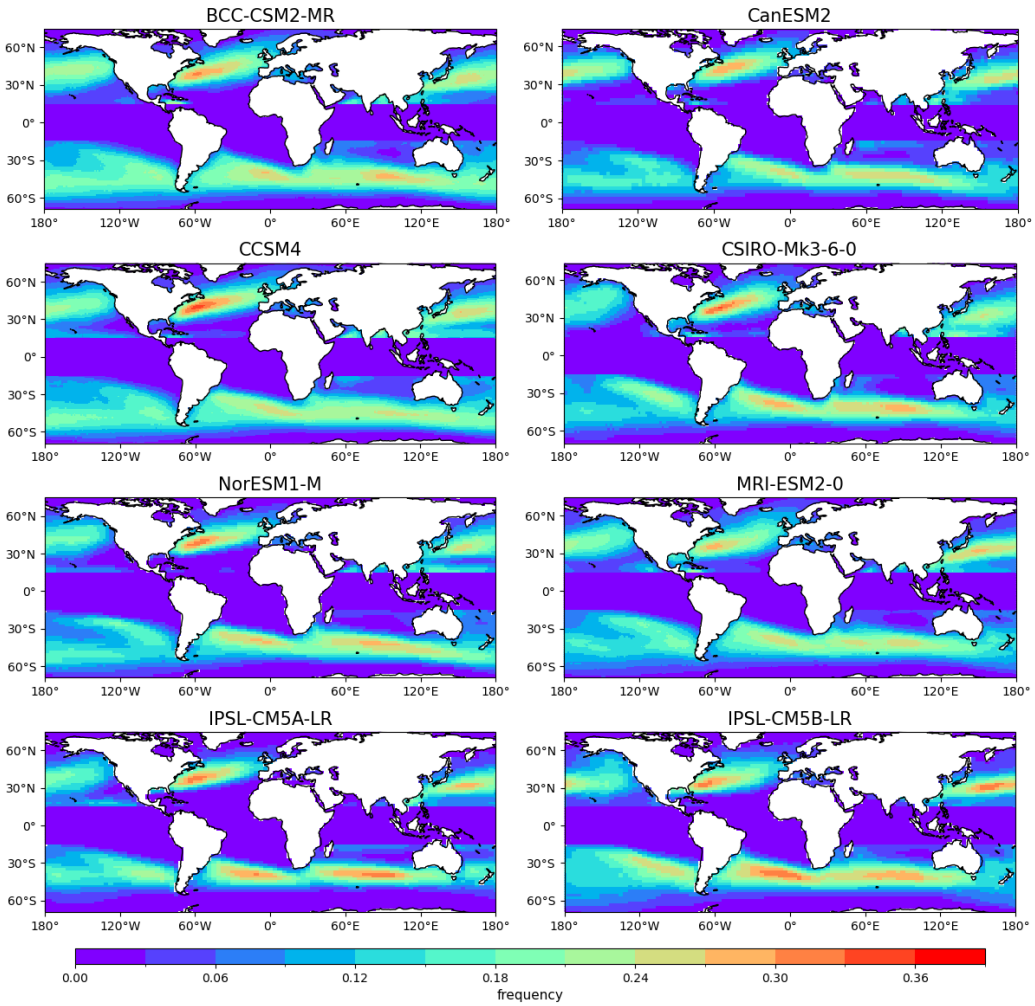


Fig. 3 is updated accordingly in the revised manuscript. The following discussion has also been added to the main text: “Another example is the AR frequency distribution over the Indian Ocean for BCC-CSM-MR (Fig. 3g-i) and IPSL-CM5A-LR (Fig. 3j-l) models. Even though, when compared to ERA5, both models show significant spatial correlation in Fig. 2 (r=0.99 and r=0.82 respectively), the spatial bias pattern in IPSL-CM5A-LR exhibits a more apparent latitudinal shift than in BCC-CSM-MR.”

Figure 4, the hatching obscures the colors; it's recommended to place asterisks next to the numbers or bold the numbers to replace the hatching. This is just a suggestion, in Figure 4, such comparisons could benefit from providing a global ocean basin average shape for each model, which would make deviations in latitude and longitude more intuitive.

Thank you for your comments and suggestions. To prevent the hatching obscuring the colors, we revised the plot with reduced density of the hatching. After various visualization trials, we decided to use hatching over placement of asterisks for sake of figure consistency. We’ve also made a plot showing the global ocean basin average shape for each model as below, which has been added to the supplementary documentation of the revised manuscript.



It is a suggestion only. Regarding Fig. S1, why not interpolate or downscale to the same resolution before comparison? It will still prove the difference in original data resolution. Different grid resolutions will inevitably introduce boundary issues. Fig. S1, the default color scheme makes the land boundaries unclear and needs to be adjusted. Fig. S2, the specific meanings of ARCONNECT and TECA are unclear.

Thank you for your suggestion. We interpolated/downscaled the AR tags to the same resolution as ERA5 for comparison only when it was needed, e.g., for calculating spatial correlation. Other metrics can be calculated directly with the models' original resolution. We agree that comparisons are more intuitive when the data are on the same grid, however, interpolation can have subtle influence on the conclusions (e.g., Ullrich and Zarzycki, 2017). Since metrics are normally computed by end users on the native grid, it is preferable to use direct model output rather than interpolating the models before their use. So, when the metric is based on domain averaged quantities, we did not interpolate model data before calculating the metric.

For figure S1, the full name of ARCONNECT and TECA are added to the figure caption as "ARCONNECT (Atmospheric River-CONNected objECT; Shearer et al. 2020), (b) TECA-BARD (Toolkit for Extreme Climate Analysis; O'Brien, Risser, et al., 2020),". The coastline boundary color is changed for improving land-sea contrast as suggested.

Figure 5, please consider that those with red-green color blindness may have difficulty distinguishing between these two colors.
Thank you for your suggestion. We have changed the color scheme and updated the figure.

Line #325-#330, could you provide some explanation in this section of the results? For example, the characteristics of the models?
Thank you for your comments. We have added the following text to the end of the paragraph: "These differences may arise from various different characteristics of the models, such as dynamical core (e.g., finite volume in CCSM4, T63 triangular spectral truncation in CanESM2, spectral-transform in ERA5), grid resolution (see supplementary Table S1), and the role of data assimilation (Buizza et al. 2018) in the ERA5 system."

Line #379-#383, Here, it would be better to explain how the different thresholds for tagging the moisture field contribute to the differences in AR shapes.
Thank you for your suggestion. We have added the following text to the paragraph: "Such a difference is attributable to the different thresholds for tagging the moisture field in the two ARDTs. The results presented here are obtained from the default criteria, i.e. in TE, ARs are tagged when the Laplacian of the IVT <= -20000, while Mundhenk uses a static 250 kg m^{-1} s^{-1} threshold on the IVT field. We might expect different results by altering these threshold numbers."

Language issues:

Line #78, the abbreviation ARDTs should be defined the first time it appears.
Revised as suggested. ARDT defined in line 38.

Line #120, the abbreviation ARDT should be defined the first time it appears.

Revised as suggested

Line #124, the TE ARDT should be defined the first time it appears?

Revised as suggested

Line #130, “Mundhenk_v3 tags” could be replaced with “fixed-relative (Mundhenk_v3 tags)”
Revised as suggested

Line #136, in general, the full name of a climate model abbreviation should be provided the first time it appears.
We’ve added a table in the supplementary material that includes the modeling center, model names and model resolutions.