Response to Reviewer 2's comment

This manuscript presents a framework and results from a new benchmarking tool for decadal prediction experiments that is based on the PCMDI metrics package. The new benchmarking tool fills a critical gap in the ecosystem of standardized benchmarking software by being engineered to handle aspects of prediction datasets, such as forecast lead time. With this tool, there is a significant opportunity to understand how initialization impacts seasonal prediction, as well as how quickly long-term mean biases in climate models emerge in shorter decadal-scale experiments. The initial application of the tool to DCPP experiments shows several interesting themes. Long-term mean biases in temperature and precipitation tend to emerge during the decadal simulations for the three primary climate variables considered in this study: surface air temperature, precipitation, and sea ice extent. Trend biases also emerge in decadal predictions in some key processes, including wintertime mid-latitude precipitation in the Northern Hemisphere. The decadal prediction models also struggle to represent seasonal trends in Antarctic sea ice extent, but the results of the benchmarking analysis suggest that biases are reduced at shorter timescales, benefiting from the model initialization. Overall, the results suggest that much more work is needed to improve the fidelity of precipitation simulation. The manuscript is well written, and the results demonstrate the novelty and utility of this new benchmarking approach. After some revision, this manuscript and benchmarking tools will be important contributions to the field.

We appreciate the reviewer's time and thoughtful comments. Please find our point-to-point responses in the following text.

Major Comments:

1. The manuscript is a bit light in some key areas for context. The second-to-last paragraph of the introduction falls short of demonstrating what software and approaches already exist, and why this approach is novel. I would recommend a more thorough review of past efforts and existing tools to better frame the new work. Similarly, benchmarking diagnostics differ distinctly from process-based diagnostics, where the latter are more suited towards understanding why the models perform the way that they do. This is an important caveat that needs to be covered in both the introduction and the conclusions. Benchmarking diagnostics are sometimes suggestive of possible reasons for biases in climate variables, but are not definitive.

:We agree with the reviewer's point. We have modified the Introduction as follows:

(Lines 46–60:)

Over the past decades, the climate research community has developed a wide range of diagnostic frameworks and standardized metrics to evaluate long-term climate simulations, particularly for the CMIP historical experiments. These efforts include benchmarking-oriented diagnostics that quantify systematic model performance using standardized metrics and intercomparison frameworks (e.g., Gleckler et al., 2008; Eyring et al., 2020) and process-based diagnostics that investigate the physical mechanisms underlying model biases and variability (e.g., Maloney et al. 2019; Planton et al. 2021). Several community tools have been developed as reusable software packages, including the Program for Climate Model Diagnosis and Intercomparison (PCMDI) Metrics Package (PMP; Lee et al., 2014), the Earth System Model Evaluation Tool (ESMValTool; Eyring et al., 2020), and related CMIP diagnostics frameworks (Maher et al., 2025; Hassler et al., 2026).

Benchmarking diagnostics provide a standardized and reproducible framework for comparing model performance. However, most existing evaluation frameworks primarily target uninitialized historical simulations and long-term climate projections. These frameworks' current capabilities do not fully address the unique characteristics of initialized decadal predictions, where the forecast drifts, the time-evolving biases, and the lead-time-dependent prediction skill require their own customized and specialized evaluation strategies. To address this gap, we propose a new evaluation framework that could assess multi-model initialized decadal prediction systems in a systematic and comprehensive way, leveraging the established community tool for the CMIP evaluation, using some of the PMP's approach and capabilities.

2. The introduction mentions the importance of model drift in the introduction (third-to-last paragraph), but it is largely ignored throughout the rest of the manuscript. The methods section would benefit from a more in-depth discussion of model drifts in decadal simulations and how the benchmarking tool handles them.

:We thank the reviewer for pointing out the need for more description of forecast drift and bias treatment in the original manuscript. In the original analysis, lead-time-dependent mean climatological biases were removed before calculating metrics, following standard decadal prediction verification practices (Goddard et al. 2013). We improved the manuscript to include further details of the methodology. In the revised manuscript, this clarification is added to Section 2.3. We also expanded the discussion of model drift and biases in the Discussion section as follows:

(Lines 119–126:) 2.3 Measure of prediction skill and spread

Initialized decadal predictions are known to exhibit systematic forecast drift as the simulations diverge from the observation-constrained initial state and approach the model's preferred climatology. Following the standard procedure used in decadal prediction studies (Goddard et al. 2013; Choi and Son 2022), the mean climatological bias for the period 1981–2010 is removed for each lead month before calculating the metrics.

In contrast, long-term trend biases are intentionally retained in the current framework. Unlike the mean climatological drift, the trend biases may evolve nonlinearly in lead time, reflecting the complex interactions between the externally forced signals, internally generated variability, and model adjustment processes. Rather than correcting these effects explicitly, this study compares evaluation metrics between initialized decadal hindcasts and uninitialized historical simulations.

(In the Discussion, Lines 414–429)

Applying this portrait-based framework reveals several robust characteristics of decadal climate predictions. Surface air temperature biases exhibit substantial inter-model spread and a pronounced evolution with lead time, indicating a clear drift toward each model's biased climatology despite initialization. This gradual evolution likely reflects the influence of slowly varying processes associated with climate feedbacks, ocean memory, and model drift toward their preferred coupled equilibrium states. In contrast, precipitation biases show a more consistent and systematic spatial pattern across models. The relatively stable structure of these biases suggests that precipitation errors are primarily driven by higher-frequency atmospheric variability and localized convective processes that establish systematic biases early in the forecast period. This uniform structure implies common structural deficiencies in model physics, such as convection, cloud processes, and land–atmosphere coupling.

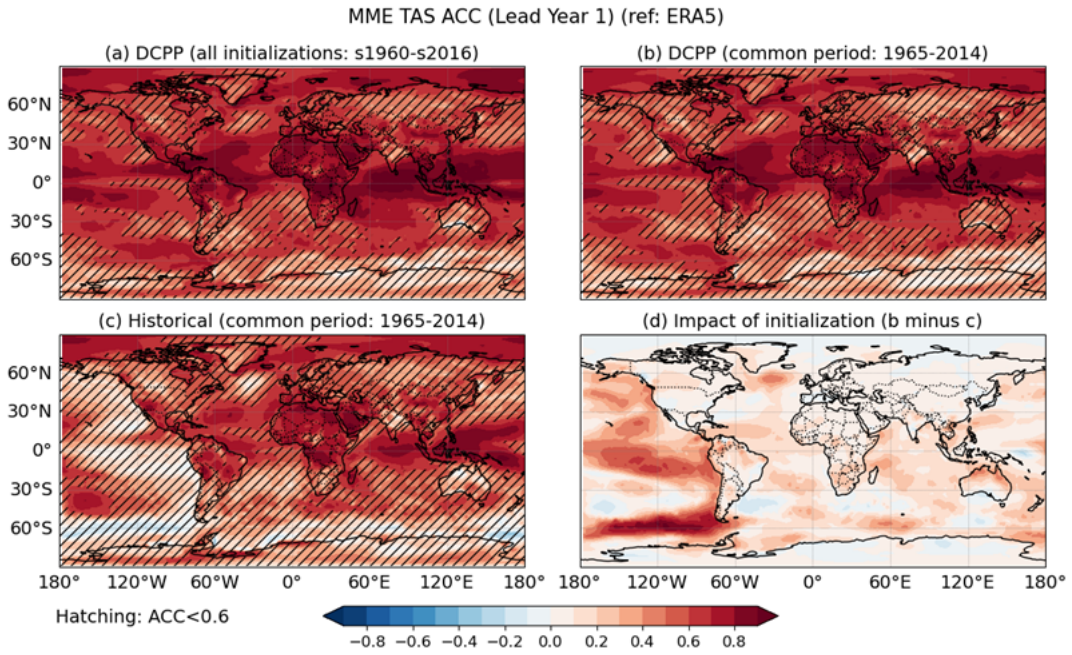
Trend-bias diagnostics also reveal systematic deficiencies in representing long-term climate changes regionally. Notably, the persistent underestimation of observed precipitation trends in the Northern Hemisphere mid-latitudes, even at short lead times, suggests that important mechanisms governing regional hydroclimate change may be inadequately represented in current prediction systems. These results underscore the importance of evaluating both mean-state biases and long-term trend accuracy when assessing decadal climate predictions. Finally, the results highlight the strong coupling between temperature and sea-ice predictions. Inter-model differences in sea ice extent bias generally scale with temperature bias, which emphasizes the role of polar feedback processes.

3. Statistical significance is largely absent from the benchmarking metrics. The benchmarking tool would benefit greatly by incorporating measures of statistical significance, for example, by shading/stippling on portrait plots and maps.

:We thank the reviewer for this valuable suggestion. We agree that incorporating statistical significance measures would strengthen the interpretability of the benchmarking diagnostics further. However, as the current portrait plots display regionally averaged grid-point metrics, it would not be very straightforward to show different statistical significance for each grid cell. Considering the current framework primarily focuses on providing a standardized, scalable visualization platform that compares lead-time-dependent biases and skill metrics across multiple initialized prediction systems using PMP's existing capabilities, we have decided to keep the portrait plots as they are. As an alternative, to incorporate the reviewer's point, we added hatching to the multi-model ensemble ACC maps in Fig. 10 to indicate regions with relatively low practical prediction skill (ACC < 0.6). This is a commonly used empirical threshold in climate prediction studies. In the revised manuscript, we have added a discussion about the importance of

incorporating more rigorous statistical significance diagnostics in the future development of the framework.

(In the last paragraph, Lines 439–440) Future extensions of the framework should incorporate statistical significance diagnostics and uncertainty quantification for lead-time-dependent skill metrics and trend analyses.



Revised Figure 10: This example figure illustrates the spatial distribution of the MME at LY1 of TAS ACC. (a) ACC calculated from all available DCPP experiments initialized from 1960 to 2016. (b and c) Same as (a) but for the common evaluation period of 1965–2014 from DCPP and HIST simulations, respectively. Regions with ACC below 0.6 are hatched to indicate limited practical prediction skill. (d) The difference between (b) and (c) represents the added skill due to initialization. Red regions indicate areas where initialization improves prediction skill.

Minor Comments:

Lines 104-105: The regional approach is fine, but the comment about regridding errors is curious. Ocean model grid remapping is not an intractable problem.

:We thank the reviewer for the comment. We agree that ocean grid remapping is a well-established procedure and did not intend to suggest that it is intractable. We have revised the

sentence to clarify that the regional averaging approach was primarily adopted to enable consistent comparison across heterogeneous native ocean grids without requiring additional remapping procedures as follows:

(Lines 110–112:) This regional averaging approach further enables consistent comparison across heterogeneous native ocean grids without requiring an additional remapping procedure.

Section 2.4: This section on the display interface should be expanded. More details on how it is done, why it is important, and the design choices made would be helpful.

: Following to the reviewer's point, we have expanded Section 2.4 in the revised manuscript as follows:

(Lines 154–166:) 2.4 Interactive visualization

Unlike uninitialized simulations (e.g., historical experiments), initialized decadal predictions exhibit evolution in model biases and prediction skill that depends on lead time. Consequently, evaluation outputs are substantially larger and more multidimensional, considering lead times, models, and metrics simultaneously. Therefore, static figures alone are not very efficient to explore and interpret the full evaluation results.

To address this challenge, in this study, we have developed an HTML-based interactive visualization framework following the PMP approach. We used the Python library Bokeh (https://bokeh.org/) to enable flexible exploration of the benchmarking diagnostics. The interface allows users to dynamically compare models, lead times, and diagnostic metrics through interactive selection and navigation tools. With this design, researchers can rapidly detect systematic model behaviors, lead-time drift, and differences between models—insights that are tough to catch using conventional static plots.

The framework supports reproducible and scalable evaluation workflows for large multi-model prediction archives such as CMIP6 DCPP. By leveraging browser-based interactive graphics, the visualization outputs can be easily shared, archived, and extended to future prediction systems and additional diagnostic metrics. All interactive figures are available at https://pcmdi.llnl.gov/metrics/dcpp (last access: 18 May 2026).

Figures 1,2,5,6,9,12,13: The “LY” labels might be better suited along the bottom of the plot. They are a bit lost with the plot titles.

:We appreciate the reviewer's suggestion to position the "LY" labels at the bottom. While this would be ideal for standard static layouts, these visualizations were optimized for a scrolling web-interface where top-alignment ensures the labels are immediately visible. To maintain consistency with the live web tool, we have opted to leave the figures as they are. Acknowledging the reviewer’s point, we have added the following note to Fig. 1 caption to clarify this web-interface design choice for the reader.

“Note that labels are top-aligned to optimize visibility within the scrolling web interface.”

Figure 10: It is difficult to see the yellow contours in these panels.

: Thank you for pointing it out. We changed the contours to hatching to indicate regions with relatively low practical prediction skill (ACC < 0.6), which is a commonly used empirical threshold in climate prediction studies.