Response to Reviewer 1's comment

This manuscript presents a multi-model evaluation framework for initialized decadal climate prediction, implemented within the PCMDI Metrics Package, featuring two diagnostic tools: (1) a model-by-lead-time portrait plot and (2) an interactive HTML visualization platform, to assess how biases and prediction skill evolve with forecast lead time across temperature, precipitation, and sea ice. Application of the framework reveals that temperature biases drift toward each model's climatology over time, precipitation biases reflect systematic model physics errors, and sea-ice skill degrades rapidly with lead time while remaining closely coupled to temperature prediction quality. Overall, this work offers great practical value to the decadal predictability community; the publicly accessible diagnostics provide a useful reference that researchers can readily draw on for their own work. It is completely understandable that a single manuscript could not provide detailed mechanistic explanations for every systematic bias identified, and I expect this framework will serve as a springboard for many future studies in this area.

We thank the reviewer's time for reviewing this paper and sharing the invaluable perspective. We are glad that the importance of the framework was acknowledged. Please find our point-to-point responses in the following text.

Minor comments:

1. The authors state in Lines 149–150 that forecasts are expected to drift toward the model's biased climatology as lead time increases. However, Figure 1 shows that tropical TAS biases in most models actually decrease with lead time rather than grow, which is somewhat counterintuitive, since one would expect small biases at short lead times that would then amplify as the forecast drifts toward the model climatology. Could the authors comment on this?

:We thank the reviewer for this insightful comment. We have added a brief discussion in the revised manuscript to clarify that forecast drift does not always imply monotonically increasing bias amplitude as follows:

(Lines 176–187:)

Although some models exhibit clear lead-time-dependent bias growth (e.g., polar regions in EC-Earth3 and HadGEM3-GC31-MM), the evolution of forecast bias is not always visually pronounced in Fig. 1. This partly reflects the substantial inter-model spread represented using a common color scale, but also the rapid adjustment of initialized forecasts toward the model climatology, particularly in the ocean component, where mean biases can develop within the first several months of integration (Ma et al. 2021).

As lead time increases, initialized forecasts generally tend to drift from the observed initial state toward the model's preferred coupled equilibrium state. However, the evolution of forecast bias and skill is not necessarily monotonic, as it can be affected by initialization

shocks and adjustment processes during the forecast period (e.g., Mulholland et al. 2015). Additionally, coupled ocean–atmosphere feedbacks and internally generated ocean variability may partially compensate for inherited model biases. Consequently, some models exhibit even decreasing biases with lead time, especially for tropical TAS (e.g., CanESM5, CMCC-CM2-SR5, FGOALS-f3-L in Fig. 1a), rather than the monotonic growth expected from climatological drift alone.

2. Line 160: It would be helpful to include a brief discussion of the physical reasons behind the tendency for models to exhibit a wet bias in the tropics and a dry bias in the mid-latitudes, rather than relying solely on the brief attribution to the summer ITCZ at Line 196.

:We appreciate the reviewer's suggestion. We have added a discussion of the physical mechanisms and references in the revised manuscript as follows:

(Lines 190–197:) Instead, Fig. 1b showcases more of the known pattern of systematic precipitation biases in climate models than the substantial lead-time-dependent variations. Most models exhibit wet biases in the tropics and dry biases in the mid-latitudes. Tropical wet biases are commonly associated with the excessive double Intertropical Convergence Zone (ITCZ), which has been linked to multiple factors, including deficiencies in dynamical circulation and thermodynamic processes (Zhang et al., 2015; Samanta et al., 2019). In contrast, dry biases in the mid-latitudes are often related to land–atmosphere feedbacks associated with summer warm biases (Mueller and Seneviratne, 2014; Lin et al., 2017), as well as atmospheric circulation biases associated with storm-track position and intensity (Priestley et al., 2020; Schemm, 2023).

3. Figure 1: Comparing TAS and PR results, TAS biases appear to evolve with lead time while PR biases remain largely constant across all initializations. A short comment on why precipitation biases are more stable with lead time than temperature biases would be great.

:We thank the reviewer for this helpful observation. We have added a discussion to clarify this distinction in the revised manuscript as follows:

(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.