We thank the reviewer for his/her helpful comments. In this document, we answer each point raised by the reviewer. The original reviewer's comments are given in **black**, our answers in **blue**. A "track changes" version of the revised manuscript highlighting all changes is available.

Reviewer #2

In the paper, the authors introduce ICONEval, an open-source evaluation framework built on top of the Earth System Model Evaluation Tool (ESMValTool) designed for rapid, reproducible, and automated performance monitoring during the development of Earth System Models (ESMs). ICONEval introduces parallel processing and HTML summary reports categorized into three diagnostic pillars: sanity checks against observational bounds, fundamental physical consistency checks (e.g., mass conservation), and multi-domain climate fidelity diagnostics. The tool is demonstrated using the ICON-XPP model and is positioned as a highly efficient benchmarking utility for comparing model configurations, testing updated parameterizations, and assessing emerging hybrid machine learning-enhanced climate models.
This manuscript is well-suited for the journal's scope, as it introduces an open-source evaluation framework (ICONEval) that directly advances our ability to benchmark, monitor, and improve ESMs, including emerging hybrid machine learning configurations. The paper is well-prepared, presenting a highly structured and logical three-pillar approach-spanning sanity checks, physical consistency, and climate fidelity-that is clearly articulated and thoroughly demonstrated using the ICON-XPP model. Given the increasing complexity of modern climate modeling, the development of reproducible, automated, and customized tools for specific models is highly relevant to the journal's audience and represents a timely, high-quality contribution to the community.
I suggest accepting the manuscript subject to the authors addressing the following a few specific comments.

We also thank Reviewer #2 for helping us to improve the manuscript.

Specific comments

Line 145 and Figure 2: Still, it is not very clear to me that how the "reasonable" upper and lower limits (red lines in the figure) were defined, in particular with using multiple reference datasets. For example, for minimum, was the minimum monthly value across all years of global spatial average taken from each reference dataset, then minimum value among them was taken as the "reasonable minimum lower limit?" I suggest further clarifying it in the manuscript.

The reviewer is right. When multiple reference datasets are available, the minimum and maximum global monthly mean values are obtained as the minimum and maximum values across all global values. We clarified this in the manuscript as follows:

> ""Reasonable" upper and lower limits for global means are defined using the global minimum and maximum monthly values across all available years of multiple observational and reanalysis datasets. This is equivalent to the lowest and highest monthly global average value observed in any of the reference datasets."

Figure 8a: While I understand there could be some artifact due to difference in the data frequency interval (i.e., ERA5 1 hourly while ICON-XPP 6 hourly), the sudden magnitude change over the pacific seems odd, which could have been resulted from incorrect local time handling. Could you please double check on this?

We agree with the reviewer that particularly the stripe over the Pacific Ocean looks strange. We double-checked the calculation and believe everything is correct. When converting UTC to local solar time, the temporal frequency of the input data is preserved. For example, hourly input data will be transformed into hourly output data. For this, a location's exact LST will be put into corresponding bins defined by the bounds of the input time coordinate (in this example the bin size is 1 hour). This means that the lower the time resolution of the input data, the larger the resulting errors. The errors can reach up to time frequency / 2 explaining the bands visible in figure 8. For this reason, a time resolution of 1 hour or better is recommended for this diagnostic. We added the following clarification to the text:

> "When converting from UTC to local solar time (LST), the temporal frequency of the input data is kept. A location's exact LST will be put into corresponding bins defined by the bounds of the input time coordinate. This means that the lower the time resolution of the input data, the larger the resulting errors that can reach up to time frequency / 2. It is therefore recommended to use this diagnostic with hourly or higher temporal resolution."

Figure 9. In addition to showing the difference field only, I suggest showing the full field from each data (ICON-XPP and ERA5) as well, so the pattern could be easily compared, like in Figure 10.

Following the suggestion of the reviewer we extended figure 9. The revised version now also contains the full fields from ICON-XPP and ERA5 in addition to the differences between the two.

Figure 11. In the Historical experiment, year-to-year comparison between model and observation does not say much, as their matching is not even expected. Although the authors have described this in the manuscript (line 275-276), I wonder if it would be beneficial to re-emphasize this point in the figure caption so advise readers what to focus on the figure. One additional option could be comparing the standard deviation or variance values obtained from each time series as an indicator for the intensity of the variability, or trend line. The same comment applies to Figs. 12 and 14 as well. For Fig. 11, I am not sure how much useful information could be derived from the CMIP6 results in the background.

We added the following reminder to the caption of figure 11:

> "Historical ESM simulations are not expected to replicate the timing of distinct observed ENSO events, but should have a comparable amplitude and frequency of the interannual variability."

Following the suggestion of the reviewer, we added the temporal standard deviations of the examples model (ICON-XPP) and the reference dataset (ESACCI-SST) to the figure as a measure of the variability.

We would prefer to keep the CMIP6 models in the figure to provide an envelope for the amplitude range that similar models show. This helps to quickly assess with this consistency check whether potential deviations from the observed amplitude are within the typical range of other ESMs. We clarified this in the manuscript by adding the following paragraph:

> "The ensemble of CMIP6 models included in this figure provide an envelope for possible anomaly amplitudes that can be expected from current ESMs. As a measure for the strength of the internal variability, the temporal standard deviations of the example model and the reference dataset are given in the top of the panel."

Similarly, we also added the temporal standard deviation to figures 12 and 14 and a reminder that historical ESM simulations are not expected to reproduce distinct events to the figure captions.

We think this does not apply to the mean seasonal cycles shown in figure 13, which therefore remains unchanged.