We thank Jaydeep Singh for his 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 #1

This manuscript by Lauer et al., entitled "An ESMValTool-based framework for sanity checks, physical consistency and climate fidelity during model development – ICONEval v1.0," presents ICONEval, an open-source wrapper around ESMValTool that can perform three sets of checks of model-development: sanity checks, physical-consistency checks, and climate-fidelity diagnostics. The capabilities of the developed tool, are demonstrated on a single ICON-XPP historical simulation. The work addresses a real and timely need of efficient and reproducible evaluation tool for model development. Also, this tool is designed to support emerging applications such as ML-enhanced ESMs and CMIP7. In its current form, however, the manuscript has a few weaknesses that should be resolved before publication. For these reasons, I recommend major revision, and the following are the comments.

We thank Reviewer #1 for providing helpful comments to improve the manuscript.

Major Comments

1. The Introduction provides a general motivation for model evaluation, but it does not sufficiently explain the specific need for ICONEval or its focus on ICON/ICON-XPP. The manuscript moves quickly from a broad discussion of ESM development and ESMValTool to the presentation of ICONEval, without clearly identifying the gap in existing evaluation workflows that this framework is intended to address. The authors should better explain why ICON-XPP requires a dedicated evaluation framework, for example, in relation to its role in CMIP7, its coupled configuration, its development workflow, or the practical challenges of rapidly assessing new model versions and parameterizations. In addition, the three central categories, sanity checks, physical consistency checks, and climate-fidelity diagnostics, can be defined more clearly in the Introduction. Some of the descriptive material currently in Sections 2.1 and 2.2 could be drawn forward to motivate the framework, while the detailed descriptions remain in Section 2.

We clarified our motivation to focus on ICON-XPP and to introduce ICONEval by describing the gap in existing workflows by adding the following paragraph to section 2.3:

> "Initially, we focus on ICON-XPP, a model that will contribute to CMIP7 and is used in operational climate prediction (Müller et al., 2025), with an active development history. For example, several data driven approaches for ICON-XPP aiming at improving selected parameterizations with machine-learning based methods are being developed such as radiation (Hafner et al., 2026), convection (Heuer et al., 2026), cloud cover (Grundner et al., 2025), gravity waves (Haslauer et al., 2026) or cloud microphysics (Sarauer et al., 2025). Such model development activities require continuous performance monitoring and basic evaluation including sanity checks and physical consistency checks

already during the development phase to identify potential problems early on. Particularly during the often-needed retuning of the model when replacing a core parameterization, an efficient and easy to use tool is needed that also allows for checking a larger number of model simulations. This gap is filled by ICONEval, which allows to run a whole suite of diagnostics with a single command line enabling automatic testing e.g. right after a simulation is finished. In order to maximize the user-friendliness, the results are then also visualized. ICONEval provides for this easy-to-read overview pages linking to the individual results accessible via a simple web browser."

Following the suggestion of the reviewer, we also added a more concise definition of each of the three diagnostic categories to introduction:

> "The sanity checks can be used to assess the global representation of selected variables in an ESM simulation to ensure that the values fall within the expected range such as observed minima and maxima. The physical consistency checks can be applied to verify that a model's equations, parameterizations, and numerical methods do not violate any known fundamental laws of physics. The climate fidelity diagnostics are used to verify that the observed basic characteristics of the atmosphere, the ocean and the land surface such as geographical distributions, seasonal and diurnal cycles or mean states are represented correctly in a model simulation."

2. The paper's specific contribution would be clearer if manuscript distinguish more clearly between functionality newly provided by ICONEval and functionality inherited from, or already available within, the underlying ESMValTool framework. At present, the two are sometimes discussed interchangeably, and it should also be made clearer which limitations of current ESMValTool usage are addressed by ICONEval. In addition, the manuscript highlights potential applications to hybrid machine-learning-enhanced (MLe) Earth system models and to community benchmarking standards such as ClimateBench. However, the examples presented are limited to a physics-based ICON-XPP simulation, so these capabilities are not demonstrated. The authors should either present them more explicitly as intended or future applications, including corresponding revisions to the abstract, or add a brief illustrative example showing how ICONEval is used in an MLe or benchmarking context. Finally, because the implementation and demonstration are currently ICON specific, this scope should be stated more explicitly in the title and abstract.

We agree with the reviewer that the new functionalities of ICONEval and old and new functionalities of ESMValTool need to be more clearly distinguished. We clarified this in the revised version of the manuscript by adding the following paragraph to the section 2.3:

> "While all diagnostics presented in this paper can in principle also be done with the latest ESMValTool version, the task of developing a model requires an efficient and easy-to-use approach to quickly assess also a large number of simulations. For this, ICONEval provides a ready-made suite of diagnostics that we found useful during model development including model tuning. ICONEval is specifically tailored towards basic automated testing and performance assessment of ICON simulations including a user-friendly visualization of the results."

The reviewer brings up the point that while application of ICONEval to hybrid (MLe) models is mentioned, only examples from the physics-based ICON are shown. We would like to note that the output of ICON-MLe versions including e.g. ML-based convection, cloud microphysics, cloud cover or radiation, is identical to the physics-based setup. In our opinion, showing more than one model version in the examples would therefore bring no additional insights but only lengthen the then required description of the other ICON versions and model experiments. Several different of these simulations would be needed as none of these simulations covers all domains addressed in ICONEval (atmosphere, ocean, land). We therefore think a single model run to illustrate the examples is sufficient and prefer to keep this single, well documented physics-based model simulation as an example as it provides output for all domains covered here (atmosphere, ocean, land).

The reviewer is right that the current implementation and demonstration of ICONEval is ICON specific. The workflow and the features presented can also be used for different models as long as the model output is either CMORized or the model is supported directly by ESMValTool. We clarified this in the abstract by adding the following paragraph:

> "While ICONEval is focused on the ICON model, the underlying ESMValTool diagnostics can also be used for different models as long as the model output is following the CMOR (Climate Model Output Rewriter) standard or the model is supported directly by ESMValTool."

3. Section 4.2 / Fig. 3 /Line 345: This is a useful physical-consistency check, but its interpretation should be clarified. Clausius-Clapeyron describes saturation vapor pressure; a 6-7% K-1 scaling of column water vapor additionally assumes approximately constant relative humidity. Therefore, describing the prw-T2m relation as "close to the Clausius Clapeyron equation" is imprecise. Please describe the regression as an empirical annual mean temperature-moisture relationship over the selected period, rather than a direct warming-response scaling. Finally, since the result depends on the analysis period, ERA5 and ICON-XPP should be compared over the same period (1993-2014); the longer ERA5 record can be shown separately for context.

As suggested, we harmonized the time periods shown in figure 3 and rephrased the description of the regression as "an empirical annual mean temperature-moisture relationship over the selected period".

Minor comments

- Line 32: renanalysis → reanalysis.

Corrected, thanks for spotting this.

- Line 41: 'of essential climate variables are helpful...' add few variables variables?

We added "selected" before "essential climate variables".

- Line 45: It is suggested to add the description of ClimateBench.

We added the following brief description of ClimateBench to the introduction: "Climate Bench is an open-source machine learning dataset and benchmarking framework to emulate complex Earth system models for different emission scenarios."

- Line 50: "In the following, the tools, methods and datasets used are briefly described", add the paragraph at the end of the Introduction section regarding the flow of the manuscript. This line here is redundant.

As suggested, we moved this line to the end of the introduction.

- Line 52: ESMValTool has been already defined in earlier sections.

We removed "ESMValTool" from the parentheses.

- Lines 83-84: Since the details of ICON-XPP are already given in the published Müller et al. (2025), the statement "More detail about the final ICON-XPP configuration for CMIP7 can be found in Pham et al. (in preparation)" should be modified or removed, as an "in preparation" reference is not retrievable by readers.

We removed the sentence referring to "Pham et al. (in preparation)".

- In Table 1, the varaible 'surface downwelling longwave radiation (rlus)' has to be checked for MERRA2 and ERA5. Isn't it rlds? Please check the surface pressure for ERA5, psl used for sea level pressure not for surface pressure.

Thanks for spotting this. "rlus" is the "surface upwelling longwave radiation", "psl" in the table should be "ps". This has been corrected.

- Check the range of Global mean TOA net downward radiation (166.3 - 231.8 W m-2). This values is very inconsistent.

This is the range for the surface (and not TOA as wrongly indicated in table 2) net downward radiation. We corrected this in the table.

- Is the CLARA-A2.1 in Table 1 and CLARA-AVHRR in Table 2 is similar?

Yes, this is the same dataset. We changed "CLARA-A2.1" to "CLARA-AVHRR" in table 1 to be consistent throughout the paper.

- Global net moisture flux into the atmosphere units in Table 2 should be mass per unit time.

Thanks for spotting this. We corrected the unit from "kg" to "kg s^{-1}".

- Line 111: 'listed in 2' → listed in Table 2?

We added the missing "Tab." Before "2".

- Lines 103-105: The opening of Section 3 contains a useful conceptual definition of sanity checks ("...the global representation of selected variables...the temporal evolution of global means, minima, and maxima can be assessed"). A concise definition of each of the three categories should be introduced where they are first named in the Introduction. The methodological detail on how the bounds are derived (lines 104-105) can remain in Section 3.

Following the suggestion of the reviewer, we added a paragraph to the introduction defining each of the three diagnostic categories more precisely (see text changes in reply to major comment #1.

- Figure 3/Line 185: The axis labels ("Delta 2m temperature [K]", "Delta water vapor [%]") use different notation from the text, which defines the regression variables as ΔT2m (in K) and Δprw (in %). Please make the figure and text consistent and define ΔT2m and Δprw in the figure caption for clarity.

We updated the labels and the caption of figure 3 as suggested.

- Line 235: biogeogemistry → biogeochemistry.

Corrected.

- Lines 244-248: The sentence beginning "These diagnostics were selected since they illustrate..." and ending "...temperature distributions (e.g. Bock et al., 2020)" is very long and please rephrase it as two or three shorter sentences.

We split the sentence into two.

- Line 273: the the El Niño... → the duplicated

Thanks for spotting. Corrected.

- Line 309: ICONEVal? Is it ICONEval? The parenthesis can also be removed.

Corrected ICONEVal to ICONEval (line 76), parentheses removed (line 309).

- Section 5.3: Please replace "lai" with "LAI" (or set it in italics as "lai") and apply this consistently throughout the section.

We capitalized LAI throughout the manuscript.

- Line 356: new ESMValTool → ICONEval

Changed as suggested.

- Line 386: The example website provided to showcase ICONEval's output (the swift.dkrz.de link in the Code and Data Availability section) is not publicly accessible (returning an HTTP 401 (Unauthorized)) response. Since the reporting and publishing capability is central to the contribution but is currently described only in text via this external link, please ensure the example site is made publicly accessible without credentials.

We double-checked the link and the site should be publicly accessible via the link provided. We suspect that a recent system downtime due to security issues might have led to the access error described by the reviewer.

General / language

In several parts of the manuscript, information is unnecessarily placed in parentheses, which interrupts the flow and makes some sentences harder to read. Please revise these passages to integrate the parenthetical content into the main sentence where appropriate, and reduce the use of parentheses overall.

We went through the manuscript and removed parentheses where possible, i.e. in lines 19, 30, 64, 93, 97, 120, 121, 156, 157, 210, 211, 232, 241, 253, 257, 295, 303, and the captions of figures 1, 2, 15, 17 and of table 1.

References

Grundner, A., Beucler, T., Savre, J., Lauer, A., Schlund, M., and Eyring, V.: Reduced cloud cover errors in a hybrid AI-climate model through equation discovery and automatic tuning, Scientific Reports, 15, https://doi.org/https://doi.org/10.1038/s41598-025-29155-3, 2025.

Hafner, K., Shamekh, S., Bertoli, G., Lauer, A., Pincus, R., Savre, J., and Eyring, V.: Representing subgrid-scale cloud effects in a radiation parameterization using machine learning: MLe-radiation v1.0, Geoscientific Model Development, 19, 3875-3891, https://doi.org/10.5194/gmd-19-3875-2026, 2026.

Haslauer, E., Schwabe, M., Dörnbrack, A., Gerber, E., Rapp, M., Å½agar, N., and Eyring, V.: Interpretable neural networks to predict momentum fluxes of orographic gravity waves, Machine Learning: Earth, 2, https://doi.org/10.1088/3049-4753/ae7df2, 2026.

Heuer, H., Beucler, T., Schwabe, M., Savre, J., Schlund, M., and Eyring, V.: Beyond the Training Data: Confidence-Guided Mixing of Parameterizations in a Hybrid AI-Climate Model, Journal of Advances in Modeling Earth Systems, 18, e2025MS005, https://doi.org/https://doi.org/10.1029/2025MS005544, e2025MS005544 2025MS005544, 2026.

Müller, W. A., Lorenz, S., Pham, T. V., Schneidereit, A., Brokopf, R., Brovkin, V., Brüggemann, N., Chegini, F., Dommenget, D., Fröhlich, K., Früh, B., Gayler, V., Haak, H., Hagemann, S., Hanke, M., Ilyina, T., Jungclaus, J., Köhler, M., Korn, P., Kornblueh, L., Kroll, C. A., Krüger, J., Castro-Morales, K., Niemeier, U., Pohlmann, H., Polkova, I., Potthast, R., Riddick, T., Schlund, M., Stacke, T., Wirth, R., Yu, D., and Marotzke, J.: The ICON-based Earth System Model for climate predictions and projections (ICON XPP v1.0), Geoscientific Model Development, 18, 9385-9415, https://doi.org/10.5194/gmd-18-9385-2025, 2025.

Sarauer, E., Schwabe, M., Weiss, P., Lauer, A., Stier, P., and Eyring, V.: A physics-informed machine learning parameterization for cloud microphysics in ICON, Environmental Data Science, 4, e40, https://doi.org/10.1017/eds.2025.10016, 2025.