

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.

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.
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.
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.

Minor comments

Line 32: reanalysis -> reanalysis.

Line 41: ‘ of essential climate variables are helpful...’ add few variables variables?

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

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.

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

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.

In Table 1, the variable ‘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.

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

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

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

Line 111: ‘listed in 2’ -> listed in Table 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.

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 ΔT_{2m} (in K) and Δprw (in %). Please make the figure and text consistent and define ΔT_{2m} and Δprw in the figure caption for clarity.

Line 235: biogeogemistry -> biogeochemistry.

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.

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

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

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

Line 356: new ESMValTool -> ICONEval

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.

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.