https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

An ESMValTool-based framework for sanity checks, physical consistency and climate fidelity during model development – ICONEval v1.0

Axel Lauer[1], Manuel Schlund[1], Lisa Bock[1], Birgit Hassler[1], Gunnar Behrens[2,*], Bettina Gier[2,1], Lukas Lindenlaub[2,1], Stephan Lorenz[3], Jan-Hendrik Malles[2,1], Wolfgang A. Müller[3], Trang v. Pham[4], Katja Weigel[2,1], Guang Zeng[4], and Veronika Eyring[1,2]

[1]Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany
[2]University of Bremen, Institute of Environmental Physics (IUP), Bremen, Germany
[3]Max-Planck Institute for Meteorology, Hamburg, Germany
[4]Deutscher Wetterdienst, Offenbach am Main, Germany
[*]Now at: Department of Environmental and Energy Sciences, Chalmers University of Technology, Gothenburg, Sweden

Correspondence: Axel Lauer (axel.lauer@dlr.de)

Abstract. Continuous evaluation and performance monitoring during the development of Earth System Models (ESMs) are essential to identify potential problems early, such as unrealistic behavior of climate-relevant quantities, insufficient skill in reproducing the observed basic climate state, or violations of physical laws. The latter is particularly important for the emerging class of hybrid machine learning (ML) enhanced ESMs, where data-driven components are integrated with physics-based model formulations. ESMs used for projections of future climate continue to increase in complexity and resolution. Efficient and user-friendly tools such as the Earth System Model Evaluation Tool (ESMValTool) can therefore greatly support the assessment of a model. So far, ESMValTool focused primarily on providing a broad collection of community-developed evaluation diagnostics and recipes, allowing users to perform a large variety of rather detailed assessments across different domains. A main application of the tool was the assessment of multiple ESMs, in particular those participating in the coupled model intercomparison project (CMIP). Here, we introduce ICONEval, an open-source evaluation framework using ESMValTool that complements existing capabilities by enabling rapid, reproducible, and physically informed assessments of model performance, also during development. ICONEval provides efficient parallel processing of ESMValTool recipes and can generate HTML summary reports allowing to easily automatize and visualize evaluation and monitoring of performance during model development. The new capabilities are grouped into three complementary categories: (1) sanity checks, (2) physical consistency checks, and (3) climate fidelity diagnostics. The sanity checks assess whether global mean values of climate-relevant variables are within the bounds derived from observational and reanalysis datasets. The physical consistency checks aim to identify potential violations of constraints imposed by fundamental physics such as conservation of total air mass, realistic variability of atmospheric water vapor with temperature or the temperature dependence of the cloud ice fraction. The climate fidelity diagnostics assess important climate variables from different ESM components (atmosphere, ocean, and land). Here, we demonstrate this extension of the ESMValTool capabilities by applying the new diagnostics to a historical simulation performed with the ICON-XPP model as an illustrative example. The three-step assessment presented here can be efficiently used

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

to compare different model configurations or versions, for example when testing new or updated parameterizations, including hybrid ML-enhanced (MLe) ESMs, also supporting emerging community benchmarking standards such as ClimateBench.

1 Introduction

Earth system models (ESMs) in combination with Earth observations are important tools not only to improve our understanding of present-day climate but also to project climate change under different future scenarios, and in turn inform global climate policy developments. For this, climate models have been continuously improved and extended to the complex state-of-the-art ESMs participating in the latest (seventh) phase of the Coupled Model Intercomparison Project (CMIP7) (Dunne et al., 2025). Particularly when changing or extending components of a model, working on model parameterizations or when optimizing a model's configuration including adjustment of only weakly or unconstrained parameters (model tuning), model evaluation is an essential element during model development. Model evaluation usually consists of comparing model output with Earth observations, renanalysis data or other suitable datasets such as the output of an ensemble of state-of-the-art models to assess a model's performance for a given diagnostic or metric. This serves as a quality control mechanism as well as guidance for identifying potential further model improvements. As the models grow increasingly complex and use a higher resolution, evaluation tools play a key role in the assessment of model performance by allowing for an efficient, reproducible and yet user-friendly way of analyzing even large sets of data intense simulations, consistent tracking of model changes over time, and objective comparison across different model configurations or versions. One of these tools is the Earth System Model Evaluation Tool (ESMValTool, see section 2.1), an open-source framework that enables rapid evaluation of model performance. ESMValTool so far focused mainly on the evaluation of multiple ESMs participating in CMIP, also allowing the benchmarking of model performance across CMIP phases (Eyring et al., 2021; Bock et al., 2020). During model development, sanity checks, physical consistency checks, and climate fidelity diagnostics of essential climate variables are helpful, particularly for assessments of hybrid machine learning (ML) enhanced Earth system models (Eyring et al., 2024b), which replace physical parameterizations with ML for example for convection (Heuer et al., 2024), cloud cover (Grundner et al., 2025) or cloud microphysics (Sarauer et al., 2025). Here, we present ICONEval that focuses on the evaluation of ESMs during model development providing diagnostics that can directly support community benchmarks such as ClimateBench (Watson-Parris et al., 2022). The new capabilities are grouped into three categories covering the following aspects: (1) sanity checks, (2) physical consistency checks, and (3) climate fidelity diagnostics. ICONEval can be configured to visualize output on an easily accessible website, which facilitates sharing of the evaluation results.

2 Methods and datasets

In the following, the tools, methods and datasets used are briefly described.

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

2.1 ESMValTool

Earth System Model Evaluation Tool (ESMValTool; Eyring et al. (2020); Lauer et al. (2020); Righi et al. (2020); Weigel et al. (2021); Schlund et al. (2023); Lauer et al. (2025); Schlund et al. (2025)) is an open-source community-developed diagnostics and performance metrics tool for the evaluation and analysis of climate models and Earth System Models (ESMs). ESMValTool allows for a comparison of single or multiple models against predecessor versions and observations. The aim of ESMValTool is to take model evaluation to the next level by facilitating analysis of many different ESM components, providing well-documented source code and scientific background of implemented diagnostics. Traceability and reproducibility of the results are ensured by providing detailed provenance records for all outputs.

ESMValTool is by now a well-established tool that has been used in numerous European projects resulting in more than 60 peer-reviewed publications (e.g., Tebaldi et al., 2021; Meehl et al., 2020; Gier et al., 2020; Bock et al., 2020). The tool has been used in several chapters of the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC AR6; IPCC (2021)) and has been selected by the CMIP7 Model Benchmarking Task Team as one of the model benchmarking and evaluation tools for the rapid evaluation framework (REF; Hoffman et al. (2025)).

The ESMValTool software package provides a large collection of “recipes” (configuration files defining input data, preprocessing steps, and diagnostics to be applied) and associated analysis scripts for a large range of scientific analyses. A set of new recipes has recently been added focusing on basic sanity and consistency checks as well as a basic evaluation across the modeling domains atmosphere, ocean and land during model development, e.g. when testing new parameterizations. With this, ESMValTool can now be used to efficiently assess new model runs and check whether a model version under development seems on the right path. Application of these diagnostics as a basis for monitoring a running model simulation is also possible.

2.2 ICON-XPP model

The eXtended Predictions and Projections (XPP) version of the ICOsahedral Nonhydrostatic (ICON) model framework is a coupled Earth system model configuration. An important aim is to provide a model platform for contributions to CMIP7 (Dunne et al., 2025). ICON-XPP includes atmosphere, ocean, land, river, sea ice and interactive carbon components in a fully coupled modeling system. It is designed to bridge numerical weather prediction and climate modeling by using a unified process representation for applications ranging from months to long-term climate projections (Müller et al., 2025).

In order to illustrate ICONEVal’s new ESMValTool-based capabilities presented in this paper, we use one of ICON-XPP’s "historical" simulations performed within the CMIP Diagnostic, Evaluation and Characterization of Klima (DECK) model experiment setup that is often used for improving and comparing coupled Earth system models (Eyring et al., 2016). The historical experiment is driven by historical forcing from CMIP7 and is used to analyze the present-day evolution of climate. The atmospheric component of the ICON-XPP simulation used here has a horizontal resolution of approximately 80 km (R2B5) and 130 vertical levels (L130), the ocean model is run at a resolution of about 20 km (R2B7) with 72 vertical levels (L72). The configuration of the ICON-XPP model used in this historical experiment is an improved version of Müller et al. (2025).

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

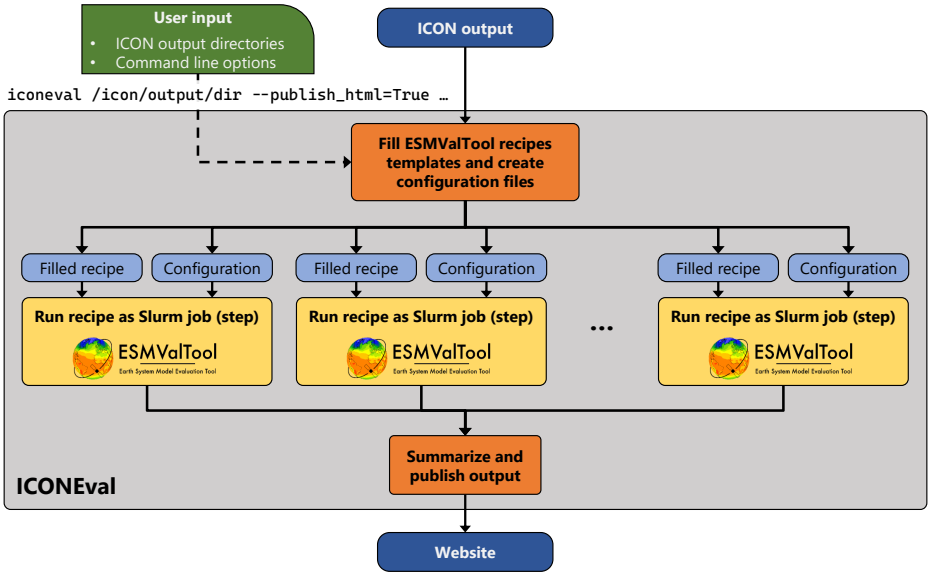

Figure 1. Schematic overview of ICONEval. Based on ICON output (one or multiple simulations), ICONEval fills ESMValTool recipes and creates corresponding configuration files. These recipes are run in parallel by ESMValTool using Slurm job (steps). Finally, results are summarized and (if desired) published to a website for easy access.

Improvement was done in the model physics as well as model parameter tuning in preparation for CMIP7 experiments. More detail about the final ICON-XPP configuration for CMIP7 can be found in Pham et al. (in preparation).

2.3 ICON model evaluation with ICONEval

Here, we present ICONEval, an open-source evaluation framework initially tailored to the evaluation of the ICON Earth system model (see Sect. 2.2) and its ML-enhanced configurations during the development phase. As a demonstration, ICONEval has been integrated into the HPC environment at the Deutsches Klimarechenzentrum (DKRZ), which uses the widespread Simple Linux Utility for Resource Management (Slurm) for managing jobs and a Swift Object Storage that can be used to make the output of ICONEval easily accessible also from outside the HPC environment. ICONEval is a wrapper around ESMValTool that allows running a set of evaluation tasks on one or more given ICON simulations with a single command line call. For this, ICONEval fills ESMValTool recipe templates with the necessary information on the model simulations, runs these recipes in parallel as Slurm jobs (or job steps, if already run within a parent job), and creates a summary HTML file to conveniently visualize the evaluation results in a web browser. Optionally, these results can be published to an available web server for easy access also from outside of the HPC environment or for sharing the results. The workflow of ICONEval is schematically shown in Fig. 1. ICONEval provides an extensive set of predefined evaluation tasks that consist of the diagnostics presented in this study, but also contain additional ones as well as detailed comparisons to other state-of-the-art Earth system models (currently from the CMIP6 project) similar to the diagnostics described by Lauer et al. (2025).

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

2.4 Datasets

Table 1 provides an overview of all datasets used as reference datasets for the basic model evaluation, sanity checks and consistency checks of model simulations including the variables used and main references for the datasets.

Table 1: Overview of the reference datasets used for the climate fidelity diagnostics as well as the consistency and sanity checks implemented in ESMValTool (in alphabetical order).

Dataset	Variable(s)	Reference(s)
CALIPSO-ICECLOUD	mass fraction of cloud ice (cli)	NASA/LARC/SD/ASDC (2018)
CERES-EBAF	TOA incident shortwave radiation (rsdt), TOA net downward radiation (rtmt), TOA outgoing clear-sky longwave radiation (rlutcs), TOA outgoing clear-sky shortwave radiation (rsutcs), TOA outgoing longwave radiation (rlut), TOA outgoing shortwave radiation (rsut)	NASA/LARC/SD/ASDC (2022); Loeb et al. (2009, 2012)
CLARA-A2.1	cloud ice water path (clivi), cloud liquid water path (lwp), total cloud area fraction (clt)	Karlsson et al. (2017, 2020)
CloudSat	cloud ice water path (clivi), cloud liquid water path (lwp), mass fraction of cloud liquid water (clw)	Stephens et al. (2002, 2018)
CM SAF/CCI COMBI	water vapor path (prw)	Schröder et al. (2023)
EN4	ocean potential temperature (thetao)	Good et al. (2013)
ERA5	air temperature (ta), eastward wind (ua), evaporation including sublimation and transpiration (evspsbl), geopotential height (zg), near-surface temperature (tas), precipitation (pr), specific humidity (hus), surface downwelling longwave radiation (rlus), surface pressure (psl), surface temperature (tS), TOA outgoing longwave radiation (rlut), total cloud area fraction (clt), water vapor path (prw)	Hersbach et al. (2020)
ERA-Interim	surface downward eastward wind stress (tauu), surface downward northward wind stress (tauv)	Dee et al. (2011)
ESACCI-CLOUD	cloud ice water path (clivi), cloud liquid water path (lwp), TOA incident shortwave radiation (rsdt), TOA net downward radiation (rtmt), TOA outgoing clear-sky longwave radiation (rlutcs), TOA outgoing clear-sky shortwave radiation (rsutcs), TOA outgoing longwave radiation (rlut), TOA outgoing shortwave radiation (rsut), total cloud area fraction (clt)	Stengel et al. (2020)

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

ESACCI-LANDCOVER	bare soil percentage area coverage (baresoilFrac), grass cover percentage (grassFrac), tree cover percentage (treeFrac), shrub cover percentage (shrubFrac), crop cover percentage (cropFrac)	Harper et al. (2023)
ESACCI-SST	sea surface temperature (tos)	Embury et al. (2024)
GPCP-SG	precipitation (pr)	Adler et al. (2018)
HadCRUT5	near-surface temperature (tas)	Morice et al. (2021)
HadISST	sea ice area fraction (siconc)	Rayner et al. (2003)
IAP	ocean potential temperature (thetao)	Cheng et al. (2024)
ISCCP-FH	surface downwelling longwave radiation (rlds), TOA incident shortwave radiation (rsdt), TOA net downward radiation (rtmt), TOA outgoing clear-sky longwave radiation (rlutcs), TOA outgoing clear-sky shortwave radiation (rsutcs), TOA outgoing longwave radiation (rlut), TOA outgoing shortwave radiation (rsut)	Rossow et al. (2016); Young et al. (2018)
JRA-55	near-surface temperature (tas), water vapor path (prw)	Kobayashi et al. (2015)
LAI3g	leaf area index (lai)	Zhu et al. (2013)
MAC-LWP	cloud liquid water path (lwp)	Elsaesser et al. (2017)
MERRA2	convective precipitation (prc) evaporation including sublimation and transpiration (evspsbl), near-surface temperature (tas), precipitation (pr), surface downward eastward wind stress (tauu), surface downward northward wind stress (tauv), surface downwelling longwave radiation (rlus), surface flux of latent heat (hfls), surface flux of sensible heat (hfss), surface pressure (ps) water vapor path (prw)	Gelaro et al. (2017)
MODIS	cloud ice water path (clivi), cloud liquid water path (lwp), total cloud area fraction (clt)	Platnick et al. (2003)
MTE	gross primary production (gpp)	Jung et al. (2011)
NOAA-CIRES-20CR-V2	surface downward eastward wind stress (tauu), surface downward northward wind stress (tauv)	Compo et al. (2011)
ORAS5	ocean mixed layer depth (mlotst)	Zuo et al. (2019)
OSI-450	sea ice area fraction (siconc)	OSI SAF (2022)
PATMOS-x	total cloud area fraction (clt)	Heidinger et al. (2014)
RAPID	Atlantic meridional overturning circulation at 26.5°N (AMOC)	Moat et al. (2026)

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

TROPFLUX	surface downward eastward wind stress (tauu), surface downward northward wind stress (tauv)	Praveen Kumar et al. (2012, 2013)
WOA	sea surface salinity (sos), sea surface temperature (tos)	Boyer et al. (2018)

3 Sanity checks

With sanity checks, the global representation of selected variables in an ESM simulation, such as the temporal evolution of global means, minima, and maxima can be assessed. "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. Consequently, these limits are only meaningful for a climate that is similar to the one described by the observations, roughly from the 1980s to now. The sanity checks cover the following categories: radiation and energy, moisture and precipitation, clouds, and temperature and wind stress. As illustrated in the example Fig. 2, the sanity plots display the model output time series in blue (solid line) alongside the observational range in red, marking the observed minimum and maximum monthly mean values of the corresponding variable. These bounds are precomputed for runtime efficiency, with the references listed in 2. The diagnostic also displays the minimum and maximum monthly mean values across all model grid cells as dashed lines. This allows for checking that a variable is within physically reasonable limits at all grid cells, e.g. for the variable "total cloud cover" between 0 and 100 %. Values falling outside these observational and physically reasonable bounds indicate that the simulation should be re-examined.

In the following, the sanity checks summarized in Table 2 are briefly introduced by category.

Table 2: Overview of the sanity checks implemented in ESMValTool.

Sanity check	Reference dataset(s)	Observational range (min, max)
Radiation and energy		
Global mean TOA absorbed solar radiation	CERES-EBAF, ESACCI-CLOUD, ISCCP-FH	215.4 - 248.7 $\mathrm{W\,m^{-2}}$
Global average surface flux of latent heat	MERRA2	76.4 - 92.3 $\mathrm{W\,m^{-2}}$
Global average surface flux of sensible heat	MERRA2	15.8 - 21.5 $\mathrm{W\,m^{-2}}$
Global mean surface downwelling longwave radiation	ERA5, ISCCP-FH, MERRA2	319.2 - 360.4 $\mathrm{W\,m^{-2}}$
Global mean TOA outgoing longwave radiation	CERES-EBAF, ESACCI-CLOUD, ISCCP-FH, ERA5	226.4 - 246.3 $\mathrm{W\,m^{-2}}$

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

Global mean TOA net downward radiation	ESACCI-CLOUD, ISCCP-FH	166.3 - 231.8 $\mathrm{W\,m^{-2}}$
Global mean TOA outgoing shortwave radiation	CERES-EBAF, ESACCI-CLOUD, ISCCP-FH	91.1 - 128.8 $\mathrm{W\,m^{-2}}$
Moisture and precipitation		
Global net moisture flux into the atmosphere	ERA5, MERRA2	-7.3·10^8 - 4.4·10^8 kg
Global average total precipitation rate	MERRA2, GPCP-SG, ERA5	2.50 - 3.21 $\mathrm{mm\,day^{-1}}$
Global average convective precipitation rate	MERRA2	0.64 - 0.96 $\mathrm{mm\,day^{-1}}$
Global average water vapor path	ERA5, CM SAF/CCI COMBI, MERRA2	22.3 - 28.7 $\mathrm{kg\,m^{-2}}$
Clouds		
Global mean cloud ice water path	ESACCI-CLOUD, CLARA-AVHRR, CloudSat, MODIS	29 - 93 $\mathrm{g\,m^{-2}}$
Global mean total cloud cover	ESACCI-CLOUD, CLARA-AVHRR, PATMOS-x, MODIS, ERA5	59 - 75 %
Global mean TOA longwave cloud radiative effect	CERES-EBAF, ESACCI-CLOUD, ISCCP-FH	23.6 - 30.5 $\mathrm{W\,m^{-2}}$
Global mean cloud liquid water path	ESACCI-CLOUD, CLARA-AVHRR, CloudSat, MAC-LWP, MODIS	23 - 157 $\mathrm{g\,m^{-2}}$
Global mean TOA net cloud radiative effect	CERES-EBAF, ESACCI-CLOUD, ISCCP-FH	-47.1 - -12.3 $\mathrm{W\,m^{-2}}$
Global mean TOA shortwave cloud radiative effect	CERES-EBAF, ESACCI-CLOUD, ISCCP-FH	-73.5 - -40.6 $\mathrm{W\,m^{-2}}$
Temperature and wind stress		
Global average near-surface temperature	ERA5, HadCRUT5, MERRA2	283.9 - 293.7 K
Global mean surface downward eastward wind stress	ERA-Interim, MERRA2, NOAA-CIRES-20CR-V2, TROPFLUX	-0.053 - 0.020 Pa
Global mean surface downward northward wind stress	ERA-Interim, MERRA2, NOAA-CIRES-20CR-V2, TROPFLUX	-0.014 - 0.028 Pa

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

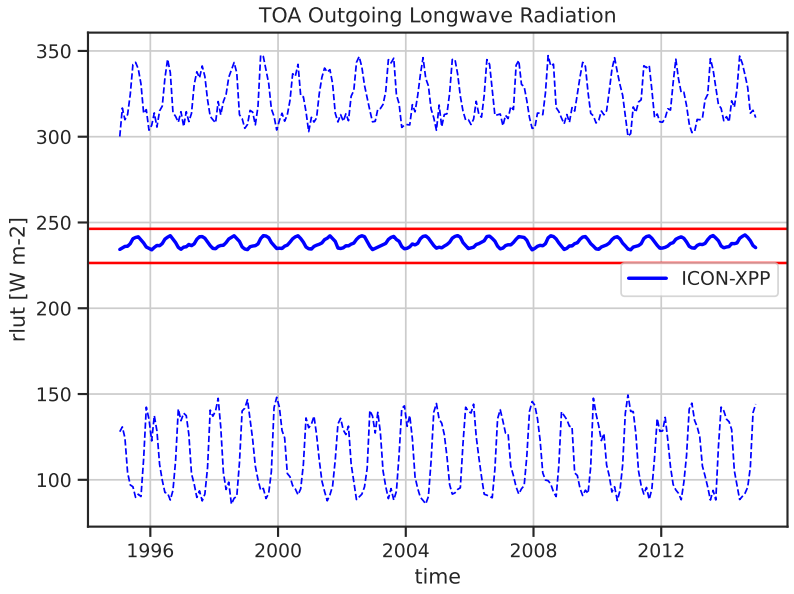

Figure 2. Time series of global monthly mean (solid line) and minimum/maximum (across all grid cells, dashed lines) TOA outgoing longwave radiation. Red horizontal lines show "reasonable" upper and lower limits for the global mean calculated from minimum and maximum global monthly mean values found in reference datasets (observations, reanalyses) across all months and all reference datasets.

3.1 Radiation and energy

Radiation and energy fluxes are the core processes that ensure the model is built on sound physical principles and can accurately simulate the Earth's climate system. Climate models include representations of physical processes, such as how radiation interacts with gases, clouds, aerosols, and surfaces. Checking fluxes tells us if those representations are accurate. The energy budget of an ESM must roughly balance incoming solar radiation with outgoing longwave radiation (plus storage in the Earth system). Significant deviations could indicate errors in radiative transfer or missing feedbacks (e.g. cloud albedo) and would require further investigation. Demonstrating that the model accurately simulates energy fluxes is essential to building confidence in its ability to predict future climate changes (Wild, 2020). Satellite instruments provide global measurements of radiation fluxes at the top of the atmosphere (TOA) and at the surface. Here, these are the primary benchmarks for model evaluation.

We included the TOA absorbed solar, longwave (see Fig. 2), shortwave and net radiation, the surface downwelling longwave radiation and the latent and sensible heat flux in the set of variables for the sanity checks regarding radiation and energy (see Table 2).

3.2 Moisture and precipitation

Checking the global moisture flux, precipitation rates (total and convective), and water vapor path (see Table 2) are critical for validating climate models. These metrics are the basis for checking that the model accurately conserves water and energy,

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

fundamental requirements for a realistic simulation of the Earth system. Discrepancies reveal issues with how the model handles evaporation, cloud formation, convection, or atmospheric moisture transport. By comparing model outputs to satellite and other observational data, these processes can be refined if needed, building confidence in the model's ability to predict future climate changes and hydrological patterns.

3.3 Clouds

Clouds play an essential role in climate as key components in the hydrological cycle and by reflecting substantial amounts of incoming solar radiation and by absorbing outgoing longwave radiation, the so-called cloud radiative effect (cre). A basic set of cloud variables is therefore also included in the sanity checks. Key variables investigated here include global mean cloud ice and liquid water paths, total cloud cover, and the TOA radiative effects in both, shortwave and longwave spectra (see Table 2). These checks assess how accurately the model simulates basic cloud properties – their amount, cloud phase (liquid/ice), and their impact on the Earth's energy balance. Comparing modeled values with satellite observations checks whether the model correctly captures the clouds' role in reflecting incoming solar radiation, trapping outgoing infrared radiation, and ultimately determining the net radiative forcing exerted by clouds. Simulating realistic clouds is a crucial prerequisite for trustworthy climate models, as clouds remain a significant source of uncertainty in climate projections and strongly influence global temperatures and precipitation patterns. As the uncertainties in observational datasets of cloud properties are typically large, it is important to include different observations to display a possible range.

3.4 Temperature and wind stress

Near-surface temperature validation can be used as a proxy for a model's energy balance and its representation of key processes like greenhouse gas effects and land-atmosphere interactions. Simultaneously, monitoring surface wind stress can be used to assess a model's representation of the atmospheric forcing on ocean currents and large-scale circulation patterns like the trade winds. Accurate representation of both, temperature and wind stress, is essential for simulating realistic climate states, predicting regional climate variability, and a correct ocean-atmosphere coupling.

4 Physical consistency checks

ESMs are built on known fundamental laws of physics. Physical consistency checks ensure that the model's equations, parameterizations, and numerical methods correctly represent these laws. This is of importance as simplified representations of subgrid processes (e.g. newly developed data driven parameterizations of convection or cloud microphysics) might violate some physical constraints. Complying with basic physical consistency checks can be seen as a necessary (even though not sufficient) precondition to allow for realistic future projections. The set of different physical consistency checks available in ESMValTool is summarized in Table 3.

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

Table 3. Overview of physical consistency checks, focused on atmospheric parameters, implemented in ESMValTool.

Consistency check	Reference dataset(s)	Observational range
Total mass of air	ERA5, MERRA2	$5.0220 \cdot 10^{19} \pm 0.0005 \cdot 10^{19}$ kg
Conservation of total air mass calculated as relative annual anomalies	ERA5, MERRA2	± 0.02 %
Total mass of water vapor	ERA5, CM SAF/CCI COMBI, MERRA2	$1.12 \cdot 10^{16}$ - $1.42 \cdot 10^{16}$ kg
Relative change in global annual mean water vapor mass per degree of global average near-surface temperature change (1993-2021)	ERA5, JRA-55, MERRA2	6.5-7.5 % K^{-1}
Dependence of 3-dim cloud ice fraction on temperature	CloudSat, CALIPSO-ICECLOUD, ERA5	-
Frequency distribution of tropical and mid-latitude 3-dim tropospheric lapse rates	ERA5	-

The implemented consistency checks summarized in Table 3 are focused on atmospheric parameters and should be seen as a starting point. They can be extended further depending on the focus of the model development efforts. In the following, the available consistency checks for the atmosphere are briefly introduced.

4.1 Mass conservation

The mass of air is usually treated as a conserved quantity in climate models, i.e. the total mass of air should remain constant during a climate model simulation (Trenberth and Smith, 2005). A violation of this rule would lead to an increase or decrease in the global average surface pressure, which is used as a proxy for the total mass of air. Here, this mass conservation is checked by calculating the anomalies of the surface pressure (p_s) integrated over the whole surface area of the globe. In order to obtain easier to read numbers, the anomalies are calculated as relative anomalies using the whole time period as reference period. Ideally, these anomalies in total air mass should remain close to 0 %. The reanalysis datasets investigated as a reference (ERA5 and MERRA2) show maximum fluctuations of about 0.02 %.

The amount of water vapor in the atmosphere is highly variable in space and time (Trenberth et al., 2007) but the global total mass of water vapor remains approximately balanced through evaporation and precipitation. A time series of the total amount of water vapor in the atmosphere, here calculated as the global sum of the vertically integrated amount of water vapor per unit area (water vapor path), is therefore expected to show only small variations in its amplitude throughout a year.

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

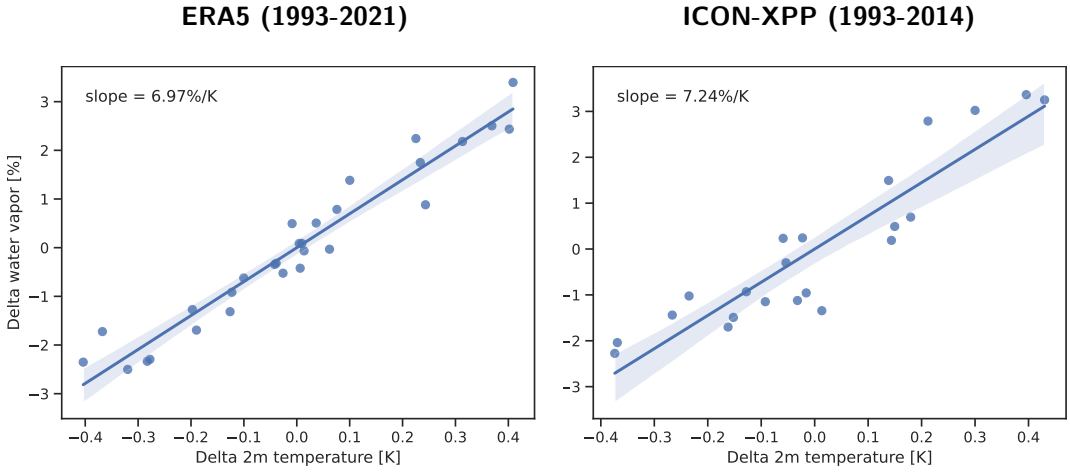

Figure 3. Linear correlation of global annual mean anomalies in 2-m temperature (x-axis) and relative global annual mean anomalies in water vapor columns (y-axis). Left: ERA5 over the time period 1993-2021, right: ICON-XPP over the time period 1993-2014.

4.2 Temperature dependence of water vapor

The amount of atmospheric water vapor depends strongly on temperature. This relation is close to the Clausius–Clapeyron equation, which describes the saturation vapor pressure of water changes with temperature. This means that if relative humidity in the lower troposphere stays roughly constant, the total column water vapor (prw) would be expected to increase by about 6–7 % per degree of warming (Boer, 1993). As proxies for the amount of atmospheric water vapor, we use the global mean water vapor path and for temperature the global mean 2-m temperature (T_{2m}). Wan et al. (2024) showed that while the relationship between total precipitable water and T_{2m} does not necessarily scale with the Clausius-Clapeyron on regional scale, it aligns with the equation in the 15-55°N latitude band (6-8 % per degree warming). Here, we calculate anomalies in global annual average T_{2m} and prw for each year. Following Wan et al. (2024), we use the period 1993-2021 for the analysis, the anomalies are calculated using the whole 29-year period as reference period. In contrast to Wan et al. (2024), we do not calculate separate trends for T_{2m} and prw, but calculate the slope of a regression line with $x = \Delta T_{2m}$ (in K) and $y = \Delta prw$ (in %). This approach is simple yet robust enough for a consistency check of the relation between temperature and water vapor on global scale. We would like to note that the actual values depend on the time period analyzed. Similar to Wan et al. (2024) we find higher values in the more recent time period 1993-2021 compared with the maximum time period available for the reanalysis datasets investigated (ERA5, JRA-55, MERRA2). An example of this diagnostic is shown in Fig. 3 comparing results from ERA5 for the time period 1993-2021 with the ICON-XPP simulation over the time period 1993-2014 (latest year in the ICON-XPP example simulation is 2014).

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

4.3 Cloud ice fraction

The fraction of cloud water that is frozen particularly depends on temperature. Typically, the cloud ice fraction reaches about 100 % at the threshold temperature for homogeneous freezing at roughly -38°C (e.g. Atkinson et al., 2016) resulting in pure ice clouds. Above the melting temperature of 0°C, all remaining cloud ice particles begin to melt resulting in pure liquid water clouds, i.e. the cloud ice fraction is approaching zero. In between these two threshold temperatures, typically mixed-phase clouds exist with an ice fraction increasing steadily between 0 % at 0°C and 100 % at -38°C. This consistency diagnostic allows to check whether a model can reproduce the qualitative relationship between temperature and cloud ice fraction obtained from vertically resolved measurements or measurement-based data of cloud ice (CALIPSO-ICECLOUD), cloud liquid water content (CLOUDSAT-L2) and temperature (ERA5). We would like to point out, that we are using monthly means resulting in a blurring of the ice fraction and leading e.g. to the possibility of small ice fractions even above 0°C. Retrieval of the vertically resolved cloud ice and cloud liquid water content from space is subject to large uncertainties. For example, a clear distinction between cloud particles and precipitation is challenging (e.g. Waliser et al., 2009) and the instruments can show saturation effects or a limited sensitivity to certain cloud types (e.g. Marchand et al., 2008). The shape of the cloud ice fraction from observations is therefore only used for a qualitative comparison with the model data. For the calculation of the average cloud ice fraction - temperature relationship, all 3-dim grid cells are used and binned into 20 temperature bins. Grid cells with a cloud water mass concentration below $10^{-6}\ \mathrm{kg\,kg^{-1}}$ (sum of liquid + ice) are not taken into account. Figure 4 shows an example of the cloud ice fraction from ICON-XPP compared with the reference datasets CALIPSO-ICECLOUD, CLOUDSAT-L2, ERA5.

4.4 Tropospheric lapse rates

The lapse rate is defined as the vertical temperature gradient $\Gamma = -\frac{dT}{dz}$ (typically in $\mathrm{K\,km^{-1}}$) and is a measure for the atmospheric stability. In dry conditions, the lapse rate amounts about 9.8 $\mathrm{K\,km^{-1}}$ (dry adiabatic conditions). The lapse rate depends on the humidity and temperature with warm, humid conditions typically leading to smaller lapse rates than dry cold conditions. On average, lapse rates in the tropical troposphere above the boundary layer are expected to be smaller than in mid- or high-latitudes (e.g. Stone and Carlson, 1979). Here we analyze probability density functions (PDFs) of the 3-dim tropospheric lapse rates above the boundary layer in the Tropics (30°S-30°N), Northern Hemisphere mid-latitudes (40°-60°N) and Southern Hemisphere mid-latitudes (60°-40°S) in the altitude range 850 hPa to 250 hPa. For comparability reasons, the model and reference data are interpolated to the same vertical levels (here: 850, 825, 800, 775, 750, 700, 650, 600, 550, 500, 450, 400, 350, 300, 250 hPa). In the example shown in Fig. 5, the peak of the lapse rate PDFs are around 5 $\mathrm{K\,km^{-1}}$ in the Tropics and 6.5 $\mathrm{K\,km^{-1}}$ in mid-latitudes.

5 Evaluation of climate fidelity

Before new simulations can be used for scientific analyses, it is important to make sure that the basic characteristics of the atmosphere, the ocean and the land surface are represented correctly. This is done by comparing the simulations to observational

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

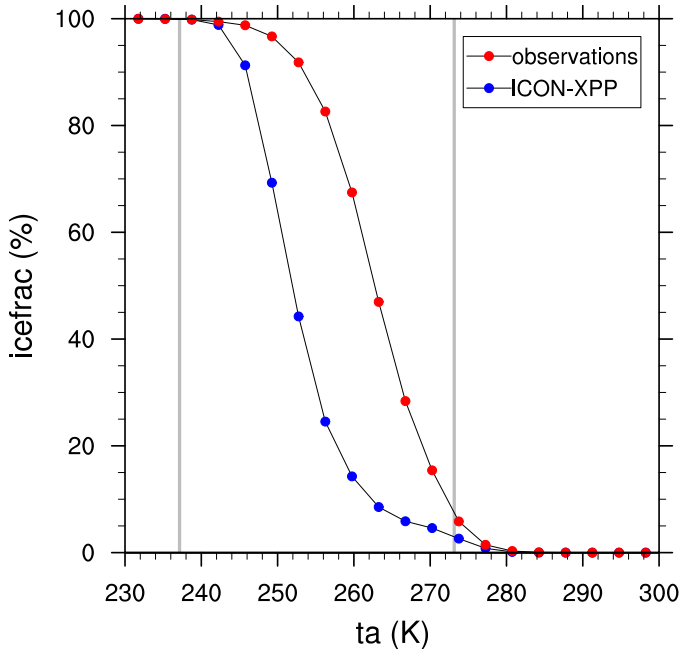

Figure 4. Global average fraction of the cloud ice water content (in %) calculated from monthly means of 3-dim cloud ice and cloud liquid water content over one year (2014) binned by air temperature. The ICON-XPP model is shown as blue dots, the reference line from the combined datasets of CLOUDSAT-L2, CALIPSO-ICECLOUD and ERA5 in red. The gray vertical lines illustrate the temperatures of homogeneous freezing (-38°C) and the melting temperature of ice (0°C).

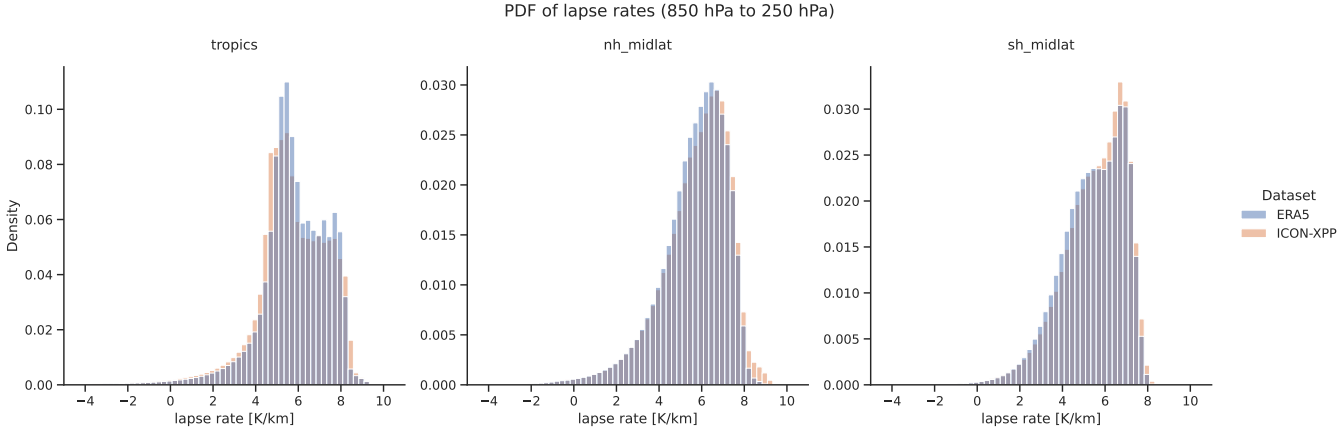

Figure 5. Probability density functions of the 3-dim lapse rates from ERA5 (blue) and ICON-XPP (orange) in the altitude range 850 hPa to 250 hPa calculated from monthly mean temperatures over the time period 1995-2014. From left to right: Tropics (30°S-30°N), Northern Hemisphere mid-latitudes (40°N-60°N) and Southern Hemisphere mid-latitudes (60°S-40°S).

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

data records (Hassler et al., 2026) and is therefore only applicable for simulations spanning the most recent decades. We created diagnostics that can be used to evaluate climatologies, annual and diurnal cycles, geographical distributions and time series of different dataset aggregations. The diagnostics are flexible enough to allow their use for many different variables without the need to adjust anything in the code. Due to their relatively simple nature, they solely fall into the "variability and biases" evaluation approach, a classification introduced by Hassler et al. (2026), and do not aim for an in-depth and complex analysis, which requires more specialized and process-oriented diagnostics. However, these simple diagnostics provide an extremely valuable first check of a simulation and a quick assessment of its quality and usefulness. As a particular goal of ICONEval is to accelerate development of hybrid ML-enhanced (MLe) ESMs that focus on reducing systematic model errors (Eyring et al., 2024b, a), we also include precipitation extremes, diurnal precipitation cycle and double-ITCZ (intertropical convergence zone).

Table 4 provides an overview of the climate fidelity diagnostics available sorted by model component (atmosphere, ocean, land). These are briefly introduced in the following and can be extended depending on the focus of the model development including new components such as sea ice, land ice, biogeogemistry, etc.

Table 4: Overview of the climate fidelity diagnostics implemented in ESMValTool for first checks during model development.

Diagnostic	Variable(s)	Reference dataset(s)
Atmosphere		
Geographical distributions (annual or seasonal means)	near-surface temperature, surface temperature, TOA radiative fluxes (longwave, shortwave), surface fluxes (latent, sensible), precipitation, cloud properties (ice water path, liquid water path, cloud fraction), water vapor path	CERES-EBAF, CLARA-AVHRR, CM SAF/CCI COMBI, ERA5, ESACCI-CLOUD, GPCP-SG
Annual or seasonal zonal means (latitude vs. pressure)	temperature, eastward wind (u-component), specific humidity	ERA5
Time series (optionally with a regression line as a trend estimate) of globally averaged monthly means	near-surface temperature, total cloud cover, TOA radiative fluxes (outgoing shortwave radiation, TOA outgoing longwave radiation, TOA net downward radiation), water vapor path, precipitation	ERA5, ESACCI-CLOUD, CERES-EBAF, CM SAF/CCI COMBI, GPCP-SG
Probability density of daily or sub-daily values	precipitation rate	ERA5
Hour of daily maximum	precipitation	ERA5
Ocean		

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

Seasonal cycle (Southern Ocean, Labrador Sea, Weddell Sea)	ocean mixed layer depth	ORAS5
Seasonal cycle (Northern and Southern Hemisphere)	sea ice area	OSI-450
Time series (Southern Hemisphere February and Northern Hemisphere September, optionally including a regression line as a trend estimate)	sea ice area	OSI-450
Global geographical distribution (annual or seasonal means)	sea ice area, sea surface temperature, sea surface salinity, mixed layer thickness	HadISST, WOA
Time series	Nino3.4 index, global average sea surface temperature, ocean heat content (OHC) in different layers (e.g. 0-100 m, 0-300 m, 0-2000 m), Atlantic meridional overturning circulation (AMOC)	ESACCI-SST, EN4/IAP, RAPID
Zonal means	sea surface temperature, sea surface salinity	ESACCI-SST, WOA
Land		
Global geographical distributions	leaf area index, surface temperature, evaporation including sublimation and transpiration	LAI3g, ERA5
Seasonal cycle and timeseries	leaf area index, gross primary production	LAI3g, MTE
Barplots	land cover fractions (bare soil, grass, trees, shrubs, crops)	ESACCI-LANDCOVER

5.1 Atmosphere

A main evaluation focus for the atmosphere is on parameters that provide general information about the global climate, such as surface temperature, precipitation, radiation, zonal wind speed, and cloud-relevant parameters including water vapor. Most of these variables are of high interest for future climate estimates with their impact on human population, and therefore their magnitude and distribution in historical simulations need to be well understood.

Three different types of analyses for atmospheric variables are available (all with comparisons to different observational datasets): (1) mean geographical distributions (seasonal of annual means), (2) zonal mean (latitude vs. pressure) aggregations for three dimensional variables (seasonal or annual means), and (3) time series of globally averaged variables (see Table 4).

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

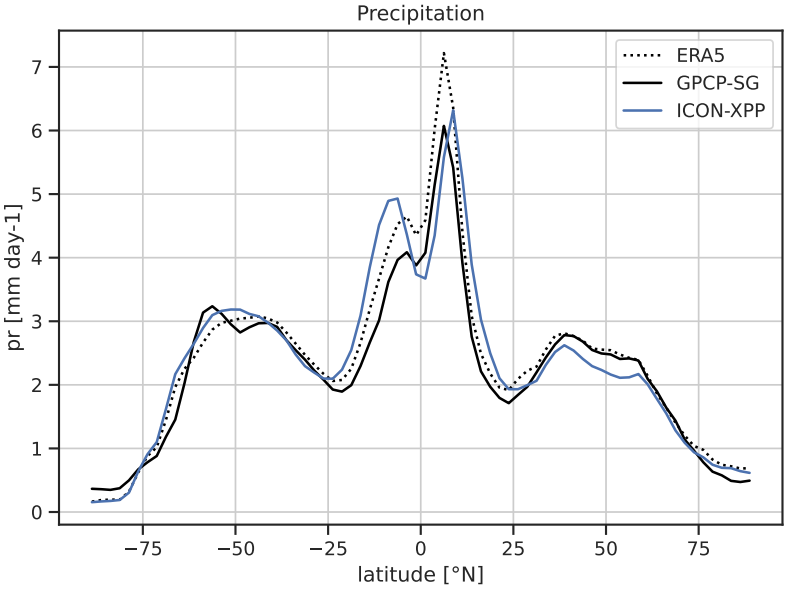

Figure 6. Zonally averaged multi-year annual mean precipitation from ICON-XPP (blue solid line) compared with GPCP-SG (black solid line) and ERA5 (black dotted line).

The time series provide the option to fit a regression line as an estimate for the trend over the specified time range. These diagnostics were selected since they illustrate some well-known and long-standing biases of climate models, like the double ITCZ (Intertropical Convergence Zone) seen in precipitation fields (two zonally elongated narrow belts of high precipitation in the Tropics, south and north of the equator, which are present in model simulations but not in observations (Tian and Dong, 2020)) or the cold bias close to the tropopause as seen in vertically resolved temperature distributions (e.g. Bock et al., 2020). Both biases can also be seen in the examples shown in Figs. 6 and 9, respectively.

For precipitation, two additional diagnostics are available to investigate the simulated sub-daily results, (1) the probability density of the precipitation rates at the surface over a specific region (the example shown in Fig. 7 focuses on the Tropics, which are defined here as latitude belt 30°S-30°N) and (2) the hour of daily maximum precipitation rate over a specific region (here, also an example for the Tropics is shown, Fig. 8).

The histogram shown in Fig. 7 can be used to investigate the frequency of extreme precipitation events. For a fair comparison, both model and reference data (here ERA5) are regridded to the same spatial and temporal grid (here: 1°x1° and 6-hourly values).

The diagnostic calculating the hour of daily maximum precipitation (exemplary shown in Fig. 8) can be used to check that basic features of the diurnal cycle such as the land-sea contrast in the timing of the daily precipitation peaks are in agreement with observations. By default, the maximum in the diurnal cycle of precipitation is obtained by applying a discrete Fourier transform (DFT) to the data and using the peak of the first component (MDTF, 2019). Alternatively, a 12-hour and a 24-hour

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

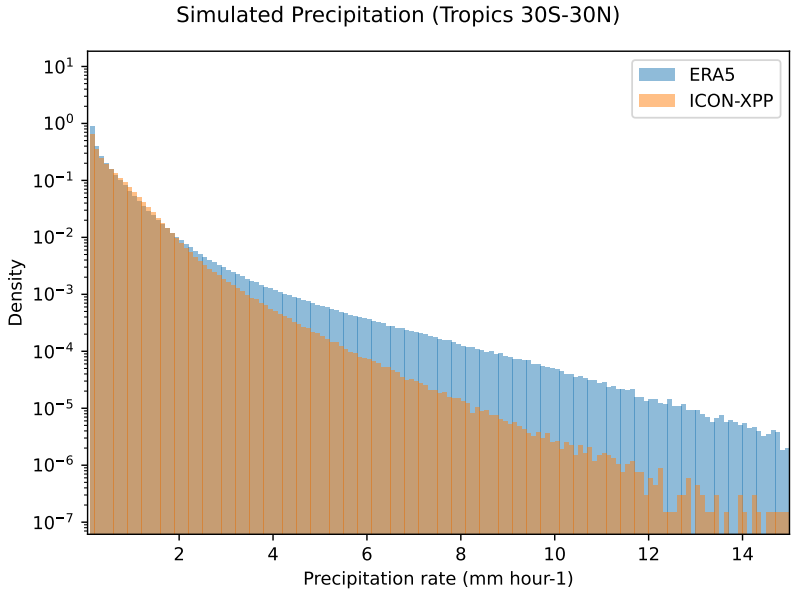

Figure 7. Probability density of the precipitation rates in the Tropics (30°S-30°N) calculated over the time period 2010-2014 from ICON-XPP (orange) compared with ERA5 (blue). All data have been regridded to the same regular 1°x1° lat-lon grid and averaged to 6-hourly intervals.

harmonic fit to the input data can be calculated using the peak of the 24-hour harmonic as hour of the daily maximum (Dai, 2024). For this diagnostic, sub-daily model output is required, ideally hourly values. While 3-hourly and 6-hourly output can be used with the diagnostic, the coarse time resolution can result in artifacts such as the vertical stripes that can be seen in Fig. 8.

5.2 Ocean

Ocean variables of interest can be compared to reanalyses, observations or data from other ESMs. The sea surface temperature is one of the key variables of interest during model development as it determines, amongst other things, the exchange of sensible and latent heat between the ocean and the atmosphere to a large degree. Figure 10 shows the averaged sea surface temperature from ICON-XPP compared with the WOA dataset for the period 1981 to 2010. In the example, the sea surface temperatures from the model and from WOA are in relatively good agreement but biases exist in particular along the western boundary currents, in upwelling regions, in the Southern Ocean and in the North Atlantic Warming Hole (e.g. Kramer et al., 2025) region south of Greenland. ESMValTool includes also diagnostics to evaluate simulated general modes of climate variability like the the El Niño–Southern Oscillation (ENSO). Here, we show the time series of the monthly mean sea surface temperature anomalies in the Niño 3.4 region, which is a sanity check of an ESM's ability to represent interannual variability caused by ENSO (Fig. 11). ESMs are not expected to replicate the timing of distinct observed El Niño and La Niña events, nonetheless they should have a comparable amplitude and frequency of the interannual variability in the Niño 3.4 region.

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

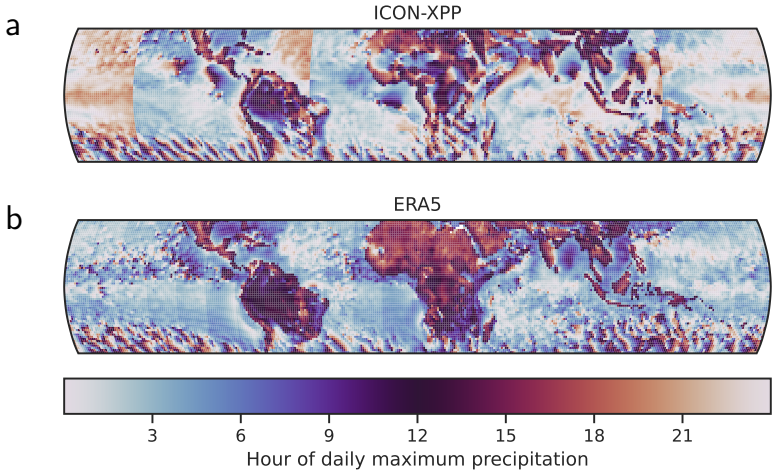

Figure 8. Hour of daily maximum precipitation (local time) in the Tropics (30°S-30°N) for June-July-August calculated over the time period 2010-2014 from (a) ICON-XPP compared with (b) ERA5. The ERA5 data have a time resolution of 1 hour, the ICON-XPP of 6 hours. A time resolution of less than hourly is not optimal and might introduce some artifacts as can be seen by the striped appearance of the ICON-XPP panel.

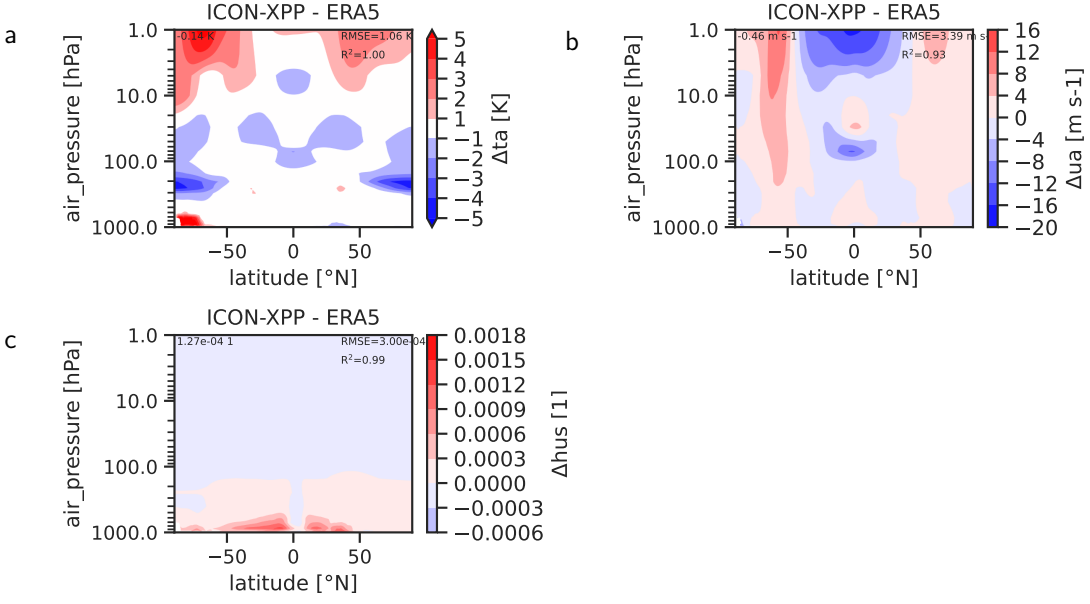

Figure 9. Zonal multi-year annual mean bias in (a) temperature, (b) zonal wind speed, and (c) specific humidity from ICON-XPP compared with ERA5.

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

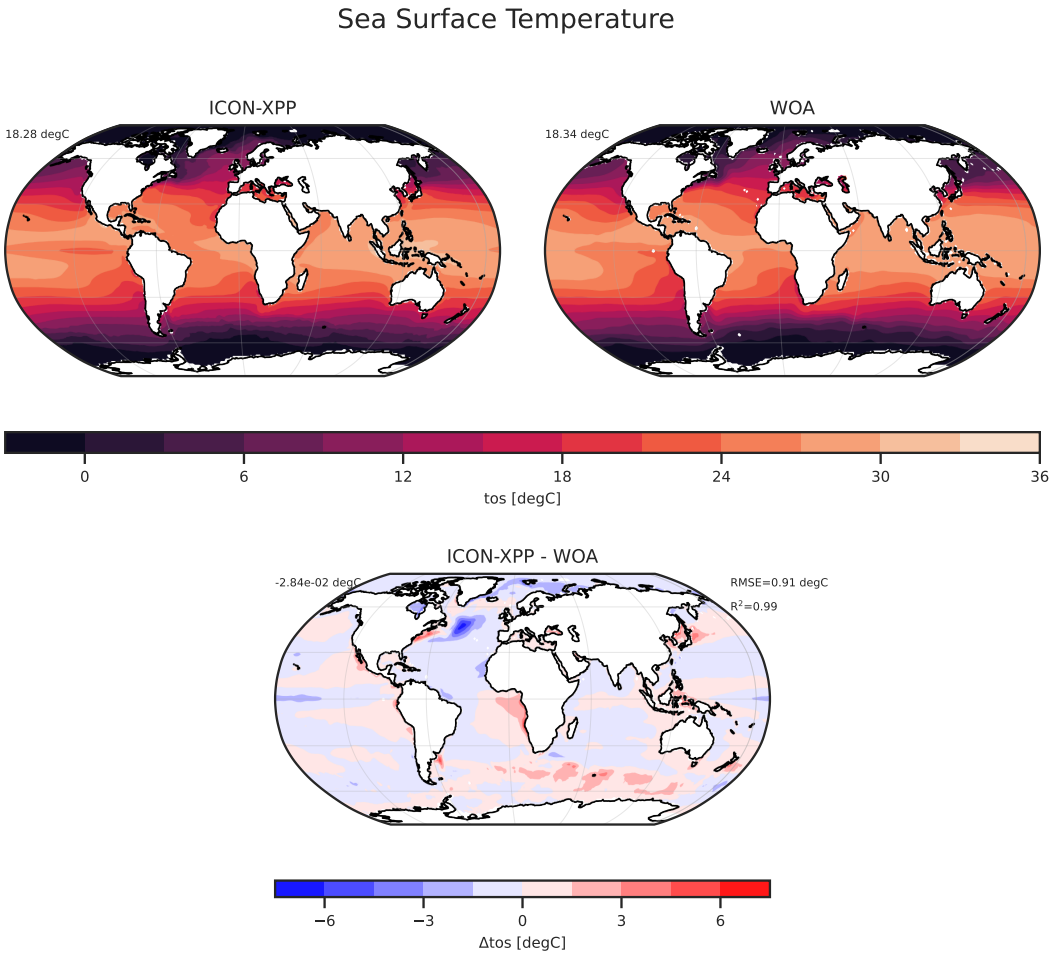

Figure 10. Map plot of the average sea surface temperature for ICON-XPP for the period 1981 to 2010 (left) and WOA for the same time period (right), as well as their difference (below).

Simulating realistic sea ice fields compared to observations is another key aspect during ocean model development. CMIP6 models exhibit pronounced deviations of sea ice area and long-term trends compared to observations especially around Antarctica (Roach et al., 2020), e.g. the Antarctic "sea ice paradox" from the beginning of satellite measurements until 2016. These
280 uncertainties in simulated historical sea ice fields result in low confidence in 21st-century sea ice projections (Fox-Kemper et al., 2021). Recent eddy-permitting simulations indicate some improvement (e.g. Rackow et al., 2022). Thus, sea ice area and extent pose as a suitable target to check the validity of the simulations during model development since they are influenced by e.g. air-sea interactions, ocean and atmospheric circulation and surface albedo. As a proxy for the long-term evolution of sea ice, the sea ice area in the month coinciding with the annual minimum in sea ice area can be used. In the Northern Hemisphere

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

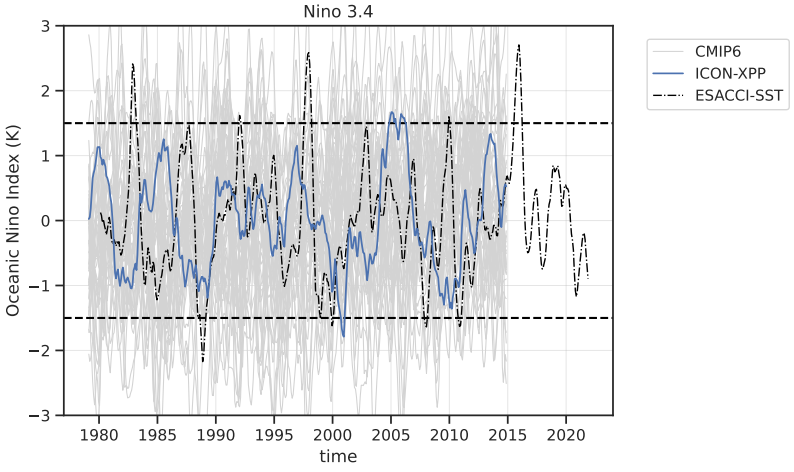

Figure 11. Time series of SST anomalies in the Niño 3.4 region (5°N–5°S and 170°W-120°W). The black dash-dotted line shows the ESACCI-SST product, the blue line the ICON-XPP results and the thin gray lines results from the CMIP6 model ensemble. The horizontal dashed lines indicate the thresholds of strong El Niño and La Niña events as used by, e.g., the NOAA Climate Prediction Center.

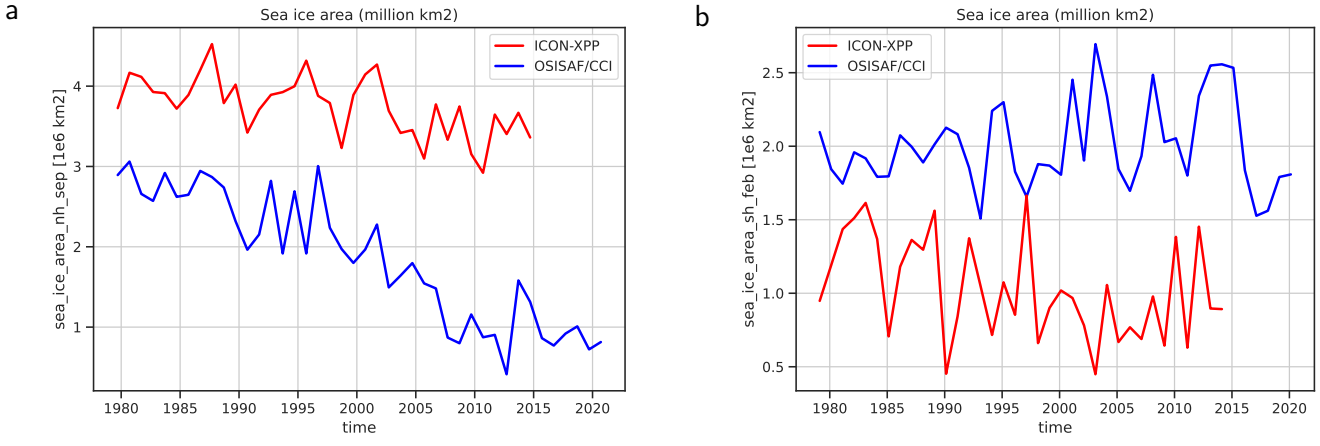

Figure 12. Time series of (a) Northern Hemisphere September and (b) Southern Hemisphere February sea ice area in million km^2 from ICON-XPP for the period 1979 to 2014 (red) compared with OSI-450 (blue) for the period 1979 to 2020.

(NH), this is typically September, in the Southern Hemisphere (SH) in February. Fig. 12 shows a time series of the simulated NH and SH sea ice area from ICON-XPP and from the OSI-450 sea ice observational product during the respective months.

Another key metric for the validity of an ocean model during development is the simulated oceanic mixed layer. The mixed layer depth (MLD) determines key processes in the ocean, e.g. deep and mode water formation (Danabasoglu et al., 2014), carbon uptake (Llort et al., 2019) or simulated oxygen minimum zones (Busecke et al., 2022) and is strongly influenced by air-sea interactions. Regions with a pronounced deep water and mode water formation due to a strong seasonal variability of

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

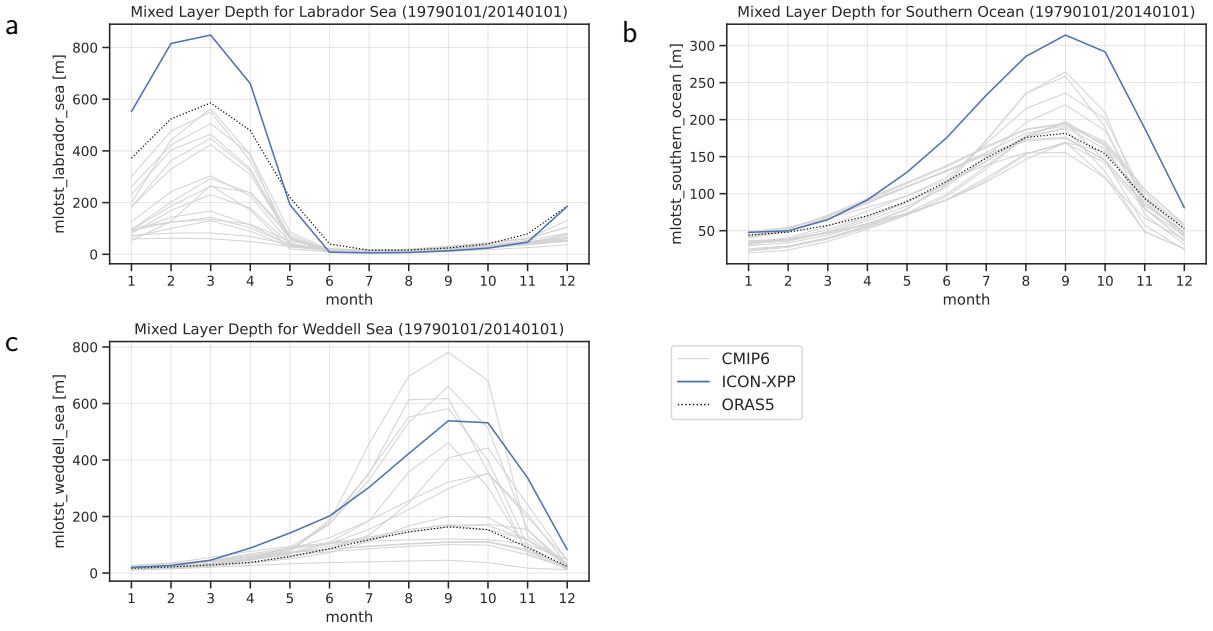

Figure 13. Seasonal cycle of the ocean mixed layer depth defined by sigma-T for ICON-XPP (blue line) compared with results from the CMIP6 ensemble (gray lines) and ORAS5 (black dashed line) for (a) Labrador Sea, (b) Southern Ocean, and (c) Weddell Sea for the period 1979 to 2014.

MLD are prone to biases in the simulated MLD compared to observations. These biases are dependent on the ocean models' ability to resolve mesoscale dynamics (Treguier et al., 2023) but also on their sub-mesoscale mixing schemes (Uchida et al., 2026): ESMs often underestimate the MLD in zones with pronounced deep water formation, but also eddy-permitting and eddy-resolving models may exhibit biases due to relying on sub-mesoscale mixing schemes. Thus, the seasonal evolution of MLD in deep and mode water forming regions (see Table 4) is a key metric during ocean model development (e.g. for tuning sub-mesoscale mixing schemes).

Figure 13 shows the mean seasonal cycle of MLD in the Labrador Sea, Southern Ocean, and Weddell Sea in ICON-XPP compared to CMIP6 models and the ORAS5 ocean reanalysis. The maximum MLD is found during winter in Southern and Northern Hemisphere, when deep and mode water masses are formed. In summer the positive surface heat flux in deep and mode water forming regions leads to a minimum MLD due to an increased stratification of the upper ocean.

An important feature of the ocean circulation related to the MLD variability in the Labrador Sea and other areas with mode and deep water formation in the North Atlantic is the Atlantic Meridional Overturning Circulation (AMOC, Danabasoglu et al. (2014)). The AMOC drives the (inter-hemispheric) northward transport of heat and salt at the surface (Zhang et al., 2019). In addition, several studies investigated a simulated AMOC collapse as a potential tipping point of the Earth system in climate projections (McKay et al., 2022). Thus, simulating the observed historical AMOC strength accurately is a key metric during ocean model development. Figure 14 shows the time series of the historical AMOC strength, defined as the maximum of the

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

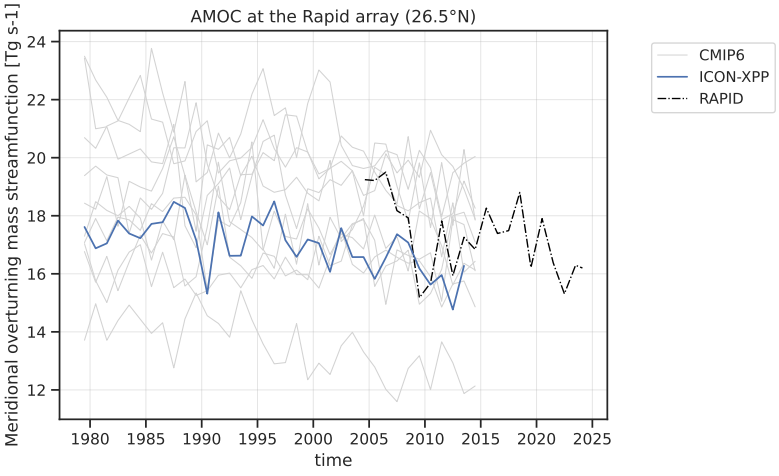

Figure 14. Time series of the AMOC strength calculated as the maximum of the meridional mass stream function at 26.5°N. The black dash-dotted line shows the RAPID observational data, the blue line the ICON-XPP results and the thin gray lines examples from the CMIP6 model ensemble.

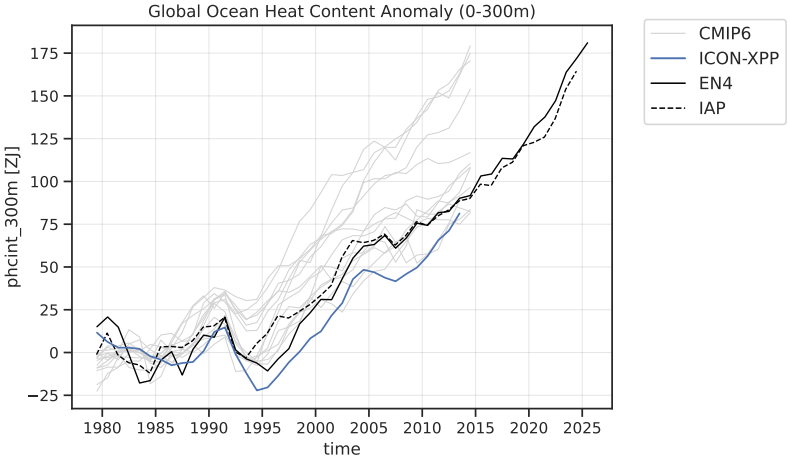

Figure 15. Time series of ocean heat content anomaly for the upper 300 m (relative to the 1995-2014 average) in zettajoules (ZJ). The black lines show the reference datasets, the blue line the ICON-XPP results and the thin gray lines results from the CMIP6 model ensemble.

meridional mass stream function at 26.5°N in the Atlantic, for ICON-XPP, CMIP6 models and plotted against the observed AMOC strength from the RAPID array (Cunningham et al., 2007) from 2004 onward.

ICONEval (and ESMValTool) can be further used to monitor key oceanic climate metrics like the ocean heat content (OHC) during model development. The OHC is influenced by atmosphere-ocean interactions as well as by the MLD. As an example, Figure 15 shows the OHC anomalies of the upper 300 m from ICON-XPP, some CMIP6 models and the EN4 and IAP reference datasets.

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

5.3 Land

An important parameter for the land component is the surface temperature as this is one of the key parameters for coupling the land component with the atmosphere. Another important parameter is land evaporation, which plays a key role in the water cycle and land-atmosphere interactions.

The leaf area index (lai, unitless) is a measure of the canopy structure and used by models to calculate the photosynthetic uptake of carbon of the total canopy (Park and Jeong, 2021). Maps of its geographical distribution can be used to identify areas with atypical growth. Figure 16 shows a 10-year average of lai based on the example ICON-XPP simulation and LAI3g observations for the same period used as a reference. Highest lai values can be found in tropical rain forest regions, for both datasets. The difference map in the bottom panel of Fig. 16 shows an overestimation in the simulated lai by 0 to 2 for many regions in the Southern Hemisphere.

The seasonal cycle of lai is expected to follow the growth season and to be particularly well defined in the Northern Hemisphere. The summer peak in July/August can be observed in the bottom right panel of Fig. 16, which shows a monthly NH climatology from ICON-XPP together with the LAI3g reference data. The generally higher lai values in the NH climatology of the LAI3g dataset are caused by missing values in arid regions. ICON-XPP in contrast, provides full spatial coverage with generally small lai in arid regions.

To check the validity of the land model, the coverage of different land types can be compared to observations, as done in Fig. 17 for the example of crop area. This is particularly of interest for models with a dynamic vegetation. While this is important for checking how well a model does compared to observations, this is also highly relevant for regional modeling and impact assessment.

6 Summary and conclusions

As the output of Earth system models continues to grow in complexity and resolution, efficient and user-friendly tools can be a great support to assess, document and compare model performance across different model versions or configurations. This paper presents recent extensions of ESMValTool, and introduces ICONEval as an ESMValTool-based framework for rapid and physical consistency evaluation of ESMs. ICONEval is a wrapper designed to complement ESMValTool capabilities by providing a simplified access to a set of predefined evaluation tasks including sanity checks, physical consistency checks, and climate fidelity diagnostics of essential climate variables. The sanity checks are used for a first and very basic quality control to verify that the global mean values of key climate parameters of historical simulations are within reasonable ranges. These ranges are derived from upper and lower limits, i.e. maximum and minimum, of these parameters across multiple observational or reanalysis datasets and all monthly means available used for comparison. In addition, minimum and maximum values across all grid cells are analyzed to check that all values are physically reasonable, e.g. that mass mixing ratios are positive or maximum total cloud fraction does not exceed 100 %. Aspects checked include radiation and energy fluxes, atmospheric moisture and precipitation, integral cloud properties and cloud radiative effects, and near-surface temperature and wind stress. The physical consistency checks analyze compliance with fundamental physical laws and thermodynamic relationships. The

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

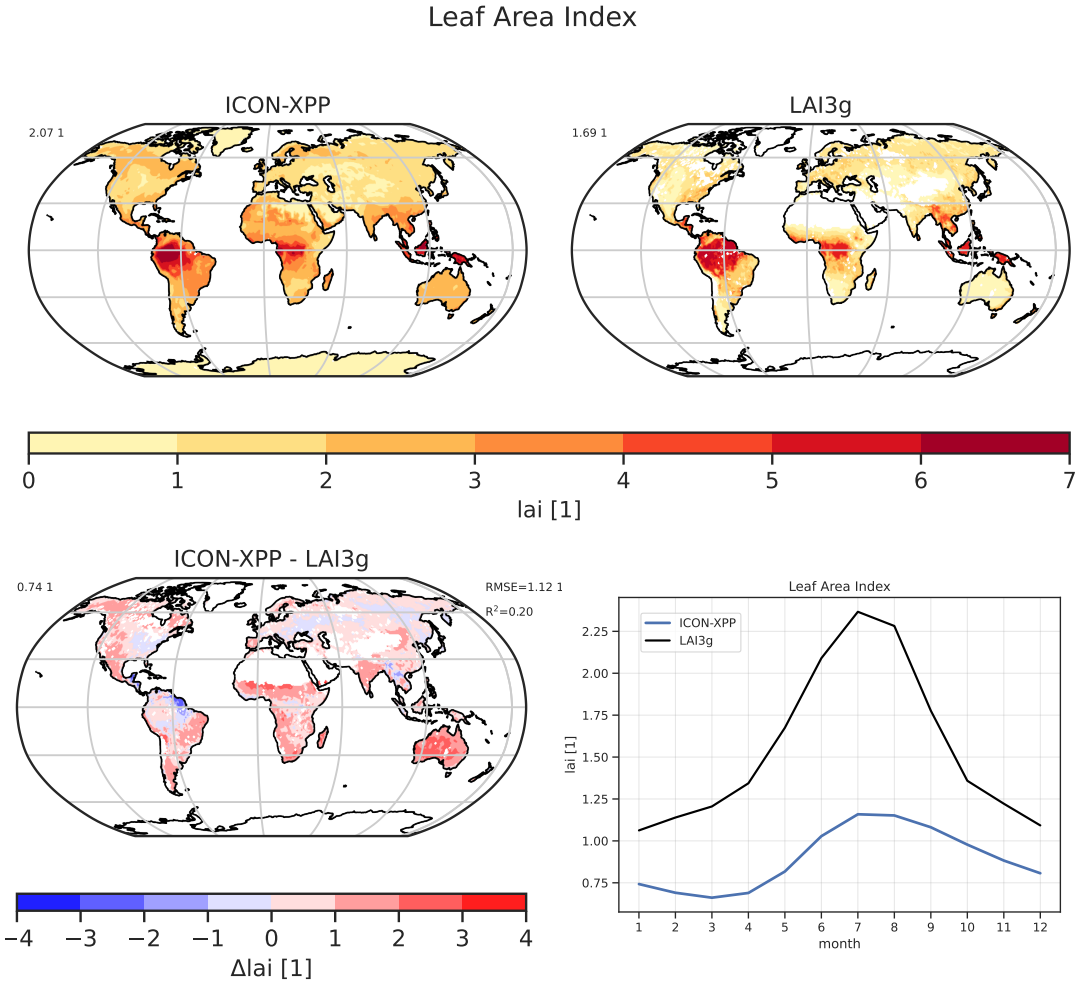

Figure 16. Average Leaf Area Index (lai) from ICON-XPP averaged over the period 1995-2004 (top left), from the LAI3g observational dataset over the same period (top right) and their difference (bottom left). The average seasonal cycles averaged over the Northern Hemisphere are shown in the bottom right.

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

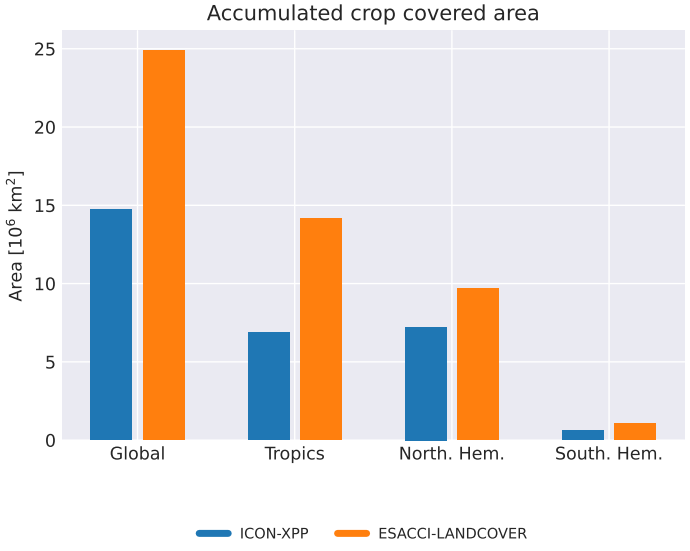

Figure 17. Total land area (in million km^2) covered by crops from ICON-XPP (blue) in comparison with the ESACCI-LANDCOVER observational dataset (orange) for the regions (from left to right) global, Tropics, Northern Hemisphere, Southern Hemisphere.

checks implemented include e.g. conservation of the total mass of air, variability of global water vapor mass with temperature (Clausius–Clapeyron), the temperature dependence of the cloud ice fraction, and tropospheric lapse rate distributions for the Tropics and mid-latitudes. The consistency checks aim at identifying potential violations of physical constraints that could be introduced when implementing new or updating existing parameterizations. The third category of assessment recipes consists of selected basic model evaluation diagnostics that can be used to assess relevant key climate parameters across atmosphere, ocean and land in historical simulations, e.g. the time period for which observations are available. The diagnostics include evaluation of climatologies, geographical distributions, seasonal and diurnal cycles, and time series averaged over selected regions. Atmospheric parameters include e.g. temperature, precipitation, radiation, and humidity. Ocean variables include sea surface temperature, sea surface salinity, sea ice area, mixed layer depth, ocean heat content and selected ENSO indices. Land variables included are leaf area index, gross primary production, evaporation, surface temperature and land cover type. For all realms it is easily possible to expand the available recipes with additional diagnostics and observations. The new ESMVal-Tool capabilities are exemplarily demonstrated by applying the new diagnostics to a historical simulation from the ICON-XPP model. In addition to the expansions of recipes and the implementation of different checks, ICONEval automates the execution of predefined ESMValTool recipe templates and can run a large number of diagnostics in parallel. This allows for producing results quickly and thus allowing multiple assessments of a models' performance also during run-time of a simulation. The ICONEval framework further provides integrated visualization and reporting capabilities, including automatically generated HTML summary pages that can be published to a web server for interactive inspection via a standard web browser and easy

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

sharing with collaborators outside the used HPC environment. This enables a continuous and user-friendly assessment of a model's performance during model development to quickly identify major problems such as unrealistic or physically inconsistent results. While initially focusing on evaluating ICON-XPP, extensions for other models are underway.

The extensions of ESMValTool together with the presented ICONEval framework provide a basis for an efficient, multi-step check for physical consistency and model performance across different ESM components tailored to the needs of modern climate model development. The implementation in ESMValTool allows for a relatively easy and straight-forward extension to additional components such as land-ice, biogeochemistry, etc. The described approach can support model developers during extension of the increasingly complex models and implementation of new parameterizations including machine learning–based schemes. This also supports emerging benchmarking efforts such as ClimateBench (Watson-Parris et al., 2022), where a consistent evaluation infrastructure is essential for placing data-driven and hybrid models on equal footing with physics-based ESMs. Sanity checks allow for quick identification of unreasonable results, physical consistency checks provide a quality control to prevent potential violations of fundamental physical laws such as conservation of mass or known thermodynamic relationships before assessing the models' performance and suitability for specific scientific applications in more detail. In combination with the benchmarking capabilities of ESMValTool for single models introduced in Lauer et al. (2025) and further extended and applied in the framwork of the CMIP7 REF (Hoffman et al., 2025), deviations of a model simulation from observational and reanalysis datasets used as reference can be put into context of other simulations with the same model but different model configurations or results from other ESMs such as the CMIP6 model ensemble.

Code and data availability. ESMValTool v2 is released under the Apache License, VERSION 2.0. The latest release of ESMValTool v2 is publicly available on Zenodo at https://doi.org/10.5281/zenodo.3401363 (Andela et al., 2026a). The source code of the ESMValCore package, which is installed as a dependency of ESMValTool v2, is also publicly available on Zenodo at https://doi.org/10.5281/zenodo.3387139 (Andela et al., 2026b). ESMValTool and ESMValCore are developed on the GitHub repositories available at https://github.com/ESMValGroup (last access: 21 April 2026). ICONEval is developed open-source on GitHub at https://github.com/EyringMLClimateGroup/ICONEval (last access: 21 April 2026). Its latest release is publicly available on Zenodo at https://doi.org/10.5281/zenodo.18937450 (Schlund and Bock, 2026). An example website that showcases parts of ICONEval's output can be found at https://swift.dkrz.de/v1/dkrz_4eefb34f-8803-415a-bd70-9c455db9a iconeval/iconeval_example/index.html (last access: 21 April 2026). The ICON-XPP model simulation used as an example in this study is carried out with the latest ICON Open Source Release available at that time (version October 2025). This is available to the public through https://gitlab.dkrz.de/icon/icon-model/-/commits/icon-2025.10-1-public?ref_type=tags (last access: 21 April 2026). CMIP6 data are available freely and publicly from the Earth System Grid Federation (ESGF) and can be retrieved by ESMValTool automatically (see https://docs.esmvaltool.org/projects/ESMValCore/en/latest/quickstart/configure.html#data-sources (last access: 21 April 2026) for detailed guidelines on this). All observations/reanalysis data used are described in Sect.2.4. The observational/reanalysis datasets are not distributed with ESMValTool, which is restricted to the code as open source software, but ESMValTool provides a collection of scripts with downloading and processing instructions to recreate all observational/reanalysis datasets used in this publication.

All data used to create the figures of this paper are available on Zenodo at https://doi.org/10.5281/zenodo.19664576 (Lauer et al., 2026).

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

Author contributions. AL, LB, BH, and MS developed the details on the key diagnostics presented in this study and contributed to the implementation of the ESMValTool extensions. VE and MS developed the concept and led the strategic development of ICONEval. MS led the technical implementation of ICONEval. Additional coding contributions were provided by BG, GB, LL, J-HM, and KW. WM, SL, TP, and GZ performed the ICON-XPP simulation and contributed to the preparation and processing of the model output used in the examples. All authors contributed to the writing and editing of the manuscript.

Competing interests. One author is a member of the editorial board of the journal Geoscientific Model Development.

Acknowledgements. The development of ESMValTool is supported by several projects. The development of ICONEval and the diagnostic development of ESMValTool v2 for this paper received funding from the European Research Council (ERC) Synergy Grant "Understanding and Modeling the Earth System with Machine Learning (USMILE) under the Horizon 2020 research and innovation programme (Grant agreement No. 855187) and additionally from European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 101003536 (ESM2025 – Earth System Models for the Future). Further support was received by the Bundesministerium für Forschung, Technologie und Raumfahrt within the project CAP7 (German Contribution to ClimAte Projections for CMIP7, grant no. 01LP2401C). The European Eddy-Rich ESMs (EERIE) project (Grant Agreement No 101081383) funded by the European Union and the Collaborative Research Centre TRR 181 "Energy Transfers in Atmosphere and Ocean" (project no. 274762653) funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) supported the development of diagnostics for ESMValTool. VE was additionally supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through the Gottfried Wilhelm Leibniz Prize awarded to VE (reference no. EY 22/2-1). We acknowledge the World Climate Research Program's (WCRP's) Working Group on Coupled Modelling (WGCM), which is responsible for CMIP, and we thank the climate modeling groups for producing and making available their model output in the framework of ESGF. The CMIP data of this study were replicated and made available for this study by the Deutsches Klimarechenzentrum (DKRZ). This work used resources of the Deutsches Klimarechenzentrum (DKRZ) granted by its Scientific Steering Committee (WLA) under project IDs bd0854 and bd1179.

This manuscript contains modified Copernicus Climate Change Service (2017) information with ERA5 and ERA-Interim data retrieved from the Climate Data Store. Data are published under a Creative Commons Attribution 4.0 International (CC-BY 4.0; https://creativecommons.org/licenses/by/4.0/). Neither the European Commission nor ECMWF is responsible for any use that may be made of the Copernicus Information or Data it contains). ECMWF does not accept any liability whatsoever for any error or omission in the data, their availability, or for any loss or damage arising from their use. CALIPSO-ICECLOUD and CERES-EBAF data were obtained from the NASA Langley Research Center Atmospheric Science Data Center. We thank the teams for their efforts in providing these data. Global Precipitation Climatology Project (GPCP) Monthly Analysis Product data used are provided by the NOAA PSL, Boulder, Colorado, USA. CLARA-AHRR data are provided by the EUMETSAT Satellite Application Facility on Climate Monitoring (CM SAF). The combined microwave and near-infrared imager based product COMBI (CM SAF/CCI TCWV-global) was initiated, funded, and provided by the Water Vapour project of the ESA CCI, with contributions from Brockmann Consult, Spectral Earth, Deutscher Wetterdienst, and CM SAF. We acknowledge the global sea ice concentration climate data record (OSI-450) provided by EUMETSAT OSI SAF, copyright EUMETSAT. CloudSat-L2 data have been obtained from the CloudSat Data Processing Center run by the Cooperative Institute for Research in the Atmosphere (CIRA) at Colorado State University. ORAS5 data were provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). The ESA Climate Change

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

Initiative (CCI) and Cloud_cci project are kindly acknowledged. Data from the ESA CCI Sea Surface Temperature project are provided via the Centre for Environmental Data Analysis (CEDA). We also acknowledge the CCI Land Cover project for providing the Maps of Plant Functional Type Fractional Cover. The MAC-LWP dataset is provided by the Goddard Earth Sciences Data and Information Services Center (GES DISC) funded by NASA's Science Mission Directorate (SMD). The AVHRR Clouds Properties (PATMOS-x) CDR was acquired from NOAA's National Climatic Data Center. This CDR was originally developed by A. Heidinger and colleagues for NOAA's CDR Program. HadCRUT5 and HadISST v1.1 data were obtained from http://www.metoffice.gov.uk/hadobs/ and are copyright British Crown Copyright, Met Office 2020, provided under an Open Government License, http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/. The ISCCP-FH Radiative Flux Profile Product used in this study was developed by Y. Zhang and W. Rossow and obtained from the NOAA National Centers for Environmental Information (NCEI). We acknowledge the Japanese 55-year Reanalysis (JRA-55) provided by the Japan Meteorological Agency (JMA). The MERRA-2 data used in this study have been provided by the Global Modeling and Assimilation Office (GMAO) at NASA Goddard Space Flight Center through the Goddard Earth Sciences Data and Information Services Center (GES DISC). The Twentieth Century Reanalysis Project dataset (NOAA-CIRES-20CR-V2) is provided by the U.S. Department of Energy, Office of Science Biological and Environmental Research (BER), the National Oceanic and Atmospheric Administration Climate Program Office, the National Energy Research Scientific Computing Center and the Oak Ridge Leadership Computing Facility. The ocean potential temperature data were provided by the Institute of Atmospheric Physics (IAP), Chinese Academy of Sciences (CAS). We thank the IAP/CAS ocean data team for providing the gridded ocean temperature and heat content products. The LAI3g data were provided by the NASA Earth Exchange (NEX). We thank the GIMMS team at Boston University for their efforts in producing and sharing this long-term global dataset. The MODIS (MYD08_M3) dataset was acquired from the Level-1 & Atmosphere Archive and Distribution System (LAADS) Distributed Active Archive Center (DAAC), located in the Goddard Space Flight Center in Greenbelt, Maryland. The MTE (Model Tree Ensemble) datasets were provided by the Max Planck Institute for Biogeochemistry (MPI-BGC). We thank the Biogeochemical Integration (BGI) department for their efforts in producing and sharing these global gridded products. Data from the RAPID AMOC observing project is funded by the Natural Environment Research Council, U.S. National Science Foundation (NSF) with support from NOAA. The TropFlux data were provided by the Indian National Centre for Ocean Information Services (ESSO-INCOIS) (INCOIS TropFlux Portal). TropFlux is a collaborative product between LOCEAN/IPSL (France) and CSIR-NIO (India), supported by the Institut de Recherche pour le Développement (IRD). World Ocean Atlas (WOA) data were provided by the NOAA National Centers for Environmental Information (NCEI). We thank the scientists, technicians, and programmers who collected and processed the ocean profile data. Support was also provided by the International Oceanographic Data and Information Exchange (IODE) and the World Data Service for Oceanography. Last but not least, we thank Franziska Winterstein (DLR) for her helpful comments on the manuscript.

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

References

Adler, R. F., Sapiano, M. R. P., Huffman, George J. nd Wang, J.-J., Gu, G., Bolvin, D., Chiu, L., chneider, U., Becker, A., Nelkin, E., Xie, P., erraro, R., and Shin, D.-B.: The Global Precipitation Climatology Project (GPCP) Monthly Analysis (New Version 2.3) and a Review of 2017 Global Precipitation, Atmosphere, 9, https://doi.org/10.3390/atmos9040138, 2018.

Andela, B., Broetz, B., de Mora, L., Drost, N., Eyring, V., Koldunov, N., Lauer, A., Mueller, B., Predoi, V., Righi, M., Schlund, M., Vegas-Regidor, J., Zimmermann, K., Adeniyi, K., Castellani, G., Arnone, E., Bellprat, O., Berg, P., Billows, C., Blockley, E., Bock, L., Bodas-Salcedo, A., Caron, L.-P., Carvalhais, N., Cionni, I., Cortesi, N., Corti, S., Crezee, B., Davin, E. L., Davini, P., Deser, C., Diblen, F., Docquier, D., Dreyer, L., Ehbrecht, C., Earnshaw, P., Geddes, T., Gier, B., Gillett, E., Gonzalez-Reviriego, N., Goodman, P., Hagemann, S., Hall, S., Hardacre, C., von Hardenberg, J., Hassler, B., Heuer, H., Hogan, E., Hunter, A., Kadow, C., Kindermann, S., Koirala, S., Kuehbacher, B., Lledó, L., Lejeune, Q., Lembo, V., Little, B., Loosveldt-Tomas, S., Lorenz, R., Lovato, T., Lucarini, V., Malinina, E., Massonnet, F., Mohr, C. W., Amarjiit, P., Parsons, N., Pérez-Zanón, N., Phillips, A., Proft, M., Russell, J., Sandstad, M., Sellar, A., Senftleben, D., Serva, F., Sillmann, J., Stacke, T., Storkey, D., Swaminathan, R., Tomkins, K., Torralba, V., Weigel, K., Sarauer, E., Schulze, K., Roberts, C., Kalverla, P., Alidoost, S., Verhoeven, S., Vreede, B., Smeets, S., Soares Siqueira, A., Kazeroni, R., Potter, J., Winterstein, F., Beucher, R., Kraft, J., Ruhe, L., Bonnet, P., Munday, G., Chun, F., and Ellis, H.: ESMValTool, https://doi.org/10.5281/zenodo.3401363, 2026a.

Andela, B., Broetz, B., de Mora, L., Drost, N., Eyring, V., Koldunov, N., Lauer, A., Predoi, V., Righi, M., Schlund, M., Vegas-Regidor, J., Zimmermann, K., Bock, L., Diblen, F., Dreyer, L., Earnshaw, P., Hassler, B., Little, B., Loosveldt-Tomas, S., Smeets, S., Camphuijsen, J., Gier, B. K., Weigel, K., Hauser, M., Kalverla, P., Galytska, E., Cos-Espuña, P., Pelupessy, I., Koirala, S., Stacke, T., Alidoost, S., Jury, M., Sénési, S., Crocker, T., Vreede, B., Soares Siqueira, A., Kazeroni, R., Hohn, D., Bauer, J., Beucher, R., Benke, J., Martin-Martinez, E., Cammarano, D., Yousong, Z., Malinina, E., Garcia Perdomo, K., and Lenhardt, J.: ESMValCore, https://doi.org/10.5281/zenodo.3387139, 2026b.

Atkinson, J. D., Murray, B. J., and O'Sullivan, D.: Rate of Homogenous Nucleation of Ice in Supercooled Water, The Journal of Physical Chemistry A, 120, https://doi.org/10.1021/acs.jpca.6b03843, 2016.

Bock, L., Lauer, A., Schlund, M., Barreiro, M., Bellouin, N., Jones, C., Meehl, G. A., Predoi, V., Roberts, M. J., and Eyring, V.: Quantifying Progress Across Different CMIP Phases With the ESMValTool, Journal of Geophysical Research: Atmospheres, 125, e2019JD032 321, https://doi.org/https://doi.org/10.1029/2019JD032321, e2019JD032321 2019JD032321, 2020.

Boer, G.: Climate change and the regulation of the surface moisture and energy budgets, Climate Dynamics, 8, 225—239, https://doi.org/10.1007/BF00198617, 1993.

Boyer, T. P., Garcia, H. E., Locarnini, R. A., Zweng, M. M., Mishonov, A. V., Reagan, J. R., Weathers, K. A., Baranova, O. K., Paver, C. R., Seidov, D., and Smolyar, I. V.: World Ocean Atlas 2018. [temperature, salinity, oxygen, nutrients], NOAA National Centers for Environmental Information, https://www.ncei.noaa.gov/archive/accession/NCEI-WOA18, accessed: 2021-03-01, 2018.

Busecke, J. J. M., Resplandy, L., Ditkovsky, S. J., and John, J. G.: Diverging Fates of the Pacific Ocean Oxygen Minimum Zone and Its Core in a Warming World, AGU Advances, 3, e2021AV000 470, https://doi.org/https://doi.org/10.1029/2021AV000470, e2021AV000470 2021AV000470, 2022.

Cheng, L., Pan, Y., Tan, Z., Zheng, H., Zhu, Y., Wei, W., Du, J., Yuan, H., Li, G., Ye, H., Gouretski, V., Li, Y., Trenberth, K. E., Abraham, J., Jin, Y., Reseghetti, F., Lin, X., Zhang, B., Chen, G., Mann, M. E., and Zhu, J.: IAPv4 ocean temperature and ocean heat content gridded dataset, Earth System Science Data, 16, 3517–3546, https://doi.org/10.5194/essd-16-3517-2024, 2024.

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

Compo, G., Whitaker, J., Sardeshmukh, P., Matsui, N., Allan, R., Yin, X., Gleason, B., Vose, R., Rutledge, G., Bessemoulin, P., Brönnimann, S., Brunet, M., Crouthamel, R., Grant, A., Groisman, P., Jones, P., Kruk, M., Kruger, A., Marshall, G., Maugeri, M., Mok, H., Nordli, O., Ross, T., Trigo, R., Wang, X., Woodruff, S., and Worley, S.: The Twentieth Century Reanalysis Project, Quarterly J. Roy. Meteorol. Soc., 137, 1–28, https://doi.org/10.1002/qj.776, 2011.

Cunningham, S. A., Kanzow, T., Rayner, D., Baringer, M. O., Johns, W. E., Marotzke, J., Longworth, H. R., Grant, E. M., Hirschi, J. J.-M., Beal, L. M., Meinen, C. S., and Bryden, H. L.: Temporal Variability of the Atlantic Meridional Overturning Circulation at 26.5°N, Science, 317, 935–938, https://doi.org/10.1126/science.1141304, 2007.

Dai, A.: The diurnal cycle from observations and ERA5 in precipitation, clouds, boundary layer height, buoyancy, and surface fluxes, Climate Dynamics, 62, 5879–5908, https://doi.org/10.1007/s00382-024-07182-6, 2024.

Danabasoglu, G., Yeager, S. G., Bailey, D., Behrens, E., Bentsen, M., Bi, D., Biastoch, A., Böning, C., Bozec, A., Canuto, V. M., Cassou, C., Chassignet, E., Coward, A. C., Danilov, S., Diansky, N., Drange, H., Farneti, R., Fernandez, E., Fogli, P. G., Forget, G., Fujii, Y., Griffies, S. M., Gusev, A., Heimbach, P., Howard, A., Jung, T., Kelley, M., Large, W. G., Leboissetier, A., Lu, J., Madec, G., Marsland, S. J., Masina, S., Navarra, A., George Nurser, A., Pirani, A., y Mélia, D. S., Samuels, B. L., Scheinert, M., Sidorenko, D., Treguier, A.-M., Tsujino, H., Uotila, P., Valcke, S., Voldoire, A., and Wang, Q.: North Atlantic simulations in Coordinated Ocean-ice Reference Experiments phase II (CORE-II). Part I: Mean states, Ocean Modelling, 73, 76–107, https://doi.org/https://doi.org/10.1016/j.ocemod.2013.10.005, 2014.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Quarterly Journal of the Royal Meteorological Society, 137, 553–597, https://doi.org/https://doi.org/10.1002/qj.828, 2011.

Dunne, J. P., Hewitt, H. T., Arblaster, J. M., Bonou, F., Boucher, O., Cavazos, T., Dingley, B., Durack, P. J., Hassler, B., Juckes, M., Miyakawa, T., Mizielinski, M., Naik, V., Nicholls, Z., O'Rourke, E., Pincus, R., Sanderson, B. M., Simpson, I. R., and Taylor, K. E.: An evolving Coupled Model Intercomparison Project phase 7 (CMIP7) and Fast Track in support of future climate assessment, Geoscientific Model Development, 18, 6671–6700, https://doi.org/10.5194/gmd-18-6671-2025, 2025.

Elsaesser, G. S., O'Dell, C. W., Lebsock, M. D., Bennartz, R., Greenwald, T. J., and Wentz, F. J.: The Multisensor Advanced Climatology of Liquid Water Path (MAC-LWP), Journal of Climate, 30, 10 193–10 210, https://doi.org/10.1175/JCLI-D-16-0902.1, 2017.

Embury, O., Merchant, C. J., Good, S. A., Rayner, N. A., Høyer, J. L., Atkinson, C., Block, T., Alerskans, E., Pearson, K. J., Worsfold, M., McCarroll, N., and Donlon, C.: Satellite-based time-series of sea-surface temperature since 1980 for climate applications, Scientific Data, 11, 5419–5454, https://doi.org/10.1038/s41597-024-03147-w, 2024.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geoscientific Model Development, 9, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, 2016.

Eyring, V., Bock, L., Lauer, A., Righi, M., Schlund, M., Andela, B., Arnone, E., Bellprat, O., Brötz, B., Caron, L.-P., Carvalhais, N., Cionni, I., Cortesi, N., Crezee, B., Davin, E. L., Davini, P., Debeire, K., de Mora, L., Deser, C., Docquier, D., Earnshaw, P., Ehbrecht, C., Gier, B. K., Gonzalez-Reviriego, N., Goodman, P., Hagemann, S., Hardiman, S., Hassler, B., Hunter, A., Kadow, C., Kindermann, S., Koirala, S., Koldunov, N., Lejeune, Q., Lembo, V., Lovato, T., Lucarini, V., Massonnet, F., Müller, B., Pandde, A., Pérez-Zanón, N., Phillips, A., Predoi, V., Russell, J., Sellar, A., Serva, F., Stacke, T., Swaminathan, R., Torralba, V., Vegas-Regidor, J., von Hardenberg, J., Weigel,

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

K., and Zimmermann, K.: Earth System Model Evaluation Tool (ESMValTool) v2.0 – an extended set of large-scale diagnostics for quasi-operational and comprehensive evaluation of Earth system models in CMIP, Geoscientific Model Development, 13, 3383–3438, https://doi.org/10.5194/gmd-13-3383-2020, 2020.

Eyring, V., Gillett, N., Achuta Rao, K., Barimalala, R., Barreiro Parrillo, M., Bellouin, N., Cassou, C., Durack, P., Kosaka, Y., McGregor, S., Min, S., Morgenstern, O., and Sun, Y.: Human Influence on the Climate System, in: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, edited by Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J. B. R., Maycock, T. K., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., book section 3, pp. 423–551, Cambridge University Press, Cambridge, UK and New York, NY, USA, https://doi.org/10.1017/9781009157896.005, 2021.

Eyring, V., Collins, W., Gentine, P., Barnes, E., Barreiro, M., Beucler, T., Bocquet, M., Bretherton, C., Christensen, H., Dagon, K., Gagne, D., Hall, D., Hammerling, D., Hoyer, S., Iglesias-Suarez, F., Lopez-Gomez, I., McGraw, M., Meehl, G., Molina, M., Monteleoni, C., Mueller, J., Pritchard, M., Rolnick, D., Runge, J., Stier, P., Watt-Meyer, O., Weigel, K., Yu, R., and Zanna, L.: Pushing the frontiers in climate modelling and analysis with machine learning, Nature Climate Change, 14, 916—928, https://doi.org/10.1038/s41558-024-02095-y, 2024a.

Eyring, V., Gentine, P., Camps-Valls, G., Lawrence, D. M., and Reichstein, M.: AI-empowered next-generation multiscale climate modelling for mitigation and adaptation, Nature Geoscience, 17, 963–971, https://doi.org/10.1038/s41561-024-01527-w, 2024b.

Fox-Kemper, B., Hewitt, H., Xiao, C., Aðalgeirsdóttir, G., Drijfhout, S., Edwards, T., Golledge, N., Hemer, M., Kopp, R., Krinner, G., Mix, A., Notz, D., Nowicki, S., Nurhati, I., Ruiz, L., Sallée, J.-B., Slangen, A., and Yu, Y.: Ocean, Cryosphere and Sea Level Change, in: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, edited by Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J. B. R., Maycock, T. K., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., pp. 1211–1361, Cambridge University Press, Cambridge, UK and New York, NY, USA, https://doi.org/10.1017/9781009157896.011, section: 9 Type: Book Section, 2021.

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., da Silva, A. M., Gu, W., Kim, G.-K., Koster, R., Lucchesi, R., Merkova, D., Nielsen, J. E., Partyka, G., Pawson, S., Putman, W., Rienecker, M., Schubert, S. D., Sienkiewicz, M., and Zhao, B.: The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2), Journal of Climate, 30, 5419–5454, https://doi.org/10.1175/JCLI-D-16-0758.1, 2017.

Gier, B. K., Buchwitz, M., Reuter, M., Cox, P. M., Friedlingstein, P., and Eyring, V.: Spatially resolved evaluation of Earth system models with satellite column-averaged CO_2, Biogeosciences, 17, 6115–6144, https://doi.org/10.5194/bg-17-6115-2020, 2020.

Good, S. A., Martin, M. J., and Rayner, N. A.: EN4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates, Journal of Geophysical Research: Oceans, 118, 6704–6716, https://doi.org/https://doi.org/10.1002/2013JC009067, 2013.

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.

Harper, K. L., Lamarche, C., Hartley, A., Peylin, P., Ottlé, C., Bastrikov, V., San Martín, R., Bohnenstengel, S. I., Kirches, G., Boettcher, M., Shevchuk, R., Brockmann, C., and Defourny, P.: A 29-year time series of annual 300m resolution plant-functional-type maps for climate models, Earth System Science Data, 15, 1465–1499, https://doi.org/10.5194/essd-15-1465-2023, 2023.

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

Hassler, B., Hoffman, F., Beadling, R., Blockley, E., Huang, B., Lee, J., Lembo, V., Lewis, J., Lu, J., Madaus, L., Malinina, E., Medeiros, B., Pokam, W., Scoccimarro, E., and Swaminathan, R.: Systematic Benchmarking of Climate Models: Methodologies, Applications, and New Directions, Reviews of Geophysics, 64, e2025RG000 891, https://doi.org/https://doi.org/10.1029/2025RG000891, 2026.

Heidinger, A. K., Foster, M. J., Walther, A., and Zhao, X. T.: The Pathfinder Atmospheres–Extended AVHRR Climate Dataset, Bulletin of the American Meteorological Society, 95, 909–922, https://doi.org/10.1175/BAMS-D-12-00246.1, 2014.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/https://doi.org/10.1002/qj.3803, 2020.

Heuer, H., Schwabe, M., Gentine, P., Giorgetta, M. A., and Eyring, V.: Interpretable Multiscale Machine Learning-Based Parameterizations of Convection for ICON, Journal of Advances in Modeling Earth Systems, 16, e2024MS004 398, https://doi.org/https://doi.org/10.1029/2024MS004398, e2024MS004398 2024MS004398, 2024.

Hoffman, F. M., Hassler, B., Swaminathan, R., Lewis, J., Andela, B., Collier, N., Hegedűs, D., Lee, J., Pascoe, C., Pflüger, M., Stockhause, M., Ullrich, P., Xu, M., Bock, L., Chun, F., Gier, B. K., Kelley, D. I., Lauer, A., Lenhardt, J., Schlund, M., Sreeush, M. G., Weigel, K., Blockley, E., Beadling, R., Beucher, R., Dugassa, D. D., Lembo, V., Lu, J., Brands, S., Tjiputra, J., Malinina, E., Mederios, B., Scoccimarro, E., Walton, J., Kershaw, P., Marquez, A. L., Roberts, M. J., O'Rourke, E., Dingley, E., Turner, B., Hewitt, H., and Dunne, J. P.: Rapid Evaluation Framework for the CMIP7 Assessment Fast Track, EGUsphere, 2025, 1–57, https://doi.org/10.5194/egusphere-2025-2685, 2025.

IPCC: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, vol. In Press, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, https://doi.org/10.1017/9781009157896, 2021.

Jung, M., Reichstein, M., Margolis, H. A., Cescatti, A., Richardson, A. D., Arain, M. A., Arneth, A., Bernhofer, C., Bonal, D., Chen, J., Gianelle, D., Gobron, N., Kiely, G., Kutsch, W., Lasslop, G., Law, B. E., Lindroth, A., Merbold, L., Montagnani, L., Moors, E. J., Papale, D., Sottocornola, M., Vaccari, F., and Williams, C.: Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations, Journal of Geophysical Research: Biogeosciences, 116, https://doi.org/10.1029/2010jg001566, 2011.

Karlsson, K.-G., Anttila, K., Trentmann, J., Stengel, M., Fokke Meirink, J., Devasthale, A., Hanschmann, T., Kothe, S., Jääskeläinen, E., Sedlar, J., Benas, N., van Zadelhoff, G.-J., Schlundt, C., Stein, D., Finkensieper, S., Håkansson, N., and Hollmann, R.: CLARA-A2: the second edition of the CM SAF cloud and radiation data record from 34 years of global AVHRR data, Atmospheric Chemistry and Physics, 17, 5809–5828, https://doi.org/10.5194/acp-17-5809-2017, 2017.

Karlsson, K.-G., Anttila, K., Trentmann, J., Stengel, M., Solodovnik, I., Meirink, J. F., Devasthale, A., Kothe, S., Jääskeläinen, E., Sedlar, J., Benas, N., van Zadelhoff, G.-J., Stein, D., Finkensieper, S., Håkansson, N., Hollmann, R., Kaiser, J., and Werscheck, M.: CLARA-A2.1: CM SAF cLoud, Albedo and surface RAdiation dataset from AVHRR data - Edition 2.1, https://doi.org/10.5676/EUM_SAF_CM/CLARA_AVHRR/V002_01, 2020.

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., Onogi, K., Kamahori, H., Kobayashi, C., Endo, H., Miyaoka, K., and Takahashi, K.: The JRA-55 Reanalysis: General Specifications and Basic Characteristics, Journal of the Meteorological Society of Japan. Ser. II, 93, 5–48, https://doi.org/10.2151/jmsj.2015-001, 2015.

Kramer, S. M., Karnauskas, K. B., Elling, M. T., Zhang, L., Liu, H., Chen, Y., Amaya, D. J., Nazarenko, L., Yang, W., Vecchi, G. A., Meegan-Kumar, D., Baldwin, J. W., and Samanta, D.: ColdBlobMIP: A Multi-Model Assessment of the Atmospheric Response to the North Atlantic Warming Hole, Geophysical Research Letters, 52, e2025GL117 784, https://doi.org/https://doi.org/10.1029/2025GL117784, e2025GL117784 2025GL117784, 2025.

Lauer, A., Eyring, V., Bellprat, O., Bock, L., Gier, B. K., Hunter, A., Lorenz, R., Pérez-Zanón, N., Righi, M., Schlund, M., Senftleben, D., Weigel, K., and Zechlau, S.: Earth System Model Evaluation Tool (ESMValTool) v2.0 – diagnostics for emergent constraints and future projections from Earth system models in CMIP, Geoscientific Model Development, 13, 4205–4228, https://doi.org/10.5194/gmd-13-4205-2020, 2020.

Lauer, A., Bock, L., Hassler, B., Jöckel, P., Ruhe, L., and Schlund, M.: Monitoring and benchmarking Earth system model simulations with ESMValTool v2.12.0, Geoscientific Model Development, 18, 1169–1188, https://doi.org/10.5194/gmd-18-1169-2025, 2025.

Lauer, A., Schlund, M., Bock, L., Hassler, B., Behrens, G., Gier, B., Lindenlaub, L., Lorenz, S., Malles, J.-H., Müller, W. A., van Pham, T., Weigel, K., Zeng, G., and Eyring, V.: An ESMValTool-based framework for sanity checks, physical consistency and climate fidelity during model development – ICONEval v1.0, https://doi.org/10.5281/zenodo.19664576, 2026.

Llort, J., Lévy, M., Sallée, J. B., and Tagliabue, A.: Nonmonotonic Response of Primary Production and Export to Changes in Mixed-Layer Depth in the Southern Ocean, Geophysical Research Letters, 46, 3368–3377, https://doi.org/10.1029/2018gl081788, 2019.

Loeb, N. G., Wielicki, B. A., Doelling, D. R., Smith, G. L., Keyes, D. F., Kato, S., Manalo-Smith, N., and Wong, T.: Toward Optimal Closure of the Earth's Top-of-Atmosphere Radiation Budget, Journal of Climate, 22, 748–766, https://doi.org/10.1175/2008JCLI2637.1, 2009.

Loeb, N. G., Lyman, J. M., Johnson, G. C., Allan, R. P., Doelling, D. R., Wong, T., Soden, B. J., and Stephens, G. L.: Observed changes in top-of-the-atmosphere radiation and upper-ocean heating consistent within uncertainty, Nature Geoscience, 5, 110–113, https://doi.org/10.1038/ngeo1375, 2012.

Marchand, R., Mace, G. G., Ackerman, T., and Stephens, G.: Hydrometeor Detection Using Cloudsat—An Earth-Orbiting 94-GHz Cloud Radar, Journal of Atmospheric and Oceanic Technology, 25, 519 – 533, https://doi.org/10.1175/2007JTECHA1006.1, 2008.

McKay, D. I. A., Staal, A., Abrams, J. F., Winkelmann, R., Sakschewski, B., Loriani, S., Fetzer, I., Cornell, S. E., Rockström, J., and Lenton, T. M.: Exceeding 1.5°C global warming could trigger multiple climate tipping points, Science, 377, eabn7950, https://doi.org/10.1126/science.abn7950, 2022.

MDTF, N. M. D. T. F.: Phase and Amplitude of Precipitation Diurnal Cycle, https://mdtf-diagnostics.readthedocs.io/en/latest/sphinx_pods/precip_diurnal_cycle.html, last access: 17 March 2026, 2019.

Meehl, G. A., Senior, C. A., Eyring, V., Flato, G., Lamarque, J.-F., Stouffer, R. J., Taylor, K. E., and Schlund, M.: Context for interpreting equilibrium climate sensitivity and transient climate response from the CMIP6 Earth system models, Science Advances, 6, eaba1981, https://doi.org/10.1126/sciadv.aba1981, 2020.

Moat, B., Smeed, D., Rayner, D., Johns, W., Smith, R., Volkov, D., Elipot, S., Petit, T., Kajtar, J., Baringer, M., and Collins, J.: Atlantic meridional overturning circulation observed by the RAPID-MOCHA-WBTS (RAPID-Meridional Overturning Circulation and Heatflux Array-Western Boundary Time Series) array at 26N from 2004 to 2024 (v2024.1a)., https://doi.org/10.5285/48d0bf43-0598-ceb2-e063-7086abc062f1, 2026.

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

Morice, C. P., Kennedy, J. J., Rayner, N. A., Winn, J. P., Hogan, E., Killick, R. E., Dunn, R. J. H., Osborn, T. J., Jones, P. D., and Simpson, I. R.: An Updated Assessment of Near-Surface Temperature Change From 1850: The HadCRUT5 Data Set, Journal of Geophysical Research: Atmospheres, 126, e2019JD032 361, https://doi.org/https://doi.org/10.1029/2019JD032361, e2019JD032361 2019JD032361, 2021.

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.

Müller, W. A., Früh, B., Korn, P., Potthast, R., Baehr, J., Bettems, J.-M., Bölöni, G., Brienen, S., Fröhlich, K., Helmert, J., Jungclaus, J., Köhler, M., Lorenz, S., Schneidereit, A., Schnur, R., Schulz, J.-P., Schlemmer, L., Sgoff, C., Pham, T. V., Pohlmann, H., Vogel, B., Vogel, H., Wirth, R., Zaehle, S., Zängl, G., Stevens, B., and Marotzke, J.: ICON: Toward Vertically Integrated Model Configurations for Numerical Weather Prediction, Climate Predictions, and Projections, Bulletin of the American Meteorological Society, 106, E1017 – E1031, https://doi.org/10.1175/BAMS-D-24-0042.1, 2025.

NASA/LARC/SD/ASDC: CALIPSO Lidar Level 3 Ice Cloud Data, Standard V1-00, https://doi.org/10.5067/CALIOP/CALIPSO/L3_ICE_CLOUD-STANDARD-V1-00, 2018.

NASA/LARC/SD/ASDC: CERES Energy Balanced and Filled (EBAF) TOA Monthly means data in netCDF Edition4.2, https://doi.org/10.5067/TERRA-AQUA-NOAA20/CERES/EBAF-TOA_L3B004.2, 2022.

OSI SAF: Global Sea Ice Concentration Climate Data Record v3.0 - Multimission, https://doi.org/10.15770/EUM_SAF_OSI_0013, 2022.

Park, H. and Jeong, S.: Leaf area index in Earth system models: how the key variable of vegetation seasonality works in climate projections, Environmental Research Letters, 16, 034 027, https://doi.org/10.1088/1748-9326/abe2cf, 2021.

Platnick, S., King, M., Ackerman, S., Menzel, W., Baum, B., Riedi, J., and Frey, R.: The MODIS cloud products: algorithms and examples from Terra, IEEE Transactions on Geoscience and Remote Sensing, 41, 459–473, https://doi.org/10.1109/TGRS.2002.808301, 2003.

Praveen Kumar, B., Vialard, J., Lengaigne, M., Murty, V. S. N., and McPhaden, M. J.: TropFlux: air-sea fluxes for the global tropical oceans—description and evaluation, Climate Dynamics, 38, 1521–1543, https://doi.org/10.1007/s00382-011-1115-0, 2012.

Praveen Kumar, B., Vialard, J., Lengaigne, M., Murty, V. S. N., McPhaden, M. J., Cronin, M. F., Pinsard, F., and Gopala Reddy, K.: TropFlux wind stresses over the tropical oceans: evaluation and comparison with other products, Climate Dynamics, 40, 2049–2071, https://doi.org/10.1007/s00382-012-1455-4, 2013.

Rackow, T., Danilov, S., Goessling, H. F., Hellmer, H. H., Sein, D. V., Semmler, T., Sidorenko, D., and Jung, T.: Delayed Antarctic sea-ice decline in high-resolution climate change simulations, Nature Communications, 13, https://doi.org/10.1038/s41467-022-28259-y, 2022.

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., Kent, E. C., and Kaplan, A.: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century, Journal of Geophysical Research: Atmospheres, 108, https://doi.org/https://doi.org/10.1029/2002JD002670, 2003.

Righi, M., Andela, B., Eyring, V., Lauer, A., Predoi, V., Schlund, M., Vegas-Regidor, J., Bock, L., Brötz, B., de Mora, L., Diblen, F., Dreyer, L., Drost, N., Earnshaw, P., Hassler, B., Koldunov, N., Little, B., Loosveldt Tomas, S., and Zimmermann, K.: Earth System Model Evaluation Tool (ESMValTool) v2.0 – technical overview, Geoscientific Model Development, 13, 1179–1199, https://doi.org/10.5194/gmd-13-1179-2020, 2020.

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

Roach, L. A., Dörr, J., Holmes, C. R., Massonnet, F., Blockley, E. W., Notz, D., Rackow, T., Raphael, M. N., O'Farrell, S. P., Bailey, D. A., and Bitz, C. M.: Antarctic Sea Ice Area in CMIP6, Geophysical Research Letters, 47, https://doi.org/10.1029/2019gl086729, 2020.

Rossow, W. B., Walker, A., Golea, V., Knapp, K. R., Young, A., Inamdar, A., Hankins, B., and Program, N. C. D. R.: International Satellite Cloud Climatology Project Climate Data Record, H-Series, https://doi.org/doi:10.7289/V5QZ281S, 2016.

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.

Schlund, M. and Bock, L.: ICONEval, https://doi.org/10.5281/zenodo.18937451, 2026.

Schlund, M., Hassler, B., Lauer, A., Andela, B., Jöckel, P., Kazeroni, R., Loosveldt Tomas, S., Medeiros, B., Predoi, V., Sénési, S., Servonnat, J., Stacke, T., Vegas-Regidor, J., Zimmermann, K., and Eyring, V.: Evaluation of native Earth system model output with ESMValTool v2.6.0, Geoscientific Model Development, 16, 315–333, https://doi.org/10.5194/gmd-16-315-2023, 2023.

Schlund, M., Andela, B., Benke, J., Comer, R., Hassler, B., Hogan, E., Kalverla, P., Lauer, A., Little, B., Loosveldt Tomas, S., Nattino, F., Peglar, P., Predoi, V., Smeets, S., Worsley, S., Yeo, M., and Zimmermann, K.: Advanced climate model evaluation with ESMValTool v2.11.0 using parallel, out-of-core, and distributed computing, Geoscientific Model Development, 18, 4009–4021, https://doi.org/10.5194/gmd-18-4009-2025, 2025.

Schröder, M., Danne, O., Falk, U., Niedorf, A., Preusker, R., Trent, T., Brockmann, C., Fischer, J., Hegglin, M., Hollmann, R., and Pinnock, S.: A combined high resolution global TCWV product from microwave and near infrared imagers - COMBI, https://doi.org/10.5676/EUM_SAF_CM/COMBI/V001, 2023.

Stengel, M., Stapelberg, S., Sus, O., Finkensieper, S., Würzler, B., Philipp, D., Hollmann, R., Poulsen, C., Christensen, M., and McGarragh, G.: Cloud_cci Advanced Very High Resolution Radiometer post meridiem (AVHRR-PM) dataset version 3: 35-year climatology of global cloud and radiation properties, Earth System Science Data, 12, 41–60, https://doi.org/10.5194/essd-12-41-2020, 2020.

Stephens, G., Winker, D., Pelon, J., Trepte, C., Vane, D., Yuhas, C., L'Ecuyer, T., and Lebsock, M.: CloudSat and CALIPSO within the A-Train: Ten Years of Actively Observing the Earth System, Bulletin of the American Meteorological Society, 99, 569–581, https://doi.org/10.1175/BAMS-D-16-0324.1, 2018.

Stephens, G. L., Vane, D. G., Boain, R. J., Mace, G. G., Sassen, K., Wang, Z., Illingworth, A. J., O'connor, E. J., Rossow, W. B., Durden, S. L., Miller, S. D., Austin, R. T., Benedetti, A., Mitrescu, C., and Team, C. S.: THE CLOUDSAT MISSION AND THE A-TRAIN: A New Dimension of Space-Based Observations of Clouds and Precipitation, Bulletin of the American Meteorological Society, 83, 1771–1790, https://doi.org/10.1175/BAMS-83-12-1771, 2002.

Stone, P. H. and Carlson, J. H.: Atmospheric Lapse Rate Regimes and Their Parameterization, Journal of Atmospheric Sciences, 36, 415 – 423, https://doi.org/10.1175/1520-0469(1979)036<0415:ALRRAT>2.0.CO;2, 1979.

Tebaldi, C., Debeire, K., Eyring, V., Fischer, E., Fyfe, J., Friedlingstein, P., Knutti, R., Lowe, J., O'Neill, B., Sanderson, B., van Vuuren, D., Riahi, K., Meinshausen, M., Nicholls, Z., Tokarska, K. B., Hurtt, G., Kriegler, E., Lamarque, J.-F., Meehl, G., Moss, R., Bauer, S. E., Boucher, O., Brovkin, V., Byun, Y.-H., Dix, M., Gualdi, S., Guo, H., John, J. G., Kharin, S., Kim, Y., Koshiro, T., Ma, L., Olivié, D., Panickal, S., Qiao, F., Rong, X., Rosenbloom, N., Schupfner, M., Séférian, R., Sellar, A., Semmler, T., Shi, X., Song, Z., Steger, C., Stouffer, R., Swart, N., Tachiiri, K., Tang, Q., Tatebe, H., Voldoire, A., Volodin, E., Wyser, K., Xin, X., Yang, S., Yu, Y., and Ziehn, T.: Climate model projections from the Scenario Model Intercomparison Project (ScenarioMIP) of CMIP6, Earth System Dynamics, 12, 253–293, https://doi.org/10.5194/esd-12-253-2021, 2021.

Tian, B. and Dong, X.: The Double-ITCZ Bias in CMIP3, CMIP5, and CMIP6 Models Based on Annual Mean Precipitation, Geophysical Research Letters, 47, e2020GL087 232, https://doi.org/https://doi.org/10.1029/2020GL087232, e2020GL087232 2020GL087232, 2020.

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

Treguier, A. M., de Boyer Montégut, C., Bozec, A., Chassignet, E. P., Fox-Kemper, B., McC. Hogg, A., Iovino, D., Kiss, A. E., Le Sommer, J., Li, Y., Lin, P., Lique, C., Liu, H., Serazin, G., Sidorenko, D., Wang, Q., Xu, X., and Yeager, S.: The mixed-layer depth in the Ocean Model Intercomparison Project (OMIP): impact of resolving mesoscale eddies, Geoscientific Model Development, 16, 3849–3872, https://doi.org/10.5194/gmd-16-3849-2023, 2023.

Trenberth, K. E. and Smith, L.: The Mass of the Atmosphere: A Constraint on Global Analyses, Journal of Climate, 18, 864 – 875, https://doi.org/10.1175/JCLI-3299.1, 2005.

Trenberth, K. E., Smith, L., Qian, T., Dai, A., and Fasullo, J.: Estimates of the Global Water Budget and Its Annual Cycle Using Observational and Model Data, Journal of Hydrometeorology, 8, 758 – 769, https://doi.org/10.1175/JHM600.1, 2007.

Uchida, T., Bodner, A., Reichl, B. G., Adcroft, A. J., Fox-Kemper, B., Ilicak, M., Bentsen, M., Marques, G. M., and Large, W. G.: Representation of Surface Mixed-Layer Eddies Affects the Large-Scale Ventilation of the Global Ocean, Geophysical Research Letters, 53, e2025GL116 872, https://doi.org/https://doi.org/10.1029/2025GL116872, e2025GL116872 2025GL116872, 2026.

Waliser, D. E., Li, J.-L. F., Woods, C. P., Austin, R. T., Bacmeister, J., Chern, J., Del Genio, A., Jiang, J. H., Kuang, Z., Meng, H., Minnis, P., Platnick, S., Rossow, W. B., Stephens, G. L., Sun-Mack, S., Tao, W.-K., Tompkins, A. M., Vane, D. G., Walker, C., and Wu, D.: Cloud ice: A climate model challenge with signs and expectations of progress, Journal of Geophysical Research: Atmospheres, 114, https://doi.org/10.1029/2008JD010015, 2009.

Wan, N., Lin, X., Pielke Sr., R. A., Zeng, X., and Nelson, A. M.: Global total precipitable water variations and trends over the period 1958–2021, Hydrology and Earth System Sciences, 28, 2123–2137, https://doi.org/10.5194/hess-28-2123-2024, 2024.

Watson-Parris, D., Rao, Y., Olivié, D., Seland, Ø., Nowack, P., Camps-Valls, G., Stier, P., Bouabid, S., Dewey, M., Fons, E., Gonzalez, J., Harder, P., Jeggle, K., Lenhardt, J., Manshausen, P., Novitasari, M., Ricard, L., and Roesch, C.: ClimateBench v1.0: A Benchmark for Data-Driven Climate Projections, Journal of Advances in Modeling Earth Systems, 14, e2021MS002 954, https://doi.org/https://doi.org/10.1029/2021MS002954, e2021MS002954 2021MS002954, 2022.

Weigel, K., Bock, L., Gier, B. K., Lauer, A., Righi, M., Schlund, M., Adeniyi, K., Andela, B., Arnone, E., Berg, P., Caron, L.-P., Cionni, I., Corti, S., Drost, N., Hunter, A., Lledó, L., Mohr, C. W., Paçal, A., Pérez-Zanón, N., Predoi, V., Sandstad, M., Sillmann, J., Sterl, A., Vegas-Regidor, J., von Hardenberg, J., and Eyring, V.: Earth System Model Evaluation Tool (ESMValTool) v2.0 – diagnostics for extreme events, regional and impact evaluation, and analysis of Earth system models in CMIP, Geoscientific Model Development, 14, 3159–3184, https://doi.org/10.5194/gmd-14-3159-2021, 2021.

Wild, M.: The global energy balance as represented in CMIP6 climate models, Climate Dynamics, 55, 553—-577, https://doi.org/10.1007/s00382-020-05282-7, 2020.

Young, A. H., Knapp, K. R., Inamdar, A., Hankins, W., and Rossow, W. B.: The International Satellite Cloud Climatology Project H-Series climate data record product, Earth System Science Data, 10, 583–593, https://doi.org/10.5194/essd-10-583-2018, 2018.

Zhang, R., Sutton, R., Danabasoglu, G., Kwon, Y.-O., Marsh, R., Yeager, S. G., Amrhein, D. E., and Little, C. M.: A Review of the Role of the Atlantic Meridional Overturning Circulation in Atlantic Multidecadal Variability and Associated Climate Impacts, Reviews of Geophysics, 57, 316–375, https://doi.org/https://doi.org/10.1029/2019RG000644, 2019.

Zhu, Z., Bi, J., Pan, Y., Ganguly, S., Anav, A., Xu, L., Samanta, A., Piao, S., Nemani, R., and Myneni, R.: Global Data Sets of Vegetation Leaf Area Index (LAI)3g and Fraction of Photosynthetically Active Radiation (FPAR)3g Derived from Global Inventory Modeling and Mapping Studies (GIMMS) Normalized Difference Vegetation Index (NDVI3g) for the Period 1981 to 2011, Remote Sensing, 5, 927–948, https://doi.org/10.3390/rs5020927, 2013.

https://doi.org/10.5194/egusphere-2026-2288
Preprint. Discussion started: 1 June 2026

Zuo, H., Balmaseda, M. A., Tietsche, S., Mogensen, K., and Mayer, M.: The ECMWF operational ensemble reanalysis–analysis system for ocean and sea ice: a description of the system and assessment, Ocean Science, 15, 779–808, https://doi.org/10.5194/os-15-779-2019, 2019.