https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

Monitoring and benchmarking Earth system model simulations with ESMValTool v2.12.0

Axel Lauer[1], Lisa Bock[1], Birgit Hassler[1], Patrick Jöckel[1], Lukas Ruhe[2], Manuel Schlund[1]

[1]Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany
[2]University of Bremen, Institute of Environmental Physics (IUP), Bremen, Germany

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

Abstract. Earth system models (ESMs) are important tools to improve our understanding of present-day climate and to project climate change under different plausible future scenarios. For this, ESMs are continuously improved and extended resulting in more complex models. Particularly during the model development phase, it is important to continuously monitor how well the historical climate is reproduced and to systematically analyze, evaluate, understand, and document possible shortcomings. For this, putting model biases relative to observations into the context of deviations shown by other state-of-the-art models greatly helps to assess which biases need to be addressed with higher priority. Here, we introduce the new capability of the open-source community-developed Earth System Model Evaluation Tool (ESMValTool) to monitor running or benchmark existing simulations with observations in the context of results from the Coupled Model Intercomparison Project (CMIP). To benchmark model output, ESMValTool calculates metrics such as the root-mean-square error, the Pearson correlation coefficient, or the Earth mover's distance relative to reference datasets. This is directly compared to the same metric calculated for an ensemble of models such as the one provided by CMIP6, which provides a statistical measure for the range of values that can be considered typical for state-of-the-art ESMs. Results are displayed in different types of plots such as map plots or time series with different techniques such as stippling (maps) or shading (time series) used to visualize the typical range of values for a given metric from the model ensemble used for comparison. Automatic downloading of CMIP results from the Earth System Grid Federation (ESGF) makes application of ESMValTool for benchmarking of individual model simulations, for example in preparation of CMIP7, easy and very user friendly.

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

1 Introduction

Earth System Models (ESMs) are complex numerical representations of the Earth system including not only interactions between the physical components such as atmospheric, oceanic, land and sea ice dynamics, but also climate relevant chemical and biological processes. In the past years, ESMs became essential tools to better understand the human impact on the climate system and to project future climate change under different emission scenarios.

For this, ESMs are continuously developed and improved with new processes added and existing processes described in more detail. As with any model development activity, a thorough evaluation of new model results is a fundamental prerequisite to assess the model's performance, and thus the model's suitability for a given scientific application (fitness for purpose). Evaluating ESMs has become quite complex as there is a growing multitude of relevant parameters from different Earth system components that typically require a team of scientists with different expertise to fully assess all details. Furthermore, evaluation of some parameters such as, for instance, biogeochemical components might suffer from a lack of global observations that are suitable for a comparison with the model results as these are very hard to obtain.

One possibility to quickly assess deviations from observations present in a new model simulation is to put them into perspective by comparing the biases with the ones obtained from a large number of other state-of-the-art ESMs. For this, for example results from the Coupled Model Intercomparison Phase 5 and 6 (Eyring et al., 2016; Taylor et al., 2012) can be used to get an overview of which biases can be considered "acceptable for now", and which would need more attention and more detailed analysis and comparisons with observations. The same approach can be used to monitor running model simulations to identify significant problems early on. In this article, we exemplary demonstrate the new capability of the Earth System Model Evaluation Tool (ESMValTool) to obtain a broad overview by benchmarking a given model simulation with CMIP results using different relevant diagnostics such as climatologies, seasonal and diurnal cycles, or geographical distributions. The examples are meant as a starting point and can be extended easily and applied to different components of the Earth system.

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

2 Methods

2.1 Earth System Model Evaluation Tool

The Earth System Model Evaluation Tool (ESMValTool) is an open-source community-developed diagnostics and performance metrics tool for the evaluation and analysis of Earth system models with Earth observations (Righi et al., 2020; Eyring et al., 2020; Lauer et al., 2020; Weigel et al., 2021). ESMValTool has been developed into a well-tested and well-documented tool that facilitates analysis across different Earth system components (e.g. atmosphere, ocean, land, sea ice).

While originally designed to facilitate a comprehensive and rapid evaluation of models participating in the Coupled Model Intercomparison Project, the tool can now also be used to analyze some output from regional models, a large variety of gridded observational data, and reanalysis datasets. Recent improvements include the possibility to read and process operational output of selected models produced by running a model through its standard workflow, without the requirement of applying further post-processing steps as well as the strongly improved capability to handle unstructured grids (Schlund et al., 2023).

ESMValTool allows for consistent processing of all model and observational datasets such as, for instance, regridding to common grids, masking of land/sea and missing values, vertical interpolation, etc. This allows for a fair comparison of all diagnostics and metrics calculated for individual models with each other. With the recently added features of being able to specify model datasets with wildcards and automatic download of datasets from the Earth System Grid Federation (ESGF), ESMValTool is well suited to provide the context for comparing model deviations from observations with each other in an easy and convenient way. This allows to check a large set of parameters and provides the flexibility to extend existing benchmarking "recipes" easily. Recipes are configuration files for ESMValTool that define all input data, preprocessing steps, and diagnostics or metrics to be applied. All examples shown in this publication can be reproduced with ESMValTool version 2.12.0 using the recipes "recipe_lauer24gmd_fig*.yml" that are available on Zenodo (doi: 10.5281/zenodo.11198445).

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

2.2 Available metrics

For the purpose of assessing the general performance of a new model simulation and to quickly identify potential problems that require more attention, a number of metrics such as, for instance, bias or root-mean-square error, are available that can be applied over one or multiple dimension coordinates of a dataset. These dimensions include longitude, latitude and time, and for parameters that are vertically resolved, also a vertical coordinate such as pressure or altitude. For example, consider two 3-dimensional datasets (model and reference) with dimensions time, latitude, and longitude. If a metric is applied over the time dimension, the result is a 2-dimensional map with dimensions latitude and longitude; if a metric is applied over the horizontal dimensions latitude and longitude, the result is a 1-dimensional time series with dimension time.

For all metrics, an unweighted and weighted version exists. In the latter case, each point (in time and/or space) that enters the metric calculation is weighted with a factor w_i (details on the calculations are given in the corresponding sections below). In case a metric is calculated over time, the weights are defined as the length of the time intervals. If a metric is calculated over geographical coordinates (latitude and/or longitude), the grid box area size of each grid cell is used as weights. If a metric is calculated over time and geographical coordinates, the weights are calculated as the product of the above. Weights are normalized, i.e., $\sum_{i=1}^{N} w_i = 1$ (N: number of data points).

The following sections give an overview of the metrics that are available.

2.2.1 Absolute and relative BIAS

The absolute BIAS metric calculates the difference between a given dataset X and a reference dataset R (e.g. observations) as

$$BIAS_{abs} = X - R \tag{1}$$

The relative BIAS is obtained by dividing by the reference dataset R:

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

$$BIAS_{rel} = \frac{X - R}{R} \tag{2}$$

In order to avoid spurious values as a result of very small values of R, an optional threshold to mask values close to zero in the denominator can be provided.

2.2.2 Weighted and unweighted RMSE

The average root-mean-square error between a dataset X and a reference dataset R is calculated as

$$RMSE_{unweighted} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - R_i)^2} \tag{3}$$

Here, N gives the number of coordinate values over all dimensions over which the metric is applied. Optionally, the individual values can be weighted with normalized weights w_i:

$$RMSE_{weighted} = \sqrt{\sum_{i=1}^{N} w_i(X_i - R_i)^2} \tag{4}$$

A smaller RMSE corresponds to a better performance. More information on the weights is given at the beginning of Section 2.2.

2.2.3 Weighted and unweighted Pearson correlation coefficient

The Pearson correlation coefficient (r) measures the linear correlation between two datasets and is defined as the ratio between the covariance of two variables and the product of their standard deviations:

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

$$r_{unweighted} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(R_i - \bar{R})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{N}(R_i - \bar{R})^2}} \quad (5)$$

Here, $\bar{X}$ and $\bar{R}$ denote the average of the dataset X and R, respectively, over the selected dimension coordinate. Similar to the RMSE, the weighted *r* considers normalized weights w_i:

$$r_{weighted} = \frac{\sum_{i=1}^{N}[w_i(X_i - \bar{X})(R_i - \bar{R})]}{\sqrt{\sum_{i=1}^{N}(w_i(X_i - \bar{X})^2)}\sqrt{\sum_{i=1}^{N}(w_i(R_i - \bar{R})^2)}} \quad (6)$$

A larger *r* corresponds to a better performance. Again, more information on these weights is given in the beginning of Section 2.2.

2.2.4 Weighted and unweighted Earth mover's distance

The Earth mover's distance (EMD), also known as first-order Wasserstein metric W_1, is a metric to measure the similarity between two probability distributions of datasets *X* and *R* (Rubner et al., 2000). It can be understood as the minimum amount of work needed to transform one distribution into the other. This concept is often explained using the analogy of moving piles of earth, where the EMD quantifies the cost required to move the earth from one pile to another, with the cost being proportional to the amount of earth moved and the distance it has travelled. Recently, the EMD has gained more attention for applications in climate science such as an evaluation of the performance of climate models (e.g.Vissio et al., 2020). Here, we implement the EMD similar to Vissio et al. (2020) but for 1-dimensional distributions (i.e., to one variable at a time) and focusing on the W_1 metric (i.e., the EMD) only. First, we use data binning over all dimensions over which the EMD is calculated to get normalized probability mass functions $p_x(x_i)$ and $p_r(r_i)$ with *n* bins. Here, x_i and r_i are the bin centers of *X* and *R*, respectively. The bins range from the minimum to the maximum value of the data calculated over both the dataset and reference dataset; thus, $x_i = r_i$ for all *i*. For the weighted EMD, each value only contributes with its associated weight *w* to the bin count; for the unweighted EMD, each value contributes with equal weight. Details on the

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

weighting is given at the beginning of Section 2.2 With these probability mass functions, the EMD can be expressed as

$$EMD = \min_{\gamma \in \mathbb{R}_+^{n \times n}} \sum_{i,j}^{n} \gamma_{ij} |x_i - r_j| \quad \text{with} \quad \sum_{j}^{n} \gamma_{ij} = \mathrm{p_x}(\mathrm{x_i}); \sum_{i}^{n} \gamma_{ij} = \mathrm{p_r}(\mathrm{r_j}) \tag{7}$$

Here, γ is the joint probability distribution of x and r (also called "optimal transport matrix") with marginals p_x and p_r that minimizes the transportation cost. The EMD is not sensitive to the number of bins n and provides robust results even with small values of n (Vissio et al., 2020; Vissio and Lucarini, 2018). The default value in ESMValTool is n=100, but that can be changed by the user if desired. Since the EMD is a true metric in the mathematical sense, smaller values of EMD correspond to a better performance.

2.3 Datasets

In the following, all observational datasets used as a reference for the examples below are briefly described, listed in alphabetical order. For more details, we refer to the references given in the subsections.

2.3.1 Observational data

CERES-EBAF

The Clouds and the Earth's Radiant Energy System (CERES) Energy Balanced and Filled (EBAF) Ed4.2 dataset (Kato et al., 2018; Loeb et al., 2018) provides global monthly mean top-of-atmosphere (TOA) longwave (LW), shortwave (SW), and net radiative fluxes under clear-sky and all-sky conditions. CERES instruments are on board NASA's Terra and Aqua satellites. These are used to calculate the TOA longwave (lwcre) and shortwave (swcre) cloud radiative effect as differences between the TOA all-sky and clear-sky radiative fluxes. The dataset covers the time period 2001-2022 on a global 1° x 1° grid.

ERA5

ERA5 is the fifth generation reanalysis of the European Centre for Medium-range Weather Forecasts (ECMWF) (Hersbach et al., 2020) replacing the widely used ERA-Interim reanalysis (Dee et al., 2011).

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

ERA5 uses a four-dimensional variational (4D-Var) data assimilation scheme and Cycle 41r2 of the Integrated Forecasting System (IFS) (Copernicus Climate Change Service, 2017). Here, we use ERA5 data served on the Copernicus Climate Change Service Climate Data Store (CDS) that are interpolated to a horizontal resolution of 0.25° x 0.25° and in case of 3-dim variables to 37 pressure levels ranging from 1000 hPa near the surface to 1 hPa (Ecmwf, 2020).

GPCP-SG

The Global Precipitation Climatology Project (GPCP) is a community-based analysis of precipitation that covers the satellite era from 1979 to present. The data are produced by merging different data sources including passive microwave-based rainfall retrievals from satellites (SSMI, SSMIS), infrared rainfall estimates from geostationary (GOES, Meteosat, GMS, MTSat) and polar-orbiting satellites (TOVS, AIRS), and surface rain gauges (Adler et al., 2018; Adler et al., 2003). Here, we use version 2.3 of GPCP-SG that provides monthly mean precipitation rates on a global 2.5° x 2.5° grid from January 1979 to present. GPCP-SG is widely used as a reference dataset for precipitation (e.g. Bock et al., 2020; Eyring et al., 2021; Hassler and Lauer, 2021; M. Nützel et al., 2023).

HadCRUT5

The Met Office Hadley Centre/Climatic Research Unit global surface temperature dataset HadCRUT5 contains monthly averaged near-surface temperature anomalies on a regular 5° x 5° grid from 1850 to near-present. HadCRUT5 combines sea surface temperature measurements from ships and buoys and near-surface air temperature measurements from weather stations over land. There are two versions of HadCRUT5 available, a version representing temperature anomalies for the measurement locations ("non-infilled") and a second version for which a statistical method has been applied for a more complete data coverage ("analysis") (Morice et al., 2021). Here, we use the ensemble mean of the "analysis" version of the dataset. HadCRUT5 is widely used as a reference dataset for near-surface temperature (e.g. Eyring et al., 2021; Uribe et al., 2022).

HadISST

The Hadley Centre Sea Ice and Sea Surface Temperature dataset (HadISST) provides a combination of monthly globally-complete fields of SST and sea ice concentration on a 1° x 1° grid from 1870 to date.

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

The SST data are taken from the Met Office Marine Data Bank (MDB) with input from the International Comprehensive Ocean-Atmosphere Data Set (ICOADS) where there are no data from MDB available (Rayner et al., 2003). For the example shown below, we use HadISST version 1.1 monthly average sea surface temperature.

ISCCP-FH

The International Satellite Cloud Climatology Project radiative flux profile dataset (ISCCP-FH; Zhang and Rossow (2023)) provides radiative flux profiles with a global resolution of 1°x1° at 3-hourly and monthly intervals. ISCCP-FH data that are available over the time period July 1983 through June 2017 are based on ISCCP H-series products derived from different geostationary and polar-orbiting satellite imagers (Young et al., 2018). Here, we use the monthly means TOA clear-sky and all-sky radiative fluxes to calculate the shortwave and longwave cloud radiative effects for comparison with the models.

2.3.2 Model data

CMIP6

In this study we use data from models participating in the latest phase of the Coupled model Intercomparison Project (CMIP6, Eyring et al., 2016) for putting model deviations from observations into the context of current ESMs. For this, we use results from the "historical" simulations, for which forcings due to natural causes such as volcanic eruptions and solar variability as well as human factors such as CO_2 and aerosol concentrations or land use were prescribed for the time period 1850-2014. For the examples shown in this article, we use only one ensemble member (typically the first member "r1i1p1f1") per model as the inter-model spread is typically much larger than the inter-model spread given by different ensemble members from the same model (e.g. Lauer et al., 2023). Table 1 provides an overview of the CMIP6 models used.

Table 1 List of CMIP6 models providing data from the historical simulation that are compared with an example simulation from the EMAC model (see below) and put into the context of current ESMs. If more than one ensemble member is available, only the first ensemble member (typically "r1i1p1f1") is used.

Model name	Institute(s)	Scientific reference(s)

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

ACCESS-CM2	Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australian Research Council Centre of Excellence for Climate System Science (ARCCSS)	Bi et al. (2020)
ACCESS-ESM1-5	CSIRO, ARCCSS	Ziehn et al. (2020)
AWI-CM-1-1-MR	Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research (AWI), Germany	Semmler et al. (2020)
AWI-ESM-1-1-LR	AWI	Rackow et al. (2018); Sidorenko et al. (2015)
BCC-CSM2-MR	Beijing Climate Center, China	Wu et al. (2019)
BCC-ESM1	Meteorological Administration, China	Wu et al. (2020)
CAMS-CSM1-0	Chinese Academy of Meteorological Sciences (CAMS), China	Rong et al. (2018)
CanESM5	Canadian Center for Atmospheric Research (CARE), Canada	Swart et al. (2019)
CanESM5-CanOE	CARE	Swart et al. (2019)
CESM2-FV2	National Science Foundation (NSF), Department of Energy (DOE), National Center for Atmospheric Research (NCAR), USA	Danabasoglu et al. (2020)
CESM2	NSF, DOE, NCAR	Danabasoglu et al. (2020)
CESM2-WACCM	NSF, DOE, NCAR	Gettelman et al. (2019); Danabasoglu et al. (2020)
CESM2-WACCM-FV2	NSF, DOE, NCAR	Gettelman et al. (2019); Danabasoglu et al. (2020)

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

CIESM	Department of Earth System Science, Tsinghua University, China	Lin et al. (2020)
CNRM-CM6-1-HR	Météo-France/Centre National de Recherches Météorologiques (CNRM) and Centre Européen de Recherches et de Formation Avancée en Calcul Scientifique (CERFACS), France	Voldoire et al. (2019)
CNRM-ESM2-1	CNRM, CERFACS	Séférian et al. (2019)
FGOALS-f3-L	CAMS	Guo et al. (2020)
FGOALS-g3	CAMS	Li et al. (2020)
FIO-ESM-2-0	First Institute of Oceanography, Ministry of Natural Resources (FIO), China, Qingdao National Laboratory for Marine Science and Technology (QNLM), China	Bao et al. (2020)
GFDL-ESM4	National Oceanic and Atmospheric Administration (NOAA) /Geophysical Fluid Dynamics Laboratory (GFDL), USA	Dunne et al. (2020)
GISS-E2-1-G	National Aeronautics and Space Administration (NASA), Goddard Institute for Space Studies (GISS), USA	Rind et al. (2020)
GISS-E2-1-H	NASA, GISS	Rind et al. (2020)
HadGEM3-GC31-LL	Met Office Hadley Centre (MOHC), UK	Williams et al. (2018); Kuhlbrodt et al. (2018)
HadGEM3-GC31-MM	MOHC	Williams et al. (2018); Kuhlbrodt et al. (2018)
INM-CM4-8	Institute for Numerical Mathematics (INM), Russian Academy of Science (RAS), Russia	Volodin et al. (2018)
INM-CM5-0	INM, RAS	Volodin et al. (2017)

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

ISPL-CM6A-LR	L'Institut Pierre-Simon Laplace (IPSL), France	Boucher et al. (2020)
KACE-1-0-G	National Institute of Meteorological Sciences/Korea Meteorological Administration, Climate Research Division, Republic of Korea	Lee et al. (2020)
MCM-UA-1-0	Department of Geosciences, University of Arizona, USA	Delworth et al. (2002)
MIROC6	Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Atmosphere and Ocean Research Institute (AORI), University of Tokyo, and National Institute for Environmental Studies (NIES), Japan	Tatebe et al. (2019)
MIROC-ES2L	JAMSTEC, AORI, NIES	Hajima et al. (2020)
MPI-ESM-1-2-HAM	HAMMOZ-Consortium: ETH Zurich, Switzerland; Max Planck Institut für Meteorologie (MPIM), Germany; ForschungszentrumJülich, Germany; University of Oxford, UK; Finnish Meteorological Institute, Finland; Leibniz Institute for Tropospheric Research, Germany; Center for Climate Systems Modeling (C2SM) at ETH Zurich, Switzerland	Mauritsen et al. (2019)
MPI-ESM1-2-HR	MPIM	Muller et al. (2018)
MPI-ESM1-2-LR	MPIM	Mauritsen et al. (2019)
MRI-ESM2-0	Meteorological Research Institute (MRI), Japan	Yukimoto et al. (2019)
NESM3	Nanjing University of Information Science and Technology, China	Cao et al. (2018)
NorESM2-LM	NorESM Climate modeling Consortium (NCC), Norway	Seland et al. (2020)
NorESM2-MM	NCC	Seland et al. (2020)

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

SAM0-UNICON*	Seoul National University, Republic of Korea	Park et al. (2019)
UKESM1-0-LL	MOHC	Sellar et al. (2019)

EMAC

The ECHAM/MESSy Atmospheric Chemistry (EMAC) model is a chemistry-climate model (Jöckel et al., 2010) that has been widely used for various studies in atmospheric sciences including, for instance, tropospheric and stratospheric ozone (e.g. Dietmüller et al., 2021; Mertens et al., 2021), climate impact of contrails and emissions from aviation (e.g. Frömming et al., 2021; Matthes et al., 2021) and the effects of transport on atmosphere and climate (e.g. Hendricks et al., 2018; Righi et al., 2015). EMAC uses the second version of the Modular Earth Submodel System (MESSy2) to link submodels for various physical and chemical processes to the host model. Here, the 5th generation of the European Centre Hamburg general circulation model (ECHAM5; Roeckner et al. (2006)) is used as host model.

In this study, we use an EMAC simulation with deliberately erroneous prescribed sea surface temperatures (SSTs) to showcase application of the new ESMValTool features to monitor and benchmark model simulations during the model development phase with results from established global climate models. While a comparison of results from coupled historical CMIP6 simulations with an AMIP-style simulation in which SSTs and sea ice concentrations are prescribed from observations is of course not completely fair for a real model benchmarking or monitoring of a simulation, this approach allows us to showcase the new ESMValTool features with a simulation in which something goes wrong after a few simulation years. For this, the SST fields are set to zonally averaged monthly values of the observed global average SST after the first five years of model simulation (see Figure 1). Such an error does not necessarily show up in time series of global mean near-surface temperature but can be identified when using other metrics.

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

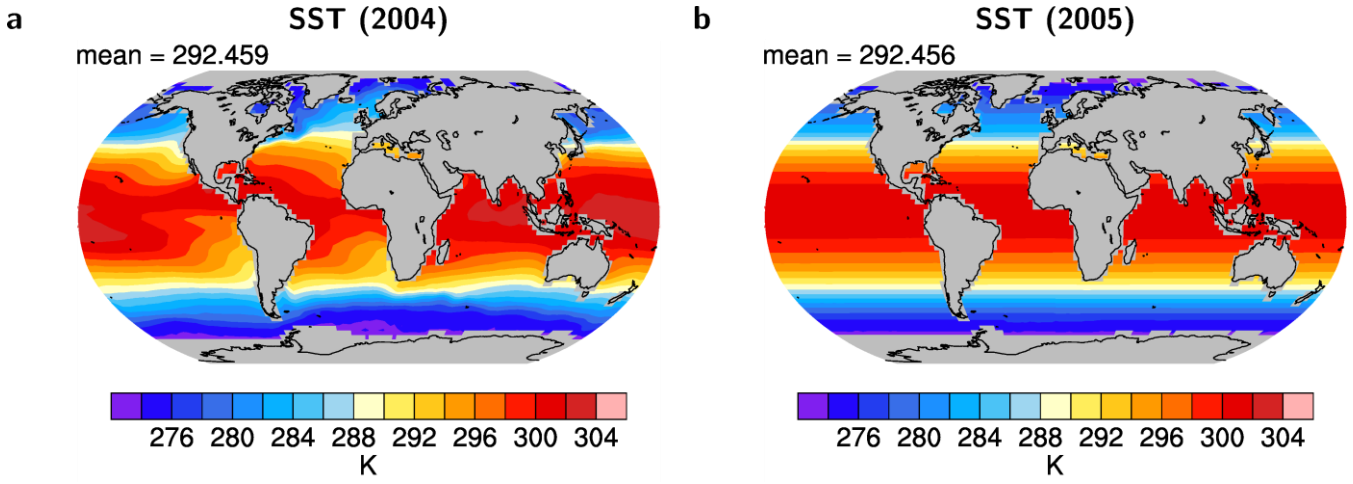


Figure 1 Annual mean of the prescribed sea surface temperatures (SSTs) for the EMAC simulation (a) before (year 2004) and (b) after (year 2005) the deliberately introduced "error".

3 Monitoring and benchmarking of ESM simulations

In the following, we show examples of how the new ESMValTool capabilities can be used to monitor and benchmark model simulations to detect problems during runtime and to assess whether the performance of a model simulation is within the range of what could be expected from current state-of-the-art ESMs (here: CMIP6). The variables and reference datasets used in the examples are listed in Table 2.

Table 2 Variables and reference datasets used.

Variable	Description	Reference dataset(s)
tas	near-surface air temperature (K)	HadCRUT5, ERA5
tas_land	same as tas but over land grid cells only (K)	HadCRUT5, ERA5
sst	sea surface temperature (K)	HadISST, ERA5
pr	precipitation (mm day^{-1})	GPCP-SG, ERA5
psl	air pressure at sea level (Pa)	ERA5
ta	air temperature (K)	ERA5
rlut	TOA outgoing longwave radiation (W m^{-2})	CERES-EBAF, ISCCP-FH
rsut	TOA outgoing shortwave radiation (W m^{-2})	CERES-EBAF, ISCCP-FH

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

lwcre	TOA longwave cloud radiative effect (W m^{-2})	CERES-EBAF, ISCCP-FH
swcre	TOA shortwave cloud radiative effect (W m^{-2})	CERES-EBAF, ISCCP-FH

3.1 Time series

Time series of climate relevant quantities or their anomalies relative to a given reference period averaged over a specific region or the entire globe are a common approach to evaluate model results with one or several reference datasets (e.g., Bock et al., 2020; Yazdandoost et al., 2021; Wang et al., 2023). As an example, Figure 2a shows a time series of global average anomalies in near-surface temperature. In addition to the EMAC model results (red line) and the observational reference data from HadCRUT5 (black line), also the CMIP6 results (Table 1) are shown as thin gray lines. The figure shows that the first five years of the EMAC simulation are rather at the high end of the CMIP6 results with the temperature anomalies frequently exceeding the 90 % percentile of the CMIP6 results. In the beginning of the year 2005, there is the sudden temperature drop when the deliberate error in the SST fields is introduced resulting in the EMAC simulation being at the low end of the CMIP6 range with temperature anomalies frequently being below the CMIP6 10 % percentile. Figure 2b shows a time series of the global average (area weighted) root mean square errors in simulated near-surface temperature from EMAC (red line). The 10 % and 90 % percentile range of the RMSE values from the individual CMIP6 models is shown as light blue shading. The "error" in the geographical distribution of the sea surface temperatures introduced in 2005 is not obvious in this time series as the performance of this EMAC simulation is within the range of what could be expected from a coupled CMIP6 model. This shows that monitoring of model simulations typically requires assessing several variables.

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

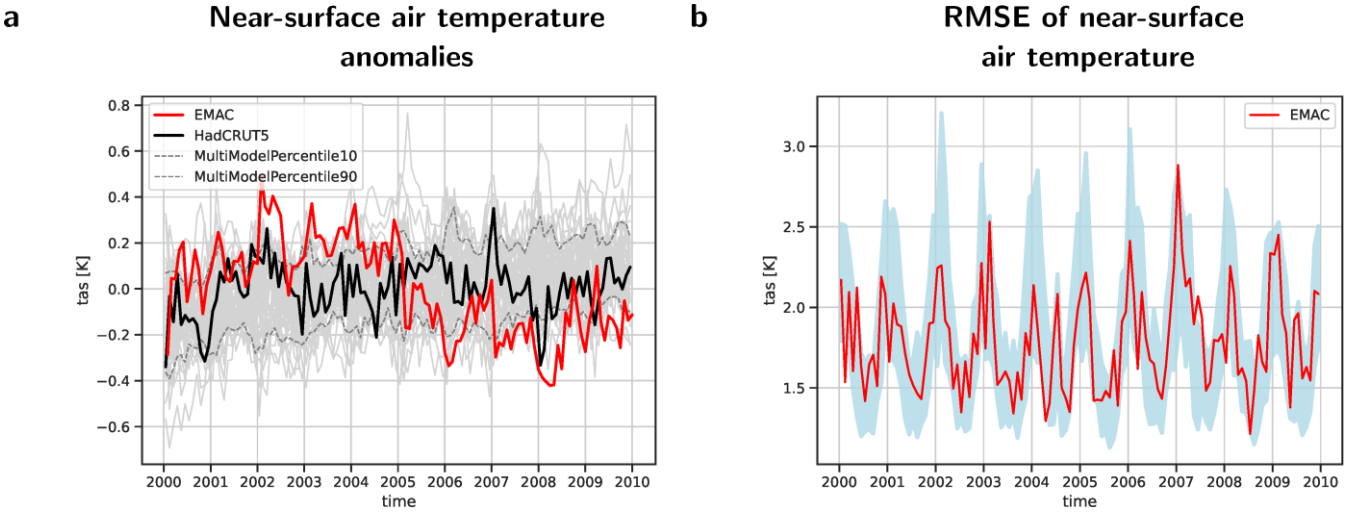


Figure 2 (a) Time series from 2000 through 2009 of global average monthly mean temperature anomalies (reference period 2000-2009) of the near-surface temperature in K from a simulation of EMAC (red) and the reference dataset HadCRUT5 (black). The thin gray lines show 43 individual CMIP6 models used for comparison, the dashed gray lines show the 10 % and 90 % percentiles of these CMIP6 models. (b) Same as (a) but for area-weighted RMSE of the near-surface air temperature. The light blue shading shows the range of the 10 % to 90 % percentiles of RMSE values from the ensemble of 43 CMIP6 models used for comparison.

3.2 Diurnal and seasonal cycle

A further commonly used metric for model evaluation is the comparison of the seasonal cycle of a specific variable, calculated for the whole globe or again for a pre-defined region. Figure 3a shows the multi-year global mean seasonal cycle of near-surface air temperature for a suite of CMIP6 models, the HadCRUT5 observations, and the specifically created EMAC simulation that has been described in Section 2.3.2. The CMIP6 model simulations and the HadCRUT5 data are averaged over the time period 2000-2009, whereas the EMAC simulation is split in the two five-year periods without and with the erroneous SSTs, 2000-2004 (red line) and 2005-2009 (dark blue line) respectively. Similar to Figure 2a, also Figure 3a indicates the 10 % and 90 % percentile range with gray dashed lines. Both five-year means of the EMAC simulation are well within the CMIP6 10 % and 90 % percentile range throughout the whole year, but the EMAC simulation period with the correct SSTs is slightly closer to the HadCRUT5 data than the simulation period with the erroneous SSTs. While this is positively noted, it itself is not a clear indication that a problem occurred with the latter five-year period of the EMAC simulation. Figure 3b shows then

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

the area-weighted RMSE values for the global mean seasonal cycle of near-surface air temperature. The blue shading depicts the 10 % to 90 % percentile range of the CMIP6 models used for the comparison. The earlier five-year period of the EMAC simulation (2000-2004, red line) is in most months below the blue shaded area, which means that with correct SST fields the example EMAC simulation can reproduce the seasonal cycle of near-surface temperature better than most CMIP6 models (smaller RMSE = better performance). With the erroneous SSTs, however, the RMSE values for the annual cycle become larger, which means that the agreement of the seasonal cycle of near-surface air temperature with the reference dataset decreased for that period of the EMAC simulation. They are still located within the blue-shaded area, but agreement is less good than for the earlier period (red line). Again, this metric alone would not allow the clear detection of a faulty simulation, but it would be clear that in "normal" simulations EMAC is performing clearly better than most CMIP6 models when looking at the RMSE of near-surface air temperature, and the clear decrease in performance could be an indicator that something might be problematic with a new simulation.

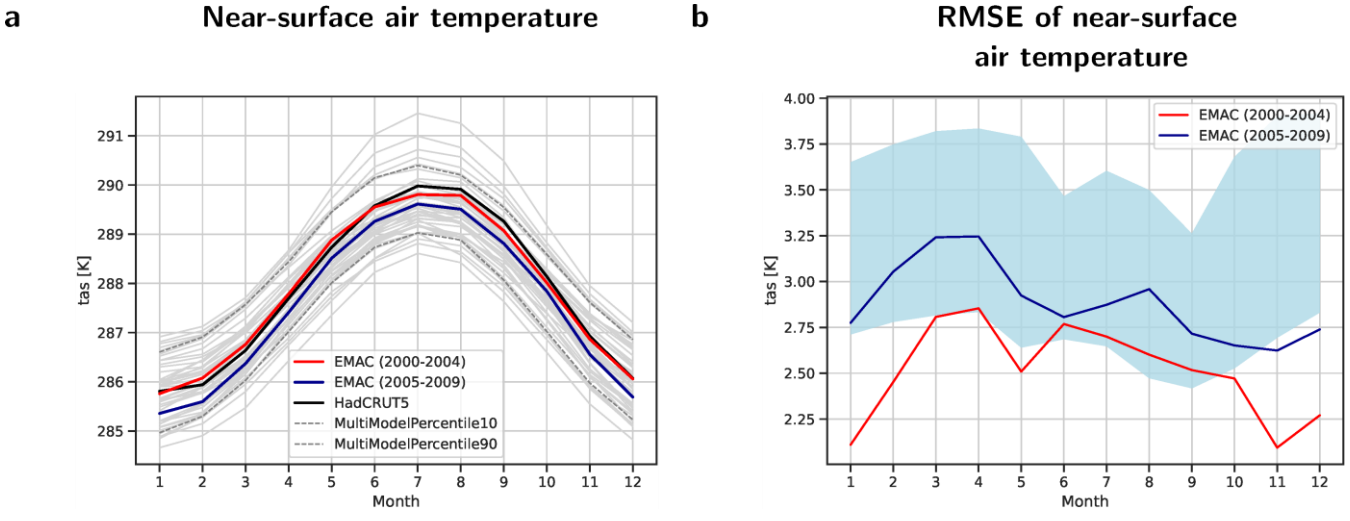


Figure 3 (a) Multi-year global mean of the seasonal cycle of near-surface air temperature in K from a simulation of EMAC averaged over the time period 2000-2004 (red) and 2005-2009 (dark blue) and the reference dataset HadCRUT5 (2000-2009, black). The thin gray lines show 43 individual CMIP6 models (2000-2009) used for comparison, the dashed gray lines show the 10 % and 90 % percentiles of these CMIP6 models. (b) Same as (a) but for area-weighted RMSE of near-surface temperature. The light blue

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

shading shows the range of the 10 % to 90 % percentiles of RMSE values from the ensemble of 43 CMIP6 models used for comparison.

A further capability implemented in ESMValTool is to intercompare the diurnal cycle of a variable, for example precipitation (see Figure 4). The basic structure of the graphs is identical to Figure 3 regarding the shown EMAC simulation, the CMIP6 simulations and their spread. The example results show, however, the precipitation averaged only over the tropical ocean instead of a global mean and averaged over only two years (2004-2005). ERA5 has been used as reference dataset. Both years of the EMAC simulation show a reduced amplitude of the average diurnal cycle of precipitation over the tropical ocean compared to ERA5 and most of the CMIP6 ensemble (Figure 4a), with 2004 (from the "correct" period) being even further away from the reference compared to 2005 (from the erroneous period). Figure 4b shows the RMSE of the diurnal cycle of precipitation over the tropical ocean. The blue shaded region indicates again the 10 % to 90 % percentile range of the CMIP6 models. The year 2004 of the EMAC simulation is fully enclosed by the CMIP6 percentile range, whereas 2005 is for most hours of the day above the CMIP6 percentile range. This reversal of which EMAC simulation year performs better compared with ERA5 suggests, that some kind of error compensation takes place when calculating the mean values (Figure 4a), while this is not the case when calculating the RMSE value at each grid cell for a given time of the day and then averaging afterwards (Figure 4b). Similar to the metric shown in Figure 3, the comparison of the diurnal cycle of precipitation alone might not be able to correctly identify erroneous simulations, but also this metric could give an indication that something might not be correct with a new simulation, if it is possible to compare it to a "baseline" simulation of the same model that has been labeled as "correct".

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

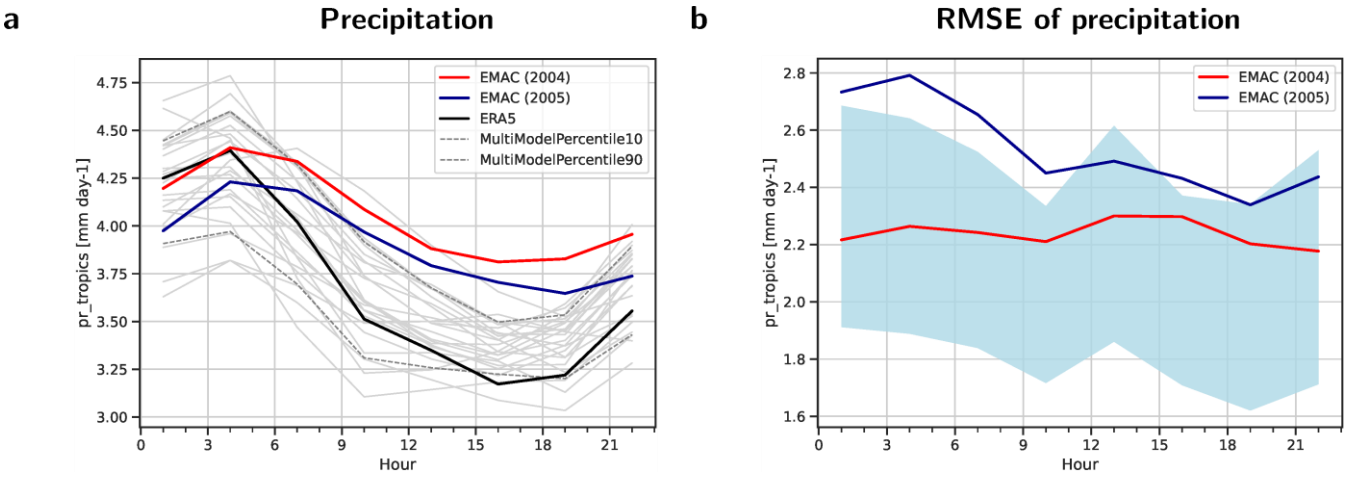


Figure 4 Annual mean diurnal cycle of precipitation averaged over the tropical ocean (ocean grid cells in the latitude belt 30° S to 30° N) from a simulation of EMAC averaged over the year 2004 (red) and over 2005 (dark blue) compared with ERA5 data (2004-2005, black). The thin gray lines show 22 individual CMIP6 models used for comparison (2004-2005), the dashed gray lines show the 10 % and 90 % percentiles of these CMIP6 models. (b) Same as (a) but for area-weighted RMSE of precipitation. The light blue shading shows the range of the 10 % to 90 % percentiles of RMSE values from the ensemble of 22 CMIP6 models used for comparison.

3.3 Geographical distribution

Figure 5 shows an example of how the RMSE of the time series of monthly mean precipitation at each grid cell from a given simulation can be compared with the range of RMSE values from the CMIP6 models. As a reference, GPCP-SG data are used (Sect. 2.3.1). The stippled grid cells denote areas at which the RMSE value of the given simulation is below the 90 % percentile of RMSE values from the CMIP6 models. This threshold can be set depending on what is considered OK during model development or model benchmarking, allowing to focus on the non-stippled areas showing larger deviations. Figure 5a shows the RMSE of the precipitation time series of the EMAC simulation for the period 2000–2004. In this figure, non-stippled areas are mainly found in the tropical East Pacific and Indian Ocean highlighting the regions that show larger RMSE values than most of the CMIP6 models and that might need further investigation during model development, or that perform worse than what could be expected from a state-of-the-art model. As a result of the deliberately introduced error in the geographical SST distribution in 2005, these areas are much larger in the second half (2005–2009) of the EMAC simulation (see Figure 5b) and now cover most of the tropical oceans. When applied to the monitoring of a running simulation,

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

this increase in areas performing less well than the majority of CMIP6 models can be a first indication of problems related to deep convection, which requires further investigation.

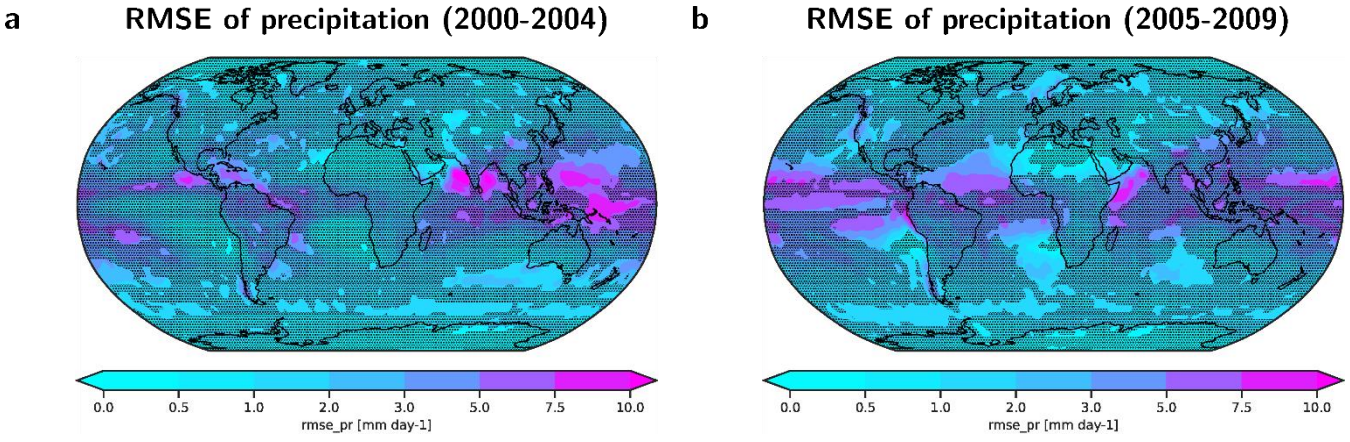


Figure 5 5-year annual mean area-weighted RMSE of the precipitation rate in mm day^{-1} from a simulation of EMAC compared with GPCP-SG data. (a) Average over the time period 2000-2004, (b) average over the time period 2005-2009. The stippled areas mask grid cells where the RMSE is smaller than the 90 % percentile of RMSE values from an ensemble of 39 CMIP6 models.

3.4 Zonal averages

For 3-dimensional variables such as air temperature, a comparison of zonally averaged fields with reference data is an easy and common way to evaluate a model simulation. For this, the absolute or relative bias can be used as a measure of how well the model simulation reproduces the reference data. In Figure 6, the absolute bias of the EMAC example simulation compared with ERA5 data for the zonally averaged 3-dimensional air temperature is shown. Here, the stippling indicates that the absolute value of the bias $|bias|$ is smaller than the maximum of the absolute 10 % and the absolute 90 % percentiles, $|p10|$ and $|p90|$, respectively, of the bias values from the CMIP6 ensemble for this grid cell. By using the criteria $|BIAS| \leq max(|p10|, |p90|)$, positive and negative bias values are given the same importance when assessing the model performance. Depending on the aim of the model development and the percentiles selected for this comparison, all non-stippled bias values outside of this range can be regarded as below par performance and might require further investigation and possibly continued model improvements or model tuning. When monitoring a running simulation, the strong increase in the grid cells that are marked as below-average performance between the first (Figure 6a) and the second simulation time period (Figure

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

6b) is a first hint that there might be an unexpected problem in the simulation that occurred during run-time.

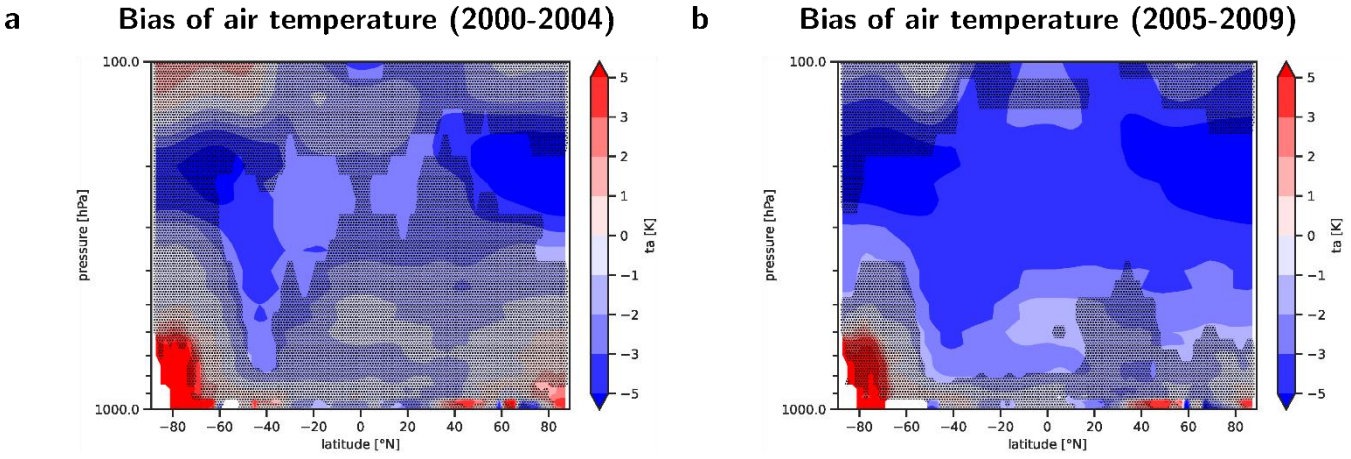


Figure 6 5-year annual mean bias of the zonally averaged temperature in K from a historical simulation of the EMAC model compared with ERA5 reanalysis data. (a) Average over the time period 2000-2004, (b) average over the time period 2005-2009. The stippled areas mask grid cells where the absolute BIAS ($|\boldsymbol{BIAS}|$) is smaller than the maximum of the absolute 10 % ($|\boldsymbol{p10}|$) and the absolute 90 % ($|\boldsymbol{p90}|$) percentiles from an ensemble of 38 CMIP6 models, i.e. $|\boldsymbol{BIAS}| \leq \boldsymbol{max}(|\boldsymbol{p10}|, |\boldsymbol{p90}|)$.

3.5 Box plots

The summary plots for different variables as shown in Figure 7 offer the possibility to quickly get a first overview on model performance. It can either be used as a starting point for a more in-depth evaluation of individual variables or climate parameters with observations, or as one possible summary of overall model performance. For every diagnostic field considered, model performance is compared to one reference dataset (see Table 2, first dataset), and the quality of the simulation is summarized in a single number such as RMSD (Figure 7a,b), Pearson's correlation coefficient (Figure 7c,d) or EMD (Figure 7e,f) computed over the time averaged global maps.

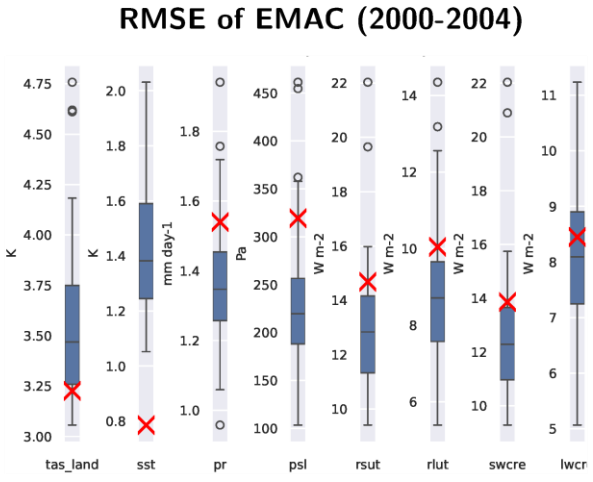
By simultaneously assessing a number of different performance indices, the general model improvements can then be quantified and compared with the CMIP6 ensemble. In our example EMAC simulation, the SSTs are prescribed; thus, we see a significantly better performance in SST than the CMIP ensemble of coupled (historical) simulations especially regarding the RMSE (Figure 7a) and the Pearson's correlation coefficient (Figure 7c). For the other variables, the EMAC example shows often a slightly worse

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

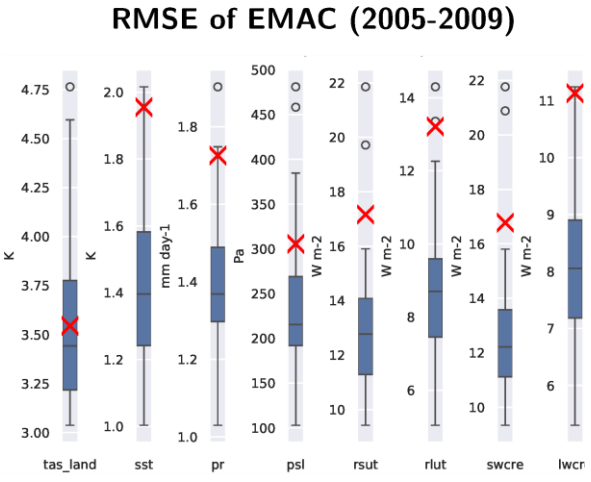
performance than 75 % percentile of the CMIP6 models, but mostly lies still in the range of the CMIP6 models. This changes when we look at the second time period (Figure 7b), where we can see a significant decrease of model performance regarding RMSE for all variables. Furthermore, it can be seen that the decrease of performance in the second time period is most prominent for the SSTs especially in the RMSE and correlation pattern values (Figure 7b,d). This is a clear hint that detailed diagnostics for this variable (e.g. see Figure 2) would be helpful in order to quickly identify the error in the simulation.

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

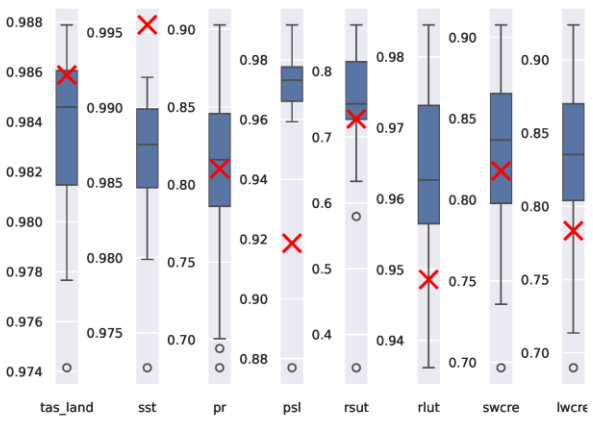
a



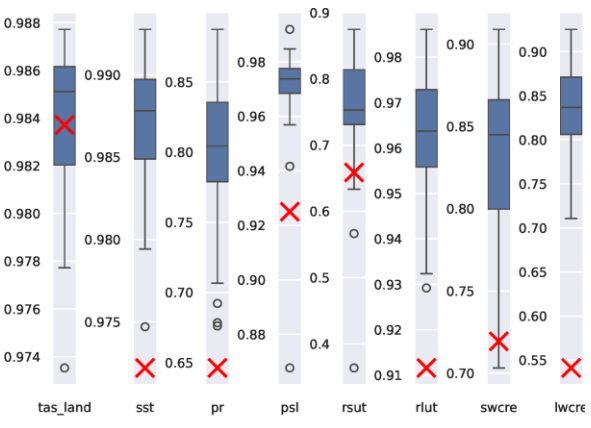
b



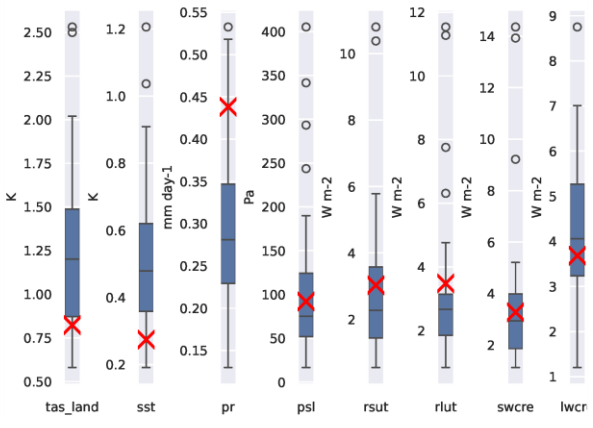
c



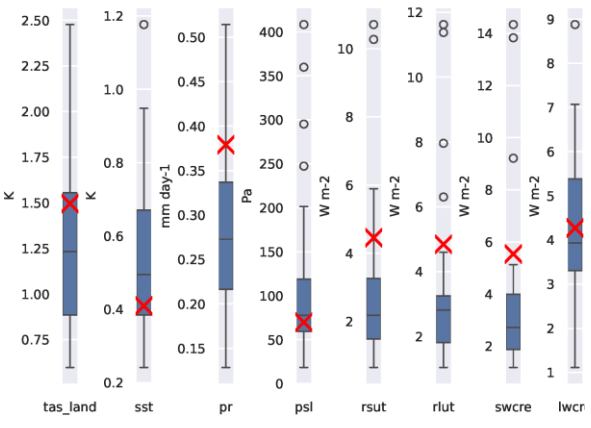
d



e



f



https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

Figure 7 (a, b) Global area-weighted RMSE (smaller=better), (c, d) weighted Pearson's correlation coefficient (higher=better) and (e, f) weighted Earth mover's distance (smaller=better) of the geographical pattern of 5-year means of different variables from a simulation of EMAC (red cross) in comparison to the CMIP6 ensemble (boxplot). The left column shows the results for the time period 2000-2004, the right column for 2005-2009. Reference datasets for calculating the three metrics are: near-surface temperature (tas): HadCRUT5, surface temperature (ts): HadISST, precipitation (pr): GPCP-SG, air pressure at sea level (psl): ERA5, shortwave (rsut) longwave (rlut) radiative fluxes at TOA and shortwave (swcre) and longwave (lwcre) cloud radiative effects: CERES-EBAF. Each box indicates the range from the first quartile to the third quartile, the vertical lines show the median, and the whiskers the minimum and maximum values, excluding the outliers. Outliers are defined as being outside 1.5 times the interquartile range.

3.6 Portrait diagram

Portrait diagrams (Gleckler et al., 2008) can be used to visualize model performance across different variables relative to one or multiple reference datasets. Unlike box plots, portrait diagrams show the performance of each model individually; thus, they provide a convenient way to benchmark each element in an ensemble of models. Figure 8 shows an example of a portrait diagram for the same set of variables as used in the box plots (see Figure 7). The horizontal axis shows the different models (left: CMIP6 models; right: the EMAC simulation for two different time periods) and the vertical axis the different variables. The colors correspond to the relative RMSE (relative to the median RMSE across all models) of the different models and variables, where red corresponds to a higher RMSE (= worse performance), and blue to a lower RMSE (= better performance) than the median. For variables where the box is split into two triangles, an alternative dataset is provided in the lower right triangle (see Table 2 for an overview of variables and reference datasets used). The effect of the deliberately introduced error in the EMAC simulation is clearly visible on the right side of the portrait diagram: as expected, the wrong SST pattern starting in 2005 leads to a sharp decline in the relative RMSE in SST from dark blue colors (i.e., very good performance) in 2000–2004 to dark red colors (i.e., very bad performance) in 2005–2009. However, the error is not only visible in the SST: across all variables, the later period (2005–2009) of the EMAC simulation shows a higher relative RMSE (i.e., worse performance) than the corresponding early period (2000–2004). In addition to RMSE, also the metrics EMD or Pearson correlation coefficient could be used (see Sect. 2.2).

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

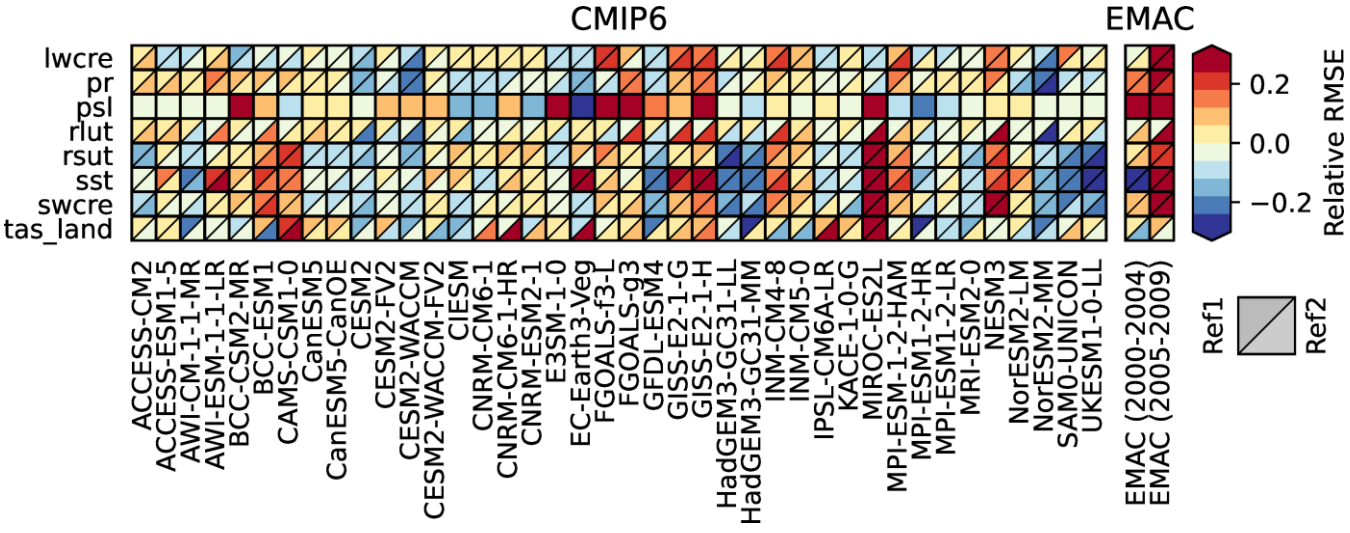


Figure 8 Portrait diagram showing the relative space-time root-mean-square error (RMSE) calculated from the seasonal cycle of the datasets. The seasonal cycle is averaged over the years 2000-2009 (CMIP6 models) and over the time periods 2000-2004 and 2005-2009 for the EMAC simulation. The figure shows the relative performance with blue shading indicating a better, and red shading indicating a worse performance than the median RMSE of all models. The lower right triangle shows the relative RMSE with respect to the reference dataset (Ref1), the upper left triangle with respect to an alternative reference dataset (Ref2). Using RMSE as a metric (as shown) gives a portrait diagram similar to Gleckler et al. (2008). Other metrics are available.

4 Summary and discussion

In this paper, we introduce the newly extended capability of the Earth System Model Evaluation Tool to benchmark and monitor climate model simulations across a wide range of different Earth system components. The new framework allows to put common performance metrics calculated for a given model simulation into the context of results from an ensemble of state-of-the-art climate models such as the ones participating in the Coupled Model Intercomparison Project Phase 6. Putting the performance of a model simulation into such a context allows to quickly assess whether, for instance, the values obtained for metrics such as bias or pattern correlation for a variable are within the typical range of model errors or might need further, more detailed investigation. This is particularly of help during model development or when monitoring a simulation to identify possible problems already during run-time as this allows a large number of variables to be assessed without the need for detailed expert knowledge on each single

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

quantity. This is also of help when automatizing monitoring of running simulations. For this, the numerical output of ESMValTool in the form of netCDF files could be used to summarize the results from the different metrics with a e.g. dash board displayed on a web site showing a green, yellow or red traffic light for each quantity tested depending on the results. The percentiles for the metric obtained from the model ensemble used for comparison can be used as thresholds to flag quantities that are outside the range of typical model errors and thus in need of further inspection. A possible application for these new model benchmarking and monitoring capabilities of ESMValTool would be the assessment of new model simulations during the preparation phase for CMIP7.

As shown in Section 3 particularly for model development, these metrics are most effective if there are already results from a well-tuned, well-understood "baseline" simulation of the same model available. With the results for this "baseline" simulation being known the evaluation and benchmarking of a new simulation can be done already quite effectively with few simulation years since the deviation from the "baseline" become apparent quickly for many relevant atmospheric variables. For the examples shown in this paper, for instance, we found that five model years are usually already sufficient for this kind of first assessment.

The possibility to use wildcards in recipes when specifying the model datasets (available since ESMValTool version 2.8.0) used to provide context for comparison in combination with the feature to download any data that are missing locally but that are available on the ESGF automatically (available since ESMValTool version 2.4.0) makes application of ESMValTool for model benchmarking and monitoring very easy and user friendly. Examples of how to use the new capabilities of ESMValTool for benchmarking and monitoring include time series, seasonal and diurnal cycles as well as map plots, box plots and portrait diagrams for any 2-dimensional variable including individual levels or, for instance, zonally averages of 3-dimensional variables that can be shown as latitude-height plots.

The benchmarking and monitoring diagnostics introduced in this paper currently support absolute and relative bias, Pearson's correlation coefficient, root mean square error and Earth mover's distance as metrics. All of these metrics can be calculated as unweighted or weighted, e.g. by using the area size of the grid cells as weights. As all of these basic metrics are implemented in the form of a generic

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

preprocessing function of ESMValTool, adding new metrics is straightforward and new metrics can then be used by all diagnostics with little to no additional effort.

5 Code availability Statement

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.3387139. 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.3401363. ESMValTool and ESMValCore are developed on the GitHub repositories available at https://github.com/ESMValGroup.

6 Data Availability Statement

CMIP6 data are available freely and publicly from the Earth System Grid Federation (ESGF) and can be retrieved by ESMValTool automatically by setting the configuration option 'search_esgf' to 'when_missing' or 'always'. All observations/reanalysis data used are described in Sect. 2.3.1. The observational/reanalysis datasets are not distributed with ESMValTool that 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. The EMAC data used as an example in this study are available on Zenodo at http://doi.org/10.5281/zenodo.11198445.

7 Author contributions

AL and BH developed the concept for this work. AL, LB, LR and MS contributed to coding the ESMValTool extensions presented, PJ designed and performed the EMAC simulation used as an example. All authors contributed to the writing and editing of the paper.

8 Competing interests

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

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

9 Acknowledgements

The development of ESMValTool is supported by several projects. The diagnostic development of ESMValTool v2 for this paper received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 101003536 (ESM2025 – Earth System Models for the Future) and 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). 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 modelling 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 DKRZ. This work used resources of the Deutsches Klimarechenzentrum (DKRZ) granted by its Scientific Steering Committee (WLA) under project IDs bd0854, id0853 and bd1179.

This manuscript contains modified Copernicus Climate Change Service (2017) information with ERA5 data retrieved from the Climate Data Store (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.

CERES-EBAF data were obtained from the NASA Langley Research Center Atmospheric Science Data Center.

Global Precipitation Climatology Project (GPCP) Monthly Analysis Product data used are provided by the NOAA PSL, Boulder, Colorado, USA, downloaded from their website at https://psl.noaa.gov (last access 9 May 2023).

HadCRUT5 data were obtained from http://www.metoffice.gov.uk/hadobs/hadcrut5 on 6 Dec 2021 and are © British Crown Copyright, Met Office 2020, provided under an Open Government License, http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/.

HadISST v1.1 data were obtained from https://www.metoffice.gov.uk/hadobs/hadisst/ (last access 20 January 2023) and are © British Crown Copyright, Met Office, 2007, provided under a Non-Commercial

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

Government Licence http://www.nationalarchives.gov.uk/doc/non-commercial-government-licence/version/2/.

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) (https://www.ncei.noaa.gov).

We would like to thank Franziska Winterstein (DLR) for helpful comments on the manuscript.

10 References

Adler, R. F., Sapiano, M. R. P., Huffman, G. J., Wang, J. J., Gu, G. J., Bolvin, D., Chiu, L., Schneider, U., Becker, A., Nelkin, E., Xie, P. P., Ferraro, 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-Basel, 9, 10.3390/atmos9040138, 2018.

Adler, R. F., Huffman, G. J., Chang, A., Ferraro, R., Xie, P. P., Janowiak, J., Rudolf, B., Schneider, U., Curtis, S., Bolvin, D., Gruber, A., Susskind, J., Arkin, P., and Nelkin, E.: The version-2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979-present), J Hydrometeorol, 4, 1147-1167, 10.1175/1525-7541(2003)004<1147:Tvgpcp>2.0.Co;2, 2003.

Bao, Y., Song, Z. Y., and Qiao, F. L.: FIO-ESM Version 2.0: Model Description and Evaluation, J Geophys Res-Oceans, 125, ARTN e2019JC016036, 10.1029/2019JC016036, 2020.

Bi, D. H., Dix, M., Marsland, S., O'Farrell, S., Sullivan, A., Bodman, R., Law, R., Harman, I., Srbinovsky, J., Rashid, H. A., Dobrohotoff, P., Mackallah, C., Yan, H. L., Hirst, A., Savita, A., Dias, F. B., Woodhouse, M., Fiedler, R., and Heerdegen, A.: Configuration and spin-up of ACCESS-CM2, the new generation Australian Community Climate and Earth System Simulator Coupled Model, J So Hemisph Earth, 70, 225-251, 10.1071/Es19040, 2020.

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, J Geophys Res-Atmos, 125, 2020.

Boucher, O., J. Servonnat, A. L. Albright, O. Aumont, Y. Balkanski, V. Bastrikov, S. Bekki, R. Bonnet, S. Bony, L. Bopp, P. Braconnot, P. Brockmann, P. Cadule, A. Caubel, F. Cheruy, A. Cozic, D. Cugnet, F. D'Andrea, P. Davini, S. Denvil, J. Deshayes, A. Ducharne, J.-L. Dufresne, C. Ethé, L. Fairhead, L. Falletti, M.-A. Foujols, S. Gardoll, G. Gastineau, J. Ghattas, J.-Y. Grandpeix, B. Guenet, L. Guez, E. Guilyardi, M. Guimberteau, D. Hauglustaine, F. Hourdin, A. Idelkadi, S. Joussaume, M. Kageyama, A. Khadre-Traoré, M. Khodri, G. Krinner, N. Lebas, G. Levavasseur, C. Lévy, F. Lott, T. Lurton, S. Luyssaert, G. Madec, J.-B.e Madeleine, F. Maignan, M. Marchand, O. Marti, L. Mellul, Y. Meurdesoif, J. Mignot, I. Musat, C. Ottlé, P. Peylin, Y. Planton, J. Polcher, C. Rio, C. Rousset, P. Sepulchre, A. Sima, D. Swingedouw, R. Thieblemont, M. Vancoppenolle, J. Vial, J. Vialard, N. Viovy, and Vuichard, N.: Presentation and evaluation of the IPSL-CM6A-LR climate model, Journal of Advances in Modeling Earth System, 10.1029/2019MS002010, 2020.

Cao, J., Wang, B., Yang, Y. M., Ma, L. B., Li, J., Sun, B., Bao, Y., He, J., Zhou, X., and Wu, L. G.: The NUIST Earth System Model (NESM) version 3: description and preliminary evaluation, Geosci Model Dev, 11, 2975-2993, 2018.

ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate: https://confluence.ecmwf.int/display/CKB/ERA5%3A+data+documentation, last access: 2 November 2021.

Danabasoglu, G., Lamarque, J. F., Bacmeister, J., Bailey, D. A., DuVivier, A. K., Edwards, J., Emmons, L. K., Fasullo, J., Garcia, R., Gettelman, A., Hannay, C., Holland, M. M., Large, W. G., Lauritzen, P. H., Lawrence, D. M., Lenaerts, J.

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

T. M., Lindsay, K., Lipscomb, W. H., Mills, M. J., Neale, R., Oleson, K. W., Otto-Bliesner, B., Phillips, A. S., Sacks, W., Tilmes, S., van Kampenhout, L., Vertenstein, M., Bertini, A., Dennis, J., Deser, C., Fischer, C., Fox-Kemper, B., Kay, J. E., Kinnison, D., Kushner, P. J., Larson, V. E., Long, M. C., Mickelson, S., Moore, J. K., Nienhouse, E., Polvani, L., Rasch, P. J., and Strand, W. G.: The Community Earth System Model Version 2 (CESM2), J Adv Model Earth Sy, 12, 2020.

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., Holm, E. V., Isaksen, L., Kallberg, P., Kohler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J. J., Park, B. K., Peubey, C., de Rosnay, P., Tavolato, C., Thepaut, J. N., and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Q J Roy Meteor Soc, 137, 553-597, 10.1002/qj.828, 2011.

Delworth, T. L., Stouffer, R. J., Dixon, K. W., Spelman, M. J., Knutson, T. R., Broccoli, A. J., Kushner, P. J., and Wetherald, R. T.: Review of simulations of climate variability and change with the GFDL R30 coupled climate model, Clim Dynam, 19, 555-574, 10.1007/s00382-002-0249-5, 2002.

Dietmüller, S., Garny, H., Eichinger, R., and Ball, W. T.: Analysis of recent lower-stratospheric ozone trends in chemistry climate models, Atmos Chem Phys, 21, 6811-6837, 10.5194/acp-21-6811-2021, 2021.

Dunne, J. P., Horowitz, L. W., Adcroft, A. J., Ginoux, P., Held, I. M., John, J. G., Krasting, J. P., Malyshev, S., Naik, V., Paulot, F., Shevliakova, E., Stock, C. A., Zadeh, N., Balaji, V., Blanton, C., Dunne, K. A., Dupuis, C., Durachta, J., Dussin, R., Gauthier, P. P. G., Griffies, S. M., Guo, H., Hallberg, R. W., Harrison, M., He, J., Hurlin, W., McHugh, C., Menzel, R., Milly, P. C. D., Nikonov, S., Paynter, D. J., Ploshay, J., Radhakrishnan, A., Rand, K., Reichl, B. G., Robinson, T., Schwarzkopf, D. M., Sentman, L. T., Underwood, S., Vahlenkamp, H., Winton, M., Wittenberg, A. T., Wyman, B., Zeng, Y., and Zhao, M.: The GFDL Earth System Model Version 4.1 (GFDL-ESM 4.1): Overall Coupled Model Description and Simulation Characteristics, J Adv Model Earth Sy, 12, 2020.

ERA5 data documentation: https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels-monthly-means?tab=overview, last access: 20 July 2020.

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, Geosci Model Dev, 9, 1937-1958, 10.5194/gmd-9-1937-2016, 2016.

Eyring, V., Gillett, N. P., Rao, K. M. A., Barimalala, R., Parrillo, M. B., Bellouin, N., Cassou, C., Durack, P. J., 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, 423-552, 10.1017/9781009157896.005, 2021.

Eyring, V., Bock, L., Lauer, A., Righi, M., Schlund, M., Andela, B., Arnone, E., Bellprat, O., Brotz, 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., Muller, B., Pandde, A., Perez-Zanon, N., Phillips, A., Predoi, V., Russell, J., Sellar, A., Serva, F., Stacke, T., Swaminathan, R., Torralba, V., Vegas-Regidor, J., von Hardenberg, J., Weigel, 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, Geosci Model Dev, 13, 3383-3438, 10.5194/gmd-13-3383-2020, 2020.

Frömming, C., Grewe, V., Brinkop, S., Jöckel, P., Haslerud, A. S., Rosanka, S., van Manen, J., and Matthes, S.: Influence of weather situation on non-CO2 aviation climate effects: the REACT4C climate change functions, Atmos Chem Phys, 21, 9151-9172, 10.5194/acp-21-9151-2021, 2021.

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

Gettelman, A., Mills, M. J., Kinnison, D. E., Garcia, R. R., Smith, A. K., Marsh, D. R., Tilmes, S., Vitt, F., Bardeen, C. G., McInerny, J., Liu, H. L., Solomon, S. C., Polvani, L. M., Emmons, L. K., Lamarque, J. F., Richter, J. H., Glanville, A. S., Bacmeister, J. T., Phillips, A. S., Neale, R. B., Simpson, I. R., DuVivier, A. K., Hodzic, A., and Randel, W. J.: The Whole Atmosphere Community Climate Model Version 6 (WACCM6), J Geophys Res -Atmos, 2019.

Gleckler, P. J., Taylor, K. E., and Doutriaux, C.: Performance metrics for climate models, J Geophys Res-Atmos, 113, 10.1029/2007jd008972, 2008.

Guo, Y. Y., Yu, Y. Q., Lin, P. F., Liu, H. L., He, B., Bao, Q., Zhao, S. W., and Wang, X. W.: Overview of the CMIP6 Historical Experiment Datasets with the Climate System Model CAS FGOALS-f3-L, Adv Atmos Sci, 37, 1057-1066, 10.1007/s00376-020-2004-4, 2020.

Hajima, T., Watanabe, M., Yamamoto, A., Tatebe, H., Noguchi, M. A., Abe, M., Ohgaito, R., Ito, A., Yamazaki, D., Okajima, H., Ito, A., Takata, K., Ogochi, K., Watanabe, S., and Kawamiya, M.: Development of the MIROC-ES2L Earth system model and the evaluation of biogeochemical processes and feedbacks, Geosci Model Dev, 13, 2197-2244, 10.5194/gmd-13-2197-2020, 2020.

Hassler, B. and Lauer, A.: Comparison of Reanalysis and Observational Precipitation Datasets Including ERA5 and WFDE5, Atmosphere-Basel, 12, 10.3390/atmos12111462, 2021.

Hendricks, J., Righi, M., Dahlmann, K., Gottschaldt, K. D., Grewe, V., Ponater, M., Sausen, R., Heinrichs, D., Winkler, C., Wolfermann, A., Kampffmeyer, T., Friedrich, R., Klötzke, M., and Kugler, U.: Quantifying the climate impact of emissions from land-based transport in Germany, Transport Res D-Tr E, 65, 825-845, 10.1016/j.trd.2017.06.003, 2018.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horanyi, A., Munoz-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., Holm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thepaut, J. N.: The ERA5 global reanalysis, Q J Roy Meteor Soc, 146, 1999-2049, 10.1002/qj.3803, 2020.

Jöckel, P., Kerkweg, A., Pozzer, A., Sander, R., Tost, H., Riede, H., Baumgaertner, A., Gromov, S., and Kern, B.: Development cycle 2 of the Modular Earth Submodel System (MESSy2), Geosci Model Dev, 3, 717-752, 10.5194/gmd-3-717-2010, 2010.

Kato, S., Rose, F. G., Rutan, D. A., Thorsen, T. J., Loeb, N. G., Doelling, D. R., Huang, X. L., Smith, W. L., Su, W. Y., and Ham, S. H.: Surface Irradiances of Edition 4.0 Clouds and the Earth's Radiant Energy System (CERES) Energy Balanced and Filled (EBAF) Data Product, J Climate, 31, 4501-4527, 10.1175/Jcli-D-17-0523.1, 2018.

Kuhlbrodt, T., Jones, C. G., Sellar, A., Storkey, D., Blockley, E., Stringer, M., Hill, R., Graham, T., Ridley, J., Blaker, A., Calvert, D., Copsey, D., Ellis, R., Hewitt, H., Hyder, P., Ineson, S., Mulcahy, J., Siahaan, A., and Walton, J.: The Low-Resolution Version of HadGEM3 GC3.1: Development and Evaluation for Global Climate, J Adv Model Earth Sy, 10, 2865-2888, 10.1029/2018ms001370, 2018.

Lauer, A., Bock, L., Hassler, B., Schröder, M., and Stengel, M.: Cloud climatologies from global climate models – a comparison of CMIP5 and CMIP6 models with satellite data, J Climate, 36, 281-311, 10.1175/JCLI-D-22-0181.1, 2023.

Lauer, A., Eyring, V., Bellprat, O., Bock, L., Gier, B. K., Hunter, A., Lorenz, R., Perez-Zanon, 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, Geosci Model Dev, 13, 4205-4228, 10.5194/gmd-13-4205-2020, 2020.

Lee, J., Kim, J., Sun, M. A., Kim, B. H., Moon, H., Sung, H. M., Kim, J., and Byun, Y. H.: Evaluation of the Korea Meteorological Administration Advanced Community Earth-System model (K-ACE), Asia-Pac J Atmos Sci, 56, 381-395, 2020.

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

Li, L. J., Yu, Y. Q., Tang, Y. L., Lin, P. F., Xie, J. B., Song, M. R., Dong, L., Zhou, T. J., Liu, L., Wang, L., Pu, Y., Chen, X. L., Chen, L., Xie, Z. H., Liu, H. B., Zhang, L. X., Huang, X., Feng, T., Zheng, W. P., Xia, K., Liu, H. L., Liu, J. P., Wang, Y., Wang, L. H., Jia, B. H., Xie, F., Wang, B., Zhao, S. W., Yu, Z. P., Zhao, B. W., and Wei, J. L.: The Flexible Global Ocean-Atmosphere-Land System Model Grid-Point Version 3 (FGOALS-g3): Description and Evaluation, J Adv Model Earth Sy, 12, 10.1029/2019MS002012, 2020.

Lin, Y. L., Huang, X. M., Liang, Y. S., Qin, Y., Xu, S. M., Huang, W. Y., Xu, F. H., Liu, L., Wang, Y., Peng, Y. R., Wang, L. N., Xue, W., Fu, H. H., Zhang, G. J., Wang, B., Li, R. Z., Zhang, C., Lu, H., Yang, K., Luo, Y., Bai, Y. Q., Song, Z. Y., Wang, M. Q., Zhao, W. J., Zhang, F., Xu, J. H., Zhao, X., Lu, C. S., Chen, Y. Z., Luo, Y. Q., Hu, Y., Tang, Q., Chen, D. X., Yang, G. W., and Gong, P.: Community Integrated Earth System Model (CIESM): Description and Evaluation, J Adv Model Earth Sy, 12, 10.1029/2019MS002036, 2020.

Loeb, N. G., Doelling, D. R., Wang, H. L., Su, W. Y., Nguyen, C., Corbett, J. G., Liang, L. S., Mitrescu, C., Rose, F. G., and Kato, S.: Clouds and the Earth's Radiant Energy System (CERES) Energy Balanced and Filled (EBAF) Top-of-Atmosphere (TOA) Edition-4.0 Data Product, J Climate, 31, 895-918, 10.1175/Jcli-D-17-0208.1, 2018.

M. Nützel, L. Stecher, P. Jöckel, F. Winterstein, M. Dameris, M. Ponater, P. Graf, and Kunze, M.: Updating the radiation infrastructure in MESSy (based on MESSy version 2.55), EGUsphere [preprint], 2023, 1-44, 10.5194/egusphere-2023-2140, 2023.

Matthes, S., Lim, L., Burkhardt, U., Dahlmann, K., Dietmüller, S., Grewe, V., Haslerud, A. S., Hendricks, J., Owen, B., Pitari, G., Righi, M., and Skowron, A.: Mitigation of Non-CO2 Aviation's Climate Impact by Changing Cruise Altitudes, Aerospace-Basel, 8, 10.3390/aerospace8020036, 2021.

Mauritsen, T., Bader, J., Becker, T., Behrens, J., Bittner, M., Brokopf, R., Brovkin, V., Claussen, M., Crueger, T., Esch, M., Fast, I., Fiedler, S., Flaeschner, D., Gayler, V., Giorgetta, M., Goll, D. S., Haak, H., Hagemann, S., Hedemann, C., Hohenegger, C., Ilyina, T., Jahns, T., Jimenez-de-la-Cuesta, D., Jungclaus, J., Kleinen, T., Kloster, S., Kracher, D., Kinne, S., Kleberg, D., Lasslop, G., Kornblueh, L., Marotzke, J., Matei, D., Meraner, K., Mikolajewicz, U., Modali, K., Mobis, B., Muller, W. A., Nabel, J. E. M. S., Nam, C. C. W., Notz, D., Nyawira, S. S., Paulsen, H., Peters, K., Pincus, R., Pohlmann, H., Pongratz, J., Popp, M., Raddatz, T. J., Rast, S., Redler, R., Reick, C. H., Rohrschneider, T., Schemann, V., Schmidt, H., Schnur, R., Schulzweida, U., Six, K. D., Stein, L., Stemmler, I., Stevens, B., von Storch, J. S., Tian, F. X., Voigt, A., Vrese, P., Wieners, K. H., Wilkenskjeld, S., Winkler, A., and Roeckner, E.: Developments in the MPI-M Earth System Model version 1.2 (MPI-ESM1.2) and Its Response to Increasing CO2, J Adv Model Earth Sy, 11, 998-1038, 2019.

Mertens, M., Jöckel, P., Matthes, S., Nützel, M., Grewe, V., and Sausen, R.: COVID-19 induced lower-tropospheric ozone changes, Environ Res Lett, 16, 10.1088/1748-9326/abf191, 2021.

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, J Geophys Res-Atmos, 126, 10.1029/2019JD032361, 2021.

Muller, W. A., Jungclaus, J. H., Mauritsen, T., Baehr, J., Bittner, M., Budich, R., Bunzel, F., Esch, M., Ghosh, R., Haak, H., Ilyina, T., Kleine, T., Kornblueh, L., Li, H., Modali, K., Notz, D., Pohlmann, H., Roeckner, E., Stemmler, I., Tian, F., and Marotzke, J.: A Higher-resolution Version of the Max Planck Institute Earth System Model (MPI-ESM1.2-HR), J Adv Model Earth Sy, 10, 1383-1413, 2018.

Park, S., Shin, J., Kim, S., Oh, E., and Kim, Y.: Global Climate Simulated by the Seoul National University Atmosphere Model Version 0 with a Unified Convection Scheme (SAM0-UNICON), J Climate, 32, 2917-2949, 2019.

Rackow, T., Goessling, H. F., Jung, T., Sidorenko, D., Semmler, T., Barbi, D., and Handorf, D.: Towards multi-resolution global climate modeling with ECHAM6-FESOM. Part II: climate variability, Clim Dynam, 50, 2369-2394, 2018.

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, J Geophys Res-Atmos, 108, 10.1029/2002jd002670, 2003.

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

Righi, M., Hendricks, J., and Sausen, R.: The global impact of the transport sectors on atmospheric aerosol in 2030-Part 1: Land transport and shipping, Atmos Chem Phys, 15, 633-651, 10.5194/acp-15-633-2015, 2015.

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, Geosci. Model Dev., 13, 1179-1199, 10.5194/gmd-13-1179-2020, 2020.

Rind, D., Orbe, C., Jonas, J., Nazarenko, L., Zhou, T., Kelley, M., Lacis, A., Shindell, D., Faluvegi, G., Romanou, A., Russell, G., Tausnev, N., Bauer, M., and Schmidt, G.: GISS Model E2.2: A Climate Model Optimized for the Middle Atmosphere-Model Structure, Climatology, Variability, and Climate Sensitivity, J Geophys Res-Atmos, 125, 2020.

Roeckner, E., Brokopf, R., Esch, M., Giorgetta, M., Hagemann, S., Kornblueh, L., Manzini, E., Schlese, U., and Schulzweida, U.: Sensitivity of simulated climate to horizontal and vertical resolution in the ECHAM5 atmosphere model, J Climate, 19, 3771-3791, 2006.

Rong, X. Y., Li, J., Chen, H. M., Xin, Y. F., Su, J. Z., Hua, L. J., Zhou, T. J., Qi, Y. J., Zhang, Z. Q., Zhang, G., and Li, J. D.: The CAMS Climate System Model and a Basic Evaluation of Its Climatology and Climate Variability Simulation, J Meteorol Res-Prc, 32, 839-861, 2018.

Rubner, Y., Tomasi, C., and Guibas, L. J.: The Earth Mover's Distance as a metric for image retrieval, Int J Comput Vision, 40, 99-121, Doi 10.1023/A:1026543900054, 2000.

Schlund, M., Hassler, B., Lauer, A., Andela, B., Jöckel, P., Kazeroni, R., Tomas, S. L., 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, Geosci Model Dev, 16, 315-333, 10.5194/gmd-16-315-2023, 2023.

Séférian, R., Nabat, P., Michou, M., Saint-Martin, D., Voldoire, A., Colin, J., Decharme, B., Delire, C., Berthet, S., Chevallier, M., Sénési, S., Franchisteguy, L., Vial, J., Mallet, M., Joetzjer, E., Geoffroy, O., Guérémy, J.-F., Moine, M.-P., Msadek, R., Ribes, A., Rocher, M., Roehrig, R., Salas-y-Mélia, D., Sanchez, E., Terray, L., Valcke, S., Waldman, R., Aumont, O., Bopp, L., Deshayes, J., Éthé, C., and Madec, G.: Evaluation of CNRM Earth System Model, CNRM-ESM2-1: Role of Earth System Processes in Present-Day and Future Climate, J Adv Model Earth Sy, n/a, 10.1029/2019ms001791, 2019.

Seland, O., Bentsen, M., Olivie, D., Toniazzo, T., Gjermundsen, A., Graff, L. S., Debernard, J. B., Gupta, A. K., He, Y. C., Kirkevag, A., Schwinger, J., Tjiputra, J., Aas, K. S., Bethke, I., Fan, Y. C., Griesfeller, J., Grini, A., Guo, C. C., Ilicak, M., Karset, I. H. H., Landgren, O., Liakka, J., Moseid, K. O., Nummelin, A., Spensberger, C., Tang, H., Zhang, Z. S., Heinze, C., Iversen, T., and Schulz, M.: Overview of the Norwegian Earth System Model (NorESM2) and key climate response of CMIP6 DECK, historical, and scenario simulations, Geosci Model Dev, 13, 6165-6200, 2020.

Sellar, A. A., Jones, C. G., Mulcahy, J., Tang, Y., Yool, A., Wiltshire, A., O'connor, F. M., Stringer, M., Hill, R., and Palmieri, J.: UKESM1: Description and evaluation of the UK Earth System Model, J Adv Model Earth Sy, 4513-4558, 10.1029/2019MS001739, 2019.

Semmler, T., Danilov, S., Gierz, P., Goessling, H. F., Hegewald, J., Hinrichs, C., Koldunov, N., Khosravi, N., Mu, L. J., Rackow, T., Sein, D. V., Sidorenko, D., Wang, Q., and Jung, T.: Simulations for CMIP6 With the AWI Climate Model AWI-CM-1-1, J Adv Model Earth Sy, 12, 10.1029/2019MS002009, 2020.

Sidorenko, D., Rackow, T., Jung, T., Semmler, T., Barbi, D., Danilov, S., Dethloff, K., Dorn, W., Fieg, K., Goessling, H., Handorf, D., Harig, S., Hiller, W., Juricke, S., Losch, M., Schroter, J., Sein, D. V., and Wang, Q.: Towards multi-resolution global climate modeling with ECHAM6-FESOM. Part I: model formulation and mean climate, Clim Dynam, 44, 757-780, 2015.

Swart, N. C., Cole, J. N. S., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., Anstey, J., Arora, V., Christian, J. R., Hanna, S., Jiao, Y. J., Lee, W. G., Majaess, F., Saenko, O. A., Seiler, C., Seinen, C., Shao, A., Sigmond, M., Solheim,

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

L., von Salzen, K., Yang, D., and Winter, B.: The Canadian Earth System Model version 5 (CanESM5.0.3), Geosci Model Dev, 12, 4823-4873, 2019.

Tatebe, H., Ogura, T., Nitta, T., Komuro, Y., Ogochi, K., Takemura, T., Sudo, K., Sekiguchi, M., Abe, M., Saito, F., Chikira, M., Watanabe, S., Mori, M., Hirota, N., Kawatani, Y., Mochizuki, T., Yoshimura, K., Takata, K., O'ishi, R., Yamazaki, D., Suzuki, T., Kurogi, M., Kataoka, T., Watanabe, M., and Kimoto, M.: Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in MIROC6, Geosci Model Dev, 12, 2727-2765, 2019.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An Overview of Cmip5 and the Experiment Design, B Am Meteorol Soc, 93, 485-498, 10.1175/Bams-D-11-00094.1, 2012.

Uribe, A., Bender, F. A. M., and Mauritsen, T.: Observed and CMIP6 Modeled Internal Variability Feedbacks and Their Relation to Forced Climate Feedbacks, Geophys Res Lett, 49, 10.1029/2022GL100075, 2022.

Vissio, G. and Lucarini, V.: Evaluating a stochastic parametrization for a fast-slow system using the Wasserstein distance, Nonlinear Proc Geoph, 25, 413-427, 10.5194/npg-25-413-2018, 2018.

Vissio, G., Lembo, V., Lucarini, V., and Ghil, M.: Evaluating the Performance of Climate Models Based on Wasserstein Distance, Geophys Res Lett, 47, 10.1029/2020GL089385, 2020.

Voldoire, A., Saint-Martin, D., Senesi, S., Decharme, B., Alias, A., Chevallier, M., Colin, J., Gueremy, J. F., Michou, M., Moine, M. P., Nabat, P., Roehrig, R., Melia, D. S. Y., Seferian, R., Valcke, S., Beau, I., Belamari, S., Berthet, S., Cassou, C., Cattiaux, J., Deshayes, J., Douville, H., Ethe, C., Franchisteguy, L., Geoffroy, O., Levy, C., Madec, G., Meurdesoif, Y., Msadek, R., Ribes, A., Sanchez-Gomez, E., Terray, L., and Waldman, R.: Evaluation of CMIP6 DECK Experiments With CNRM-CM6-1, J Adv Model Earth Sy, 11, 2177-2213, 2019.

Volodin, E. M., Mortikov, E. V., Kostrykin, S. V., Galin, V. Y., Lykosov, V. N., Gritsun, A. S., Diansky, N. A., Gusev, A. V., and Yakovlev, N. G.: Simulation of modern climate with the new version of the INM RAS climate model, Izv Atmos Ocean Phy+, 53, 142-155, 10.1134/S0001433817020128, 2017.

Volodin, E. M., Mortikov, E. V., Kostrykin, S. V., Galin, V. Y., Lykossov, V. N., Gritsun, A. S., Diansky, N. A., Gusev, A. V., Iakovlev, N. G., Shestakova, A. A., and Emelina, S. V.: Simulation of the modern climate using the INM-CM48 climate model, Russ J Numer Anal M, 33, 367-374, 10.1515/rnam-2018-0032, 2018.

Wang, X., Lu, R., Wang, S. Y., Chen, R. T., Chen, Z. Q., Hui, F. M., Huang, H. B., and Cheng, X.: Assessing CMIP6 simulations of Arctic sea ice drift: Role of near-surface wind and surface ocean current in model performance, Adv Clim Chang Res, 14, 691-706, 10.1016/j.accre.2023.09.005, 2023.

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., Lledo, L., Mohr, C. W., Pacal, A., Perez-Zanon, 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, Geosci Model Dev, 14, 3159-3184, 10.5194/gmd-14-3159-2021, 2021.

Williams, K. D., Copsey, D., Blockley, E. W., Bodas-Salcedo, A., Calvert, D., Comer, R., Davis, P., Graham, T., Hewitt, H. T., Hill, R., Hyder, P., Ineson, S., Johns, T. C., Keen, A. B., Lee, R. W., Megann, A., Milton, S. F., Rae, J. G. L., Roberts, M. J., Scaife, A. A., Schiemann, R., Storkey, D., Thorpe, L., Watterson, I. G., Walters, D. N., West, A., Wood, R. A., Woollings, T., and Xavier, P. K.: The Met Office Global Coupled Model 3.0 and 3.1 (GC3.0 and GC3.1) Configurations, J Adv Model Earth Sy, 10, 357-380, 10.1002/2017ms001115, 2018.

Wu, T. W., Zhang, F., Zhang, J., Jie, W. H., Zhang, Y. W., Wu, F. H., Li, L., Yan, J. H., Liu, X. H., Lu, X., Tan, H. Y., Zhang, L., Wang, J., and Hu, A. X.: Beijing Climate Center Earth System Model version 1 (BCC-ESM1): model description and evaluation of aerosol simulations, Geosci Model Dev, 13, 977-1005, 10.5194/gmd-13-977-2020, 2020.

Wu, T. W., Lu, Y. X., Fang, Y. J., Xin, X. G., Li, L., Li, W. P., Jie, W. H., Zhang, J., Liu, Y. M., Zhang, L., Zhang, F., Zhang, Y. W., Wu, F. H., Li, J. L., Chu, M., Wang, Z. Z., Shi, X. L., Liu, X. W., Wei, M., Huang, A. N., Zhang, Y. C., and

https://doi.org/10.5194/egusphere-2024-1518
Preprint. Discussion started: 26 June 2024

Liu, X. H.: The Beijing Climate Center Climate System Model (BCC-CSM): the main progress from CMIP5 to CMIP6, Geosci Model Dev, 12, 1573-1600, 2019.

Yazdandoost, F., Moradian, S., Izadi, A., and Aghakouchak, A.: Evaluation of CMIP6 precipitation simulations across different climatic zones: Uncertainty and model intercomparison, Atmos Res, 250, 10.1016/j.atmosres.2020.105369, 2021.

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 Syst Sci Data, 10, 583-593, 10.5194/essd-10-583-2018, 2018.

Yukimoto, S., Kawai, H., Koshiro, T., Oshima, N., Yoshida, K., Urakawa, S., Tsujino, H., Deushi, M., Tanaka, T., Hosaka, M., Yabu, S., Yoshimura, H., Shindo, E., Mizuta, R., Obata, A., Adachi, Y., and Ishii, M.: The Meteorological Research Institute Earth System Model Version 2.0, MRI-ESM2.0: Description and Basic Evaluation of the Physical Component, J Meteorol Soc Jpn, 97, 931-965, 2019.

Zhang, Y. C. and Rossow, W. B.: Global Radiative Flux Profile Data Set: Revised and Extended, J Geophys Res-Atmos, 128, 10.1029/2022JD037340, 2023.

Ziehn, T., Chamberlain, M. A., Law, R. M., Lenton, A., Bodman, R. W., Dix, M., Stevens, L., Wang, Y. P., and Srbinovsky, J.: The Australian Earth System Model: ACCESS-ESM1.5, J So Hemisph Earth, 70, 193-214, 10.1071/Es19035, 2020.