https://doi.org/10.5194/egusphere-2025-4744
Preprint. Discussion started: 6 October 2025

Developing Guidelines for Working with Multi-Model Ensembles in CMIP

Anja Katzenberger[1,2], Jhayron S. Perez-Carrasquilla[3], Keighan Gemmell[4], Evgenia Galytska[5,6], Christine Leclerc[7], Punya P[8], Indrani Roy[9], Arianna Varuolo-Clarke[10,11], Milica Tošić[12], Nina Črnivec[13]

[1] Potsdam Institute for Climate Impact Research, Potsdam, 14473, Germany

[2] Institute of Physics and Astronomy, Potsdam University, Potsdam, 14469, Germany

[3] Atmospheric and Oceanic Science Department, University of Maryland, College Park, 20740, United States

[4] Department of Chemistry, The University of British Columbia, Vancouver, V6T 1Z4, Canada

[5] University of Bremen, Institute of Environmental Physics, Bremen, Germany

[6] Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany

[7] Department of Geography, Simon Fraser University, Burnaby, V5A 1S6, Canada

[8] Department of Earth and Space Sciences, Indian Institute of Space Science and Technology, Trivandrum, 695547, India

[9] University College London (UCL), Earth Science Department, Gower Street, London, WC1E 6BT, UK

[10] Cooperative Programs for the Advancement of Earth System Science, University Corporation for Atmospheric Research, Boulder, CO

[11] Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO

[12] Faculty of Physics, University of Belgrade, Belgrade, 11000, Serbia

[13] Department of Atmospheric Science, Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, 1000, Slovenia

Correspondence to: Anja Katzenberger (anja.katzenberger@pik-potsdam.de)

Abstract. Earth System Models (ESMs) are the key tool for studying the climate under changing conditions. Over recent decades, it has been established to not only rely on projections of a single model but to combine various ESMs in multi-model ensembles (MMEs) to improve robustness and quantify the uncertainty of the projections. The data access for MME studies has been fundamentally facilitated by the World Climate Research Programme's Coupled Model Intercomparison Project (CMIP) - a collaborative effort bringing together ESMs from modelling communities all over the world. Despite the CMIP standardisation processes, addressing specific research questions using MMEs requires unique ensemble design, analysis, and interpretation choices. Based on the collective expertise within the Fresh Eyes on CMIP initiative, mainly composed of early-career researchers engaged in CMIP, we have identified common issues and questions encountered while working with climate MMEs. In this project, we provide a comprehensive literature review addressing these questions. We provide statistics tracing the development of the climate MMEs analysis field throughout the last decades, and, synthesising existing studies, we outline guidelines regarding model evaluation, model dependence, weighting methods, and uncertainty treatment. We summarize a collection of useful resources for MME studies, we review common questions and strategies, and finally, we outline emerging scientific trends, such as the integration of machine learning (ML) techniques, single model initial-condition large ensembles (SMILES), and computational resource considerations. We thereby strive to support researchers working with climate MMEs particularly in the upcoming 7th phase of CMIP.

1 Introduction

The Earth system models (ESMs), whose data is provided by the World Climate Research Programme (WCRP) Coupled Model Intercomparison Project (CMIP), are the key tool for making future climate projections. These projections are essential for informing communities and policy-makers, helping develop both mitigation and adaptation strategies to climate change at the global and regional scales (Meehl et al., 2000). Starting from the seminal work of Manabe and Hasselmann (e.g., Manabe and Strickler, 1964; Manabe and Wetherald, 1967; Manabe and Bryan, 1969; Hasselmann, 1976), who were awarded the 2021 Nobel Prize in Physics for laying the foundation of climate modelling, climate models have continuously evolved over decades. During this process, models have become progressively more complex encapsulating processes related to aerosols, atmospheric chemistry, the carbon cycle, and ocean biogeochemistry (IPCC, AR4, AR5, AR6). This development of ESMs has been going "hand in hand" with advances in Earth system observations, high-resolution numerical models giving valuable insight into smaller-scale phenomena (e.g., detailed radiative transfer models, cloud-resolving models, large-eddy simulations), and growing computational power (e.g. Gettelman et al., 2022; Schneider et al., 2017) allowing horizontal and vertical model resolution to steadily improve. Concurrently, the ESM simulation output data has been steadily increasing (Williams et al., 2016) and is stored at the Earth System Grid Federation (ESGF) central repository (Cinquini et al., 2012).

The main components of an ESM are models describing the atmosphere, ocean, cryosphere, land, and increasingly, the carbon cycle and other biogeochemical processes. Each component involves a variety of interacting phenomena occurring at a wide range of spatial and temporal scales (e.g. Gettelman et al., 2022). For instance, the atmospheric component involves phenomena spanning from micro-scale events, such as formation of cloud droplets on aerosol particles, to global-scale dynamics like planetary Rossby waves. In all ESMs, the continuous behavior of the atmosphere is first discretized in space and time via the so-called "model dynamical core," which encompasses the governing equations that capture the resolved (grid-scale) phenomena as well as the physical parameterization schemes for representing unresolved (subgrid-scale) processes. Various

ESMs thereby generally differ in the choice of computational grids (e.g., latitude-longitude structured grid, icosahedral grid, variable resolution cube-sphere grid), numerical methods for solving the dynamical core equations, as well as in physical parameterization schemes.

In summary, each ESM is an attempt to represent a multitude of highly complex, nonlinear processes, and what is even more difficult, the synchronized interplay among them. Within each of the model components, there are processes that are well represented by known and proved physical laws. However, our current knowledge of how the Earth system operates is still limited. Many processes are represented in models through parameterizations — relationships used to approximate behaviour of unresolved or poorly understood phenomena. While some parameterizations are based on well-established physical theory, others, particularly those related to clouds or turbulence, remain subject to substantial uncertainty. In addition to our incomplete knowledge about the climate system, there are also computational limitations that hinder the fidelity of the models to represent certain relevant processes. The decisions made at modeling centers in response to these limiting factors make each model a unique imperfect idealization of the Earth system, and depending on the processes of interest to the end user, some idealizations may be more suitable than others. To account for this model uncertainty, models are combined in multi-model ensembles (MMEs).

Besides the possibility to quantify uncertainty and increase robustness, MMEs have been found to generally outperform the projections of individual models. Inspired by the findings within the weather forecasting community, where numerous studies have shown that ensemble forecasts are more reliable than individual forecasts (Doblas-Reyes et al., 2003; Krishnamurti et al., 1999), studies in the climate context also analysed the potential benefits from working with MMEs. In climate model evaluation, the MME has proven to outperform individual models in numerous studies e.g. regarding the mean (Gleckler et al., 2008; Knutti et al., 2010a; Lambert and Boer, 2001; Palmer et al., 2005; Phillips and Gleckler, 2006; Pincus et al., 2008; Reichler and Kim, 2008) or variability (Zhang et al., 2007), further strengthening the motivation to use MMEs.

Given these benefits, MME studies have become an established tool for climate studies addressing a broad range of research questions. In the process, they also became the standard method to analyse and present results in the Assessment Reports (ARs) of the Intergovernmental Panel on Climate Change (IPCC) where the state-of-the-art knowledge on climate change is reviewed. For researchers, MMEs provide an efficient way to get an overview of general tendencies for specific questions. Also for non-experts, presenting results in a synthesised format as e.g. in the context of MME also facilitates accessibility and interpretation (Knutti et al., 2010a), underlining the benefits of MMEs for the users.

Since the beginning of large-scale atmospheric modelling in the 1950s, such intercomparison among models has been carried out. Initially, this intercomparison was mostly performed for numerical weather prediction as computational resources limited the intercomparison of studies in the climate studies, and a clear experimental strategy was lacking (Gates, 1992). Since the 1970s, the Working Group on Numerical Experimentation (WGNE), supporting the World Climate Research Programme, has organised several intercomparison projects among climate models (Gates, 1992). The first international systematic intercomparison framework for climate models was established in 1990 in the context of the Atmospheric Model Intercomparison Project (AMIP; Gates, 1992). In the early 1990s, the Intergovernmental Panel on Climate Change (IPCC) provided an intercomparison of atmospheric models in their first assessment report (AR; Gates, 1992). Räisänen (1997) advocated the need for quantitative model comparison and raised the thought that the agreement between models can indirectly serve as a measure for the reliability of the simulations. Accordingly, Räisänen and Palmer (2001) introduced a probabilistic perspective on multi-model ensemble projections. The authors quantified the probability of specific climate events happening based on 17 coupled atmosphere-ocean general circulation models (AOGCMs). Contemporaneously, AMIP was followed by the Coupled Model Intercomparison Project (CMIP), which also incorporated results from AOGCMs (Meehl et al., 2000). While the first phase of CMIP was limited to control runs, new standardised scenarios were incorporated throughout the phases of CMIP with an increasing number of international model centres contributing simulations. Also, in recent CMIP generations, a variety of supporting experiments is conducted (e.g. Eyring et al., 2016), including paleoclimate runs (simulations of the 'distant past'), historical runs (simulations of the 'recent past'), control runs to study natural variability, as well as various developmental runs such as AMIP experiments. In AMIP simulations, for example, various modelling centres use prescribed global sea surface temperature (SST) fields which enables the intercomparison of the atmospheric model component across various ESMs, while excluding effects of differing ocean models. Finally, future climate change experiments are performed for various greenhouse gas emission scenarios such as abrupt carbon dioxide doubling or quadrupling to derive equilibrium climate sensitivity (measure of how much the Earth's climate system will warm under a doubling of atmospheric CO2 concentration) as well as for multiple "shared socioeconomic pathways (SSPs)" (O'Neill et al., 2017; Riahi et al., 2017). The latter denote diverse scenarios of evolution of the global society (including population, economy, and technology) which thus lead to differing emissions of greenhouse gases (CO_2, CH_4, NO_2) and other air pollutants until the end of the 21st century and are associated with different climate change mitigation and adaptation policies and challenges (IPCC, AR6).

The availability of standardised climate model outputs facilitated model intercomparison and has naturally inspired the use of multi-model ensembles (MMEs) since the beginning of the 2000s (Tebaldi and Knutti, 2007). Consequently, the AR3 of the IPCC (2001) presented many results based on MME means, accompanied by measures of inter-model variability (Tebaldi and Knutti, 2007). In the AR4 of IPCC (2007), model projections were only included if the models were successors from previous generations, thus a model selection *de facto* has taken place (Knutti et al., 2010b) . To support IPCC lead authors for the AR5 and later, a "Good Practice Guidance Paper" was published in 2010, summarising current recommendations for the work with MMEs (Knutti et al., 2010b).

In the meantime, numerous studies have proposed diverse methods for MME studies (e.g., in the context of model selection or model weighting). However, for individual researchers whose main focus is often on the specific atmospheric or ocean problem that they study, it is challenging to have an overview of these studies. There is still a lack of guidelines on how to combine models within MMEs (Herger et al., 2018). The design of MME studies involves a set of decisions related to model selection, weighting, and uncertainty measures. Each of these decisions requires careful consideration of a broad range of aspects and often entails compromises that differ depending on the research question, as the advantages and disadvantages are highly dependent on the individual study's details. We acknowledge that this individuality makes it challenging and sometimes even impossible to establish universally applicable guidelines for MME studies. However, we believe it is valuable to give an overview of the key aspects to consider, and in some cases, present approaches that the Fresh Eyes on CMIP community has found to be useful. With this, we hope to support researchers that have newly entered the field of MME studies, but also to provide an overview of existing resources and approaches for more experienced MME researchers, particularly for (but not restricted to) the upcoming 7th phase of CMIP.

While the focus of this paper is on the challenges associated with working with climate MMEs, it should be pointed out there are other types of climate ensembles such as initial condition ensembles (ICE) and perturbed parameter ensembles (PPE) (IPCC, AR5). Similarly as in the weather forecasting community, the climate ICE is generated with a single climate model using varying initial conditions (i.e., perturbed initial state) to address the uncertainty due to natural or internal variability. If sufficiently many ensemble members are available, they are referred to as Single Model Initial-condition Large Ensembles (SMILEs). The perturbed parameter ensemble (frequently called also the perturbed physics ensemble) also compares multiple realizations from a single climate model, but in this case, a set of chosen physical parameters which are assumed to affect the quantity of interest (e.g., global mean surface temperature) is systematically varied to quantify the effect on model outcome (e.g. Eidhammer et al., 2024; Sexton et al., 2021). This enables a systematic exploration of intra-model uncertainty. Finally, the so-called grand ensembles are based on a combination (nesting) of various ensemble types - for example, PPE or MME followed by an ICE (IPCC, AR6).

In the following section, we conduct a comprehensive literature review on studies regarding model evaluation (2.1), systematic model biases (2.2), model dependence (2.3), model selection and weighting methods (2.4) and uncertainty characterization (2.5). In this context, we also provide a summary of useful tools for MME analysis (2.6). In the third section, we complement these guidelines with a collection of frequently asked questions and challenges that appear while working with MMEs based on the experience of the WCRP Fresh Eyes on CMIP community. We address these questions based on the literature. In the fourth section, we discuss emerging trends for working with MMEs such as ML, SMILEs and the necessity for more awareness of computational resources associated with MME studies.

2 Guidelines for working with MMEs

Over 84 General Circulation Models (GCMs) from at least 43 international institutes are available in the context of the CMIP network (https://wcrp-cmip.org/map/). When addressing any specific research question, the need for specific variables,

https://doi.org/10.5194/egusphere-2025-4744
Preprint. Discussion started: 6 October 2025

scenarios, resolutions or experiment participation narrows the pool of available models. However, the remaining number is often still rather large prompting the question: which of those models should be included for a specific analysis? Should all available models be utilised, or only a subset? How to identify the models that are most suitable? The choice of adequate selection criteria to distinguish between more and less suitable models for specific MME studies is central for the study design. The two primary objectives when selecting models are to firstly optimize model performance and secondly, reduce duplicated information, thus to create a subset of independent models (Herger et al., 2018). The subsection 2.1 focuses on how to perform a model evaluation and subsection 2.2. provides examples of existing model bias, while subsection 2.3 discusses model dependency. Subsection 2.4 gives an overview of selection and weighting methods and subsection 2.5 introduces the quantification of uncertainty. Subsection 2.5 lists useful tools and resources for MME analysis.

2.1 Model Evaluation

Observation datasets for model evaluation

The reference data sets are a key element of model evaluation. These are typically observations or reanalysis data derived from observations. A wide array of observational datasets used in ESM evaluation comprise paleoclimate data, measurements from ground-based stations over land, various ocean observational platforms, ships and buoys, sail drones, aircraft and balloon (in-situ) measurements, and satellite data. These observational datasets are frequently used in synergy, as they generally all have advantages and disadvantages (e.g., cover different spatial and temporal scales and time periods, are based on differing measurement techniques, have different accuracy, etc.). The paleoclimate data give insight into the state of the Earth's climate hundreds to millions of years ago and simultaneously provide valuable constraints on climate models for paleoclimate simulations, which help us understand recent and future climate change in the context of longer-term climate variability. For the more recent past, most of the reference observations originated in land in-situ measurements. It is important to keep in mind that these ground-based observations are not equally distributed around the globe (e.g., there are more land measurement stations in the Northern Hemisphere than in the Southern Hemisphere). The advent of Earth observation satellites has revolutionized the availability of global reference data sets, which are of key importance for the evaluation of global climate models. However, satellite datasets are limited to the time after the 1970s or later, depending on the variable of interest.

Moreover, model evaluation using observations is not always straightforward because observational sensors do not necessarily measure variables simulated by climate models. To ensure an "apple-to-apple comparison," observed quantities must be properly converted into model-output-like variables, or vice versa. To that end, comprehensive satellite simulation software has been developed which enables simulating what a satellite would observe flying over the model atmosphere. Also it is important to keep in mind that each observational data set is associated with observational uncertainty, e.g. due to instrument uncertainty, calibration limitations, or the interpolation procedure. Accounting for uncertainty in the observational data sets used as reference can be done by including multiple data sets. Depending on the variable of interest, commonly used reanalysis data sets are ERA5 (produced by ECMWF), MERRA-2 (produced by NASA GSFC), NCEP-NCAR reanalysis (produced by NOAA and UCAR), JRA-55 (produced by Japan Meteorological Agency). Also, it must be assured that observation and simulations have the same temporal and spatial resolution, including the horizontal grid and number of vertical levels (Simpson et al., 2025). This can be also achieved by appropriate regridding methods. However, the regridding has to be conducted with care as also conservative remapping of e.g. precipitation changes the statistical properties of the variable (Simpson et al., 2025).

Another issue to bear in mind is the problem of "model tuning", where model parameters are adjusted to best match the observational dataset, e.g., the observational dataset which is used for model evaluation was previously used for model tuning. In the case of reanalysis data, however, models are included in their creation and therefore using reanalysis data for reference is even more problematic, as the underlying data set should be independent.

Generally, there are two approaches for model evaluation. The performance-oriented approach focuses on identifying the models that perform best concerning the research question, meaning their output is closest to observations or reanalysis data. The process-oriented approach seeks models that best capture the relevant dynamics. Regardless of the chosen approach, it is essential for any research project to report on the performance of all models available before applying any ranking or weighting methods, and the selection criteria should be reported transparently (Knutti et al., 2010a). Such evaluations are sometimes already available in the literature and can be referred to. But in that case it is important to make sure that they cover the variables, scales etc. as relevant to the specific research questions that are of interest in the new study.

Performance-oriented evaluation

In weather forecasting, predictions can be verified within days as actual weather observations become available.This is not the case in climate model projections where the scales are much longer than weather scales (decades to centuries) and prevent any immediate verification. Therefore, climate model performances are evaluated with reference to past and present-day climatology (Knutti, 2010). Performance-oriented model evaluation is based on the assumption that models that performed well for the past regarding some specific climate phenomena will also perform well for the future climate.

Taylor diagrams (Taylor, 2001) serve as a very useful tool to assess model performance against observations. Such analyses help to identify better performing models, which may be more useful than others. Also outliers can be identified. Models closer to the observed standard deviation, along with higher correlation values and hence lesser root mean square errors are considered as better performing models for specific climate features (Taylor, 2001) and those can also be used for evaluating future climate. For example, the Western Pacific pattern, a prominent teleconnection pattern during the boreal winter over the North Pacific was analysed for 56 CMIP6 models using a Taylor diagram (Fig. 1, Aru et al., 2023). It depicts that the spatial correlations of the geopotential height anomalies at 500-hPa over the Western North Pacific between individual CMIP6 models and observations generally exceed 0.6. Also, in reproducing spatial patterns, the mean of the MME typically outperforms most

https://doi.org/10.5194/egusphere-2025-4744
Preprint. Discussion started: 6 October 2025

individual models, which is evidenced by a spatial root mean square deviation of 0.97. This diagram also makes it possible to identify outlier models, such as the MIROC-ES2L in this example. Finally, only the best performing models can be considered when estimating the final MME mean to improve results. This method can be used for different phenomena, e.g. for analysing the Indian Summer Monsoon (Roy et al., 2019) or for exploring seasonal mean temperature (Tang et al., 2016).

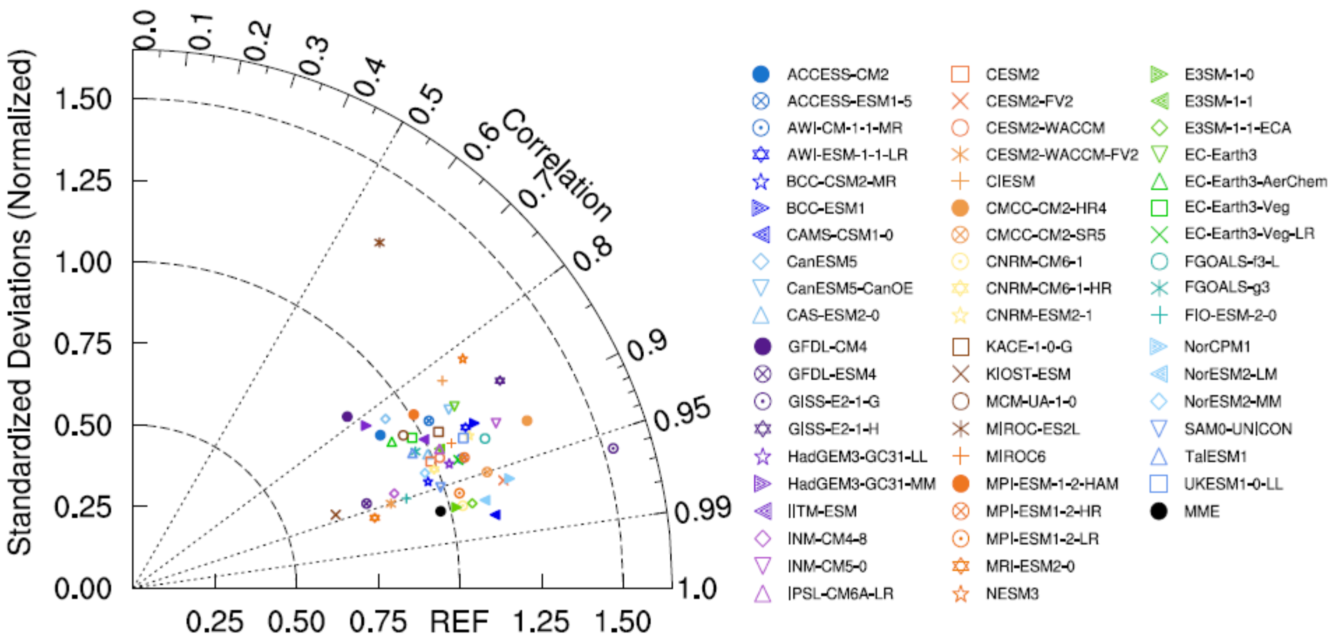

Fig. 1. Taylor diagram showing the geopotential height anomalies at 500-hPa over the Western North Pacific (20°N-80°N, 120°E-120°W) in individual CMIP6 models, MME and observations, taken from Aru et al. (2023).

It is important to remember that models are calibrated with the aim to reduce anomalies compared to observational data before becoming available in the CMIP context. During this calibration, various parameters are adjusted to reduce model bias. Consequently, improvements in overall model performance may not necessarily stem from enhanced capabilities in capturing relevant processes but optimized calibration (Knutti, 2010). On the other hand, this complex calibration procedure does not only have to compromise one individual regional pattern and the associated circulation. Thus, the calibration was not designed to optimize for specific climate phenomena, and parameters are not tuned to get as close as possible to specific variable patterns. Additionally, observational data also influence model behavior through the forcings themselves — for instance, in concentration-driven CO_2 simulations, where observed atmospheric concentrations are prescribed directly for historical simulations, rather than being computed from emissions (as in emission-driven models). This approach further constrains the model output, as the model does not simulate atmospheric CO_2 concentrations from emissions via an interactive carbon cycle. As a result, improvements in the model's output do not necessarily indicate better representation of the carbon cycle itself.

Another deficiency of this assumption is based on the fact that the climate is changing. While a reasonable performance in today's climate might serve as a reasonably good proxy to decide if the model captures current and past dynamics well, the role of specific circulation patterns and their interactions might change throughout the 21st century. In this context, (Knutti et al., 2010a) found that the model performance evaluated for the past correlates only weakly with the magnitude of the projected change in the future, illustrating that constraining models based on their performance in the past does not necessarily reduce the intermodal spread in the future. Given these pitfalls, Mendlik and Gobiet (2016) propose to only remove the severely unrealistic models alternatively. A detailed assessment on how to deal with outliers can be found in subsection 3.4. However, it remains interesting and relevant to understand which models perform best concerning a specific question and the assumption (models that perform well in the past will also do so in the future) may provide relevant insights given the lack of alternatives.

The praxis of performance-oriented model evaluation comes down to the choice of appropriate metrics. Model ranking has been found to be sensitive to this choice (Gleckler et al., 2008). However, for specific variables, it is possible that the model projections are independent of the choice of underlying metrics and ranking methods (Santer et al., 2009). Given the diversity of possible research questions, there is no single or combined performance metric that can reliably identify the "best" model independent of the research question). While this may sound disappointing since it prevents the standardization of model evaluation, it also has the advantage of reducing the effect of model convergence due to tuning (Knutti, 2010), which allows for a more reliable representation of future uncertainty and decreases the likelihood of making overconfident predictions. Generally, a metric is recommended if it's as simple as possible while at the same time being as statistically robust as possible, meaning that the dependence on specifications of the metric is rather low (Knutti et al., 2010b). Therefore, for any study, it is essential to determine the metrics that are relevant to the specific research question. One relevant aspect is the spatial and temporal scale of the phenomenon in question. For example, if the analysis is supposed to quantify extremes on a daily basis, then the performance on a daily scale should be the focus of the evaluation procedure.

A frequent challenge in climate model evaluation is determining whether models yield correct results for incorrect reasons, due to compensating errors (Eyring et al., 2016; Ivanova et al., 2016). There is a possibility that, while a model appears to accurately represent some variable, the underlying processes are not well-captured, which could mask inherent biases in the model. For example, analysing CMIP6 models, Zhao et al. (2022) reported that the cloud radiative effect reveals compensating errors between the modeled total cloud fraction and the liquid water path. These errors offset each other, resulting in a smaller net error in the cloud radiative effect. Di Luca et al. (2020a) addressed the issue of error compensation in CMIP5 simulations of hot temperature extremes by developing a new error metric called the "additive error." This metric adds up the absolute errors of four components contributing to temperature extremes: the long-term mean, seasonality, diurnal temperature range, and the local temperature anomaly on the day of the extreme. Compared to traditional bias or absolute error metrics, the additive error more sensitively captures the total error in extreme temperature estimates. Furthermore, Di Luca et al. (2020b) defined a new error estimator that aims to minimize error compensation.

Ideally, the evaluation process also allows insights on how well basic dynamic processes relevant to the research questions are reproduced in models (Knutti et al., 2010b). For a research question regarding rainfall, for example, this could mean to not only analyze the precipitation pattern, but also inspect wind patterns to see if the associated circulation is captured well. Process-oriented model evaluation specifically targets the model performance concerning such dynamics.

Process-oriented evaluation

Process-oriented evaluation of climate ESMs, particularly within CMIP, focuses on assessing how well models simulate the individual physical processes driving climate behavior. This approach shifts from traditional performance-oriented evaluation to more detailed, process-oriented metrics, critical for advancing the next generation of climate ESMs. Almost two decades ago, Eyring et al. (2005); Gleckler et al. (2008) emphasized the need to evaluate a wide range of climate processes, since accurate simulation of one aspect doesn't ensure accuracy in others. The authors recommended developing a comprehensive set of model metrics to assess important processes in climate simulations. Therefore, process-oriented evaluation identifies sources and limitations of predictability, enhancing model performance and more reliable climate projections (Eyring et al., 2016). It also fosters collaboration across modeling centers, integrating model development and evaluation efforts to ensure consistency and improve accuracy. By incorporating process-oriented analysis into diagnostic packages, evaluations become reproducible, accelerating model improvements and establishing benchmarks for progress. In the MME framework, this approach helps identify which processes contribute most to inter-model differences, providing insights into the mechanisms behind model performance. Below, we highlight examples of process-oriented analysis applied to CMIP models.

Using observations for processed-based evaluation: Ahmed and Neelin (2021) utilized the observed relationship between tropical precipitation and buoyancy as a foundation for a process-oriented analysis of CMIP6 models. They quantitatively assessed the thermodynamic sensitivities of convection across these models and applied regime-oriented diagnostics. Their findings indicated that several models exhibited excessive moisture sensitivity, potentially due to underactive convective schemes or tuning assumptions. Consequently, models with this excessive moisture sensitivity tended to have mean precipitation states biased toward grid-scale saturation.

Another example is the Indian Summer monsoon (ISM) and the El Nino Southern Oscillation (ENSO) teleconnection what was captured well in MME of CMIP5 and CMIP6 models (Katzenberger et al., 2021; Roy et al., 2017; Roy and Tedeschi, 2016). Around central northeast India, the teleconnection is strongest (Roy et al., 2017). For El Nino, there is a significant

https://doi.org/10.5194/egusphere-2025-4744
Preprint. Discussion started: 6 October 2025

deficit of rain, while for La Nina there is a significant excess rain. For the MME, the method used is simple mean ('one-model-one-vote', Knutti, 2010), instead of weighting or ranking models. Results are similar even when MME of only good models (as was identified for ISM by Jourdain et al., 2013) are considered. Anomalies in precipitation for different types of ENSO are captured well in most models and MME, agreeing with observation (see details in Roy et al., 2017). The model ensemble of ISM and SST in the Pacific showed a clear connection between Walker circulation and ISM across the central northeast India, matching observation. This region of India is the meeting point of Hadley and Walker circulation during ISM, that coupling process and teleconnection seems captured well by most CMIP models as well as MME, allowing us to understand why the teleconnection is captured well.

Using observations for a multiple diagnostic ensemble regression: Karpechko et al. (2013) developed the multiple diagnostic ensemble regression (MDER) methodology to link future climate projections with process-oriented diagnostics evaluating twentieth century processes, applying it to Antarctic ozone columns. MDER identifies key processes influencing ozone and explains variability in projected ozone across climate chemistry models (CCMs). The regression model, based on observed diagnostics, is then applied to predict future ozone and its uncertainty. Validated in a pseudo-realistic setting, MDER outperforms the unweighted Multi-Model Mean in forecasting Antarctic ozone levels. Wenzel et al. (2016) applied MDER algorithm (represented as a diagnostic in ESMValTool, see Section 2.6) to analyze the austral jet position in projections of the twenty-first century under the RCP4.5 scenario of CMIP5 simulations.The authors state that MDER reduced uncertainty in the ensemble mean projection without significantly changing the jet's long-term position.

Process-oriented evaluation to reduce model bias: Another key focus is the development of process-oriented metrics for phenomena that have a strong bias in the models, as e.g. MJO, the dominant mode of tropical intraseasonal variability. To address the reasons for these biases, a number of process-oriented diagnostics was developed to facilitate improvements in the representation of the MJO in weather and climate models (Ahn et al., 2020; Li et al., 2022; Wang et al., 2020). The first multi-model comparison study on MJO teleconnections was conducted by Ahn et al. (2017) and Henderson et al. (2017). The authors found that biases in simulating the Pacific westerly jet's position contribute to errors in MJO teleconnections, along with poor MJO representation.

Another example are low-level clouds over tropical and subtropical oceans that have been poorly simulated in multiple CMIP generations when evaluated against satellite observations in the present-day climate (e.g. Nam et al., 2012), which inhibits reliable future climate projections. Črnivec et al. (2023) and Cesana et al. (2023) introduced a qualitative approach to discriminate stratocumulus (Sc) from shallow cumulus (Cu) low-cloud regimes to evaluate their horizontal extent (cloud cover), radiative effect at the top of the atmosphere (TOA) and cloud-radiative feedbacks in CMIP5 and CMIP6 models. This approach is essential for guiding model improvements, because Sc and Cu formation and evolution are driven by a distinct interplay of coupled processes within the moist marine boundary layer (such as radiation, turbulence, convection); and Sc and Cu clouds also respond differently to global warming (Cesana and Del Genio, 2021).

Using idealization or a hierarchy of models: Another possibility is to design a model setup in order to isolate specific processes in order to test their relevance for specific phenomena. As an example, Katzenberger et al. (2024) used an aquaplanet with a circumglobal land stripe to study the meridional circulation, particularly the Hadley cell, in an idealized setup. By moving the landstripe north and southwards, changing the surface albedo, or the aerosol concentrations the role of these features for monsoon dynamics could be studied in an idealized setup - undisturbed by the complexity of the real world topography. With this method, a barrier dynamics in the surface pressure could be identified. By slowly adding different components and increasing the complexity and realism of the setup in a hierarchy of models, the contribution by these components can be identified as well, see e.g. Zhou and Xie (2018).

Identifying the role of model configurations: Another significant aspect of process-oriented model evaluation is understanding how specific characteristics are influenced by model configurations, such as resolution and parameterization schemes. Kim et al. (2018) proposed a set of diagnostics to assess how model physics affect the representation of TCs, particularly their intensity in GCMs. The findings suggest that model-specific factors, beyond large-scale environmental parameters, play a key role in shaping TC intensity, with differences in convection schemes contributing significantly to the intermodel spread. Wing et al. (2019) and Moon et al. (2020) further applied these methods, with Moon et al. (2020) showing that TC wind structures are strongly influenced by model resolution. Dirkes et al. (2023) emphasizes the necessity of applying the developed diagnostics for TC analysis in CMIP6 models.

2.2 Systematic model biases

Some systematic biases are present in the vast majority of CMIP models at the global and regional scale and might even persist over multiple CMIP generations, which requires special attention. In this section we review some long-standing biases in CMIP models and strive to discuss the origins and consequences of these systematic model biases. With this list we do not intend to provide a complete list of all bias reported, but to give some relevant examples of model biases and its background. For further details on this topic, we also recommend Simpson et al. (2025).

General evaluation: Bock et al., 2020 employed the ESMValTool (see Section 2.6 and Eyring et al., 2020; Righi et al., 2020), to quantify the progress of climate models across different CMIP phases. Their analysis revealed significant advancements from CMIP3 to CMIP6 in simulating the vertical distributions of key variables, including temperature, water vapor, and zonal wind speed. The authors also demonstrated that high-resolution models in the historical CMIP6 simulations show a notable reduction of temperature and precipitation mean biases.

Sea surface temperature (SST) and ocean model biases: The ocean accumulates more than 90% of the excess energy from the global greenhouse effect (IPCC, AR6). The oceanic global circulation gyres transport excess heat from the tropics towards the poles. Furthermore, the oceanic surface fluxes of heat and moisture enter the atmosphere and thereby affect its dynamics. The ocean component also interacts with the cryosphere and influences processes therein (IPCC, AR6). These various oceanic processes have to be properly captured in ESMs. Long-standing SST biases result in biases when simulating other key phenomena such as tropical cyclones (e.g. Dutheil et al., 2020) and extratropical cyclones (e.g., Priestley et al., 2023a). Wills et al. (2022) investigated systematic biases in the large-scale patterns of recent sea-surface temperature (SST) and sea-level pressure change and showed that CMIP5 and CMIP6 ensembles are not able to reproduce the observed trends. Luo et al. (2023), moreover, discussed the origins of Southern Ocean warm SST bias in CMIP6 models. The Southern Ocean has namely been subjected to systematic warm SST bias in several generations of CMIP models (Sen Gupta et al., 2009; Wang et al., 2014). Westen and Dijkstra (2024) recently discussed persistent climate model biases in the Atlantic Ocean's freshwater transport. These various aforementioned biases are linked to the Atlantic Meridional Overturning Circulation (AMOC), which consists of the northward flow in the upper oceanic layers and returning southward flow in the deep ocean (Luo et al., 2023; Wang et al., 2024). The AMOC is considered to be one of the major tipping elements in the global climate system (Armstrong McKay et al., 2022; Van Westen et al., 2024), which may weaken or even collapse with future global warming, thus a more reliable representation of SST/ocean model would be desirable e.g. to better foresee the future AMOC behaviour.

The Intertropical Convergence Zone (ITCZ) bias: ITCZ is a band of a zonally-oriented surface convergence zone near the equator associated with deep convective clouds and heavy precipitation (Schneider et al., 2014; Waliser and Gautier, 1993). The common problem of fully-coupled global climate models from the early stage of their development is that they simulate two ITCZs over the central and eastern Pacific and the Atlantic in both hemispheres, instead of one ITCZ over the northern hemisphere as in observations, which is referred to as the double-ITCZ bias (Adam et al., 2018; Li and Xie, 2014; Oueslati and Bellon, 2015; Tian and Dong, 2020; Xiang et al., 2017). Tian and Dong (2020), as an illustration, recently examined the double-ITCZ bias in CMIP3, CMIP5, and CMIP6 based on annual mean precipitation. They found that all three generations of CMIP models exhibit similar systematic annual MME mean precipitation errors in the tropics when evaluated against the NOAA Global Precipitation Climatology Project (GPCP; Adler et al., 2003) and the NASA Tropical Rainfall Measurement Mission (TRMM; Huffman et al., 2007) observational datasets.

Biases in extratropical cyclones: Extratropical cyclones involving weather fronts and related overall storm tracks are an important component of the climate system since they transport heat poleward and are associated with a notable amount of precipitation and severe weather in the midlatitudes (Clark and Gray, 2020; Dacre, 2020; Schultz et al., 2019). The accurate representation of extratropical cyclones, including their thermodynamics, frontal structure, and track in CMIP models, however, remains challenging and has been subjected to biases (e.g. Chang et al., 2012; Priestley et al., 2023a, b). Priestley et al. (2023a) investigated drivers of biases in the CMIP6 extratropical storm tracks in the Northern Hemisphere (NH). Even

https://doi.org/10.5194/egusphere-2025-4744
Preprint. Discussion started: 6 October 2025

though the previous work demonstrated that the representation of extratropical storm tracks in the NH has improved from CMIP5 to CMIP6, the persistent biases remain in CMIP6 (Priestley et al., 2023a). A follow-up study by Priestley et al. (2023b) investigated drivers of biases in the CMIP6 extratropical storm tracks in the Southern Hemisphere (SH). The Southern Hemisphere storm tracks have been commonly simulated too far equatorward in CMIP models during the historical period. This issue was somewhat reduced in CMIP6 compared to CMIP5, although it is still a problem.

Marine tropical/subtropical low cloud biases: Črnivec et al. (2023) analyzed 12 CMIP6 ESMs and demonstrated that they all underestimate the aerial extent of low clouds and simultaneously overestimate their radiative effect at the top of the atmosphere. This well-known issue, referred to as the "too few, too bright" tropical low-cloud bias, was already present in previous generations of climate models such as CMIP5 and CMIP3 (e.g., Nam et al., 2012, and references therein). Cesana et al. (2023), moreover, addressed how the representation of marine tropical Sc and Cu clouds and associated feedbacks in the abrupt 4xCO2 scenario changed between CMIP5 and CMIP6. They found that, collectively, CMIP6 models notably increased Sc cloud cover and slightly increased Cu cloud cover compared to their CMIP5 predecessors and are thus closer to observations. They further showed that CMIP6 models notably improved the representation of Sc feedback and slightly improved the representation of Cu feedback compared to CMIP5 models. Yet CMIP6 models still underestimate the magnitude of positive Sc and Cu feedbacks relative to observationally inferred estimates, which should drive further climate model development.

Biases in the cryosphere: The global cryosphere plays an important role in determining the planetary climate since bright ice and snow surfaces reflect a significant portion of the solar radiation back to space and cool the planet (IPCC, AR6). In a warming world, sea ice is shrinking and thinning, with both Arctic and Antarctic sea ice approaching historic lows (NASA Earth Observatory; IPCC AR6). The melting of sea ice with global surface warming implies that an increasing area of dark and absorptive ocean surface is exposed to warming sunlight, which forms one of the principal climate feedback mechanisms – namely, the sea ice albedo feedback (IPCC, AR6). It is thus pivotal to best capture the cryosphere extent, properties, and its response to global warming. To that end, Frankignoul et al. (2024) investigated Arctic September sea ice concentration biases in CMIP6 models and their relationships with other model variables. They demonstrated that CMIP6 models exhibit large biases in Arctic sea ice climatology, which seem to be related to biases in seasonal oceanic and atmospheric circulations. Notz and the Sea-Ice Model Intercomparison Project (SIMIP) Community (2020) furthermore showed that CMIP6 models still fail to simulate a plausible evolution of Arctic sea-ice area (SIA), even though CMIP6 models better capture the sensitivity of Arctic sea ice to forcing changes compared to CMIP5 and CMIP3 models. Roach et al. (2020) evaluated the Antarctic sea ice in CMIP6 and demonstrated that the mean Antarctic sea-ice area is close to satellite observations, but inter-model spread remains substantial, with summer Antarctic SIA being consistently biased low across the ensemble. Nevertheless, they found modest improvements in the simulation of sea-ice area and concentration compared to CMIP5.

Biases in extremes: Human-induced global warming is expected to intensify extreme events such as severe thunderstorms, intense precipitation, heatwaves, droughts, etc. (IPCC, AR6). Extreme weather and climate events and related hazards already cause substantial economic damage and pose a serious threat to human lives (IPCC, AR6). Therefore, it is imperative to evaluate the CMIP ensemble for climate extremes as a first step towards more reliable prediction of extreme events affecting society and ecosystems globally in the near and distant future. This endeavor is well aligned with the WCRP Grand Challenge on Weather and Climate Extremes. To that end, Kim et al. (2020) evaluated the CMIP6 multi-model ensemble for climate extreme indices defined by the WCRP Expert Team on Climate Change Detection and Indices (ETCCDI). They reported several systematic biases even with strong amplitudes, such as the cold bias in cold extremes over high-latitude regions. When comparing CMIP6 with CMIP5, Kim et al. (2020) overall found only limited improvements in model skill simulating climate temperature and precipitation extremes, implying that further work is urgently required to advance the understanding of climate extreme phenomena and their representation in climate models. Moreover, Abdelmoaty et al. (2021) found biases in CMIP6 models when simulating both the mean precipitation and its variability, and thereby emphasized shortcomings of CMIP6 models in the Arctic, Tropics, arid, and semi-arid regions.

2.3 Model dependence

Current day ESMs, including those used for CMIP, are developed by multiple modeling groups worldwide. Ideally, each ESM included in a MME should be independent of the others so there is an adequate representation of the epistemic model uncertainty within the ensemble. Historically, climate projections are derived by calculating simple averages across the MME, with the assumption that the mean is the most accurate representation of the Earth system given all the individual modeling efforts (Abramowitz et al., 2019; Knutti et al., 2010a). Assuming that all models aim to represent the real climate system independently, it is expected that all ESMs, while differing in their approaches, would still be sufficiently independent, and reflect a broad range of uncertainties in a MME. The assumption of model independence allows for the aggregation of results that should smooth out individual model biases. However, the development of these models is often not independent (Pincus et al., 2008).

Recent analysis of model errors in CMIP6 reveals an intriguing and concerning phenomenon: the number of independent climate models is smaller than the total number of models included in CMIP6 (Jun et al., 2008; Masson and Knutti, 2011; Pennell and Reichler, 2011). It has also been shown that the models included in MMEs have biases resulting from a lack of model independence (Jun et al., 2008; Knutti, 2008; Reichler and Kim, 2008; Tebaldi and Knutti, 2007), with errors across different models being correlated, which exacerbates the problem (Knutti et al., 2010a). The dynamical core for resolving grid-scale dynamics is often shared among various ESMs. Furthermore, smaller model components (e.g., physical parameterization schemes) are exchanged between various modeling groups. Although widely accepted methodologies being shared may result from confidence in their correctness, this also implies that potential inadequacies shared across most ESMs also gain more relevance within a MME context (Knutti et al., 2010b). As an illustration, the radiation scheme McICA introduced by Pincus et al. (2003) proved to be an efficient and flexible methodology to represent one-dimensional radiative transfer in a cloudy atmosphere and is thus implemented in multiple contemporary ESMs such as several US models (NSF NCAR CESM2, NOAA GFDL-CM4, DOE E3SM-1-0), the Canadian model (CanESM5), the UK model (HadGEM3), and the Norwegian model (NorESM2). Similarly, the NEMO ocean model is widely used across different modeling centers, including the UK Met Office HadGEM3 and the Norwegian NorESM2, further illustrating the sharing of key components across modeling systems.

The lack of a clear, universally accepted and unambiguous definition of model independence complicates efforts to address model dependence in MMEs. Some definitions are more abstract, focusing on the idea of whether or not a model adds novel additional information (Masson and Knutti, 2011). Others, such as the statistical framework presented by Annan and Hargreaves (2017), provide a more analytical approach to understanding model dependence, offering examples for evaluating model dependence and using their framework. Their framework argues for a rigorous mathematical approach to best capture model dependence, and ensure that MMEs accurately reflect the uncertainty inherent in climate projections. Despite the absence of an unambiguous definition of model dependence, it is clear that climate models are interdependent as shown in Figure 3.

https://doi.org/10.5194/egusphere-2025-4744
Preprint. Discussion started: 6 October 2025

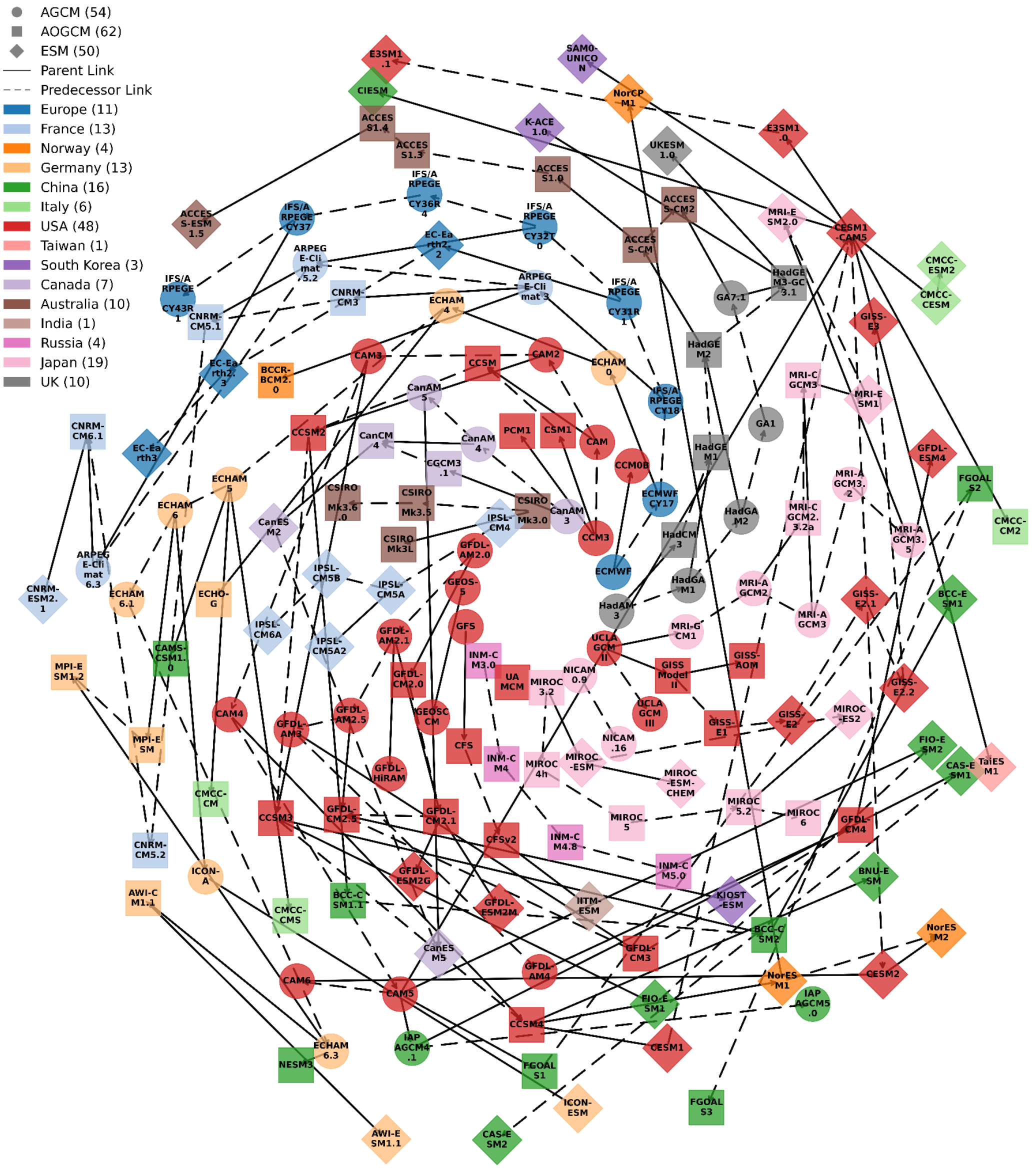

Fig. 2. Spiral plot of climate model dependencies, adapted from Kuma et al. (2023). The oldest model in any given family is in the center of the plot, spiralling out as more models are made. Model type is differentiated by shape of marker, and link type is differentiated by arrow type (solid for parent or dashed for predecessor). Models developed in different countries are assigned distinct colors. Markers indicate atmosphere general circulation models (AGCMs), atmosphere-ocean global circulation models (AOGCMs), and Earth system models (ESMs). Numbers of models from each country are indicated in brackets in the legend. ECMWF models are denoted by the country "Europe".

Despite recent advances, a generalized solution to address the issue of model dependence has not yet been widely accepted. Proposed solutions to address model dependence are specific to the problem at hand, and while some solutions such as weighting schemes (Section 2.4.1) have been proposed, there is still work to be done in representing model dependencies in ensemble means effectively. It is widely acknowledged that climate models are not independent, illustrated in Fig. 2, which leads to inherent flaws in ensemble means and giving the impression of greater model convergence than would otherwise be the case, there remains no consensus on the exact definition of independence and how this issue should be addressed in projects such as CMIP. In coming years, as updated model versions are published it will be crucial to continue developing methods to quantify and correct for model dependence, ensuring that ensemble projections are more robust and better reflect the true uncertainty in climate projections.

2.4 Model Selection and Weighting Methods

CMIP MME weighting and selection techniques are used to categorize the CMIP models based on historical model performance and independence using several metrics (Palmer et al., 2023). Model weighting is crucial for optimizing accuracy and reliability in CMIP MME projections (Strobach and Bel, 2020). Several statistical and performance-based approaches are used for MME weighting (Bhowmik and Sankarasubramanian, 2020; Brunner et al., 2020). Statistical model weighting assigns weights based on statistical properties like independence and spread, while performance-based weighting assigns weights based on their ability to reproduce observed historical climate patterns (Brunner et al., 2020). Weighting methods are used for assessing model dependence, and for uncertainty reduction. In model dependence evaluation, weighting accounts for model redundancy due to shared components. In model uncertainty evaluation, higher weights are assigned to more accurate or reliable models based on specific criteria. Model weighting for detecting model outliers are discussed specifically in Section 3.4.

2.4.1 Weighting methods to deal with model dependence

As highlighted in Section 2.3, climate models are not fully independent and a weighting scheme is needed to ensure that ensemble results reflect the true average of independent climate models. A common approach to address the issue of model dependency is by weighting models differently based on their independence from others. Sanderson et al. (2015) demonstrated a proof of concept for model weighting schemes that considers model dependence, and developed a mathematical formulation to determine model uniqueness. Knutti et al. (2017) later proposed a model weighting method that includes two distance metrics, from models to observations, and among models. Here the “effective repetition of a model” within an ensemble, outlined by Sanderson et al. (2015), is accounted for, along with the accuracy of a model with respect to observations. It is also argued by Boé (2018) that a better method of assessing model interdependencies is through code similarity, instead of through result similarity. While evaluating source code similarity is indeed challenging (due to issues such as the complexity of model architectures, differing programming languages, licensing issues and proprietary restrictions) it should offer valuable insights into shared model components and algorithms that may not be evident from model output comparisons alone. Evaluating source code similarity as well as evaluating similarity of results allows for the identification of common methodologies that may lead to correlated predictions, which highlights potential redundancies within MMEs that could skew results. Integrating both model independence and code similarity into weighting schemes can enhance the robustness of MMEs, contributing to producing more reliable and unbiased outcomes. Recent model selection methods also emphasize model independence (Snyder et al., 2024) with tools being developed that account for model dependence such as ClimSIPS (Merrifield et al., 2023).

2.4.2 Model weighting to reduce model uncertainty

Model weighting can improve the accuracy and estimate the uncertainties of CMIP multi-model ensemble projections (Merrifield et al., 2020). The weighted MME's estimates are more reliable since they consider the better-performing models and remove models with poor simulation capabilities (Shuaifeng and Xiaodong, 2022). Tang et al. (2021) compared weighted and unweighted MMEs projections in four extreme precipitation indices over the Indo-China peninsula and south China. The results indicate that weighted MMEs produce more robust results than unweighted MMEs and the reduction in uncertainty depends on the projection scenarios. Brunner et al. (2020) discovered a reduction in the projected warming when applying model weighting because some models showing high future warming have systematically lower performance weights. A Rank-based weighting approach was utilized for the CMIP6 MMEs projection and uncertainty estimation of cold surges over northern China (Shuaifeng and Xiaodong, 2022).

However, the weighting is a challenging process, as the basis for weights must be determined and by that other not yet identified but equally relevant factors may be excluded in the assessment. Also the relevance of features for phenomena may change with global warming, making it unjustified to use weights with regard to current relevance. Besides, similar models with "main-stream" results may be strengthened for the wrong reasons, while models that provide outlier results and could add valuable insights to the understanding may be wrongly penalized by low weights, see also subsection 3.4.

Most studies in the literature use simple multi-model means, thus equally weighted MMEs to project future climate change impacts (Shuaifeng and Xiaodong, 2022). However, equal weighting of MME (without any model selection) is criticized for not considering model performance (Shin et al., 2020). How the unequal weights reflect the model performance by applying a hybrid weighting scheme has been studied by Shin et al. (2020). In unequal weighting schemes, the chi-square statistics are used for the smoothening of unfairly high or low weights.

2.4.3 Model Subselection to reduce uncertainty

https://doi.org/10.5194/egusphere-2025-4744
Preprint. Discussion started: 6 October 2025

Another way to account for uncertainty is by selecting a subset of models. This can also be considered as a weighting method, which uses the weight 1 for included models, and the weight 0 for excluded models. MMEs with optimized sub-selection can reduce the computational load, produce more reliable uncertainty estimates, and make predictions more accurate (Hamed et al., 2021; Snyder et al., 2024). Herger et al. (2018) compared different sub-selection approaches such as random ensemble, performance ranking, and optimal ensemble sub-selection and found improved performance over the multi-model is possible depending on the case, meanwhile maintaining model spread and interdependence. Random ensemble is one of the model subselection techniques, in which multiple models are combined randomly without an explicit optimization strategy. Performance ranking is another subselection technique where models are ranked based on certain performance metrics such as accuracy, Q-statistics, mean square error etc. In Optimal ensemble sub-selection, a subset of models is chosen that maximizes performance.

Furthermore, Yang et al. (2020) studied the uncertainty contribution of ranking and optimal ensemble model sub-selection for the historical performance of precipitation and temperature. The results indicate that the optimal ensemble sub-selection of nine models has smaller uncertainties, indicating more accurate simulation of present and future climate patterns. Almazroui et al. (2017) have taken three categories of CMIP5 MMEs (all model ensembles, selected model ensembles, and best-performing ensembles) to evaluate the projected temperature and precipitation uncertainties. Among the three categories, the best-performing model outperformed and showed better temperature and precipitation projection over the Arabian Peninsula. Studies further used all model ensembles and selected model ensembles to explore ENSO teleconnection (Roy et al., 2018) and lightning over South/South-east Asia (Chandra et al., 2022).

Model weighting and selection can be valuable for enhancing both the accuracy and reliability of climate projections. Weighting schemes that account for model interdependence are crucial for reducing redundancy, and schemes that account for model performance can improve uncertainty estimation. By giving more weight to models that perform well and are independent from other models, MME weighting attempts to ensure projections are based on the most reliable data instead of relying on equal weighting distributions which introduces significant biases. It is important to note that past performance does not guarantee future performance, and one must always be careful of becoming overconfident in models that perform well in the past. Also, a study may be interested in the overall CMIP model performance. In this case, excluding models e.g. with outlier results by subselection of weighting is not useful.

2.5 Uncertainty Characterization

Uncertainty is inevitable when trying to predict the climate (Knutti et al., 2019). Characterizing and understanding uncertainty is essential not only for guiding model evaluation and development but also for science and risk communication, and for assessing climate change impacts (Deser et al., 2012a; Deser, 2020; Snyder et al., 2024). When using future projections from CMIP, three types of uncertainty must be dealt with (Hawkins and Sutton, 2009; Lehner et al., 2020; Simpson et al., 2021):

scenario or forcing uncertainty, natural variability uncertainty, and model uncertainty. The scenario uncertainty arises because it is not known how human emissions of greenhouse gases and other pollutants from all over the world will vary in the future, and it is accounted for by modeling different emission scenarios (O'Neill et al., 2014). The natural or internal variability uncertainty is due to the chaotic and, thus, unpredictable evolution of the climate system (Deser et al., 2012b), and it has a great impact on climate projections (Lehner and Deser, 2023). Our unique realization of the future climate is the response to the combined effect of anthropogenic forcing and internal Earth system variability. Although internal variability uncertainty cannot be reduced, it is quantifiable (Deser, 2020), and using large ensembles of a single model is helpful for this purpose (Tebaldi et al., 2021). Finally, the third type of uncertainty–model uncertainty–results from our imperfect attempts to predict the aforementioned real world realization. This uncertainty also includes the varying results that can be obtained within the same model when varying its parameters. Model uncertainty can be reduced, and the ways to interpret and quantify it need to be mindful of details about the ensemble's nature and how it is built (Knutti et al., 2019). Furthermore, an adequate treatment of uncertainty has the potential to help MMEs users with model selection and reduce computational burdens (Snyder et al., 2024).

Decomposing the total uncertainty of climate estimates into contributions from scenario, internal, and model uncertainty provides insights into projections' reliability and potential uncertainty reductions. This process is called uncertainty partitioning, and it often involves quantifying the consistency among different members of a MME (Hawkins and Sutton, 2009; Lehner et al., 2020; Woldemeskel et al., 2012; Yip et al., 2011). For long-term means of climate data, Hawkins and Sutton (2009) proposed a widely used method for uncertainty partitioning: they fit a polynomial to each model's output in the time dimension to separate the forced response from the internal variability. The variance across different model's polynomials corresponds to the model uncertainty, and the mean of the different residuals across models represents the internal variability. Finally, the scenario uncertainty is the variance across multi-model means for different forcings. This method assumes (i) that the forced response can be approximated by the polynomial and (ii) that the arithmetic sum of the different uncertainties comprises the total uncertainty. To consider the potential non-additive nature of the total uncertainty (ii), Yip et al. (2011) used analysis of variance (ANOVA)–an approach that partitions the total variance into components due to different sources of variation–to improve the uncertainty partitioning. Later, Woldemeskel et al. (2012) expanded the uncertainty quantification methodology to include also the spatial dimension, by introducing the Square Root Error Variance (SREV) method. This method has proven useful for highlighting regional differences in uncertainty. More recently, and exploiting the computational capabilities that allow running a high number of simulations using the same model, Lehner et al. (2020) overcame the assumption of the polynomial fit (i) from Hawkins and Sutton (2009), which produced significant regional biases by using several single-model large ensembles (SMILEs). The reduction of assumptions when using SMILES and subsequent improvement of results makes them a crucial tool currently to partition uncertainty in climate projections. As detailed in Section 2.3, in a multi-model ensemble, models are not entirely independent, and the lack of independence complicates the interpretation of any statistic extracted from the ensemble, including the spread or uncertainty. Consequently, the methods mentioned above often involve some weighting, which further details provided in Section 2.4.

A question that should be considered, although it can only be partially answered, is whether the MME spread is too narrow, too broad, or about right. The uncertainty may be too wide if observations are not used correctly to tune models, or if the models have extensive and diverse structural errors. The ensemble may be overly confident if the models are structurally similar but incomplete or if uncertain processes are missing. One might answer this question using observations, which is addressed using weighting methods (see Section 2.4). However, present-day uncertainty arises from different sources than future uncertainty. Present-day uncertainty results from the models' inability to fit observations, while uncertainty in the future is due to variate representations of physical processes and feedbacks (Sanderson and Knutti, 2012). Additionally, it must be considered that observations-based products, which are often used to perform model-observation comparisons, also possess significant uncertainties (e.g., Chemke and Polvani, 2019). Care should be taken when assuming that the spread (attributed to any source of uncertainty) of present-day or historical simulations will be the same in the future.

If the only tool for assigning confidence to climate change projections is a direct comparison between observations and historical simulations, then there is the risk that "good" models under this framework don't really represent well the changes under future greenhouse gas scenarios. Similarly, "bad" models that may be disregarded due to their skill relative to observations may contain useful information about some characteristics of the future changes (Hall et al., 2019). An evaluation and uncertainty reduction technique that avoids this bias is the development of emergent constraints (Hall et al., 2019). Emergent constraints, based on data from an MME, exploit the relationship between a model's representation of a present-day quantity (x) and the projected future change (Δ) in a quantity (y) using a typically linear approximation (Simpson et al., 2021). An analysis of the probability distribution function of Δy within the ensemble allows for a reduction of the uncertainty. This method has been used for assessing the uncertainty of many processes within different Earth system components (Keenan et al., 2023; Nijsse et al., 2020; Shaw et al., 2024; Simpson et al., 2021; Smith et al., 2022; Thackeray et al., 2022). ML approaches have also been used to demonstrate a potential to discover and explore emergent constraints (Nowack et al., 2020). Despite the usefulness of emergent constraints, care should also be taken when interpreting the results, since the method assumptions may produce overconfident predictions and may be vulnerable to artifacts within the model (Breul et al., 2023; Sanderson et al., 2021), similar to other uncertainty reduction methods.

While climate models exhibit high confidence in thermodynamic aspects of climate change (e.g. global temperature increase) due to robust theoretical and observational evidence, dynamic aspects, particularly related to atmospheric circulation, present significant uncertainties due to their dependency on nonlinear dynamics and feedback mechanisms (Shepherd, 2014). Model uncertainties in these two components are uncorrelated (Zappa and Shepherd, 2017), meaning that errors in one component do not influence or predict the errors in the other, so separating them allows better understanding of where the biggest uncertainties lie. Considering this, uncertainty in climate projections can be communicated through climate storylines (Shepherd et al., 2018), which show different plausible future climates, emphasising exploring and understanding physically plausible events or pathways. The storyline approach differs from traditional methods of uncertainty evaluation in climate models, which are primarily probabilistic and rely on ensembles of simulations. Traditional methods of uncertainty evaluation in climate models, such as probabilistic approaches based on multi-model ensembles, often assume that model spread adequately represents uncertainty. However, this assumption may not hold for dynamically driven climate phenomena, where MME means may obscure critical regional details with individual climate models exhibiting atmospheric circulation patterns that can differ qualitatively from the multi-model mean (Bellomo et al., 2021; Zappa and Shepherd, 2017), further complicating the understanding of future climate impacts. Instead of quantifying the likelihood of events, storylines focus on causality and go through the physical drivers and interactions that make an event possible (Shepherd et al., 2018), constructing a causal network and conditioning on specific physical assumptions. If we know thermodynamic changes are robust, the thermodynamic aspects of the observed changes are regarded as certain and the dynamic aspects as uncertain. By explicitly linking causal mechanisms to regional climate hazards, storylines are especially useful for regional climate impacts and understanding extreme events (Bevacqua et al., 2022; Shepherd, 2019; Zappa and Shepherd, 2017), improving the interpretability and usability of projections for decision-makers (Kunimitsu et al., 2023).

2.6 Available tools for MME analysis

The analysis of comprehensive CMIP datasets is greatly facilitated with the aid of various tools that have been developed within the global climate community. However, the wide range of available tools was not centrally cataloged, making it difficult to gain a clear overview of their capabilities for climate data analysis. To address this, the WCRP CMIP has undertaken an effort to compile a central repository of these tools (https://wcrp-cmip.org/tools/). This collection encompasses various data access platforms (e.g., Earth System Grid Federation, Climate Data Store, IPCC data distribution centre, PANGEO, CAVA, Climate Information Portal), which notably facilitate accessing large and complex data volumes. The collection furthermore lists handy command line operators (e.g., ncview, NCO, CDO) as well as programming languages, which are suitable for climate data analysis (such as Python, R, Julia) together with useful packages (e.g., multiple Python packages such as matplotlib, scipy, pandas, Iris, xarray, xGCM, xMIP, xclim, xCDAT, UXarray, Metpy, aospy). The repository contains several comprehensive evaluation and benchmarking tools such as ESMValTool, bgcval2, RUBISCO, PCMDI Metrics Package, AMBER, the MDTF Diagnostic Package. These evaluation tools include a set of diagnostics designed to address specific scientific focuses. For example, among various diagnostics, ESMValTool incorporates the Climate Variability Diagnostics Package (CVDP, Eyring et al., 2020; Phillips et al., 2020, 2014) that facilitates the exploration of modes of climate variability and change in models and observations (Maher et al., 2024 and Section 4.2). The source code for the CVDP package is also available in the GitHub repository: https://github.com/NCAR/CVDP-ncl. Another important initiative in process-oriented analysis is led by the Model Diagnostics Task Force (MDTF) under NOAA's Climate Program Office (CPO) Modeling, Analysis, Predictions, and Projections (MAPP) program. It promotes the development and use of process-oriented diagnostics (see Section 2.1) in climate and weather prediction models (Maloney et al., 2019; Neelin et al., 2023). Additionally, the WCRP repository includes various data analysis and visualization tools, including the IPCC WGI Interactive Atlas, Panoply, TempestExtremes, CAVA, TECA, KNMI Climate Explorer, Google Earth Engine. Figure 3 highlights some of these tools aiming to promote their usage across the wider climate community. The basic information about each tool can otherwise easily be deduced from "Tools description cards" at the CMIP website, which additionally provide links to tool websites as well as available documentation, tutorials and community support. It should finally be emphasized that the tools repository is being actively maintained and continuously updated. To enhance its utility for the broader climate science community, new contributions are highly welcomed.

While the CMIP tool repository is a key resource for many widely used climate analysis tools, it does not cover all available tool resources. Beyond this collection, the wider open-source ecosystem - especially within the Python community - provides additional tools and libraries for analyzing climate data, and at the same time is being supported by a large and active scientific community on platforms such as GitHub.

https://doi.org/10.5194/egusphere-2025-4744
Preprint. Discussion started: 6 October 2025

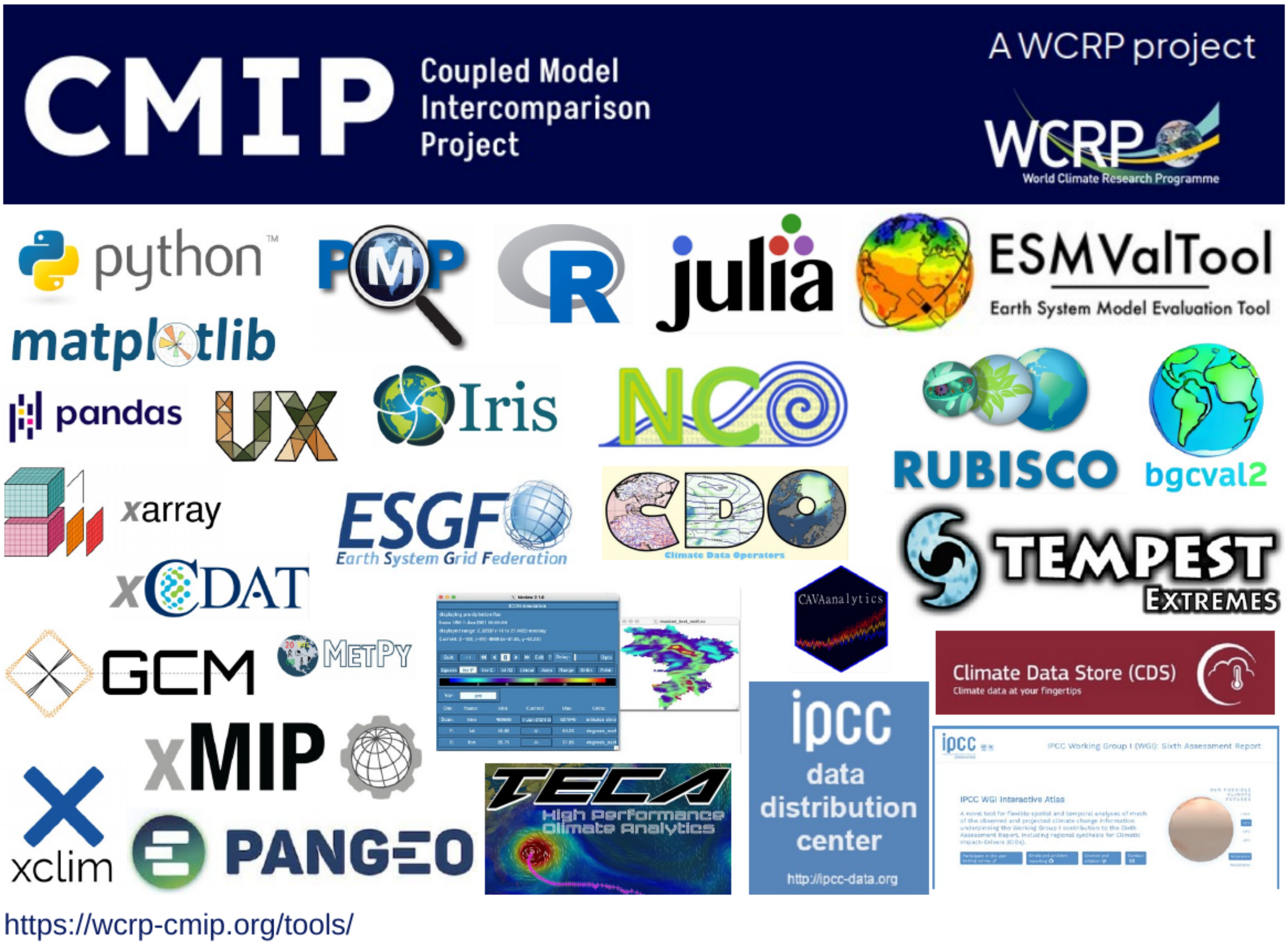

Figure 3: Collection of useful tools for using climate data available at https://wcrp-cmip.org/tools/.

3. Specific challenges and common questions

3.1 How can observations be used to improve MME projections beyond model evaluation?

Observations are integral to improving the reliability of MME projections, especially to reduce model bias and increase the physical realism of ESM simulations at the model evaluation stage (Haarsma et al., 2016). Using MME outputs in conjunction with observational datasets can also help bridge the gap between model outputs and real-world earth system processes, where such gaps exist (e.g. Tebaldi et al., 2005). Observations can also serve as ensemble members themselves when viewed as exchangeable with model simulations (Annan and Hargreaves, 2010).

Within CMIP6, activities such as the Detection and Attribution MIP (DAMIP), Polar Amplification MIP (PAMIP), and SIMIP, motivations for pairing observation data with MME simulations beyond those mentioned above exist. These include determining how anthropogenic activity contributes to climate change (Gillett et al., 2016), reducing intermodal spread/uncertainty by leveraging emergent relationships based at least in part on observations (Smith et al., 2019), and understanding how ice, air, and the ocean interact (Notz et al., 2016).

As observational datasets are subject to uncertainty and vary in reported quantities, spatial coverage, and spatial and temporal resolution, it has become common practice to consider observational uncertainty when multiple observational datasets are employed (Notz et al., 2016). This practice emerged out of the need to account for structural uncertainty in observation data ensembles to improve signal detection for subsequent comparison with model ensemble outputs (Santer et al., 2008). Observational ensembles have been paired with MMEs in studies e.g. with regard to the tropical troposphere (Santer et al., 2008) or to Antarctic sea ice (Roach et al., 2018).

3.2 How many models to include?

Any MME analysis has to face the question of how many models to include. However, determining the optimal number of models to include in an ensemble is not straightforward, as it involves balancing the trade-off between model diversity, computational cost, and the desired accuracy of the results. Increasing the number of ensemble members enhances the robustness of the results by reducing statistical uncertainty, at least as long as they are independent. At the same time, state-of-the-art climate models remain computationally expensive. Downloading and processing these large datasets, particularly in the context of major intercomparison projects like CMIP, is also a resource-intensive challenge that limits the number of models used in MME studies. These challenges raise the question how many models are actually required to form a “good” ensemble size. A similar question exists in the context of large ensembles where the number of perturbed simulations is discussed. A lot of the arguments and findings as presented in the following apply for both contexts.

Lower threshold of ensemble size: At least 5 models

If the ensemble size is too small, the inter-model variability that also serves as a proxy for natural variability may not be fully captured. This variability has the potential to lead to an underestimation of uncertainties and can consequently result in an overestimation of the models’ performance in the procedure of evaluation and an overconfident interpretation of the results. It is even possible that a too small ensemble size leads to a qualitatively different finding, as shown by an example of two or three models in a study by Milinski et al. (2020). In this study, the small subsets showed a warming after a volcanic eruption, while the actual known response would be a cooling effect. So, how many models or simulations should be used as a minimum? Several studies have shown that the error (e.g. root mean squared error when compared to reference data) is reduced

https://doi.org/10.5194/egusphere-2025-4744
Preprint. Discussion started: 6 October 2025

substantially up to about five models in different contexts (Herger et al., 2018; Knutti et al., 2010a; Mendlik and Gobiet, 2016; Milinski et al., 2020; Steinman et al., 2015). Adding further models is generally beneficial, but the improvement per additional model is much smaller. Mendlik and Gobiet (2016) find that the subset size can be reduced from 25 to 5 while still being representative for the entire ensemble. As these studies refer to different quantities and research questions, and were conducted independently, but still share five as a lower "threshold", we propose five models/simulations as an initial baseline minimum for MME studies. Depending on the research question however, the minimum number of required models might vary. It can be determined by a specific method, as explained below.

Determining specific minimum ensemble size following Milinski et al.

If feasible, an individual check for the appropriate minimum number depending on the specific research question and requirements is even better than a general minimum. A procedure for diverse research questions has been proposed by (Milinski et al., 2020). After (1) defining the research question, (2) an error metric (e.g. RMSE) as well as a maximum acceptable error has to be decided. As a next step (3), the error for randomly sampled subsets of different sizes has to be quantified. The number of required models can now be identified as the smallest subset size that has an error below the chosen threshold (4). If the identified model number is less than half of the initial sample (e.g. the identified subset included 40, thus less than 50 members, when evaluating 100 members) the estimated subset size is robust (5). While this method provides a straight-forward, rather simple method to identify the ideal number of models in an ensemble, it still requires the availability and analysis of a high number of model simulations. Consequently, this method might not be feasible for all studies. Therefore, we provide here a collection of studies that identified the optimal number of models for different research questions. It may be used as an orientation for future studies with limited capacities for the model selection process.

List of studies with identification of ideal subset sizes for different research questions

For a variable like temperature where the internal variability is rather low, 10 ensemble members can be used to sufficiently detect changes in global mean land temperature (Deser et al., 2012b). To robustly detect significant warming (at the 95% confidence level) in the 2050s relative to the 2010s, Deser et al. (2012b) only needed 1 ensemble member for nearly all locations. Alternatively, 3-6 ensemble members are needed for tropical and high latitude precipitation, while >15 ensemble members are needed for mid-latitude precipitation with 40 ensembles being a larger estimate (Deser et al., 2012a, b). When it comes to sea level pressure (SLP), they found they needed only 3-6 ensemble members in the tropics but 9-30 in the extra tropics.

The number of required models might differ in different regions, as the signal itself and the local internal variability will vary (Bittner et al., 2016). Over the ocean, less SMILE members are required (Milinski et al., 2020). Table 1 highlights a small sample of papers that have employed large ensembles for a variety of research questions.

Table 1. Examples of large ensembles used and how many models were investigated.

Variable/Metric	**No. of ensemble members**	**Study**
Aridity and risk of consecutive drought years	Two 10-member ensembles from CESM	(Lehner et al., 2017)
Precipitation and temperature	Two 10-member atmosphere only ensembles from CESM and GFDL 40 models (1 simulation each) from CMIP5 40-member CESM1 Large Ensemble 10-member GFDL Large Ensemble	(Lehner et al., 2018)
Ocean carbon uptake	38-member CESM1-LE 9 models from CMIP5	(Lovenduski et al., 2016)
Temperature and precipitation influence on near-term snow trends	40-member CESM1-LE	(Mankin and Diffenbaugh, 2015)
Irreducible uncertainty	100-member MPI Grand Ensemble	(Marotzke, 2019)
Ocean ecosystem drivers (warming, acidification, deoxygenation and perturbations to biological productivity)	30-member GFDL Ensemble	(Rodgers et al., 2015)
Ocean carbon cycle	30-member GFDL Ensemble	(Schlunegger et al., 2019)

When multiple realizations (or variants) for a given simulation are available for the same model, it is considered good practice to average all members of a model ensemble and incorporate such means into the MME (Knutti et al., 2010b).

Remarks for including more models

For specific applications, higher number of simulations are necessary, e.g. for the quantification of internal variability, more simulations are necessary because higher-order moments of the distribution need to be estimated (Milinski et al., 2020). Generally, adding further models improves the statistical robustness of the MME analysis, but it has to be remembered that the added models should at least partly be independent of the existing models as otherwise only the weight of single models is increased without any physical reason (Knutti, 2010). See Section 2.4 and Section 2.5. for more details. A too large ensemble size has also the potential to increase the spread beyond a realistic range as the inclusion of outliers becomes more probable

(Knutti, 2010). In this context, Section 3.4 provides more detail regarding the question how to deal with outliers. Another consideration becomes relevant when working with different scenarios. As the range of uncertainty increases with the number of models, the same number of models should be used for all scenarios for comparability (Knutti et al., 2010a).

3.3 What is important to consider when applying MMEs for extremes?

Extreme weather and climate events have significant impacts on human society and ecosystems, so it is essential to understand their causes and produce reliable future projections for climate change adaptation planning. In the context of using MMEs to study extreme climate events, ensembles offer both strengths and challenges.

MMEs such as CMIP or CORDEX are widely used in various studies (both global and regional) concerning climate extremes (Kim et al., 2020; Soares et al., 2023; Vogel et al., 2020; Yang et al., 2012) typically applying statistical approaches, such as probabilistic modeling, or using climate extremes indices defined by the Expert Team on Climate Change Detection and Indices (ETCCDI). Extreme Value Theory (EVT) provides a theoretical foundation for analyzing extreme events, offering statistical methods to model the tails of probability distributions (Coles, 2001; DelSole and Tippett, 2022). One widely used approach within EVT is Generalized Extreme Value (GEV) distribution analysis (Rypkema and Tuljapurkar, 2021), a statistical framework for modeling the tail of the distribution of rare events, such as extreme temperatures or precipitation. For example, studies use GEV to estimate return periods of extreme rainfall events, helping to assess how the likelihood of such events might change under future climate scenarios (Wehner, 2020). By fitting GEV to observed and modeled data, researchers can evaluate shifts in the intensity and frequency of extreme events.

A major advantage of using the mean of the MME is its ability to amplify the climate change signal by reducing noise from internal variability, making it easier to identify trends in extreme events (Intergovernmental Panel on Climate Change (IPCC), 2021), but it might not always be the best choice, particularly when examining the intensity and frequency of extreme events (Knutti et al., 2010b). Different models in a MME may have biases in how they simulate extremes, such as heatwaves, heavy precipitation, or droughts. MMEs allow for a sensitivity test for structural differences between models, helping researchers identify common trends in certain indices or events across models, increasing confidence in results where models agree. However, it should be noted that using MME's median or mean can sometimes mask the severity of local extremes, as averaging across multiple ensemble members can obscure the range of possible outcomes of individual extreme events, especially if some models predict significantly different extreme event patterns, leading to an underestimation of risks in certain regions. Uncertainties exist for hot and cold extremes, with some models deviating considerably from the multi-model average and are particularly large for precipitation extremes, where despite a general trend towards heavier precipitation and longer dry periods, several models predict opposing trends in certain locations (Sillmann et al., 2013).

It is therefore important to evaluate how well each model performs for the region or variable of interest in simulating extremes (Kim et al., 2020; Sillmann et al., 2013) and to correct for biases when possible. As discussed in Section 2.1, model evaluation is generally conducted using performance-oriented or process-oriented approaches, which tend to focus on a model's ability to capture mean climate states (mean and median performances) or large-scale circulation patterns, which may not prioritize models that best capture extreme events. Kim et al. (2020) evaluated the CMIP6 multi-model ensemble against ETCCDI climate indices and identified systematic biases, such as a persistent cold bias in cold extremes over high-latitude regions. When comparing CMIP6 models with CMIP5, they found only limited improvements in simulating temperature and precipitation extremes, highlighting the need for further advancements in the understanding and representation of extreme climate events in ESMs. More reliable predictions of climate extremes are enabled by the use of MMEs, but according to Kim et al. (2020) the choice of the methods for the assessment of these high-impact, low-frequency phenomena in the ensemble, as well as the choice of reference data is crucial for evaluating model performance.

When it comes to studying extreme climate events, uncertainty is another aspect that is important to account for. As discussed in Section 3.2, the size of an ensemble plays a key role in reducing uncertainty and a larger ensemble allows for a more comprehensive assessment of the spread of possible outcomes. Many studies of climate extremes using MMEs typically use only a single ensemble member from each model to ensure comparability (Kim et al., 2020). The limited availability of large ensembles for all models within a MME also makes this approach practical. However, using only one ensemble member per model could miss some of the variability in extreme events that larger ensemble runs could capture. Nevertheless, given the constraints on computational resources and the availability of large ensembles, this method remains a common compromise.

While increasing ensemble size can help mitigate uncertainties, it does not eliminate the challenges posed by model limitations. To address these limitations when applying MMEs for extreme weather and climate events, different methods are applied. Employing model weighting (Balhane et al., 2022) can enhance the accuracy and reliability of extreme event projections and downscaling techniques, either statistical or dynamical with the use of RCMs, can provide higher-resolution data to improve the representation of extremes in specific regions. For example, the bias-adjusted high-resolution RCM outputs in the EURO-CORDEX project showed an improvement in the simulation of extreme temperature and precipitation indices across Europe, underscoring the value of RCMs for more reliable and region-specific climate projections (Coppola et al., 2021; Dosio, 2016). Highly vulnerable regions benefit from MME based on RCMs' projections, which provide insights into future changes of local extreme events (Dosio, 2017; Tegegne et al., 2021) and help address issues such as water scarcity, food security and disaster preparedness.

3.4 How to deal with outliers?

Convergence has at times been criticized as a measure of model reliability on the grounds that it gives more weight to simulations that are more similar to the multi-model mean at the expense of sampling uncertainty over a broader probabilistic space (Tebaldi and Knutti, 2007). In particular, the initial version of the reliability ensemble average (REA) weighting method penalized outliers for diverging from the ensemble mean because convergence, which may be due in part to the genealogical similarity of models exhibiting convergence towards the ensemble mean, was used as a metric in determining the REA weight for each member of an MME (Tebaldi and Knutti, 2007). However, there is a history of privileging MME convergence within the climate science community, as in the third IPCC assessment report where two models were discarded because of extreme estimates of warming, resulting in very large climate sensitivity (Tebaldi and Knutti, 2007). Unsurprisingly, the convergence principle is still found in MME subsetting efforts (Palmer et al., 2023) and to at least partially inform MME evaluation (Amali et al., 2024). Yet, privileging models whose values cluster around an MME mean can be more or less desirable depending on the particular aims of a study. In other words, there are cases where outlier inclusion–which deemphasizes convergence–is preferred, as in the study of climate extremes. Furthermore, in some cases, excluding models based on the results of overall evaluation has been shown to have little effect on projection spread (Knutti et al., 2010a).

Before diving into the details of how outliers are or are not addressed within the recent literature, let us consider outlier detection. When defined quantitatively, outliers are commonly detected using the method employed in Sun and Archibald (2021), where such models are defined as those that exceed the 1st or 99th percentile. This method provides a statistical basis for identifying extreme deviations in model output. Another approach, used by Bracegirdle and Stephenson (2012) identifies "high-leverage" models with the *3p/N* method developed by Hoaglin and Kempthorne (1986). In this method, *p* is the number of variables considered and *N* the number of models. The value of the expression *3p/N* then serves as a high-leverage threshold for members of a given ensemble.

So, when does it make sense to privilege MME member convergence and penalize or exclude outliers? As mentioned above, it depends on the goals of a study as well as what is being studied. For example, some variables such as sea ice extent, or regions such as the poles, are prone to significant model spread with increased spread in some locations depending on the season (Bracegirdle and Stephenson, 2012). Studies focused on understanding the average state of such variables or locations may benefit from outlier penalization or exclusion. That said, depending on the variable(s) and region(s) of interest, there may be alternatives to exclusion outliers, such as the use of emergent relationships, to constrain future projections (Sansom et al., 2021).

Inaction is a form of action when it comes to outliers models, so along with “active” approaches to handling outliers, doing nothing is also considered. The main approaches seen in recent CMIP studies include: (1) exclusion, (2) penalization, (3) methods in classical (or frequentist) statistics, (4) methods in Bayesian statistics, (5) presenting results with and without outliers, and (6) including outliers. Examples of these approaches and the context in which they were applied are summarized below. Although it is common for outliers to receive some form of special treatment in ML studies, these methods are often based on statistical methods and so are not discussed separately here.

(1) Exclusion

The first approach is exclusion, which is to remove models with outlier status from an ensemble. If considered from a model weighting perspective, these models are assigned a weight of zero within an MME. This approach risks omitting simulations with realistic but rare events, but in some cases the benefits of exclusion outweigh its drawbacks. For example, Mudryk et al. (2020) identified outlier models for some seasons and regions in their study of snow cover change in the Northern hemisphere, excluding such models to achieve better agreement between observation data and CMIP6 MME projections. This study focuses on trends in snow cover change and the outlier model, which is known to have higher than expected snow cover fractions in areas of low snow mass, contributed to unrealistic conclusions about MME spread. The Swiss Climate Scenarios CH2018 (CH in the abbreviated name for this dataset is from *Confoederatio Helvetica*, the latin name for Switzerland), are another example of exclusion. These scenarios are based on EURO-CORDEX, which excludes some outlier GCMs to narrow uncertainty ranges for temperature and precipitation. So CH2018 inherits outlier exclusion from another dataset (Sørland et al., 2020). While models with outlier projections may be excluded on to improve MME alignment with observations or to reduce uncertainty, caution ought to be taken with the latter unless the model is known to be deeply flawed, as excluding projections that include information about rare but possible events can impede proper evaluation of adaptation policy options (Knutti et al. 2010).

(2) Penalization

Penalization is where an outlier model is not removed from an MME, but is given a reduced weight. This can be done through model weighting (see Section 2.4), but it has also recently been achieved through bias correction and ridge regularization. Bias correction is used to calibrate historical and future MME projections against historical observations to reduce the influence of outlier models on uncertainty ranges to lessen uncertainty. This can be seen in a study of future precipitation over Northern Europe (Moradian et al., 2023), precipitation being a high-variability variable to begin with. Ridge regularization, used in ML context, is a form of linear regression that incorporates a penalty term to reign in variables with unusually high linear correlation to protect against overfitting. In Labe and Barnes (2022), ridge regularization is applied to limit the sensitivity of an artificial neural network to outlier influence.

(3) Methods in classical (or frequentist) statistics

Among the approaches seen within classical statistics is the use of outlier insensitive methods. These are methods that retain outlier models without being disproportionately influenced by them. Such methods include taking the ensemble median instead of its mean as a measure of the MME's center. This helps ensure that the result is not overly influenced by outliers (Ge et al., 2021). Rank based tests of statistical significance can also be used. These tests are insensitive to outliers in that they are calculated based on the rank, or position of a value within a distribution, rather than the value of a particular data point within a sample (DelSole and Tippett, 2022). Similarly, when analyzing data for the presence of trends, the rank-based Mann-Kendall correlation test can be used as per the World Meteorological Organization's recommendation for working with hydrological data (Rojpratak and Supharatid, 2022).

(4) Methods in Bayesian statistics

However, including outlier models to sample uncertainty from a broader statistical space can be desirable. Toward this end, MME model weighting methods that apply Bayesian statistics have been developed. Compared to the frequentist statistics which uses a fixed population parameter to describe probability distributions, Bayesian statistics uses a conditional parameter that depends on the probability distribution of a given dataset (Clyde et al., 2022). In Shin et al. (2020), the authors define outlier models as those that generate projections that are unusually close to the hydrological variable observation data. Excessive model calibration to observations for certain regions is given as the reason for models with simulations that are very close to precipitation observations being considered outliers. They propose a Bayesian weighted average and bias correction hybrid method to reduce the influence of outliers. This method is also a form of penalization.

Xu et al. (2019) provide another example of a Bayesian approach to model weighting in the context of downscaling precipitation data to study particular watersheds, agricultural fields, or water infrastructure sites. The authors argue that statistical downscaling is often preferable to dynamic downscaling because statistical downscaling requires less computation and produces data with finer spatial and temporal resolution which is useful at the very fine spatial scale they seek to study. However, Xu et al. (2019) also point out that dynamic downscaling can underestimate extremes and be overly sensitive to outliers, along with inheriting too many features from historical observations. This team therefore adopts a Bayesian weighted average approach to MME data that preserves the benefits of dynamical downscaling while diminishing its drawbacks.

The Bayesian paradigm can also be seen in ML techniques. For example, in Sun and Archibald (2021) the authors combine data fusion–a form of post-simulation data mining–with a Bayesian neural network (a machine learning method) as an alternative to reanalysis. Sun and Archibald (2021) do this to improve future projections of surface ozone concentrations from Aerosol and Chemistry Model Intercomparison Project simulations. This study uses "aggressive" and "conservative" multi-model fusion approaches to improve surface ozone predictions. The "aggressive" approach favors observation values over simulated values in a multi-layer learning process. Conversely, the "conservative" approach favors simulated over observed value within prescribed probability distribution functions (PDFs). The conservative approach performs better when compared 1:1 with model outputs, but slightly worse overall due to reduced variability associated with weighting in this Bayesian method leading to the omission of outlier data exceeding the 1st and 99th percentiles, as per the use of prescribed PDFs in their approach.

(5) Presenting results with and without outliers

In using "aggressive" and "conservative" approaches, Sun and Archibald (2021) present results that allow and exclude outliers respectively and show that for their particular study the difference between the results for each approach is not overwhelming. Bracegirdle and Stephenson (2012) also present some of their results with and without outliers in a less recent study on how to increase precision of polar warming estimates to illustrate the sensitivity of different forms of regression to outlier inclusion.

(6) Including outliers

As mentioned at the start of this section, it can also be beneficial to include outlier models by weighting model ensemble members by RMSE skill score, as in Tegegne et al. (2020) where the authors preserve the full extent of model spread within an MME to study climate extremes (also see Section 3.3 of this article). To do this, the authors use the Katsavounidis–Kuo–Zhang (KKZ) algorithm to select ensemble members based on their ability to help represent the full range variability that exists within the sampling space for climate extreme indices recommended by World Meteorological Organization's ETCCDI. The IPCC report *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation* characterizes this approach as being capable of detecting "moderate extremes", that is to say, events that are expected to occur up to 10% of the time (Seneviratne et al., 2012). To identify models that represent more extreme events, extreme value theory, which is treated in detail in *Statistical Methods for Climate Scientists*, is needed. Researchers use EVT to identify values that lie in the tails of a probability distribution, often focussing on distribution minima or maxima (DelSole and Tippett, 2022). Including outliers offers a better estimate of worst-case scenarios.

While this discussion of how to handle outliers in MMEs covers situations in which treating outlier models in a non-democratic way may or may not be desirable, related questions to outliers and model-weighting are discussed in Section 2.4 of this article. For a discussion that touches on why outliers may or may not be found in an MME in the first place, please see Section 2.3 on model genealogy. The reader is also directed to CMIP activity articles for simulation protocols designed to help investigate the process representation basis of outlier behavior for variables of interest. The Radiative Forcing Model Intercomparison Project is an example of this (Pincus et al., 2016).

3.5 What should be considered when working with regional MMEs/downscaling?

Acquiring regional information about climate change is crucial for climate change impact, vulnerability and adaptation studies, and hence the coarse-resolution GCMs have to be downscaled (i.e., the spatial and temporal resolution of the GCM output has to be increased) for policy decisions. CMIP GCMs are internationally established sources for climate projection data. In the CMIP6 GCM projected data, each grid cell has a resolution of 100 to 250 km (Liang-Liang et al., 2022; Weigel et al., 2010). So, this coarse resolution of GCM has limitations in producing locally relevant information (Grose et al., 2023). Downscaling is a set of methods used to improve the spatial and temporal resolution of GCMs (Baño-Medina et al., 2022). Downscaled

CMIP GCM data are crucial for understanding regional climate change impacts, and it is helpful to create targeted adaptation strategies at the regional level. Downscaling is especially crucial for regions with complex topography or localized climate phenomena (Wilby and Fowler, 2010). Various downscaling techniques exist such as statistical downscaling (Gebrechorkos et al., 2023; Wootten et al., 2024), dynamical downscaling (Knutson et al., 2013; Tapiador et al., 2020) as well as novel machine-learning based approaches (Sachindra et al., 2018; Soares et al., 2024), and they have their strengths and limitations (Hall, 2014).

In dynamic downscaling, output fields from a GCM are used as input for a Regional Climate Model (RCM), which simulates climate on a limited-area domain and hence employs a finer resolution (Di Luca et al., 2015). Specifically, the WCRP COordinated Regional climate Downscaling EXperiment (CORDEX) (Giorgi, 2019; Gutowski Jr. et al., 2016) initiative unites multiple institutions from all over the world striving to best acquire regional climate change information from global climate models. The dynamic downscaling technique is highly dependent on the availability of RCMs. Moreover, dynamic downscaling can capture regional physical processes that GCMs cannot resolve (Giorgi and Gutowski, 2015). The statistical downscaling technique uses statistical relations between coarse-resolution GCM climate data and observed local climate data to generate fine-scale downscaled projections for a specific region (Oxarart and Parker, 2024), and it entirely relies on observations and data quality. ML-based downscaling methods have recently been used for high-resolution GCM simulations (Rampal et al., 2024). ML algorithms can handle non-linear, complex relations between large-scale GCM predictors and observed local climate variables. Furthermore, ML-based downscaling can handle large datasets and produce better resolution CMIP multi-variable long-term projections than traditional statistical techniques (Rampal et al., 2024).

The dynamic downscaling technique was used to derive the bias-corrected global dataset from CMIP6 and the European Centre for Medium-Range Weather Forecasts Reanalysis 5 (ERA5) dataset (Xu et al., 2021). Grose et al. (2023) used CMIP6 multimodel ensemble downscaling to provide accurate, scenario-based climate change projections for the Australian region. They developed a sparse matrix framework to apply the downscaling method to a selected group of CMIP6 models to produce optimized climate change projection results for Australia. Di Virgilio et al. (2022) studied the effects of model subselection (based on performance, independence and diversity) on dynamic downscaling. The results indicate that systematic biases in GCMs can degrade dynamic downscaling simulations.

The limitation of the dynamic downscaling method has been addressed by Liu et al. (2021) by presenting a singular value decomposition (SVD)-multi-linear regression statistical downscaling model to predict the interannual variation of East Asian winter surface air temperature at a better resolution. The study found that the pattern correlation coefficient skill of the original MME is much lower than that of the statistical downscaled prediction model, indicating that statistical downscaling can overcome the limitations of the dynamic downscaling approach. Statistical downscaling refers to a set of methodologies to determine statistical relationships between GCM climate fields and observed (local) climate patterns in combination with various bias correction techniques. Su et al. (2016) investigated the projected impacts of climate change in the Indus River Basins through one of the statistical downscaling methods, the Equidistant Cumulative Distribution Functions matching method (EDCDFm) and the regional ensemble results captured the dominant features of the temperature and precipitation variation. The statistical downscaling of extreme temperature data from the selected CMIP6 GCMs is done by Wang et al. (2016). The study found that statistically downscaled data from most of the GCMs gave the correct sign of recent trends in all the extreme temperature indices compared to the original GCM data. The Bias Correction and Spatial Downscaling (BCSD) technique is used to statistically downscale the projected daily maximum temperature over China from the selected CMIP5 GCM models. The results indicate that statistical downscaling reduces the cool bias compared to the original CMIP5 simulations (Xu and Wang, 2019). Furthermore, Wang et al. (2021) compared the spatial and temporal downscaling of the CMIP5 and CMIP6 MMEs over the Hanjiang River Basin in China. This multi-site downscaling method accurately downscaled the CMIP5-MME and CMIP6-MME precipitation.

Even though the statistical downscaling technique reduces biases in regional climate change projection, ML-based downscaling techniques can outperform existing statistical approaches (Rampal et al., 2022). For the first time, deep learning has been used for the MME downscaling of temperature and precipitation projection over Europe by Baño-Medina et al. (2022). They used different convolutional neural networks (CNNs) for downscaling, and the results were compared with the European ensemble RCM. These results indicate that deep learning-based downscaling reduces distributional biases in the historical period. Besides, Xu et al. (2020) explored the use of advanced machine-learning techniques for downscaling multiple GCM precipitation data in the Upper Han River basin. They used Multilayer Perceptron, Support Vector Machine, and Random Forest algorithms for downscaling and found that downscaled models greatly improved model performance.

CMIP multimodel ensemble downscaling can provide reliable and regionally-relevant climate projection data. Future advancements in computational methods, artificial intelligence, and hybrid approaches (combination of dynamic, statistical and ML-based downscaling) can enhance the accuracy and utility of MME downscaled datasets.

3.6 How should MME data be regridded?

Each model output is based on a specific underlying grid, often referred to as the ‘native’ grid. When combining several models with at least partly different native grids to a MME, researchers must decide on whether to keep the native grids (1) or to regrid their data to a uniform grid (2). A variety of approaches to working with data in different grids can be found in the literature that can be distinguished with these two categories. Methods that retain native grids avoid regridding altogether. Showing individual MME member results in the member’s native grid is one way to accomplish this (Quesada et al., 2017). An alternative to this is plotting the MME mean of the zonal means for each model, which allows data from different models to be combined without regridding (Boysen, 2020). Although there are cases where native grids are retained within an MME, it is more common to regrid to establish grid uniformity within an MME prior to analysis. Regridding involves several

https://doi.org/10.5194/egusphere-2025-4744
Preprint. Discussion started: 6 October 2025

considerations related to spatial and temporal dataset dimensions. For example, one must consider (a) whether it is best to adopt a coarser, intermediate, or finer grid, (b) how to interpolate, and (c) which calendar to use.

Let us consider the question of which grid resolution to choose. A range of grid resolutions are likely to exist within an MME, with one or more of those grids being at the coarse end of the range. Some studies where the direction of regridding is mentioned are silent on why (Achugbu et al., 2022; Cook et al., 2020; Gergel et al., 2024; Hong et al., 2022; Song et al., 2021; Zhao and Dai, 2021) showing that it is common in literature to not disclose the direction of or rationale behind regridding. However, Iles et al. (2020) explain that selecting a coarser grid from multiple high-resolution grids can be acceptable where studies show similar sensitivity test results for the finer and coarser high-resolution grids. In addition, Teuling et al. (2019) regrid to a coarser grid only for data visualization purposes. Iles et al. (2020) state that regridding to a finer grid has the ability to preserve localized extremes to a greater degree than lower resolution data.

Next, one must consider how to interpolate the data that is being regridded. The default interpolation method in most Python packages, for example, is bilinear. This is suitable for many, but not all, variables depending on the type of analysis that is being carried out. Table 2 provides an introduction to the available interpolation methods, which data types they should be applied to, and some examples of CMIP variables for each data type.

Table 2. Interpolation methods commonly used in climate data analysis

Interpolation method	**When to use**	**Data type**	**Example variables**
None	When no filling or averaging of the original data is desired	Categorical	treeFrac, cropFrac
Bilinear	When data point values vary smoothly across a surface	Continuous	tas, sst
First-order conservative	When fluxes must be conserved over a given area	Conservative	pr, evspsbl
Second-order conservative	When fluxes must be conserved over a given area (smoother than first-order conservative when going from coarser to finer grid)	Conservative	mrro, mrso

Nearest neighbor	When strong contrast between areas with discrete or categorical values must be maintained	Categorical	treeFrac, cropFrac
Patch	When the computation of accurate derivatives is needed	Conservative	tauu, tauv

Please note that other interpolation methods exist. Those mentioned here are simply those most commonly used in the regridding of climate data (National Center for Atmospheric Research Staff (Eds).2014).

In addition to the variety of spatial resolutions present within an MME, multiple temporal differences may also exist among members. This is because models may encode different calendars in the simulation files, which are often in netCDF format. There are close to ten calendar options (NetCDF Users Guide: NetCDF Utilities, 2025) and the best choice of calendar for a given study will depend on the study particulars and researcher preference. However, calendars should be brought into alignment during the regridding process to avoid issues when attempting to analyze MME data.

4. Outlook

4.1 Machine Learning

With the rapid production and accumulation of prodigious volumes of climate data, the development and application of automated and increasingly sophisticated analysis techniques are essential (Glymour et al., 2019; Rupe et al., 2017). ML has demonstrated great potential and has emerged as a valuable tool in enhancing ensemble approaches, especially in climate science, see Fig. 4. Over the past 5-10 years ML applications have offered significant advantages in addressing non-linear, high-dimensional, and hierarchical problems (Li et al., 2021 and references therein) and have gained significant popularity by using innovative methods such as neural networks (NN), causal inference, explainable artificial intelligence (XAI), and nonlinear multivariate emergent constraints, and have thus become increasingly competitive with traditional numerical, knowledge-based approaches (see Fig. 5 and de Burgh-Day and Leeuwenburg, 2023; Eyring et al., 2024). Owing to these properties, ML is particularly well-suited for extracting crucial dynamical and physical processes from climate models, enabling a more comprehensive exploration of the valuable information embedded within the data (Reichstein et al., 2019; Wang et al., 2018). Nevertheless, the application of ML algorithms in constructing MMEs for climate impact assessments remains in its early stages. By utilizing observational data as either a reference, benchmark or a constraint, ML offers significant potential for extracting additional insights from MMEs. In short, ML has the potential to make climate models better, faster and to reduce their high energy consumption. Below, we provide an overview of emerging ML approaches for analyzing MMEs, including downscaling and bias correction, causal discovery and process-oriented causal model evaluation, ML for climate system emulation and surrogate modeling, and promising future ML avenues.

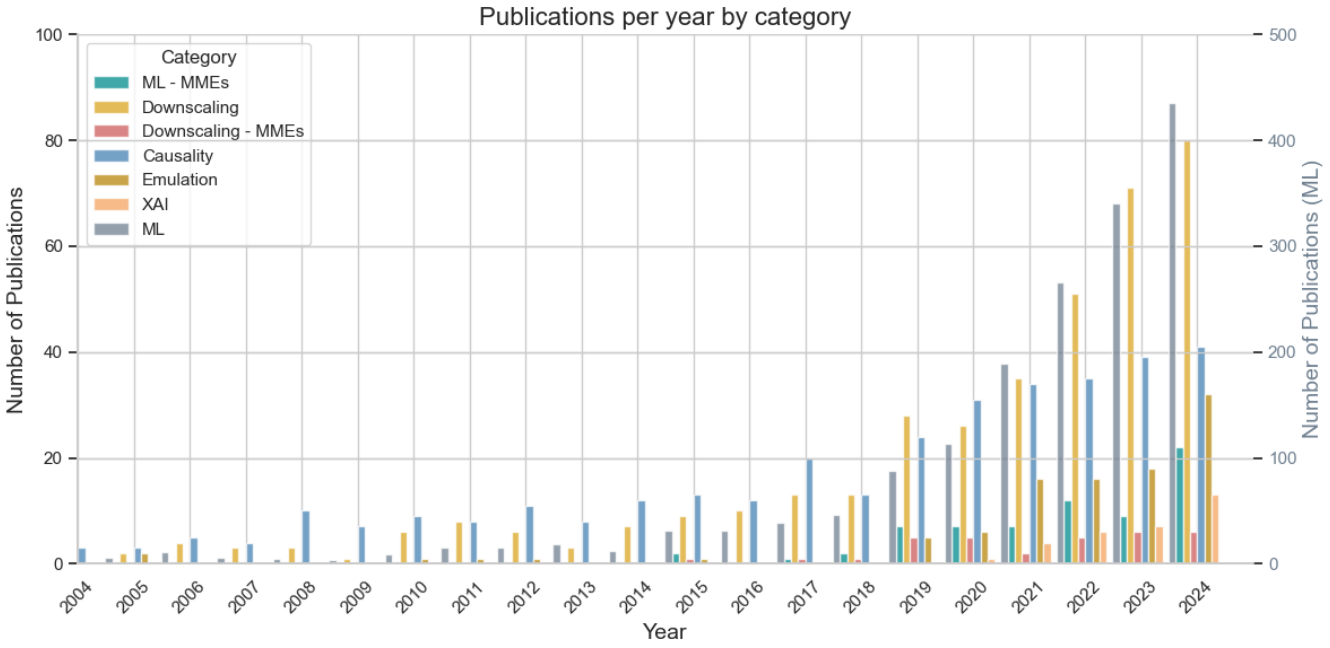

Figure 4. Number of publications per year involving different ML-related techniques and CMIP or general circulation models: ML (y-axis on the right), ML and MMEs, Downscaling, Downscaling and MMEs, Causality, Emulators, and Explainable AI (XAI). The data was extracted from the citation reports available at Web of Science (https://www.webofscience.com/wos/woscc/basic-search) using the queries provided in Appendix 1.

Downscaling and Bias Correction

ESMs have horizontal resolutions often far coarser than those needed by decision makers, and also suffer from substantial biases (Maraun et al., 2017). Recently, the capacity of ML algorithms to summarize large amounts of data and represent non-linear relationships has been exploited, mostly in a regional way, to bias-correct and downscale MME's outputs–with both processes often done simoultaneously. Multiple ML methods have been tested and compared during recent years to predict variables such as temperature and precipitation from MMEs. Some studies have tested algorithms such as random forests (RFs), support vector machines (SVMs), relevance vector machines (RVMs), and artificial neural networks (ANNs) to estimate monthly precipitation, maximum temperature, and minimum temperature (Crawford et al., 2019; Sachindra et al., 2018; Wang et al., 2018; Xu et al., 2020) both at a daily (Dey et al., 2022; Jose et al., 2022; Shetty et al., 2023; Zebarjadian et al., 2024) and yearly temporal resolution (Li et al., 2021). The targets or predictands used in these studies are commonly gridded products that have been interpolated from gauge stations, and it is a common practice to perform dimensionality reduction (e.g. performing principal component analysis on the raw data) before the training process. The domain of these studies is generally limited to the basin scale (Crawford et al., 2019; Dey et al., 2022; Jose et al., 2022; Sachindra et al., 2018; Shetty et al., 2023; Xu et al., 2020; Zebarjadian et al., 2024), although Wang et al. (2018) and Li et al. (2021) obtained good results at a country level for Australia and China, respectively. In many of these downscaling and bias correction studies, it has been found that tree-based approaches (like RFs) commonly perform better than other algorithms. Therefore, they seem to be a good baseline for future research that aims to improve bias correction or downscaling algorithms.

Although these approaches provide a practical way to leverage MME future projections and observations to obtain a “best estimate” of future quantities, there are several critical limitations to consider. First, within these methods, it is assumed that the relationships between model outputs and observations remain stationary, including model biases and errors (Maraun, 2016). However, skillful or poor model performance during the historical period does not necessarily translate into the same for the future, especially since model skill can vary depending on the specific emissions scenario that unfolds. This uncertainty cannot be captured within the historical period, which serves as the only source of information for training algorithms. As a result, such projections may become overly constrained and therefore require careful interpretation, as fundamentally wrong projections come with the danger of influencing wrong policies or eroding public trust. Potential solutions for this aspect are to use trend-preserving learning (Wang and Tian, 2024) or climate-invariant ML methods (Beucler et al., 2024).

Another critical aspect that requires further attention in future ML-based bias correction and downscaling efforts is the potential degradation of the representation temporal variability in final estimates (Shetty et al., 2023). Among the studies mentioned above, only Li et al. (2021) acknowledged that their model outputs showed a significant reduction in the amplitude of interannual variability relative to the original CMIP models. Thus, it is necessary to implement evaluation metrics for the algorithms that consider aspects such as the standard deviation of the generated time series, the frequency and persistence of extreme events, and the amplitude of different modes of variability. Most approaches aim to minimize only one error metric, which could be ignoring the skill regarding these aspects and the physics behind them. For example, the mean precipitation could be improved but the representation of the extreme events or the number of wet days may not be addressed. Algorithms that can minimize multiple loss functions simultaneously could be advantageous to preserve multiple statistical features of the fields of interest (Lin et al., 2019; Sener and Koltun, 2018; Zuluaga et al., 2013). Furthermore, ML-based approaches normally focus on predicting just one variable. Using methods that aim to predict multiple variables could help preserve inter-variable relationships (while also helping preserve different modes of variability). Finally, most bias correction or downscaling algorithms are trained to predict the outputs in one grid cell based on the nearest CMIP grid cell. This approach dismisses spatial relationships contained either within the inputs or the desired outputs. ML methods that consider the spatial relationships within the field of interest could be of use, including convolutional neural networks (Gu et al., 2018; LeCun et al., 2015; Wang and Tian, 2022, 2024). Considering spatial relationships, multiple variables, and multiple error metrics, also diminishes the impact of observational uncertainty, since physical relationships are more easily preserved, and it also reduces the risk of producing overly constrained projections. Considering the limitations of the approaches mentioned for detecting physically plausible connections, it is essential to explore additional methodologies, with causal inference being one promising option.

Causal inference for climate models

A prominent example of supervised ML is causal inference, which strives to discover the causal structure of a complex system like Earth and quantify causal effects by combining domain knowledge, ML models, and data from observations and climate model simulations (Runge et al., 2023 and references therein). Structural causal models (SCMs) have gained traction in statistics and ML for causal inference, maturing into a robust scientific approach (Runge et al., 2019). Widely adopted methods often relying on simple descriptive statistics may not accurately capture the physical mechanisms, leading to underdetermination or equifinality, where multiple incorrect models fit the data equally well (Beven and Freer, 2001). From this perspective, causal dependencies, more closely tied to physical processes, offer a more robust framework against overfitting than simple statistics. Models that reflect causal relationships observed in data are more likely to remain valid under future climate scenarios. Moreover, in the long term, integrating observational data analysis and Earth system modelling is envisioned as a robust approach. In particular, detecting similar causal connections in observations and model simulations provides an opportunity to assess model performance that indicates whether models can correctly reproduce local and remote processes in the climate system and do not simulate expected links for the wrong or unknown reasons. This framework was first introduced by Nowack et al. (2020) and was termed causal model evaluation (CME). A similar approach was proposed by Vázquez-Patiño et al. (2020) for global climate models (GCMs). In this regard, causal inference can identify weaknesses in physical models and guide their improvement, including the development of parameterization schemes. It can also optimize computationally expensive physical model experiments by determining where numerical experiments will likely yield significant results.

Another important development in this area is a causality benchmark platform **causeme.net**, which aims to advance more focused methodological research in Earth System sciences and related fields (Runge et al., 2020), with potential for valuable applications in future studies, particularly in refining approaches for MME analysis. The platform offers synthetic models replicating real data challenges for comparing causal discovery methods, such as for example spatially aggregated vector-autoregressive (SAVAR) models, which can be used to benchmark causal discovery methods for teleconnections (Tibau et al., 2022). It also encourages submissions of real or modeled datasets with well-established causal structures. Therefore, defining evaluation and comparison statistics based on causal networks is vital for building more realistic models, improving future projections, and informing policy-making (Eyring et al., 2019, 2024).

https://doi.org/10.5194/egusphere-2025-4744
Preprint. Discussion started: 6 October 2025

Process-oriented causal analysis and model evaluation

Introduced by Nowack et al. (2020), the CME framework, based on sea level pressure (SLP) data and its components as proxies for modes of variability, enhances the understanding of precipitation patterns in CMIP5 MMEs and meteorological reanalyses. This approach enables a process-oriented evaluation of models, helping to reduce uncertainties in climate projections. To facilitate the comparison of causal relationships and estimate the similarities among observed and modeled causal graphs, the authors introduced a modified asymmetric F_1 score method. The higher the score, the better the agreement between compared causal graphs (with the F_1-score ranging from 1 indicating perfect match to 0 indicating no match). Nowack et al. (2020) showed that causal graphs estimated from different ensemble members of the same model are more consistent than graphs estimated from two different models. Additionally, CME can also serve as a skill to recognize models with shared development backgrounds. Moreover, the authors state that the models with causal fingerprints similar to those in observational data are more effective in replicating significant precipitation patterns in populated regions. The authors find strong indications that CME can help reduce uncertainty in predicting rainfall changes due to climate change, as past model accuracy doesn't guarantee skill for future projections. Numerous examples demonstrate the successful application of the proposed CME in Earth system science by analyzing MMEs. For instance, Karmouche et al. (2023) analyzed Atlantic–Pacific interactions and their phase-dependent changes using the CVDP diagnostic package (see Section 2.6) and regime-oriented CME, focusing on large-ensemble CMIP6 historical model simulations and reanalyses. They highlighted the importance of large ensembles in addressing sampling issues and explained causal pathways specific to regimes that may not appear in reanalysis-based causal networks. Intra-model comparison is crucial to assess differences within the same model ensemble. The study also emphasizes the need for modeling groups to review the documentation regarding realization attributes. In the later study, Karmouche et al. (2024) separated external forcing from internal variability in Atlantic–Pacific climate connections using the CMIP6 multi-ensemble mean (MEM). The MEM, derived from models that realistically simulate the spatiotemporal characteristics of major climate variability modes, was subtracted from the used datasets. This subtraction provided an estimate of the externally forced component, which was further refined using the CME procedure. Process-oriented causal analysis was also successfully applied to study Arctic processes and their connections to the mid-latitudes (Docquier et al., 2022, 2024; Galytska et al., 2023; Kaufman et al., 2024; Kretschmer et al., 2020; Polkova* et al., 2021), subpolar gyre variability (Falkena and von der Heydt, 2024), and evaluation of climate sensitivity (Ricard et al., 2024).

The recent work of Debeire et al. (2025) built their study upon the findings of Nowack et al. (2020) to address the practical challenges of integrating CME with a novel causal multimodel weighting scheme in CMIP6 MMEs of SLP. Their study seeks to improve projections of precipitation changes over land, enhancing the ability to anticipate and respond to the consequences of climate change in populated and vulnerable areas and reduce uncertainties in multi-model climate projections, providing more robust climate change information for more effective mitigation and adaptation strategies. Similarly to Nowack et al. (2020), the authors adopted and adjusted the F_1 score definition and complemented it with a measure of

https://doi.org/10.5194/egusphere-2025-4744
Preprint. Discussion started: 6 October 2025

distance metric $1 - F_1$ score as the performance metric: smaller distance values indicate greater similarity, both in terms of performance relative to the reference graph and in terms of dependence among the models. Debeire et al. (2025) developed a new weighting scheme, termed causal weighting, inspired by the earlier works of Knutti et al. (2017) and Brunner et al. (2020) is based on both the performance and interdependence of model causal networks. They normalize a distance metric $1 - F_1$ score using the median score across all analyzed models, which enables the weighting scheme to assign higher weights to models that closely match the reference causal network (e.g., observational), signifying strong model performance while also favoring models with distinct causal structures, indicating greater independence. Similarly to Nowack et al. (2020) and Debeire et al. (2025) confirm that evaluating the SLP causal networks can identify models with similar physical cores and, consequently, similar dynamical sea-level pressure processes.

Causal (network-based) constraint for evaluation of model sensitivity

The study of Ricard et al. (2024) evaluates climate sensitivity, specifically Equilibrium Climate Sensitivity (ECS) and Transient Climate Response (TCR), using a novel network-based approach built on the analysis of SST patterns and their connectivity. The authors argue that the behavior of SST networks serves as a reliable proxy for how models respond to increased CO_2 levels. The network-based approach called netCS leverages sea surface temperature (SST) variability and teleconnections to constrain climate sensitivity estimate differences from traditional emergent constraints (EC) by relying on 2-D metric space, such as the Weighted Wasserstein Distance (WWD) and Distance Average Causal Effect (D_{ACE}). These metrics quantify the distance between simulated and observed SST patterns, focusing on fast-propagating perturbations over short time scales (up to three months). The study finds that some models may capture regional SST distributions well but fail to replicate connectivity patterns, and vice versa (see discussion to Fig. 5 in Ricard et al., 2024). This distinction is crucial for evaluating model performance over historical periods, as models that accurately reproduce past SST patterns may have better-underlying physics (if not better tuned). While this does not guarantee that those models are the best for future projections (Rasp et al., 2018; Zhu and Poulsen, 2021) it offers valuable evidence, especially when evaluation is based on climate-relevant parameters that are less influenced by tuning, such as for example detrended SST patterns. Runge et al. (2019) has previously stated that the current relationships between predictors and climate sensitivity represent actual physical processes likely to hold under future climate change. Based on their analysis, Ricard et al. (2024) defined two clusters of models that best reproduce the SST variability: low-sensitivity and high-sensitivity models, and the dominant one is persistently in the low ECS/TCR, which might suggest that the warming will be less than the average one from the models. The authors propose that causal networks, used alongside traditional ECs, provide a more reliable ranking of models for future climate projections. Ultimately, the authors recommend combining netCS with other ECs to improve the plausibility of future climate projections and provide robust estimates of ECS and TCR. The application of causal discovery algorithms helps bridge the gap between physical understanding and statistical tools, enabling more comprehensive insights into Earth system processes.

Machine Learning for Climate System Emulation

Climate model emulators, including surrogate models, are simplified representations of the complex systems included in climate models, allowing for faster computations and predictions. They can mimic the behaviour of a climate model without needing to solve the underlying equations in full. ML presents a unique opportunity to replicate parts of the climate system in novel and computationally viable parameterizations. These approaches have the potential to increase both the accuracy and efficiency of climate simulations while significantly reducing computational costs and enabling higher resolution simulations (Eyring et al., 2021; Gentine et al., 2018). Traditional climate models, which often rely on complex numerical methods, can be computationally expensive when simulating small scale processes. ML based emulation of these processes provides a computationally cheaper alternative that can capture these dynamics with, in some cases, comparable or even improved accuracy compared to observations. The success of ML emulation of the climate system varies depending on the choice of algorithm, temporal resolution, type of training data, and model complexity (Dueben and Bauer, 2018; Scher, 2018). The ML emulation of MME is of particular interest. As discussed previously, conventional MME approaches face challenges such as high computational costs and model biases, and ML-based MME frameworks could help overcome these computational costs while also reducing biases and uncertainties (Wang et al., 2018).

Efforts to overcome initial barriers of the use of ML in the climate sciences have recently gained momentum (see Figure 4). One notable initiative is ClimSim, a hybrid physics-ML dataset designed to provide high-quality data for training ML emulators of climate processes (Yu et al., 2023). These datasets have been tested for deterministic and stochastic parameters, and show promise for future climate simulations if used properly. Future studies could include using MME as training data to train novel ML emulator models. Complimenting the available data to train emulators, Lu and Ricciuto (2019) highlight an innovative approach integrating SVD, Bayesian optimization, and neural networks to create a computationally efficient surrogate model. Weber et al. (2020) provides valuable technical notes of ML, using the example of forecasting precipitation under CO_2 forcing, for creating surrogate models to overcome potential computational burdens. The continued development and advancement of ML emulators and surrogate models for climate systems, particularly in the context of MME, will require ongoing innovation in interpretability, generalization, and reliability. The remarkable computational efficiency and ability of ML emulators to replicate complex climate processes with high precision demonstrates their immense potential. However, several challenges remain, including the high cost of running models, limited diversity in training data, and the need for more robust methods to evaluate simulations. As these tools develop further, they show promise to play a transformative role in enhancing the speed, resolution, and reliability of future climate projections.

Promising Future ML Avenues

There are many avenues of promising research involving ML to process CMIP outputs. Work that aims to predict end-user variables that are not directly available in GCMs, including crop yield (Crane-Droesch, 2018; Sidhu et al., 2023; Veenadhari et al., 2014) and power generation potential (Jung et al., 2021; Nwokolo et al., 2023; Yeganeh-Bakhtiary et al., 2022), highlights the potential of AI for increasing MMEs applicability to end-users, including decision-makers and stakeholders. Explainable AI, which aims to obtain physical interpretations from the initially black-box-like ML models, is especially helpful in inferring physical changes in the Earth system based on CMIP simulations (Rader et al., 2022). Layer-wise relevance propagation (LRP), for example, has been used to provide insights into the regions and features that a neural network relies on for making predictions (Toms et al., 2020). LRP has proven to be particularly useful in climate science, allowing for the interpretability of a neural networks decision making process by visualizing heatmaps of relevant regions (Hilburn et al., 2020; Labe et al., 2024; Labe and Barnes, 2022; Sonnewald and Lguensat, 2021). This interpretability adds value to ensemble evaluation, providing critical information that can inform model weighting schemes, as discussed in Section 2.4. These types of methods, in addition to ML algorithms, are useful to move toward process-informed or process-oriented correction or downscaling of MME outputs (Maraun et al., 2017). ML also serves as an effective tool for evaluating both the performance and independence of climate models within MMEs, offering valuable potential for assessing model individuality and developing ensemble weighting metrics to address interdependencies among models (Brunner and Sippel, 2023). Given the potential that ML has to improve climate projections or help with their interpretability and applications, AI-ready databases such as ClimateSet (Kaltenborn et al., 2023) are of great help to the climate research community. Real world applications of ML based climate emulation highlight the value of this approach. For example, ML emulation models have been employed to predict crop yields (Folberth et al., 2019; Leng and Hall, 2020). CNN surrogates also show promise in modelling spatio-temporal precipitation patterns, with deeper networks offering greater accuracy, improving long-term forecasting (Weber et al., 2020).

The integration of causal discovery and deep learning (DL) presents a promising avenue for improving climate simulations (Iglesias-Suarez et al., 2024; Kyono et al., 2020; Luo et al., 2020; Russo and Toni, 2022; Wang et al., 2024; Yoon and Schaar, 2017; Zhang et al., 2023). This combination aims to enhance the stability and trustworthiness of models, particularly addressing biases and uncertainties associated with subgrid-scale processes, such as clouds and convection, which are significant contributors to climate projection uncertainties. Previous research has demonstrated DL's capability to represent small-scale processes effectively, such as deep convection, using storm-resolving model simulations (Eyring et al., 2021; Gentine et al., 2018; Grundner et al., 2022). Despite this potential, DL algorithms have faced criticism for robustness issues, poor generalization, and the reliance on spurious, non-physical relationships, particularly when conditions diverge from the training data (Brenowitz et al., 2020; Scholkopf et al., 2021; Thuy and Benoit, 2024). However, Iglesias-Suarez et al. (2024) demonstrated that causal discovery can effectively identify the physical drivers of subgrid-scale processes across different climate regimes, thereby enhancing the interpretability and reliability of DL algorithms. Their causally-informed, data-driven approach operates stably within the reference climate conditions, generating climate means and variability that closely match original simulations. Moreover, their findings suggest that causally-informed NN help prevent spurious links typically seen in traditional DL-based parameterizations, directing more focus on physical drivers. This aligns with previous work by Zhang et al., 2023 emphasizing the value of integrating domain knowledge to address the limitations of purely data-driven models. While these studies currently do not pertain to multi-model analysis, their methodologies hold significant potential for future applications in this area. The integration of causal discovery and deep learning thus represents a novel strategy that could lead to more stable and reliable climate simulations, paving the way for advancements in climate modeling methodologies.

4.2 SMILES

Using several simulations per model in MMEs

For the majority of models in CMIP5 and CMIP6, only one ensemble member is available (Milinski et al., 2020; Olonscheck and Notz, 2017). Thus, modeling groups strive to provide their best performing models, carefully calibrated to the same internationally available observational datasets. In this context, Sanderson et al. (2008) found that the standard model performed comparatively to the best-performing model. Therefore, there is an indirect incentive for modelling groups to add simulations to the CMIP MME that are less extreme, potentially leading to a MME that underestimates the uncertainties. As one consequence, the seemingly reduced uncertainty throughout different climate model generations might be at least partly originated in improved calibration and model selection rather than improvements in capturing the physical dynamics (Knutti, 2010).

To overcome this issue, including several simulations from individual models into MMEs might be the next step forward. When the ensemble size reaches 10-100 members, ICEs are referred to as SMILES (Deser et al., 2020). Olonscheck and Notz (2017) found that for annual global-mean surface air temperature and sea ice volume and area, even small ensemble sizes greater than one, as provided in CMIP5, are representative for the model's total internal variability as demonstrated by the CESM1 and MPI-ESM-LR large ensembles. The authors highlight that incorporating multiple small ensembles from different models can improve projections compared to single model ensembles, particularly for extreme events. Additionally, such ensembles can also be useful for quantifying the response uncertainty across different models. Such multi-model collection of SMILEs can be used for robust comparison of both the forced response on regional or decadal scales across models and internal variability across models (Deser et al., 2020). However, accessing and processing large data sets from various sources can be challenging and is probably a key reason why most SMILE studies so far included only one, maximum two large ensembles (Deser et al., 2020). To overcome this issue and to facilitate future usage of multi-model large ensembles, a data repository for large ensembles from CMIP5 models was created including gridded fields of key variables at daily and monthly resolution for historic and future emission scenarios, the 'Multi-Model Large Ensemble Archive (MMLEA)' (US CLIVAR, 2020). When more ensemble members are used, it is important to remember that the ensemble size available for the individual models should not influence the weight given to this model in the MME (Knutti et al., 2010a). Future studies should provide a methodological framework on how to combine SMILES and MMEs in the most productive and meaningful way.

What are SMILEs and how do we benefit from them

SMILEs represent valuable resources for studying the climate system. A SMILE consists of many simulations from a single climate model based on the same model physics and under the same external forcings, but each starting from slightly different initial states (Maher et al., 2021). Although MMEs are useful for examining the combined influence of three types of uncertainties in climate projections (model uncertainty, internal variability uncertainty, and scenario uncertainty), it remains a challenge to distinguish internal variability from the forced response with a limited number of ensemble members of each model. For addressing uncertainties related to both internal variability and unknown future pathways (scenario uncertainty), SMILEs can be very powerful (Deser et al., 2012b; Lehner et al., 2020), especially when it comes to regional detection and attribution and extreme climate events (Lehner et al., 2017; McKenna and Maycock, 2021; von Trentini et al., 2020; van der Wiel et al., 2021).

A large ensemble provides more instances of extreme events, allowing researchers to better estimate changes in their frequency, intensity and future likelihood. This is particularly important for assessing the risk and impacts of climate extremes in a changing climate, as a single realization of a model might not capture a sufficient number of examples. Such information is crucial for decision makers and policy makers in developing climate change adaptation and mitigation strategies, providing them with the data necessary to understand the full range of potential outcomes.

While large ensemble simulations are known to be important to study extreme univariate events, they are even more relevant for the analysis of compound events (such as simultaneous drought and heatwave) (Bevacqua et al., 2022, 2023; Wu et al., 2023). Compound events result from combinations of multiple weather and climate drivers, characterized by complex interactions between extreme conditions across variables, space, or time. Because of these multivariate relationships, a univariate approach for examining hazards may underestimate risks and potential changes in dependence between variables may lead to even larger uncertainties. As internal variability can obscure the detection of trends or make the estimation of event probabilities less certain, SMILEs can reduce this uncertainty by providing a larger sample size, and enabling a clearer distinction between internal variability and forced responses. Bevacqua et al. (2023) showed that attributing compound events requires larger sample sizes than univariate events, especially when the drivers are weakly correlated and have similar trends. Sampling a wide range of possible atmospheric conditions using SMILEs helps avoid underestimating the frequency and severity of compound events and provides deeper insights into their physical drivers and potential future changes.

SMILES as a way of employing MME

Given the value of integrating SMILEs into MME analysis (see Figure 5), we highlight their potential to improve uncertainty quantification and the robustness of climate projections. One challenge to employing SMILEs can be accessing the data. To address this the Multi-Model Large Ensemble Archive (MMLEA) was developed (Deser et al., 2020). The newly published

MMLEAv2 expands beyond the original MMLEA by including more models (the original included 7, the new version includes 18) and more three-dimensional variables (Maher et al., 2024). The MMLEAv2 and a suite of corresponding observational datasets have been regridded onto a 2.5° common horizontal grid, reducing data size, allowing for straightforward model-to-model comparison, and model-to-observation comparisons. An additional tool that is being published with the MMLEAv2 archive is the newest version (version 6) of the CVDP (CVDPv6; Phillips et al., 2020) mentioned in Section 2.6.

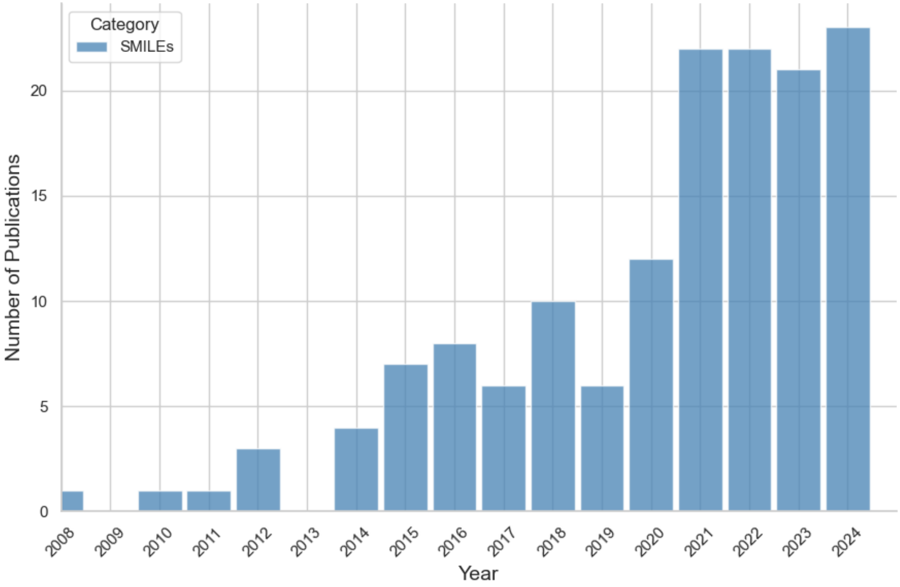

Figure 5. Number of publications per year involving SMILEs. The data was extracted from the citation report available at Web of Science (https://www.webofscience.com/wos/woscc/basic-search) for the queries provided in Appendix 1.

4.3 Computational Resources and Energy Costs

MMEs, such as CMIP6, are powerful tools for exploring past climates, assessing our current changing climate, and projecting future scenarios, but they come with significant computational and energy demands. MMEs rely on ensemble runs across multiple models or multiple versions of a single model, generating a large volume of data that requires careful management and optimization. These simulations are run on high-performance computing (HPC) platforms, which must process large amounts of data and perform calculations across many parallel cores. Simulating a century-scale global climate model with high spatial and temporal resolutions can take weeks, even on high-performance computing systems. For example, the MPI-ESM1.2 model, in its standard low-resolution configuration (approximately 200 km grid spacing), runs at around 45 years per day up to approximately 85 simulated years per physical day, which is a significant improvement over the 17 years per day achieved during CMIP5 simulations (Mauritsen et al., 2019). On the other hand, running an ultrahigh-resolution climate model in a near-global setup, with a ~1 km horizontal resolution attains a performance of approximately 0.043 simulated years per day (~15.7 simulated days per day) (Fuhrer et al., 2018). Computational performance is a key limitation when designing ESM experiments, requiring trade-offs between resolution, complexity, and the size of ensembles.

CPMIP metrics for climate modelling

The demand for computing power has continued to increase over time. Several factors contribute to this: increasing resolution, explicit resolving of complex processes within the climate system replacing parameterization, the need for larger ensemble sizes, and the associated need for more storage space for the large amounts of data (input and output). Balaji et al. (2017) introduced a universal set of metrics to evaluate HPC and ESM performance and emphasize that traditional metrics (e.g., floating point operations per second) are becoming insufficient to represent the generations of new machines and the diversity of ESMs. Given the complexity of ESMs and the diverse computational characteristics of their components, they advocated making these metrics a standard in globally coordinated modeling initiatives and proposed collecting them in the Computational Performance MIP (CPMIP). The metrics (Table 3) are intended to serve as a uniform basis for assessing the advances and technological progress of climate models and take into account the structure of ESM and production runs. The advantage is that they are universally accessible and easily collected during routine production runs without special additional tools, reflect real-world performance (rather than idealized estimates) and are designed to capture performance over the entire modeling lifecycle.

Table 3. List of metrics introduced in CPMIP, table is adapted from Acosta et al. (2024).

Metric	**Short description of the metric**
Resolution (spatial degrees of freedom)	Number of grid points per model component
Complexity	Number of prognostic variables per component
Platform	Description of the computational hardware (core count, clock speed, and double-precision operations per clock cycle)
Simulation years per day (SYPD)	Number of simulated years per day for the ESM in a 24-hour period on a given platform
Actual SYPD (ASYPD)	Actual simulated years per day for a long-running simulation on a given platform (system interruptions, queue wait time, or issues with the model workflow accounted)
Core hours per simulated year (CHSY)	Cost, measured in core hours per simulated year

Parallelization	Total number of cores allocated for the run
Joules per simulated year (JPSY)	Energy cost per simulated year
Coupling cost	Computing cost of the coupling algorithm and load imbalance
Memory bloat	Ratio of actual memory size to ideal memory size
Data output cost	Computing cost for performing input/output (I/O)
Data intensity	Measure of data produced per computing hour

Evaluating models' performances using CPMIP metrics

Each model in a MME may have different performance characteristics, and addressing these can lead to more balanced and effective use of computational resources. Recent main findings from the CPMIP (Acosta et al., 2024) represents analysis of metrics proposed by Balaji et al. (2017), collected during long, real-time model runs, from the 14 institutions that conducted a total of 33 experiments used in CMIP6 (almost 500,000 years of simulations on 14 different HPC machines). Acosta et al. (2024) extends the foundational work CPMIP by incorporating empirical data from CMIP6, emphasizing energy consumption, addressing data storage challenges, and offering strategic recommendations for future climate modeling efforts.

Improving model accuracy through higher resolutions and increased complexity in representing physical, chemical, and biological processes, which provide more detailed spatial and temporal outputs, would require immense computational resources. For example, Flato (2011) found that increasing model resolution from 200 km to 20 km demands roughly 10,000 times more computing power. As shown in the CPMIP study, institutions found that increasing model resolution tends to increase execution costs due to both the computational power required and the challenges posed by coupling independent model components like atmosphere, ocean, land and cryosphere (Acosta et al., 2024).

Kilometer-scale simulations of individual models and multi-model ensembles of these high-resolution simulations are being actively developed (Ban et al., 2021; Coppola et al., 2020; Pichelli et al., 2021; Rackow et al., 2025). Alongside these developments, coordinated intercomparisons for global storm-resolving models (GSRM) are emerging, including a recently introduced protocol for one-year simulations (Takasuka et al., 2024), aimed at extending GSRM evaluations toward climatic timescales. The increase in resolution and process detail comes with significantly higher computational demands, requiring substantial computing power and storage resources (Schär et al., 2020). To cope with the high computational and energy demands, high-resolution simulations are usually regional and provide information for different specific geographical regions (Coppola et al., 2020; Nolan and Flanagan, 2020) or rely on some simplified parameterizations (for processes such as radiation or soil interactions), as more complex and advanced schemes are computationally expensive and would significantly increase the computational load in long-term simulations. Another constraint that arises for such high-resolution modeling, is that while regional models can simulate periods up to a decade, global models are typically confined to high-resolution simulations spanning only a few weeks (Schär et al., 2020). However, this limitation is rapidly being overcome, with multi-year global simulations at such resolutions already conducted using models such as ICON in its Sapphire configuration (Hohenegger et al., 2023), the eXperimental System for High-resolution prediction on Earth-to-Local Domains (X-SHiELD) (Guendelman et al., 2024; Merlis et al., 2024), and the IFS model coupled to the Finite-volumE Sea ice-Ocean Model (Rackow et al., 2025).

ESMs are structured with a component-based architecture, which means different climate components are modular, allowing scientists to update or add new components over time. This architecture enables continuous innovation, but it also brings software engineering challenges by changing the model's computational demands, affecting aspects such as data processing, I/O operations, and network traffic (Wang and Yuan, 2020). As shown in Acosta et al. (2024) coupling components, which synchronize different processes, adds up to 5–15% overhead to execution costs.

Queue times significantly impact overall execution speed and efficiency, although they can vary across different institutions (Acosta et al., 2024). Consistent and minimal queue times are beneficial for MMEs in terms of ensuring timely completion of simulations and data availability and reducing them would allow for more simulations to be run in parallel, enhancing the overall throughput of the ensemble.

Estimated carbon footprint of climate modeling: Towards "greener" hardware

Running climate models, especially in large-scale MMEs, requires significant computational power which can have a notable carbon footprint, since HPC facilities are consuming large amounts of energy. The climate modeling community is aware of this and is exploring ways to optimize code efficiency and transition to greener energy sources to minimize the carbon impact of their research efforts. One unique aspect of CPMIP is its focus on capturing the real energy costs of running models, aiming to help climate scientists make eco-friendly decisions in computing. With the CPMIP metrics and the efforts of the Infrastructure for the European Network for Earth System Modelling Phase 3 (IS-ENES3) project (Joussaume and Budich, 2013) consortium's Carbon Footprint Group assessing the total computational energy costs of climate experiments enabled (Acosta et al., 2024) the estimation of carbon footprint related to those experiments. For 8 out of 49 institutions that were involved in CMIP6, the estimation is 1,692 t CO_2 in total (with total energy costs ranging from 0.41 TJ to 26.70 TJ). According to the International Energy Agency (IEA), the "global average energy-related carbon footprint" is ~ 4.7 t CO_2 per person and per year. For the context, given that the total emissions from CMIP6 modeling centers are estimated at 1,692 tons of CO_2, this is equivalent to the annual emissions of 360 people.

Eco-friendly hardware is increasingly becoming a consideration in HPC for climate modeling as researchers recognize the environmental impact of extensive model runs. One example of this good practice is the Energy-efficient climate simulations

https://doi.org/10.5194/egusphere-2025-4744
Preprint. Discussion started: 6 October 2025

on heterogeneous supercomputers through co-design (EECliPs) project led by German Climate Computing Centre (Deutsches Klimarechenzentrum, DKRZ) (https://www.dkrz.de/en/projects-and-partners/projects-1/eeclips), aiming to improve simulation quality with lower energy requirements of the ESM ICON (Adamidis et al., 2025). By encouraging institutions to collect the data needed to estimate their carbon footprint and adopting eco-friendly hardware and thoughtful modeling practices, the climate modeling community can reduce its carbon footprint while advancing its scientific mission.

HPC facilities: petascale and beyond

As climate models continue to evolve, HPC facilities operating at the petascale and beyond are necessary to handle the spatial and temporal resolutions required by these models, especially for simulating more complex interactions or high-impact short-term events and regional processes that require finer spatial scale and higher accuracy, as well as advanced data management systems to handle large data sets required for model validation, diagnostic analysis and impact studies. The development of exascale computing systems, capable of achieving 10^{18} floating-point operations per second, holds significant potential for advancing our understanding of the predictability boundaries in ESMs through sophisticated mathematical and statistical methods, which led to the launch of many projects aiming to develop and optimize the parallel execution on exascale systems (Adamidis et al., 2025; Taylor et al., 2023; https://www.fz-juelich.de/en/ias/jsc/projects/ifces2).

Addressing the computational and energy challenges of MMEs requires standardized performance metrics, efficient computing and eco-friendly practices. Findings from the CPMIP and performance metrics applied to CMIP6 experiments, highlight the need for better optimization of model configurations, improved coupling mechanisms, and more efficient use of HPC resources, which is particularly important as modeling centers strive to improve projections while managing resource limitations. The intercomparison reveals significant differences in computational costs between models and institutions, highlighting the need for strategic advancements in model optimization to balance scientific accuracy with practical constraints. Joint efforts are needed to integrate the latest technological advances such as AI-driven model optimization, novel HPC architectures and energy-efficient computing. Using standardized measurements of computational and energy costs across different MMEs is highly encouraged, ensuring that model performance is comparable and consistent, allowing researchers to identify areas for improvement and make informed decisions for hardware, software, and resource planning in climate modelling.

5. Concluding remarks

Climate modeling has been key to the understanding of past, present, and future changing climates. It is a dynamic field, profiting from growing computational capacities and advances as well as benefits from the increasing understanding of physical and chemical phenomena. Climate projections rely on MMEs to assess uncertainties and improve their robustness. This review synthesizes key practices, challenges, and emerging approaches in working with MMEs, drawing on the collective insights of the Fresh Eyes on CMIP community. By examining model evaluation strategies, systematic biases, model dependence, selection and weighting methods, and uncertainty quantification, we aim to support researchers in making informed choices when designing MME studies—while fully acknowledging that the diversity of research questions makes it impossible to create a set of universally transferable recommendations. We further highlight the growing relevance of ML and SMILEss, which are shaping the future of climate ensemble analysis, particularly in the context of CMIP7. Finally, we advocate for awareness of the computational costs associated with climate modeling and analyses.

Acknowledgements

We thank the wider Fresh Eyes on CMIP community and steering group, the CMIP International Project Office as well as the broader CMIP community for their valuable engagement and support, as well as for providing foundational infrastructure including communication platforms that made this work possible. We specifically thank Elisabeth Dingley and Yuhan Douglas Rao for their valuable guidance and continuous support. We greatly acknowledge the valuable feedback provided by Ranjini Swaminathan and Tomoki Miyakawa during the CMIP internal review process, which helped improve the manuscript. We acknowledge Josh Dorrington for initial help with this project. We also acknowledge Hasi Aru et al. for their figure that we included as Fig. 1 in our manuscript. NČ thanks Robert Pincus, Gregory Cesana, Andrew Ackerman, and others at Columbia CCSR and NASA GISS for introducing her to the CMIP community and for insightful discussions on various topics related to analysis and evaluation of CMIP MMEs in earlier projects. JSPC thanks Maria J. Molina for mentoring on how ML can assist the climate science community, and Isla R. Simpson for sharing her expertise in uncertainty in climate projections and emergent constraints. AK thanks Anders Levermann, Jacob Schewe, and Julia Pongratz for insightful discussions on MME design in earlier projects, which offered a valuable foundation for the present study. MT thanks Vladimir Djurdjevic for his foundational mentorship in understanding global and regional climate model ensembles, and Theodore Shepherd for insightful discussions that significantly shaped her thinking on uncertainty, storylines and the interpretation of MME data. EG thanks Veronika Eyring and Jakob Runge for useful discussions on ML and causality topics related to climate model evaluation. CL thanks Kirsten Zickfeld for valuable exchanges on result robustness and statistical analysis more generally and Alex Koch for the introduction to working with CMIP data.

CL acknowledges support from the Natural Sciences and Engineering Research Council of Canada Discovery Grant Program (grant no. RGPIN-2018-06881 awarded to K. Zickfeld. EG is funded by the Central Research Development Fund at the University of Bremen, Funding No: ZF04A/2023/FB1/Galytska Evgenia. PP acknowledges the financial support received in the form of a doctoral research fellowship from the Council of Scientific and Industrial Research (CSIR), India, Award no: 09/1187(11135)/2021-EMR-I. M. T. acknowledges support from the Science Fund of the Republic of Serbia (Grant No. 7389, Project Extreme weather events in Serbia - analysis, modelling and impacts” - EXTREMES). JSPC was supported by a University of Maryland Grand Challenges Seed Grant. NČ acknowledges support from the NOAA grant NA20OAR4310390, the NASA Modeling, Analysis, and Prediction Program number 80NSSC21K1134, ARIS Programme P1-0188, and the University of Ljubljana Grant SN-ZRD/22-27/0510, which covers the fee costs of this publication.

Author Contribution Statement

All authors conducted a literature review, contributed valuable ideas to the scientific content and study design, topic discussions, and writing of the manuscript (Abstract: AK, NČ; Introduction: NČ, AK; Subsection 2.1: EG, AK, IR, NČ; Subsection 2.2: NČ; Subsection 2.3: KG; Subsection 2.4: KG, PP; Subsection 2.5: JSPC, MT; Subsection 2.6: NČ, EG, MT; Subsection 3.1: CL; Subsection 3.2: AK, AVC; Subsection 3.3: MT; Subsection 3.4: CL; Subsection 3.5: PP; Subsection 3.6. CL; Subsection 4.1: EG, KG, JSPC; Section 4.2: AK, AVC, MT; Subsection 4.3: MT; Conclusion: AK). Final details will be provided with publication.

APPENDIX 1. Statistics of the field over past decades

Figures 5 and 6 were built using data from the Web of Science database. The queries for each category are:

Total ML:

TS=("machine learning" OR "artificial intelligence" OR "neural networks" OR "random forest" OR "decision trees" OR "deep learning" OR "supervised learning" OR "unsupervised learning") AND TS=("CMIP" OR "CMIP3" OR "CMIP5" OR "CMIP6" OR "Coupled Model Intercomparison Project" OR "climate model" OR "climate models" OR "general circulation model" OR "general circulation models" OR "Earth system model" OR "Earth system models")

ML-MME:

TS=("machine learning" OR "artificial intelligence" OR "neural networks" OR "random forest" OR "decision trees" OR "deep learning" OR "supervised learning" OR "unsupervised learning") AND TS=("CMIP" OR "CMIP3" OR "CMIP5" OR "CMIP6" OR "Coupled Model Intercomparison Project" OR "climate model" OR "climate models" OR "general circulation model" OR "general circulation models" OR "Earth system model" OR "Earth system models") AND TS=("multi-model ensemble" OR" multi-model ensembles")

ML-Downscaling:

TS=("machine learning" OR "artificial intelligence" OR "neural networks" OR "random forest" OR "decision trees" OR "deep learning" OR "supervised learning" OR "unsupervised learning") AND TS=("CMIP" OR "CMIP3" OR "CMIP5" OR "CMIP6" OR "Coupled Model Intercomparison Project" OR "climate model" OR "climate models" OR "general circulation model" OR "general circulation models" OR "Earth system model" OR "Earth system models") AND TS=("downscaling" OR "bias correction")

ML-Downscaling MME:

TS=("machine learning" OR "artificial intelligence" OR "neural networks" OR "random forest" OR "decision trees" OR "deep learning" OR "supervised learning" OR "unsupervised learning") AND TS=("CMIP" OR "CMIP3" OR "CMIP5" OR "CMIP6" OR "Coupled Model Intercomparison Project" OR "climate model" OR "climate models" OR "general circulation model" OR "general circulation models" OR "Earth system model" OR "Earth system models") AND TS=("downscaling" OR "bias correction") AND TS=("multi-model ensemble" OR" multi-model ensembles")

ML Causality:

TS=("CMIP" OR "CMIP3" OR "CMIP5" OR "CMIP6" OR "Coupled Model Intercomparison Project" OR "climate model" OR "climate models" OR "general circulation model" OR "general circulation models" OR "Earth system model" OR "Earth system models") AND TS=("causal discovery" OR "causality" OR "causal inference" OR "causal")

ML Emulators:

TS=("machine learning" OR "artificial intelligence" OR "neural networks" OR "random forest" OR "decision trees" OR "deep learning" OR "supervised learning" OR "unsupervised learning") AND TS=("CMIP" OR "CMIP3" OR "CMIP5" OR "CMIP6" OR "Coupled Model Intercomparison Project" OR "climate model" OR "climate models" OR "general circulation model" OR "general circulation models" OR "Earth system model" OR "Earth system models") AND TS=("emulation" or "surrogate" or "emulator" or "emulators" or "surrogates")

ML XAI:

TS=("machine learning" OR "artificial intelligence" OR "neural networks" OR "random forest" OR "decision trees" OR "deep learning" OR "supervised learning" OR "unsupervised learning") AND TS=("CMIP" OR "CMIP3" OR "CMIP5" OR "CMIP6" OR "Coupled Model Intercomparison Project" OR "climate model" OR "climate models" OR "general circulation model" OR "general circulation models" OR "Earth system model" OR "Earth system models") AND TS=("XAI" OR "explainable AI" OR "Layer-wise Relevance Propagation" OR "LRP" OR "Feature importance analysis" OR "feature importance")

Model Independence:

TS=("climate" OR "Earth" OR "Earth System") AND TS=("CMIP" OR "Coupled Model Intercomparison Project" OR "climate model" OR "general circulation model") AND TS=("ensemble" OR "multi-model ensemble") AND TS=("dependence" OR "independence" OR "genealogy")

SMILEs:

TS=("Multi-model ensemble" OR "coupled model intercomparison" OR "cmip") AND TS= ("large ensemble" OR "grand ensemble" OR "smile")

References

Abdelmoaty, H. M., Papalexiou, S. M., Rajulapati, C. R., and AghaKouchak, A.: Biases Beyond the Mean in CMIP6 Extreme Precipitation: A Global Investigation, Earths Future, 9, e2021EF002196, https://doi.org/10.1029/2021EF002196, 2021.

Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., Lorenz, R., Pincus, R., and Schmidt, G. A.: ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing, Earth Syst. Dyn., 10, 91–105, https://doi.org/10.5194/esd-10-91-2019, 2019.

Achugbu, I. C., Olufayo, A. A., Balogun, I. A., Adefisan, E. A., Dudhia, J., and Naabil, E.: Modeling the spatiotemporal response of dew point temperature, air temperature and rainfall to land use land cover change over West Africa, Model. Earth Syst. Environ., 8, 173–198, https://doi.org/10.1007/s40808-021-01094-8, 2022.

Acosta, M. C., Palomas, S., Paronuzzi Ticco, S. V., Utrera, G., Biercamp, J., Bretonniere, P.-A., Budich, R., Castrillo, M., Caubel, A., Doblas-Reyes, F., Epicoco, I., Fladrich, U., Joussaume, S., Kumar Gupta, A., Lawrence, B., Le Sager, P., Lister, G., Moine, M.-P., Rioual, J.-C., Valcke, S., Zadeh, N., and Balaji, V.: The computational and energy cost of simulation and storage for climate science: lessons from CMIP6, Geosci. Model Dev., 17, 3081–3098, https://doi.org/10.5194/gmd-17-3081-2024, 2024.

Adam, O., Schneider, T., and Brient, F.: Regional and seasonal variations of the double-ITCZ bias in CMIP5 models, Clim. Dyn., 51, 101–117, https://doi.org/10.1007/s00382-017-3909-1, 2018.

Adamidis, P., Pfister, E., Bockelmann, H., Zobel, D., Beismann, J.-O., and Jacob, M.: The real challenges for climate and weather modelling on its way to sustained exascale performance: a case study using ICON (v2.6.6), Geosci. Model Dev., 18, 905–919, https://doi.org/10.5194/gmd-18-905-2025, 2025.

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, https://doi.org/10.1175/1525-7541(2003)004<1147:TVGPCP>2.0.CO;2, 2003.

Ahmed, F. and Neelin, J. D.: A Process-Oriented Diagnostic to Assess Precipitation-Thermodynamic Relations and Application to CMIP6 Models, Geophys. Res. Lett., 48, e2021GL094108, https://doi.org/10.1029/2021GL094108, 2021.

Ahn, M., Daehyun, K., Sperber, K. R., Kang, I.-S., Maloney, E., Waliser, D., Hendon, H., and on behalf of WGNE MJO Task Force: MJO simulation in CMIP5 climate models: MJO skill metrics and process-oriented diagnosis, Clim. Dyn., 49, 4023–4045, https://doi.org/10.1007/s00382-017-3558-4, 2017.

Ahn, M., Kim, D., Kang, D., Lee, J., Sperber, K. R., Gleckler, P. J., Jiang, X., Ham, Y., and Kim, H.: MJO Propagation Across the Maritime Continent: Are CMIP6 Models Better Than CMIP5 Models?, Geophys. Res. Lett., 47, e2020GL087250, https://doi.org/10.1029/2020GL087250, 2020.

Almazroui, M., Saeed, S., Islam, M. N., Khalid, M. S., Alkhalaf, A. K., and Dambul, R.: Assessment of uncertainties in projected temperature and precipitation over the Arabian Peninsula: a comparison between different categories of CMIP3 models, Earth Syst. Environ., 1, 12, https://doi.org/10.1007/s41748-017-0012-z, 2017.

Amali, A. A., Schwingshackl, C., Ito, A., Barbu, A., Delire, C., Peano, D., Lawrence, D. M., Wårlind, D., Robertson, E., Davin, E. L., Shevliakova, E., Harman, I. N., Vuichard, N., Miller, P. A., Lawrence, P. J., Ziehn, T., Hajima, T., Brovkin, V., Zhang, Y., Arora, V. K., and Pongratz, J.: Biogeochemical versus biogeophysical temperature effects of historical land-use change in CMIP6, https://doi.org/10.5194/egusphere-2024-2460, 27 August 2024.

Annan, J. D. and Hargreaves, J. C.: Reliability of the CMIP3 ensemble, Geophys. Res. Lett., 37, https://doi.org/10.1029/2009GL041994, 2010.

Annan, J. D. and Hargreaves, J. C.: On the meaning of independence in climate science, Earth Syst. Dyn., 8, 211–224, https://doi.org/10.5194/esd-8-211-2017, 2017.

NetCDF Users Guide: NetCDF Utilities: https://docs.unidata.ucar.edu/nug/current/netcdf_utilities_guide.html, last access: 12 May 2025.

Aru, H., Chen, W., Chen, S., Garfinkel, C. I., Ma, T., Dong, Z., and Hu, P.: Variation in the Impact of ENSO on the Western Pacific Pattern Influenced by ENSO Amplitude in CMIP6 Simulations, J. Geophys. Res. Atmospheres, 128, e2022JD037905, https://doi.org/10.1029/2022JD037905, 2023.

Balaji, V., Maisonnave, E., Zadeh, N., Lawrence, B. N., Biercamp, J., Fladrich, U., Aloisio, G., Benson, R., Caubel, A., Durachta, J., Foujols, M.-A., Lister, G., Mocavero, S., Underwood, S., and Wright, G.: CPMIP: measurements of real computational performance of Earth system models in CMIP6, Geosci. Model Dev., 10, 19–34, https://doi.org/10.5194/gmd-10-19-2017, 2017.

Balhane, S., Driouech, F., Chafki, O., Manzanas, R., Chehbouni, A., and Moufouma-Okia, W.: Changes in mean and extreme temperature and precipitation events from different weighted multi-model ensembles over the northern half of Morocco, Clim. Dyn., 58, 389–404, https://doi.org/10.1007/s00382-021-05910-w, 2022.

Ban, N., Caillaud, C., Coppola, E., Pichelli, E., Sobolowski, S., Adinolfi, M., Ahrens, B., Alias, A., Anders, I., Bastin, S., Belušić, D., Berthou, S., Brisson, E., Cardoso, R. M., Chan, S. C., Christensen, O. B., Fernández, J., Fita, L., Frisius, T., Gašparac, G., Giorgi, F., Goergen, K., Haugen, J. E., Hodnebrog, Ø., Kartsios, S., Katragkou, E., Kendon, E. J., Keuler, K., Lavin-Gullon, A., Lenderink, G., Leutwyler, D., Lorenz, T., Maraun, D., Mercogliano, P., Milovac, J., Panitz, H.-J., Raffa, M., Remedio, A. R., Schär, C., Soares, P. M. M., Srnec, L., Steensen, B. M., Stocchi, P., Tölle, M. H., Truhetz, H., Vergara-Temprado, J., de Vries, H., Warrach-Sagi, K., Wulfmeyer, V., and Zander, M. J.: The first multi-model ensemble of regional climate simulations at kilometer-scale resolution, part I: evaluation of precipitation, Clim. Dyn., 57, 275–302, https://doi.org/10.1007/s00382-021-05708-w, 2021.

Baño-Medina, J., Manzanas, R., Cimadevilla, E., Fernández, J., González-Abad, J., Cofiño, A. S., and Gutiérrez, J. M.: Downscaling multi-model climate projection ensembles with deep learning (DeepESD): contribution to CORDEX EUR-44, Geosci. Model Dev., 15, 6747–6758, https://doi.org/10.5194/gmd-15-6747-2022, 2022.

Bellomo, K., Angeloni, M., Corti, S., and von Hardenberg, J.: Future climate change shaped by inter-model differences in Atlantic meridional overturning circulation response, Nat. Commun., 12, 3659, https://doi.org/10.1038/s41467-021-24015-w, 2021.

Beucler, T., Gentine, P., Yuval, J., Gupta, A., Peng, L., Lin, J., Yu, S., Rasp, S., Ahmed, F., O'Gorman, P. A., Neelin, J. D., Lutsko, N. J., and Pritchard, M.: Climate-invariant machine learning, Sci. Adv., 10, eadj7250, https://doi.org/10.1126/sciadv.adj7250, 2024.

Bevacqua, E., Zappa, G., Lehner, F., and Zscheischler, J.: Precipitation trends determine future occurrences of compound hot–dry events, Nat. Clim. Change, 12, 350–355, https://doi.org/10.1038/s41558-022-01309-5, 2022.

Bevacqua, E., Suarez-Gutierrez, L., Jézéquel, A., Lehner, F., Vrac, M., Yiou, P., and Zscheischler, J.: Advancing research on compound weather and climate events via large ensemble model simulations, Nat Commun, 14, 2145, https://doi.org/10.1038/s41467-023-37847-5, 2023.

Beven, K. and Freer, J.: Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, J. Hydrol., 249, 11–29, https://doi.org/10.1016/S0022-1694(01)00421-8, 2001.

Bhowmik, R. and Sankarasubramanian, A.: A performance-based multi-model combination approach to reduce uncertainty in seasonal temperature change projections, Int. J. Climatol., 41, https://doi.org/10.1002/joc.6870, 2020.

Bittner, M., Schmidt, H., Timmreck, C., and Sienz, F.: Using a large ensemble of simulations to assess the Northern Hemisphere stratospheric dynamical response to tropical volcanic eruptions and its uncertainty, Geophys. Res. Lett., 43, 9324–9332, https://doi.org/10.1002/2016GL070587, 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, J. Geophys. Res. Atmospheres, 125, e2019JD032321, https://doi.org/10.1029/2019JD032321, 2020.

Boé, J.: Interdependency in Multimodel Climate Projections: Component Replication and Result Similarity, Geophys. Res. Lett., 45, 2771–2779, https://doi.org/10.1002/2017GL076829, 2018.

Boysen, L. R.: BG - Global climate response to idealized deforestation in CMIP6 models, 2020.

Bracegirdle, T. J. and Stephenson, D. B.: Higher precision estimates of regional polar warming by ensemble regression of climate model projections, Clim. Dyn., 39, 2805–2821, https://doi.org/10.1007/s00382-012-1330-3, 2012.

Brenowitz, N. D., Henn, B., McGibbon, J., Clark, S. K., Kwa, A., Perkins, W. A., Watt-Meyer, O., and Bretherton, C. S.: Machine Learning Climate Model Dynamics: Offline versus Online Performance, https://doi.org/10.48550/ARXIV.2011.03081, 2020.

Breul, P., Ceppi, P., and Shepherd, T. G.: Revisiting the wintertime emergent constraint of the southern hemispheric midlatitude jet response to global warming, Weather Clim. Dyn., 4, 39–47, https://doi.org/10.5194/wcd-4-39-2023, 2023.

Brunner, L. and Sippel, S.: Identifying climate models based on their daily output using machine learning, Environ. Data Sci., 2, e22, https://doi.org/10.1017/eds.2023.23, 2023.

Brunner, L., Pendergrass, A. G., Lehner, F., Merrifield, A. L., Lorenz, R., and Knutti, R.: Reduced global warming from CMIP6 projections when weighting models by performance and independence, Earth Syst. Dyn., 11, 995–1012, https://doi.org/10.5194/esd-11-995-2020, 2020.

de Burgh-Day, C. O. and Leeuwenburg, T.: Machine learning for numerical weather and climate modelling: a review, Geosci. Model Dev., 16, 6433–6477, https://doi.org/10.5194/gmd-16-6433-2023, 2023.

Cesana, G. V. and Del Genio, A. D.: Observational constraint on cloud feedbacks suggests moderate climate sensitivity, Nat. Clim. Change, 11, 213–218, https://doi.org/10.1038/s41558-020-00970-y, 2021.

Cesana, G. V., Ackerman, A. S., Črnivec, N., Pincus, R., and Chepfer, H.: An observation-based method to assess tropical stratocumulus and shallow cumulus clouds and feedbacks in CMIP6 and CMIP5 models, Environ. Res. Commun., 5, 045001, https://doi.org/10.1088/2515-7620/acc78a, 2023.

Chandra, S., Kumar, P., Siingh, D., Roy, I., Victor, N. J., and Kamra, A. K.: Projection of lightning over South/South East Asia using CMIP5 models, Nat. Hazards, 114, 57–75, https://doi.org/10.1007/s11069-022-05379-8, 2022.

Chemke, R. and Polvani, L. M.: Opposite tropical circulation trends in climate models and in reanalyses, Nat. Geosci., 12, 528–532, https://doi.org/10.1038/s41561-019-0383-x, 2019.

Cinquini, L., Crichton, D., Mattmann, C., Harney, J., Shipman, G., Wang, F., Ananthakrishnan, R., Miller, N., Denvil, S., Morgan, M., Pobre, Z., Bell, G. M., Drach, B., Williams, D., Kershaw, P., Pascoe, S., Gonzalez, E., Fiore, S., and Schweitzer, R.: The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data, in: 2012 IEEE 8th International Conference on E-Science, 2012 IEEE 8th International Conference on E-Science, 1–10, https://doi.org/10.1109/eScience.2012.6404471, 2012.

Clyde, M., Çetinkaya-Rundel, M., Rundel, C., Banks, D., Chai, C., and Huang, L.: An Introduction to Bayesian Thinking, 2022.

Coles, S.: An Introduction to Statistical Modeling of Extreme Values, Springer London, London, https://doi.org/10.1007/978-1-4471-3675-0, 2001.

Cook, B. I., Mankin, J. S., Marvel, K., Williams, A. P., Smerdon, J. E., and Anchukaitis, K. J.: Twenty-First Century Drought Projections in the CMIP6 Forcing Scenarios, Earths Future, 8, e2019EF001461, https://doi.org/10.1029/2019EF001461, 2020.

Coppola, E., Sobolowski, S., Pichelli, E., Raffaele, F., Ahrens, B., Anders, I., Ban, N., Bastin, S., Belda, M., Belusic, D., Caldas-Alvarez, A., Cardoso, R. M., Davolio, S., Dobler, A., Fernandez, J., Fita, L., Fumiere, Q., Giorgi, F., Goergen, K., Güttler, I., Halenka, T., Heinzeller, D., Hodnebrog, Ø., Jacob, D., Kartsios, S., Katragkou, E., Kendon, E., Khodayar, S., Kunstmann, H., Knist, S., Lavín-Gullón, A., Lind, P., Lorenz, T., Maraun, D., Marelle, L., van Meijgaard, E., Milovac, J., Myhre, G., Panitz, H.-J., Piazza, M., Raffa, M., Raub, T., Rockel, B., Schär, C., Sieck, K., Soares, P. M. M., Somot, S., Srnec, L., Stocchi, P., Tölle, M. H., Truhetz, H., Vautard, R., de Vries, H., and Warrach-Sagi, K.: A first-of-its-kind multi-model convection permitting ensemble for investigating convective phenomena over Europe and the Mediterranean, Clim. Dyn., 55, 3–34, https://doi.org/10.1007/s00382-018-4521-8, 2020.

Coppola, E., Nogherotto, R., Ciarlo', J. M., Giorgi, F., Van Meijgaard, E., Kadygrov, N., Iles, C., Corre, L., Sandstad, M., Somot, S., Nabat, P., Vautard, R., Levavasseur, G., Schwingshackl, C., Sillmann, J., Kjellström, E., Nikulin, G., Aalbers, E., Lenderink, G., Christensen, O. B., Boberg, F., Sørland, S. L., Demory, M., Bülow, K., Teichmann, C., Warrach-Sagi, K., and Wulfmeyer, V.: Assessment of the European Climate Projections as Simulated by the Large EURO-CORDEX Regional and Global Climate Model Ensemble, J. Geophys. Res. Atmospheres, 126, e2019JD032356, https://doi.org/10.1029/2019JD032356, 2021.

Crane-Droesch, A.: Machine learning methods for crop yield prediction and climate change impact assessment in agriculture, Environ. Res. Lett., 13, 114003, https://doi.org/10.1088/1748-9326/aae159, 2018.

Crawford, J., Venkataraman, K., and Booth, J.: Developing climate model ensembles: A comparative case study, J. Hydrol., 568, 160–173, https://doi.org/10.1016/j.jhydrol.2018.10.054, 2019.

Črnivec, N., Cesana, G., and Pincus, R.: Evaluating the Representation of Tropical Stratocumulus and Shallow Cumulus Clouds As Well As Their Radiative Effects in CMIP6 Models Using Satellite Observations, J. Geophys. Res. Atmospheres, 128, e2022JD038437, https://doi.org/10.1029/2022JD038437, 2023.

https://doi.org/10.5194/egusphere-2025-4744
Preprint. Discussion started: 6 October 2025

Debeire, K., Bock, L., Nowack, P., Runge, J., and Eyring, V.: Constraining uncertainty in projected precipitation over land with causal discovery, Earth Syst. Dyn., 16, 607–630, https://doi.org/10.5194/esd-16-607-2025, 2025.

DelSole, T. and Tippett, M.: Statistical Methods for Climate Scientists, Cambridge University Press, Cambridge, https://doi.org/10.1017/9781108659055, 2022.

Deser, C.: "Certain Uncertainty: The Role of Internal Climate Variability in Projections of Regional Climate Change and Risk Management," Earths Future, 8, e2020EF001854, https://doi.org/10.1029/2020EF001854, 2020.

Deser, C., Knutti, R., Solomon, S., and Phillips, A. S.: Communication of the role of natural variability in future North American climate, Nat. Clim. Change, 2, 775–779, https://doi.org/10.1038/nclimate1562, 2012a.

Deser, C., Phillips, A., Bourdette, V., and Teng, H.: Uncertainty in climate change projections: the role of internal variability, Clim. Dyn., 38, 527–546, https://doi.org/10.1007/s00382-010-0977-x, 2012b.

Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., Fiore, A., Frankignoul, C., Fyfe, J. C., Horton, D. E., Kay, J. E., Knutti, R., Lovenduski, N. S., Marotzke, J., McKinnon, K. A., Minobe, S., Randerson, J., Screen, J. A., Simpson, I. R., and Ting, M.: Insights from Earth system model initial-condition large ensembles and future prospects, Nat. Clim. Change, 10, 277–286, https://doi.org/10.1038/s41558-020-0731-2, 2020.

Dey, A., Sahoo, D. P., Kumar, R., and Remesan, R.: A multimodel ensemble machine learning approach for CMIP6 climate model projections in an Indian River basin, Int. J. Climatol., 42, 9215–9236, https://doi.org/10.1002/joc.7813, 2022.

Di Luca, A., De Elía, R., and Laprise, R.: Challenges in the Quest for Added Value of Regional Climate Dynamical Downscaling, Curr. Clim. Change Rep., 1, 10–21, https://doi.org/10.1007/s40641-015-0003-9, 2015.

Di Luca, A., De Elía, R., Bador, M., and Argüeso, D.: Contribution of mean climate to hot temperature extremes for present and future climates, Weather Clim. Extrem., 28, 100255, https://doi.org/10.1016/j.wace.2020.100255, 2020a.

Di Luca, A., Pitman, A. J., and de Elía, R.: Decomposing Temperature Extremes Errors in CMIP5 and CMIP6 Models, Geophys. Res. Lett., 47, e2020GL088031, https://doi.org/10.1029/2020GL088031, 2020b.

Di Virgilio, G., Ji, F., Tam, E., Nishant, N., Evans, J. P., Thomas, C., Riley, M. L., Beyer, K., Grose, M. R., Narsey, S., and Delage, F.: Selecting CMIP6 GCMs for CORDEX Dynamical Downscaling: Model Performance, Independence, and Climate Change Signals, Earths Future, 10, e2021EF002625, https://doi.org/10.1029/2021EF002625, 2022.

Dirkes, C. A., Wing, A. A., Camargo, S. J., and Kim, D.: Process-Oriented Diagnosis of Tropical Cyclones in Reanalyses Using a Moist Static Energy Variance Budget, J. Clim., 36, 5293–5317, https://doi.org/10.1175/JCLI-D-22-0384.1, 2023.

Doblas-Reyes, F. J., Pavan, V., and Stephenson, D. B.: The skill of multi-model seasonal forecasts of the wintertime North Atlantic Oscillation, Clim. Dyn., 21, 501–514, https://doi.org/10.1007/s00382-003-0350-4, 2003.

Docquier, D., Vannitsem, S., Ragone, F., Wyser, K., and Liang, X. S.: Causal Links Between Arctic Sea Ice and Its Potential Drivers Based on the Rate of Information Transfer, Geophys. Res. Lett., 49, e2021GL095892, https://doi.org/10.1029/2021GL095892, 2022.

Docquier, D., Massonnet, F., Ragone, F., Sticker, A., Fichefet, T., and Vannitsem, S.: Drivers of summer Arctic sea-ice extent in CMIP6 large ensembles revealed by information flow, https://doi.org/10.21203/rs.3.rs-4434953/v1, 4 June 2024.

Dosio, A.: Projections of climate change indices of temperature and precipitation from an ensemble of bias-adjusted high-resolution EURO-CORDEX regional climate models, J. Geophys. Res. Atmospheres, 121, 5488–5511, https://doi.org/10.1002/2015JD024411, 2016.

Dosio, A.: Projection of temperature and heat waves for Africa with an ensemble of CORDEX Regional Climate Models, Clim. Dyn., 49, 493–519, https://doi.org/10.1007/s00382-016-3355-5, 2017.

Dueben, P. D. and Bauer, P.: Challenges and design choices for global weather and climate models based on machine learning, Geosci. Model Dev., 11, 3999–4009, https://doi.org/10.5194/gmd-11-3999-2018, 2018.

Eidhammer, T., Gettelman, A., Thayer-Calder, K., Watson-Parris, D., Elsaesser, G., Morrison, H., Van Lier-Walqui, M., Song, C., and McCoy, D.: An extensible perturbed parameter ensemble for the Community Atmosphere Model version 6, Geosci. Model Dev., 17, 7835–7853, https://doi.org/10.5194/gmd-17-7835-2024, 2024.

Eyring, V., Harris, N. R. P., Rex, M., Shepherd, T. G., Fahey, D. W., Amanatidis, G. T., Austin, J., Chipperfield, M. P., Dameris, M., Forster, P. M. D. F., Gettelman, A., Graf, H. F., Nagashima, T., Newman, P. A., Pawson, S., Prather, M. J., Pyle, J. A., Salawitch, R. J., Santer, B. D., and Waugh, D. W.: A Strategy for Process-Oriented Validation of Coupled Chemistry–Climate Models, Bull. Am. Meteorol. Soc., 86, 1117–1134, https://doi.org/10.1175/BAMS-86-8-1117, 2005.

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, https://doi.org/10.5194/gmd-9-1937-2016, 2016.

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., Collins, W. D., Gier, B. K., Hall, A. D., Hoffman, F. M., Hurtt, G. C., Jahn, A., Jones, C. D., Klein, S. A., Krasting, J. P., Kwiatkowski, L., Lorenz, R., Maloney, E., Meehl, G. A., Pendergrass, A. G., Pincus, R., Ruane, A. C., Russell, J. L., Sanderson, B. M., Santer, B. D., Sherwood, S. C., Simpson, I. R., Stouffer, R. J., and Williamson, M. S.: Taking climate model evaluation to the next level, Nat. Clim. Change, 9, 102–110, https://doi.org/10.1038/s41558-018-0355-y, 2019.

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, 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, https://doi.org/10.5194/gmd-13-3383-2020, 2020.

Eyring, V., Mishra, V., Griffith, G. P., Chen, L., Keenan, T., Turetsky, M. R., Brown, S., Jotzo, F., Moore, F. C., and Van Der Linden, S.: Reflections and projections on a decade of climate science, Nat. Clim. Change, 11, 279–285, https://doi.org/10.1038/s41558-021-01020-x, 2021.

Eyring, V., Collins, W. D., Gentine, P., Barnes, E. A., Barreiro, M., Beucler, T., Bocquet, M., Bretherton, C. S., Christensen, H. M., Dagon, K., Gagne, D. J., Hall, D., Hammerling, D., Hoyer, S., Iglesias-Suarez, F., Lopez-Gomez, I., McGraw, M. C., Meehl, G. A., Molina, M. J., Monteleoni, C., Mueller, J., Pritchard, M. S., 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, Nat. Clim. Change, 14, 916–928, https://doi.org/10.1038/s41558-024-02095-y, 2024.

Falkena, S. K. J. and von der Heydt, A. S.: Subpolar Gyre Variability in CMIP6 Models: Is there a Mechanism for Bistability?, https://doi.org/10.48550/ARXIV.2408.16541, 2024.

Flato, G. M.: Earth system models: an overview, WIREs Clim. Change, 2, 783–800, https://doi.org/10.1002/wcc.148, 2011.

Folberth, C., Baklanov, A., Balkovič, J., Skalský, R., Khabarov, N., and Obersteiner, M.: Spatio-temporal downscaling of gridded crop model yield estimates based on machine learning, Agric. For. Meteorol., 264, 1–15, https://doi.org/10.1016/j.agrformet.2018.09.021, 2019.

Frankignoul, C., Raillard, L., Ferster, B., and Kwon, Y.-O.: Arctic September Sea Ice Concentration Biases in CMIP6 Models and Their Relationships with Other Model Variables, J. Clim., 37, 4257–4274, https://doi.org/10.1175/JCLI-D-23-0452.1, 2024.

Fuhrer, O., Chadha, T., Hoefler, T., Kwasniewski, G., Lapillonne, X., Leutwyler, D., Lüthi, D., Osuna, C., Schär, C., Schulthess, T. C., and Vogt, H.: Near-global climate simulation at 1 km resolution: establishing a performance baseline on 4888 GPUs with COSMO 5.0, Geosci. Model Dev., 11, 1665–1681, https://doi.org/10.5194/gmd-11-1665-2018, 2018.

Galytska, E., Weigel, K., Handorf, D., Jaiser, R., Köhler, R., Runge, J., and Eyring, V.: Evaluating Causal Arctic-Midlatitude Teleconnections in CMIP6, J. Geophys. Res. Atmospheres, 128, e2022JD037978, https://doi.org/10.1029/2022JD037978, 2023.

Gates, W. L.: AN AMS CONTINUING SERIES: GLOBAL CHANGE--AMIP: The Atmospheric Model Intercomparison Project, Bull. Am. Meteorol. Soc., 73, 1962–1970, https://doi.org/10.1175/1520-0477(1992)073<1962:ATAMIP>2.0.CO;2, 1992.

Ge, F., Zhu, S., Luo, H., Zhi, X., and Wang, H.: Future changes in precipitation extremes over Southeast Asia: insights from CMIP6 multi-model ensemble, Environ. Res. Lett., 16, 024013, https://doi.org/10.1088/1748-9326/abd7ad, 2021.

Gebrechorkos, S., Leyland, J., Slater, L., Wortmann, M., Ashworth, P. J., Bennett, G. L., Boothroyd, R., Cloke, H., Delorme, P., Griffith, H., Hardy, R., Hawker, L., McLelland, S., Neal, J., Nicholas, A., Tatem, A. J., Vahidi, E., Parsons, D. R., and Darby, S. E.: A high-resolution daily global dataset of statistically downscaled CMIP6 models for climate impact analyses, Sci. Data, 10, 611, https://doi.org/10.1038/s41597-023-02528-x, 2023.

Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., and Yacalis, G.: Could Machine Learning Break the Convection Parameterization Deadlock?, Geophys. Res. Lett., 45, 5742–5751, https://doi.org/10.1029/2018GL078202, 2018.

Gergel, D. R., Malevich, S. B., McCusker, K. E., Tenezakis, E., Delgado, M. T., Fish, M. A., and Kopp, R. E.: Global Downscaled Projections for Climate Impacts Research (GDPCIR): preserving quantile trends for modeling future climate impacts, Geosci. Model Dev., 17, 191–227, https://doi.org/10.5194/gmd-17-191-2024, 2024.

Gettelman, A., Geer, A. J., Forbes, R. M., Carmichael, G. R., Feingold, G., Posselt, D. J., Stephens, G. L., Van Den Heever, S. C., Varble, A. C., and Zuidema, P.: The future of Earth system prediction: Advances in model-data fusion, Sci. Adv., 8, eabn3488, https://doi.org/10.1126/sciadv.abn3488, 2022.

Gillett, N. P., Shiogama, H., Funke, B., Hegerl, G., Knutti, R., Matthes, K., Santer, B. D., Stone, D., and Tebaldi, C.: The Detection and Attribution Model Intercomparison Project (DAMIP v1.0) contribution to CMIP6, Geosci. Model Dev., 9, 3685–3697, https://doi.org/10.5194/gmd-9-3685-2016, 2016.

Giorgi, F.: Thirty Years of Regional Climate Modeling: Where Are We and Where Are We Going next?, J. Geophys. Res.

Atmospheres, 124, 5696–5723, https://doi.org/10.1029/2018JD030094, 2019.

Giorgi, F. and Gutowski, W. J.: Regional Dynamical Downscaling and the CORDEX Initiative, Annu. Rev. Environ. Resour., 40, 467–490, https://doi.org/10.1146/annurev-environ-102014-021217, 2015.

Gleckler, P. J., Taylor, K. E., and Doutriaux, C.: Performance metrics for climate models, J. Geophys. Res. Atmospheres, 113, https://doi.org/10.1029/2007JD008972, 2008.

Glymour, C., Zhang, K., and Spirtes, P.: Review of Causal Discovery Methods Based on Graphical Models, Front. Genet., 10, 524, https://doi.org/10.3389/fgene.2019.00524, 2019.

Grose, M. R., Narsey, S., Trancoso, R., Mackallah, C., Delage, F., Dowdy, A., Di Virgilio, G., Watterson, I., Dobrohotoff, P., Rashid, H. A., Rauniyar, S., Henley, B., Thatcher, M., Syktus, J., Abramowitz, G., Evans, J. P., Su, C.-H., and Takbash, A.: A CMIP6-based multi-model downscaling ensemble to underpin climate change services in Australia, Clim. Serv., 30, 100368, https://doi.org/10.1016/j.cliser.2023.100368, 2023.

Grundner, A., Beucler, T., Gentine, P., Iglesias-Suarez, F., Giorgetta, M. A., and Eyring, V.: Deep Learning Based Cloud Cover Parameterization for ICON, J. Adv. Model. Earth Syst., 14, https://doi.org/10.1029/2021ms002959, 2022.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., and Chen, T.: Recent advances in convolutional neural networks, Pattern Recognit., 77, 354–377, https://doi.org/10.1016/j.patcog.2017.10.013, 2018.

Guendelman, I., Merlis, T. M., Cheng, K., Harris, L. M., Bretherton, C. S., Bolot, M., Zhou, L., Kaltenbaugh, A., Clark, S. K., and Fueglistaler, S.: The Precipitation Response to Warming and CO_2 Increase: A Comparison of a Global Storm Resolving Model and CMIP6 Models, Geophys. Res. Lett., 51, e2023GL107008, https://doi.org/10.1029/2023GL107008, 2024.

Gutowski Jr., W. J., Giorgi, F., Timbal, B., Frigon, A., Jacob, D., Kang, H.-S., Raghavan, K., Lee, B., Lennard, C., Nikulin, G., O'Rourke, E., Rixen, M., Solman, S., Stephenson, T., and Tangang, F.: WCRP COordinated Regional Downscaling EXperiment (CORDEX): a diagnostic MIP for CMIP6, Geosci. Model Dev., 9, 4087–4095, https://doi.org/10.5194/gmd-9-4087-2016, 2016.

Haarsma, R. J., Roberts, M. J., Vidale, P. L., Senior, C. A., Bellucci, A., Bao, Q., Chang, P., Corti, S., Fučkar, N. S., Guemas, V., von Hardenberg, J., Hazeleger, W., Kodama, C., Koenigk, T., Leung, L. R., Lu, J., Luo, J.-J., Mao, J., Mizielinski, M. S., Mizuta, R., Nobre, P., Satoh, M., Scoccimarro, E., Semmler, T., Small, J., and von Storch, J.-S.: High Resolution Model Intercomparison Project (HighResMIP v1.0) for CMIP6, Geosci. Model Dev., 9, 4185–4208, https://doi.org/10.5194/gmd-9-4185-2016, 2016.

Hall, A.: Projecting regional change, Science, 346, 1461–1462, https://doi.org/10.1126/science.aaa0629, 2014.

Hall, A., Cox, P., Huntingford, C., and Klein, S.: Progressing emergent constraints on future climate change, Nat. Clim. Change, 9, 269–278, https://doi.org/10.1038/s41558-019-0436-6, 2019.

Hamed, M. M., Nashwan, M. S., and Shahid, S.: A novel selection method of CMIP6 GCMs for robust climate projection, Int. J. Climatol., 42, 4258–4272, https://doi.org/10.1002/joc.7461, 2021.

Hawkins, E. and Sutton, R.: The Potential to Narrow Uncertainty in Regional Climate Predictions, Bull. Am. Meteorol. Soc., 90, 1095–1108, https://doi.org/10.1175/2009BAMS2607.1, 2009.

Henderson, S. A., Maloney, E. D., and Son, S.-W.: Madden–Julian Oscillation Pacific Teleconnections: The Impact of the Basic State and MJO Representation in General Circulation Models, J. Clim., 30, 4567–4587, https://doi.org/10.1175/JCLI-D-16-0789.1, 2017.

Herger, N., Abramowitz, G., Knutti, R., Angélil, O., Lehmann, K., and Sanderson, B. M.: Selecting a climate model subset to optimise key ensemble properties, Earth Syst. Dyn., 9, 135–151, https://doi.org/10.5194/esd-9-135-2018, 2018.

Hilburn, K. A., Ebert-Uphoff, I., and Miller, S. D.: Development and Interpretation of a Neural-Network-Based Synthetic Radar Reflectivity Estimator Using GOES-R Satellite Observations, https://doi.org/10.1175/JAMC-D-20-0084.1, 2020.

Hoaglin, D. C. and Kempthorne, P. J.: [Influential Observations, High Leverage Points, and Outliers in Linear Regression]: Comment, Stat. Sci., 1, https://doi.org/10.1214/ss/1177013627, 1986.

Hohenegger, C., Korn, P., Linardakis, L., Redler, R., Schnur, R., Adamidis, P., Bao, J., Bastin, S., Behravesh, M., Bergemann, M., Biercamp, J., Bockelmann, H., Brokopf, R., Brüggemann, N., Casaroli, L., Chegini, F., Datseris, G., Esch, M., George, G., Giorgetta, M., Gutjahr, O., Haak, H., Hanke, M., Ilyina, T., Jahns, T., Jungclaus, J., Kern, M., Klocke, D., Kluft, L., Kölling, T., Kornblueh, L., Kosukhin, S., Kroll, C., Lee, J., Mauritsen, T., Mehlmann, C., Mieslinger, T., Naumann, A. K., Paccini, L., Peinado, A., Praturi, D. S., Putrasahan, D., Rast, S., Riddick, T., Roeber, N., Schmidt, H., Schulzweida, U., Schütte, F., Segura, H., Shevchenko, R., Singh, V., Specht, M., Stephan, C. C., Von Storch, J.-S., Vogel, R., Wengel, C., Winkler, M., Ziemen, F., Marotzke, J., and Stevens, B.: ICON-Sapphire: simulating the components of the Earth system and their interactions at kilometer and subkilometer scales, Geosci. Model Dev., 16, 779–811, https://doi.org/10.5194/gmd-16-779-2023, 2023.

Hong, T., Wu, J., Kang, X., Yuan, M., and Duan, L.: Impacts of Different Land Use Scenarios on Future Global and Regional Climate Extremes, Atmosphere, 13, 995, https://doi.org/10.3390/atmos13060995, 2022.

Huffman, G. J., Bolvin, D. T., Nelkin, E. J., Wolff, D. B., Adler, R. F., Gu, G., Hong, Y., Bowman, K. P., and Stocker, E. F.: The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-Global, Multiyear, Combined-Sensor Precipitation Estimates at Fine Scales, J. Hydrometeorol., 8, 38–55, https://doi.org/10.1175/JHM560.1, 2007.

Iglesias-Suarez, F., Gentine, P., Solino-Fernandez, B., Beucler, T., Pritchard, M., Runge, J., and Eyring, V.: Causally-Informed Deep Learning to Improve Climate Models and Projections, J. Geophys. Res. Atmospheres, 129, e2023JD039202, https://doi.org/10.1029/2023JD039202, 2024.

Iles, C. E., Vautard, R., Strachan, J., Joussaume, S., Eggen, B. R., and Hewitt, C. D.: The benefits of increasing resolution in global and regional climate simulations for European climate extremes, Geosci. Model Dev., 13, 5583–5607, https://doi.org/10.5194/gmd-13-5583-2020, 2020.

Intergovernmental Panel On Climate Change: Climate Change 2001– The Scientific Basis: Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, 2001.

Intergovernmental Panel On Climate Change: Climate Change 2007 – The Physical Science Basis: Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change., Cambridge University Press, Cambridge, 2007.

Intergovernmental Panel On Climate Change (Ed.): Climate Change 2013 – The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, 1st ed., Cambridge University Press, https://doi.org/10.1017/CBO9781107415324, 2014.

Intergovernmental Panel on Climate Change (IPCC): Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, https://doi.org/10.1017/9781009157896, 2021.

Ivanova, D. P., Gleckler, P. J., Taylor, K. E., Durack, P. J., and Marvel, K. D.: Moving beyond the Total Sea Ice Extent in Gauging Model Biases, J. Clim., 29, 8965–8987, https://doi.org/10.1175/JCLI-D-16-0026.1, 2016.

Jose, D. M., Vincent, A. M., and Dwarakish, G. S.: Improving multiple model ensemble predictions of daily precipitation and temperature through machine learning techniques, Sci. Rep., 12, 4678, https://doi.org/10.1038/s41598-022-08786-w, 2022.

Jourdain, N. C., Gupta, A. S., Taschetto, A. S., Ummenhofer, C. C., Moise, A. F., and Ashok, K.: The Indo-Australian monsoon and its relationship to ENSO and IOD in reanalysis data and the CMIP3/CMIP5 simulations, Clim. Dyn., 41, 3073–3102, https://doi.org/10.1007/s00382-013-1676-1, 2013.

Joussaume, S. and Budich, R.: The Infrastructure Project of the European Network for Earth System Modelling: IS-ENES, in: Earth System Modelling - Volume 1, Springer Berlin Heidelberg, Berlin, Heidelberg, 5–9, https://doi.org/10.1007/978-3-642-36597-3_2, 2013.

Jun, M., Knutti, R., and Nychka, D. W.: Spatial Analysis to Quantify Numerical Model Bias and Dependence: How Many Climate Models Are There?, J. Am. Stat. Assoc., 103, 934–947, https://doi.org/10.1198/016214507000001265, 2008.

Jung, J., Han, H., Kim, K., and Kim, H. S.: Machine Learning-Based Small Hydropower Potential Prediction under Climate Change, Energies, 14, https://doi.org/10.3390/en14123643, 2021.

Kaltenborn, J., Lange, C. E. E., Ramesh, V., Brouillard, P., Gurwicz, Y., Nagda, C., Runge, J., Nowack, P., and Rolnick, D.: ClimateSet: A Large-Scale Climate Model Dataset for Machine Learning, https://doi.org/10.48550/ARXIV.2311.03721, 2023.

Karmouche, S., Galytska, E., Runge, J., Meehl, G. A., Phillips, A. S., Weigel, K., and Eyring, V.: Regime-oriented causal model evaluation of Atlantic–Pacific teleconnections in CMIP6, Earth Syst. Dyn., 14, 309–344, https://doi.org/10.5194/esd-14-309-2023, 2023.

Karmouche, S., Galytska, E., Meehl, G. A., Runge, J., Weigel, K., and Eyring, V.: Changing effects of external forcing on Atlantic–Pacific interactions, Earth Syst. Dyn., 15, 689–715, https://doi.org/10.5194/esd-15-689-2024, 2024.

Karpechko, A. Yu., Maraun, D., and Eyring, V.: Improving Antarctic Total Ozone Projections by a Process-Oriented Multiple Diagnostic Ensemble Regression, J. Atmospheric Sci., 70, 3959–3976, https://doi.org/10.1175/JAS-D-13-071.1, 2013.

Katzenberger, A., Schewe, J., Pongratz, J., and Levermann, A.: Robust increase of Indian monsoon rainfall and its variability under future warming in CMIP6 models, Earth Syst. Dyn., 12, 367–386, https://doi.org/10.5194/esd-12-367-2021, 2021.

Katzenberger, A., Petri, S., Feulner, G., and Levermann, A.: Monsoon Planet: Bimodal Rainfall Distribution due to Barrier Structure in Pressure Fields, J. Clim., 37, 1295–1315, https://doi.org/10.1175/JCLI-D-23-0055.1, 2024.

Kaufman, Z., Feldl, N., and Beaulieu, C.: Warm Arctic–Cold Eurasia pattern driven by atmospheric blocking in models and observations, Environ. Res. Clim., 3, 015006, https://doi.org/10.1088/2752-5295/ad1f40, 2024.

Keenan, T. F., Luo, X., Stocker, B. D., De Kauwe, M. G., Medlyn, B. E., Prentice, I. C., Smith, N. G., Terrer, C., Wang, H., Zhang, Y., and Zhou, S.: A constraint on historic growth in global photosynthesis due to rising CO2, Nat. Clim. Change, 13, 1376–1381, https://doi.org/10.1038/s41558-023-01867-2, 2023.

Kim, D., Moon, Y., Camargo, S. J., Wing, A. A., Sobel, A. H., Murakami, H., Vecchi, G. A., Zhao, M., and Page, E.: Process-Oriented Diagnosis of Tropical Cyclones in High-Resolution GCMs, J. Clim., 31, 1685–1702, https://doi.org/10.1175/JCLI-D-17-0269.1, 2018.

Kim, Y.-H., Min, S.-K., Zhang, X., Sillmann, J., and Sandstad, M.: Evaluation of the CMIP6 multi-model ensemble for climate extreme indices, Weather Clim. Extrem., 29, 100269, https://doi.org/10.1016/j.wace.2020.100269, 2020.

Knutson, T. R., Sirutis, J. J., Vecchi, G. A., Garner, S., Zhao, M., Kim, H.-S., Bender, M., Tuleya, R. E., Held, I. M., and Villarini, G.: Dynamical Downscaling Projections of Twenty-First-Century Atlantic Hurricane Activity: CMIP3 and CMIP5 Model-Based Scenarios, J. Clim., 26, 6591–6617, https://doi.org/10.1175/JCLI-D-12-00539.1, 2013.

Knutti, R.: Should We Believe Model Predictions of Future Climate Change?, Philos. Trans. Math. Phys. Eng. Sci., 366, 4647–4664, 2008.

Knutti, R.: The end of model democracy?: An editorial comment, Clim. Change, 102, 395–404, https://doi.org/10.1007/s10584-010-9800-2, 2010.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G. A.: Challenges in Combining Projections from Multiple Climate Models, https://doi.org/10.1175/2009JCLI3361.1, 2010a.

Knutti, R., Abramowitz, G., Collins, M., Eyring, V., Gleckler, P. J., and Hewitson, B.: Good Practice Guidance Paper on Assessing and Combining Multi Model Climate Projections, in: Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Assessing and Combining Multi Model Climate Projections [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, and P.M. Midgley (eds.)], 2010b.

Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., and Eyring, V.: A climate model projection weighting scheme accounting for performance and interdependence, Geophys. Res. Lett., 44, 1909–1918, https://doi.org/10.1002/2016GL072012, 2017.

Knutti, R., Baumberger, C., and Hirsch Hadorn, G.: Uncertainty Quantification Using Multiple Models—Prospects and Challenges, in: Computer Simulation Validation: Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives, edited by: Beisbart, C. and Saam, N. J., Springer International Publishing, Cham, 835–855, https://doi.org/10.1007/978-3-319-70766-2_34, 2019.

Kretschmer, M., Zappa, G., and Shepherd, T. G.: The role of Barents–Kara sea ice loss in projected polar vortex changes, Weather Clim. Dyn., 1, 715–730, https://doi.org/10.5194/wcd-1-715-2020, 2020.

Krishnamurti, T. N., Kishtawal, C. M., LaRow, T. E., Bachiochi, D. R., Zhang, Z., Williford, C. E., Gadgil, S., and Surendran, S.: Improved Weather and Seasonal Climate Forecasts from Multimodel Superensemble, Science, 285, 1548–1550, https://doi.org/10.1126/science.285.5433.1548, 1999.

Kuma, P., Bender, F. A.-M., and Jönsson, A. R.: Climate Model Code Genealogy and Its Relation to Climate Feedbacks and Sensitivity, J. Adv. Model. Earth Syst., 15, e2022MS003588, https://doi.org/10.1029/2022MS003588, 2023.

Kunimitsu, T., Baldissera Pacchetti, M., Ciullo, A., Sillmann, J., Shepherd, T. G., Taner, M. Ü., and van den Hurk, B.:

Representing storylines with causal networks to support decision making: Framework and example, Clim. Risk Manag., 40, 100496, https://doi.org/10.1016/j.crm.2023.100496, 2023.

Kyono, T., Zhang, Y., and van der Schaar, M.: CASTLE: Regularization via Auxiliary Causal Graph Discovery, in: Advances in Neural Information Processing Systems, 1501–1512, 2020.

Labe, Z. M. and Barnes, E. A.: Comparison of Climate Model Large Ensembles With Observations in the Arctic Using Simple Neural Networks, Earth Space Sci., 9, e2022EA002348, https://doi.org/10.1029/2022EA002348, 2022.

Labe, Z. M., Johnson, N. C., and Delworth, T. L.: Changes in United States Summer Temperatures Revealed by Explainable Neural Networks, Earths Future, 12, e2023EF003981, https://doi.org/10.1029/2023EF003981, 2024.

Lambert, S. J. and Boer, G. J.: CMIP1 evaluation and intercomparison of coupled climate models, Clim. Dyn., 17, 83–106, https://doi.org/10.1007/PL00013736, 2001.

LeCun, Y., Bengio, Y., and Hinton, G.: Deep learning, Nature, 521, 436–444, https://doi.org/10.1038/nature14539, 2015.

Lehner, F. and Deser, C.: Origin, importance, and predictive limits of internal climate variability, Environ. Res. Clim., 2, 023001, 2023.

Lehner, F., Coats, S., Stocker, T. F., Pendergrass, A. G., Sanderson, B. M., Raible, C. C., and Smerdon, J. E.: Projected drought risk in 1.5°C and 2°C warmer climates, Geophys. Res. Lett., 44, 7419–7428, https://doi.org/10.1002/2017GL074117, 2017.

Lehner, F., Deser, C., Simpson, I. R., and Terray, L.: Attributing the U.S. Southwest's recent shift into drier conditions, Geophys Res Lett, 45, 6251–61, https://doi.org/10.1029/2018GL078312, 2018.

Lehner, F., Deser, C., Maher, N., Marotzke, J., Fischer, E., Brunner, L., Knutti, R., and Hawkins, E.: Partitioning climate projection uncertainty with multiple large ensembles and CMIP5/6, Earth Syst Dyn, 11, 491–508, https://doi.org/10.5194/esd-11-491-2020, 2020.

Leng, G. and Hall, J. W.: Predicting spatial and temporal variability in crop yields: an inter-comparison of machine learning, regression and process-based models, Environ. Res. Lett., 15, 044027, https://doi.org/10.1088/1748-9326/ab7b24, 2020.

Li, G. and Xie, S.-P.: Tropical Biases in CMIP5 Multimodel Ensemble: The Excessive Equatorial Pacific Cold Tongue and Double ITCZ Problems*, J. Clim., 27, 1765–1780, https://doi.org/10.1175/JCLI-D-13-00337.1, 2014.

Li, T., Jiang, Z., Le Treut, H., Li, L., Zhao, L., and Ge, L.: Machine learning to optimize climate projection over China with multi-model ensemble simulations, Environ. Res. Lett., 16, 094028, 2021.

Li, Y., Wu, J., Luo, J.-J., and Yang, Y. M.: Evaluating the Eastward Propagation of the MJO in CMIP5 and CMIP6 Models Based on a Variety of Diagnostics, J. Clim., 35, 1719–1743, https://doi.org/10.1175/JCLI-D-21-0378.1, 2022.

Liang-Liang, L., Jian, L., and Ru-Cong, Y.: Evaluation of CMIP6 HighResMIP models in simulating precipitation over Central Asia, Adv. Clim. Change Res., 13, 1–13, https://doi.org/10.1016/j.accre.2021.09.009, 2022.

Lin, X., Zhen, H.-L., Li, Z., Zhang, Q.-F., and Kwong, S.: Pareto Multi-Task Learning, in: Advances in Neural Information Processing Systems, 2019.

Liu, Y., Fan, K., Chen, L., Ren, H.-L., Wu, Y., and Liu, C.: An operational statistical downscaling prediction model of the winter monthly temperature over China based on a multi-model ensemble, Atmospheric Res., 249, 105262, https://doi.org/10.1016/j.atmosres.2020.105262, 2021.

Lovenduski, N. S., McKinley, G. A., Fay, A. R., Lindsay, K., and Long, M. C.: Partitioning uncertainty in ocean carbon uptake projections: internal variability, emission scenario, and model structure, Glob Biogeochem Cycles, 30, 1276–87, https://doi.org/10.1002/2016GB005426, 2016.

Lu, D. and Ricciuto, D.: Efficient surrogate modeling methods for large-scale Earth system models based on machine-learning techniques, Geosci. Model Dev., 12, 1791–1807, https://doi.org/10.5194/gmd-12-1791-2019, 2019.

Luo, Y., Peng, J., and Ma, J.: When causal inference meets deep learning, Nat. Mach. Intell., 2, 426–427, https://doi.org/10.1038/s42256-020-0218-x, 2020.

Maher, N., Milinski, S., and Ludwig, R.: Large ensemble climate model simulations: introduction, overview, and future prospects for utilising multiple types of large ensemble, Earth Syst. Dyn., 12, 401–418, https://doi.org/10.5194/esd-12-401-2021, 2021.

Maher, N., Phillips, A. S., Deser, C., Wills, R. C. J., Lehner, F., Fasullo, J., Caron, J. M., Brunner, L., and Beyerle, U.: The updated Multi-Model Large Ensemble Archive and the Climate Variability Diagnostics Package: New tools for the study of climate variability and change, https://doi.org/10.5194/egusphere-2024-3684, 19 December 2024.

Maloney, E. D., Gettelman, A., Ming, Y., Neelin, J. D., Barrie, D., Mariotti, A., Chen, C.-C., Coleman, D. R. B., Kuo, Y.-H., Singh, B., Annamalai, H., Berg, A., Booth, J. F., Camargo, S. J., Dai, A., Gonzalez, A., Hafner, J., Jiang, X., Jing, X., Kim, D., Kumar, A., Moon, Y., Naud, C. M., Sobel, A. H., Suzuki, K., Wang, F., Wang, J., Wing, A. A., Xu, X., and Zhao, M.: Process-Oriented Evaluation of Climate and Weather Forecasting Models, Bull. Am. Meteorol. Soc., 100, 1665–1686, https://doi.org/10.1175/BAMS-D-18-0042.1, 2019.

Mankin, J. S. and Diffenbaugh, N. S.: Influence of temperature and precipitation variability on near-term snow trends, Clim. Dyn., 45, 1099–1116, https://doi.org/10.1007/s00382-014-2357-4, 2015.

Maraun, D.: Bias Correcting Climate Change Simulations - a Critical Review, Curr. Clim. Change Rep., 2, 211–220, https://doi.org/10.1007/s40641-016-0050-x, 2016.

Maraun, D., Shepherd, T. G., Widmann, M., Zappa, G., Walton, D., Gutiérrez, J. M., Hagemann, S., Richter, I., Soares, P. M. M., Hall, A., and Mearns, L. O.: Towards process-informed bias correction of climate change simulations, Nat. Clim. Change, 7, 764–773, https://doi.org/10.1038/nclimate3418, 2017.

Marotzke, J.: Quantifying the irreducible uncertainty in near-term climate projections, WIREs Clim. Change, 10, e563, https://doi.org/10.1002/wcc.563, 2019.

Masson, D. and Knutti, R.: Climate model genealogy, Geophys. Res. Lett., 38, https://doi.org/10.1029/2011GL046864, 2011.

Mauritsen, T., Bader, J., Becker, T., Behrens, J., Bittner, M., Brokopf, R., Brovkin, V., Claussen, M., Crueger, T., Esch, M., Fast, I., Fiedler, S., Fläschner, D., Gayler, V., Giorgetta, M., Goll, D. S., Haak, H., Hagemann, S., Hedemann, C., Hohenegger, C., Ilyina, T., Jahns, T., Jimenéz-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., Möbis, B., Müller, W. A., Nabel, J. E. M. S., Nam, C. C. W., Notz, D., Nyawira, S., Paulsen, H., Peters, K., Pincus, R., Pohlmann, H.,

https://doi.org/10.5194/egusphere-2025-4744
Preprint. Discussion started: 6 October 2025

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., Tian, F., Voigt, A., Vrese, P., Wieners, K., 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 CO_2, J. Adv. Model. Earth Syst., 11, 998–1038, https://doi.org/10.1029/2018MS001400, 2019.

McKenna, C. M. and Maycock, A. C.: Sources of Uncertainty in Multimodel Large Ensemble Projections of the Winter North Atlantic Oscillation, Geophys. Res. Lett., 48, e2021GL093258, https://doi.org/10.1029/2021GL093258, 2021.

Meehl, G. A., Boer, G. J., Covey, C., Latif, M., and Stouffer, R. J.: The Coupled Model Intercomparison Project (CMIP), Bull. Am. Meteorol. Soc., 81, 313–318, 2000.

Mendlik, T. and Gobiet, A.: Selecting climate simulations for impact studies based on multivariate patterns of climate change, Clim. Change, 135, 381–393, https://doi.org/10.1007/s10584-015-1582-0, 2016.

Merlis, T. M., Cheng, K.-Y., Guendelman, I., Harris, L., Bretherton, C. S., Bolot, M., Zhou, L., Kaltenbaugh, A., Clark, S. K., Vecchi, G. A., and Fueglistaler, S.: Climate sensitivity and relative humidity changes in global storm-resolving model simulations of climate change, Sci. Adv., 10, eadn5217, https://doi.org/10.1126/sciadv.adn5217, 2024.

Merrifield, A. L., Brunner, L., Lorenz, R., Medhaug, I., and Knutti, R.: An investigation of weighting schemes suitable for incorporating large ensembles into multi-model ensembles, Earth Syst. Dyn., 11, 807–834, https://doi.org/10.5194/esd-11-807-2020, 2020.

Merrifield, A. L., Brunner, L., Lorenz, R., Humphrey, V., and Knutti, R.: Climate model Selection by Independence, Performance, and Spread (ClimSIPS v1.0.1) for regional applications, Geosci. Model Dev., 16, 4715–4747, https://doi.org/10.5194/gmd-16-4715-2023, 2023.

Milinski, S., Maher, N., and Olonscheck, D.: How large does a large ensemble need to be?, Earth Syst. Dyn., 11, 885–901, https://doi.org/10.5194/esd-11-885-2020, 2020.

Moon, Y., Kim, D., Camargo, S. J., Wing, A. A., Sobel, A. H., Murakami, H., Reed, K. A., Scoccimarro, E., Vecchi, G. A., Wehner, M. F., Zarzycki, C. M., and Zhao, M.: Azimuthally Averaged Wind and Thermodynamic Structures of Tropical Cyclones in Global Climate Models and Their Sensitivity to Horizontal Resolution, J. Clim., 33, 1575–1595, https://doi.org/10.1175/JCLI-D-19-0172.1, 2020.

Moradian, S., Torabi Haghighi, A., Asadi, M., and Mirbagheri, S. A.: Future Changes in Precipitation Over Northern Europe Based on a Multi-model Ensemble from CMIP6: Focus on Tana River Basin, Water Resour. Manag., 37, 2447–2463, https://doi.org/10.1007/s11269-022-03272-4, 2023.

Mudryk, L., Santolaria-Otín, M., Krinner, G., Ménégoz, M., Derksen, C., Brutel-Vuilmet, C., Brady, M., and Essery, R.: Historical Northern Hemisphere snow cover trends and projected changes in the CMIP6 multi-model ensemble, The Cryosphere, 14, 2495–2514, https://doi.org/10.5194/tc-14-2495-2020, 2020.

Nam, C., Bony, S., Dufresne, J. -L., and Chepfer, H.: The ‘too few, too bright’ tropical low-cloud problem in CMIP5 models, Geophys. Res. Lett., 39, 2012GL053421, https://doi.org/10.1029/2012GL053421, 2012.

The Climate Data Guide: Regridding Overview: https://climatedataguide.ucar.edu/climate-tools/regridding-overview.

Neelin, J. D., Krasting, J. P., Radhakrishnan, A., Liptak, J., Jackson, T., Ming, Y., Dong, W., Gettelman, A., Coleman, D. R., Maloney, E. D., Wing, A. A., Kuo, Y.-H., Ahmed, F., Ullrich, P., Bitz, C. M., Neale, R. B., Ordonez, A., and Maroon, E. A.: Process-Oriented Diagnostics: Principles, Practice, Community Development, and Common Standards, Bull. Am. Meteorol. Soc., 104, E1452–E1468, https://doi.org/10.1175/BAMS-D-21-0268.1, 2023.

Nijsse, F. J. M. M., Cox, P. M., and Williamson, M. S.: Emergent constraints on transient climate response (TCR) and equilibrium climate sensitivity (ECS) from historical warming in CMIP5 and CMIP6 models, Earth Syst. Dyn., 11, 737–750, https://doi.org/10.5194/esd-11-737-2020, 2020.

Nolan, P. and Flanagan, J.: High-resolution climate projections for Ireland - a multi-model ensemble approach: 2014-CCRP-MS.23, Online version., Environmental Protection Agency, Johnstown Castle, Co. Wexford, Ireland, 1 pp., 2020.

Notz, D. and Community, S.: Arctic Sea Ice in CMIP6, Geophys. Res. Lett., 47, e2019GL086749, https://doi.org/10.1029/2019GL086749, 2020.

Notz, D., Jahn, A., Holland, M., Hunke, E., Massonnet, F., Stroeve, J., Tremblay, B., and Vancoppenolle, M.: The CMIP6 Sea-Ice Model Intercomparison Project (SIMIP): understanding sea ice through climate-model simulations, Geosci. Model Dev., 9, 3427–3446, https://doi.org/10.5194/gmd-9-3427-2016, 2016.

Nowack, P., Runge, J., Eyring, V., and Haigh, J. D.: Causal networks for climate model evaluation and constrained projections, Nat. Commun., 11, 1415, https://doi.org/10.1038/s41467-020-15195-y, 2020.

Nwokolo, S. C., Obiwulu, A. U., and Ogbulezie, J. C.: Machine learning and analytical model hybridization to assess the impact of climate change on solar PV energy production, Phys. Chem. Earth Parts ABC, 130, 103389, https://doi.org/10.1016/j.pce.2023.103389, 2023.

Olonscheck, D. and Notz, D.: Consistently Estimating Internal Climate Variability from Climate Model Simulations, J. Clim., 30, 9555–9573, https://doi.org/10.1175/JCLI-D-16-0428.1, 2017.

O’Neill, B. C., Kriegler, E., Riahi, K., Ebi, K. L., Hallegatte, S., Carter, T. R., Mathur, R., and van Vuuren, D. P.: A new scenario framework for climate change research: the concept of shared socioeconomic pathways, Clim. Change, 122, 387–400, https://doi.org/10.1007/s10584-013-0905-2, 2014.

O’Neill, B. C., Kriegler, E., Ebi, K. L., Kemp-Benedict, E., Riahi, K., Rothman, D. S., Van Ruijven, B. J., Van Vuuren, D. P., Birkmann, J., Kok, K., Levy, M., and Solecki, W.: The roads ahead: Narratives for shared socioeconomic pathways describing world futures in the 21st century, Glob. Environ. Change, 42, 169–180, https://doi.org/10.1016/j.gloenvcha.2015.01.004, 2017.

Oueslati, B. and Bellon, G.: The double ITCZ bias in CMIP5 models: interaction between SST, large-scale circulation and precipitation, Clim. Dyn., 44, 585–607, https://doi.org/10.1007/s00382-015-2468-6, 2015.

Oxarart, A. and Parker, L.: Global Climate Models and Land Management, USDA California Climate Hub, 2024.

Palmer, T. E., McSweeney, C. F., Booth, B. B. B., Priestley, M. D. K., Davini, P., Brunner, L., Borchert, L., and Menary, M. B.: Performance-based sub-selection of CMIP6 models for impact assessments in Europe, Earth Syst. Dyn., 14, 457–483, https://doi.org/10.5194/esd-14-457-2023, 2023.

Palmer, T. n, Doblas-Reyes, F. j, Hagedorn, R., and Weisheimer, A.: Probabilistic prediction of climate using multi-model ensembles: from basics to applications, Philos. Trans. R. Soc. B Biol. Sci., 360, 1991–1998, https://doi.org/10.1098/rstb.2005.1750, 2005.

Pennell, C. and Reichler, T.: On the Effective Number of Climate Models, https://doi.org/10.1175/2010JCLI3814.1, 2011.

Phillips, A., Deser, C., Fasullo, J., Schneider, D. P., and Simpson, I. R.: Assessing Climate Variability and Change in Model Large Ensembles: A User's Guide to the "Climate Variability Diagnostics Package for Large Ensembles," https://doi.org/10.5065/H7C7-F961, 2020.

Phillips, A. S., Deser, C., and Fasullo, J.: Evaluating Modes of Variability in Climate Models, Eos Trans. Am. Geophys. Union, 95, 453–455, https://doi.org/10.1002/2014EO490002, 2014.

Phillips, T. J. and Gleckler, P. J.: Evaluation of continental precipitation in 20th century climate simulations: The utility of multimodel statistics, Water Resour. Res., 42, 2005WR004313, https://doi.org/10.1029/2005WR004313, 2006.

Pichelli, E., Coppola, E., Sobolowski, S., Ban, N., Giorgi, F., Stocchi, P., Alias, A., Belušić, D., Berthou, S., Caillaud, C., Cardoso, R. M., Chan, S., Christensen, O. B., Dobler, A., de Vries, H., Goergen, K., Kendon, E. J., Keuler, K., Lenderink, G., Lorenz, T., Mishra, A. N., Panitz, H.-J., Schär, C., Soares, P. M. M., Truhetz, H., and Vergara-Temprado, J.: The first multi-model ensemble of regional climate simulations at kilometer-scale resolution part 2: historical and future simulations of precipitation, Clim. Dyn., 56, 3581–3602, https://doi.org/10.1007/s00382-021-05657-4, 2021.

Pincus, R., Barker, H. W., and Morcrette, J.: A fast, flexible, approximate technique for computing radiative transfer in inhomogeneous cloud fields, J. Geophys. Res. Atmospheres, 108, 2002JD003322, https://doi.org/10.1029/2002JD003322, 2003.

Pincus, R., Batstone, C. P., Hofmann, R. J. P., Taylor, K. E., and Glecker, P. J.: Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models, J. Geophys. Res. Atmospheres, 113, https://doi.org/10.1029/2007JD009334, 2008.

Pincus, R., Forster, P. M., and Stevens, B.: The Radiative Forcing Model Intercomparison Project (RFMIP): experimental protocol for CMIP6, Geosci. Model Dev., 9, 3447–3460, https://doi.org/10.5194/gmd-9-3447-2016, 2016.

Polkova*, I., Afargan-Gerstman, H., Domeisen, D. I. V., King, M. P., Ruggieri, P., Athanasiadis, P., Dobrynin, M., Aarnes, Ø., Kretschmer, M., and Baehr, J.: Predictors and prediction skill for marine cold-air outbreaks over the Barents Sea, Q. J. R. Meteorol. Soc., 147, 2638–2656, https://doi.org/10.1002/qj.4038, 2021.

Quesada, B., Arneth, A., and de Noblet-Ducoudré, N.: Atmospheric, radiative, and hydrologic effects of future land use and land cover changes: A global and multimodel climate picture, J. Geophys. Res. Atmospheres, 122, 5113–5131, https://doi.org/10.1002/2016JD025448, 2017.

Rackow, T., Pedruzo-Bagazgoitia, X., Becker, T., Milinski, S., Sandu, I., Aguridan, R., Bechtold, P., Beyer, S., Bidlot, J., Boussetta, S., Deconinck, W., Diamantakis, M., Dueben, P., Dutra, E., Forbes, R., Ghosh, R., Goessling, H. F., Hadade, I., Hegewald, J., Jung, T., Keeley, S., Kluft, L., Koldunov, N., Koldunov, A., Kölling, T., Kousal, J., Kühnlein, C., Maciel, P., Mogensen, K., Quintino, T., Polichtchouk, I., Reuter, B., Sármány, D., Scholz, P., Sidorenko, D., Streffing, J., Sützl, B., Takasuka, D., Tietsche, S., Valentini, M., Vannière, B., Wedi, N., Zampieri, L., and Ziemen, F.: Multi-year simulations at kilometre scale with the Integrated Forecasting System coupled to FESOM2.5 and NEMOv3.4, Geosci. Model Dev., 18, 33–69, https://doi.org/10.5194/gmd-18-33-2025, 2025.

Rader, J. K., Barnes, E. A., Ebert-Uphoff, I., and Anderson, C.: Detection of Forced Change Within Combined Climate Fields Using Explainable Neural Networks, J. Adv. Model. Earth Syst., 14, e2021MS002941, https://doi.org/10.1029/2021MS002941, 2022.

Räisänen, J.: Objective comparison of patterns of CO 2 induced climate change in coupled GCM experiments, Clim. Dyn., 13, 197–211, https://doi.org/10.1007/s003820050160, 1997.

Räisänen, J. and Palmer, T. N.: A Probability and Decision-Model Analysis of a Multimodel Ensemble of Climate Change Simulations, J. Clim., 14, 3212–3226, https://doi.org/10.1175/1520-0442(2001)014<3212:APADMA>2.0.CO;2, 2001.

Rampal, N., Gibson, P. B., Sood, A., Stuart, S., Fauchereau, N. C., Brandolino, C., Noll, B., and Meyers, T.: High-resolution downscaling with interpretable deep learning: Rainfall extremes over New Zealand, Weather Clim. Extrem., 38, 100525, https://doi.org/10.1016/j.wace.2022.100525, 2022.

Rampal, N., Hobeichi, S., Gibson, P. B., Baño-Medina, J., Abramowitz, G., Beucler, T., González-Abad, J., Chapman, W., Harder, P., and Gutiérrez, J. M.: Enhancing Regional Climate Downscaling through Advances in Machine Learning, Artif. Intell. Earth Syst., 3, 230066, https://doi.org/10.1175/AIES-D-23-0066.1, 2024.

Rasp, S., Pritchard, M. S., and Gentine, P.: Deep learning to represent subgrid processes in climate models, Proc. Natl. Acad. Sci., 115, 9684–9689, https://doi.org/10.1073/pnas.1810286115, 2018.

Reichler, T. and Kim, J.: How Well Do Coupled Models Simulate Today’s Climate?, https://doi.org/10.1175/BAMS-89-3-303, 2008.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat: Deep learning and process understanding for data-driven Earth system science, Nature, 566, 195–204, https://doi.org/10.1038/s41586-019-0912-1, 2019.

Riahi, K., van Vuuren, D. P., Kriegler, E., Edmonds, J., O’Neill, B. C., Fujimori, S., Bauer, N., Calvin, K., Dellink, R., Fricko, O., Lutz, W., Popp, A., Cuaresma, J. C., Kc, S., Leimbach, M., Jiang, L., Kram, T., Rao, S., Emmerling, J., Ebi, K., Hasegawa, T., Havlik, P., Humpenöder, F., Da Silva, L. A., Smith, S., Stehfest, E., Bosetti, V., Eom, J., Gernaat, D., Masui, T., Rogelj, J., Strefler, J., Drouet, L., Krey, V., Luderer, G., Harmsen, M., Takahashi, K., Baumstark, L., Doelman, J. C., Kainuma, M., Klimont, Z., Marangoni, G., Lotze-Campen, H., Obersteiner, M., Tabeau, A., and Tavoni, M.: The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: An overview, Glob. Environ. Change, 42, 153–168, https://doi.org/10.1016/j.gloenvcha.2016.05.009, 2017.

Ricard, L., Falasca, F., Runge, J., and Nenes, A.: network-based constraint to evaluate climate sensitivity, Nat. Commun., 15, 6942, https://doi.org/10.1038/s41467-024-50813-z, 2024.

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, https://doi.org/10.5194/gmd-13-1179-2020, 2020.

Roach, L. A., Dean, S. M., and Renwick, J. A.: Consistent biases in Antarctic sea ice concentration simulated by climate models, The Cryosphere, 12, 365–383, https://doi.org/10.5194/tc-12-365-2018, 2018.

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, Geophys. Res. Lett., 47, e2019GL086729, https://doi.org/10.1029/2019GL086729, 2020.

Rodgers, K. B., Lin, J., and Frölicher, T. L.: Emergence of multiple ocean ecosystem drivers in a large ensemble suite with an Earth system model, Biogeosciences, 12, 3301–20, https://doi.org/10.5194/bg-12-3301-2015, 2015.

Rojpratak, S. and Supharatid, S.: Regional extreme precipitation index: Evaluations and projections from the multi-model ensemble CMIP5 over Thailand, Weather Clim. Extrem., 37, 100475, https://doi.org/10.1016/j.wace.2022.100475, 2022.

Roy, I. and Tedeschi, R.: Influence of ENSO on Regional Indian Summer Monsoon Precipitation—Local Atmospheric Influences or Remote Influence from Pacific, Atmosphere, 7, 25, https://doi.org/10.3390/atmos7020025, 2016.

Roy, I., Tedeschi, R. G., and Collins, M.: ENSO teleconnections to the Indian summer monsoon in observations and models, Int. J. Climatol., 37, 1794–1813, https://doi.org/10.1002/joc.4811, 2017.

Roy, I., Gagnon, A. S., and Siingh, D.: Evaluating ENSO teleconnections using observations and CMIP5 models, Theor. Appl. Climatol., 136, 1085–1098, https://doi.org/10.1007/s00704-018-2536-z, 2018.

Roy, I., Tedeschi, R. G., and Collins, M.: ENSO teleconnections to the Indian summer monsoon under changing climate, Int. J. Climatol., 39, 3031–3042, https://doi.org/10.1002/joc.5999, 2019.

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., Van Nes, E. H., Peters, J., Quax, R., Reichstein, M., Scheffer, M., Schölkopf, B., Spirtes, P., Sugihara, G., Sun, J., Zhang, K., and Zscheischler, J.: Inferring causation from time series in Earth system sciences, Nat. Commun., 10, 2553, https://doi.org/10.1038/s41467-019-10105-3, 2019.

Runge, J., Tibau, X.-A., Bruhns, M., Muñoz-Marí, J., and Camps-Valls, G.: The Causality for Climate Competition, in: Proceedings of the NeurIPS 2019 Competition and Demonstration Track, 110–120, 2020.

Runge, J., Gerhardus, A., Varando, G., Eyring, V., and Camps-Valls, G.: Causal inference for time series, Nat. Rev. Earth Environ., 4, 487–505, https://doi.org/10.1038/s43017-023-00431-y, 2023.

Rupe, A., Crutchfield, J. P., Kashinath, K., and Prabhat: A Physics-Based Approach to Unsupervised Discovery of Coherent Structures in Spatiotemporal Systems, https://doi.org/10.48550/ARXIV.1709.03184, 2017.

Russo, F. and Toni, F.: Causal Discovery and Knowledge Injection for Contestable Neural Networks (with Appendices), https://doi.org/10.48550/ARXIV.2205.09787, 2022.

Rypkema, D. and Tuljapurkar, S.: Modeling extreme climatic events using the generalized extreme value (GEV) distribution, in: Handbook of Statistics, vol. 44, Elsevier, 39–71, https://doi.org/10.1016/bs.host.2020.12.002, 2021.

Sachindra, D. A., Ahmed, K., Rashid, Md. M., Shahid, S., and Perera, B. J. C.: Statistical downscaling of precipitation using machine learning techniques, Atmospheric Res., 212, 240–258, https://doi.org/10.1016/j.atmosres.2018.05.022, 2018.

Sanderson, B. M. and Knutti, R.: On the interpretation of constrained climate model ensembles, Geophys. Res. Lett., 39, https://doi.org/10.1029/2012GL052665, 2012.

Sanderson, B. M., Knutti, R., Aina, T., Christensen, C., Faull, N., Frame, D. J., Ingram, W. J., Piani, C., Stainforth, D. A., Stone, D. A., and Allen, M. R.: Constraints on Model Response to Greenhouse Gas Forcing and the Role of Subgrid-Scale Processes, J. Clim., 21, 2384–2400, https://doi.org/10.1175/2008JCLI1869.1, 2008.

Sanderson, B. M., Knutti, R., and Caldwell, P.: A Representative Democracy to Reduce Interdependency in a Multimodel Ensemble, https://doi.org/10.1175/JCLI-D-14-00362.1, 2015.

Sanderson, B. M., Pendergrass, A. G., Koven, C. D., Brient, F., Booth, B. B. B., Fisher, R. A., and Knutti, R.: The potential

https://doi.org/10.5194/egusphere-2025-4744
Preprint. Discussion started: 6 October 2025

for structural errors in emergent constraints, Earth Syst. Dyn., 12, 899–918, https://doi.org/10.5194/esd-12-899-2021, 2021.

Sansom, P. G., Stephenson, D. B., and Bracegirdle, T. J.: On Constraining Projections of Future Climate Using Observations and Simulations From Multiple Climate Models, J. Am. Stat. Assoc., 116, 546–557, https://doi.org/10.1080/01621459.2020.1851696, 2021.

Santer, B. D., Thorne, P. W., Haimberger, L., Taylor, K. E., Wigley, T. M. L., Lanzante, J. R., Solomon, S., Free, M., Gleckler, P. J., Jones, P. D., Karl, T. R., Klein, S. A., Mears, C., Nychka, D., Schmidt, G. A., Sherwood, S. C., and Wentz, F. J.: Consistency of modelled and observed temperature trends in the tropical troposphere, Int. J. Climatol., 28, 1703–1722, https://doi.org/10.1002/joc.1756, 2008.

Santer, B. D., Taylor, K. E., Gleckler, P. J., Bonfils, C., Barnett, T. P., Pierce, D. W., Wigley, T. M. L., Mears, C., Wentz, F. J., Brüggemann, W., Gillett, N. P., Klein, S. A., Solomon, S., Stott, P. A., and Wehner, M. F.: Incorporating model quality information in climate change detection and attribution studies, Proc. Natl. Acad. Sci., 106, 14778–14783, https://doi.org/10.1073/pnas.0901736106, 2009.

Schär, C., Fuhrer, O., Arteaga, A., Ban, N., Charpilloz, C., Di Girolamo, S., Hentgen, L., Hoefler, T., Lapillonne, X., Leutwyler, D., Osterried, K., Panosetti, D., Rüdisühli, S., Schlemmer, L., Schulthess, T. C., Sprenger, M., Ubbiali, S., and Wernli, H.: Kilometer-Scale Climate Models: Prospects and Challenges, Bull. Am. Meteorol. Soc., 101, E567–E587, https://doi.org/10.1175/BAMS-D-18-0167.1, 2020.

Scher, S.: Toward Data-Driven Weather and Climate Forecasting: Approximating a Simple General Circulation Model With Deep Learning, Geophys. Res. Lett., 45, 12,616-12,622, https://doi.org/10.1029/2018GL080704, 2018.

Schlunegger, S., Rodgers, K. B., Sarmiento, J. L., Frölicher, T. L., Dunne, J. P., Ishii, M., and Slater, R.: Emergence of anthropogenic signals in the ocean carbon cycle, Nat. Clim. Change, 9, 719–725, https://doi.org/10.1038/s41558-019-0553-2, 2019.

Schneider, T., Bischoff, T., and Haug, G. H.: Migrations and dynamics of the intertropical convergence zone, Nature, 513, 45–53, https://doi.org/10.1038/nature13636, 2014.

Schneider, T., Teixeira, J., Bretherton, C. S., Brient, F., Pressel, K. G., Schär, C., and Siebesma, A. P.: Climate goals and computing the future of clouds, Nat. Clim. Change, 7, 3–5, https://doi.org/10.1038/nclimate3190, 2017.

Scholkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y.: Toward Causal Representation Learning, Proc. IEEE, 109, 612–634, https://doi.org/10.1109/jproc.2021.3058954, 2021.

Sener, O. and Koltun, V.: Multi-Task Learning as Multi-Objective Optimization, in: Advances in Neural Information Processing Systems, 2018.

Seneviratne, S. I., Nicholls, N., Easterling, D., Goodess, C. M., Kanae, S., Kossin, J., Luo, Y., Marengo, J., McInnes, K., Rahimi, M., Reichstein, M., Sorteberg, A., Vera, C., Zhang, X., Rusticucci, M., Semenov, V., Alexander, L. V., Allen, S., Benito, G., Cavazos, T., Clague, J., Conway, D., Della-Marta, P. M., Gerber, M., Gong, S., Goswami, B. N., Hemer, M., Huggel, C., Van Den Hurk, B., Kharin, V. V., Kitoh, A., Tank, A. M. G. K., Li, G., Mason, S., McGuire, W., Van Oldenborgh, G. J., Orlowsky, B., Smith, S., Thiaw, W., Velegrakis, A., Yiou, P., Zhang, T., Zhou, T., and Zwiers, F. W.: Changes in Climate Extremes and their Impacts on the Natural Physical Environment, in: Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation, edited by: Field, C. B., Barros, V., Stocker, T. F., and Dahe, Q., Cambridge University Press, 109–230, https://doi.org/10.1017/CBO9781139177245.006, 2012.

Sexton, D. M. H., McSweeney, C. F., Rostron, J. W., Yamazaki, K., Booth, B. B. B., Murphy, J. M., Regayre, L., Johnson, J. S., and Karmalkar, A. V.: A perturbed parameter ensemble of HadGEM3-GC3.05 coupled model projections: part 1: selecting the parameter combinations, Clim. Dyn., 56, 3395–3436, https://doi.org/10.1007/s00382-021-05709-9, 2021.

Shaw, T. A., Arblaster, J. M., Birner, T., Butler, A. H., Domeisen, D. I. V., Garfinkel, C. I., Garny, H., Grise, K. M., and Karpechko, A. Yu.: Emerging Climate Change Signals in Atmospheric Circulation, AGU Adv., 5, e2024AV001297, https://doi.org/10.1029/2024AV001297, 2024.

Shepherd, T. G.: Atmospheric circulation as a source of uncertainty in climate change projections, Nat. Geosci., 7, 703–708, https://doi.org/10.1038/ngeo2253, 2014.

Shepherd, T. G.: Storyline approach to the construction of regional climate change information, Proc. R. Soc. Math. Phys. Eng. Sci., 475, 20190013, https://doi.org/10.1098/rspa.2019.0013, 2019.

Shepherd, T. G., Boyd, E., Calel, R. A., Chapman, S. C., Dessai, S., Dima-West, I. M., Fowler, H. J., James, R., Maraun, D., Martius, O., Senior, C. A., Sobel, A. H., Stainforth, D. A., Tett, S. F. B., Trenberth, K. E., Van Den Hurk, B. J. J. M., Watkins, N. W., Wilby, R. L., and Zenghelis, D. A.: Storylines: an alternative approach to representing uncertainty in physical aspects of climate change, Clim. Change, 151, 555–571, https://doi.org/10.1007/s10584-018-2317-9, 2018.

Shetty, S., Umesh, P., and Shetty, A.: The effectiveness of machine learning-based multi-model ensemble predictions of CMIP6 in Western Ghats of India, Int. J. Climatol., 43, 5029–5054, https://doi.org/10.1002/joc.8131, 2023.

Shin, Y., Lee, Y., and Park, J.-S.: A Weighting Scheme in A Multi-Model Ensemble for Bias-Corrected Climate Simulation, Atmosphere, 11, 775, https://doi.org/10.3390/atmos11080775, 2020.

Shuaifeng, S. and Xiaodong, Y.: Projected changes and uncertainty in cold surges over northern China using the CMIP6 weighted multi-model ensemble, Atmospheric Res., 278, 106334, https://doi.org/10.1016/j.atmosres.2022.106334, 2022.

Sidhu, B. S., Mehrabi, Z., Ramankutty, N., and Kandlikar, M.: How can machine learning help in understanding the impact of climate change on crop yields?, Environ. Res. Lett., 18, 024008, https://doi.org/10.1088/1748-9326/acb164, 2023.

Sillmann, J., Kharin, V. V., Zhang, X., Zwiers, F. W., and Bronaugh, D.: Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate, J. Geophys. Res. Atmospheres, 118, 1716–1733, https://doi.org/10.1002/jgrd.50203, 2013.

Simpson, I. R., McKinnon, K. A., Davenport, F. V., Tingley, M., Lehner, F., Fahad, A. A., and Chen, D.: Emergent Constraints on the Large-Scale Atmospheric Circulation and Regional Hydroclimate: Do They Still Work in CMIP6 and How Much Can They Actually Constrain the Future?, J. Clim., 34, 6355–6377, https://doi.org/10.1175/JCLI-D-21-0055.1, 2021.

Simpson, I. R., Shaw, T. A., Ceppi, P., Clement, A. C., Fischer, E., Grise, K. M., Pendergrass, A. G., Screen, J. A., Wills, R. C. J., Woollings, T., Blackport, R., Kang, J. M., and Po-Chedley, S.: Confronting Earth System Model trends with observations, Sci. Adv., 11, eadt8035, https://doi.org/10.1126/sciadv.adt8035, 2025.

Smith, D. M., Screen, J. A., Deser, C., Cohen, J., Fyfe, J. C., García-Serrano, J., Jung, T., Kattsov, V., Matei, D., Msadek, R., Peings, Y., Sigmond, M., Ukita, J., Yoon, J.-H., and Zhang, X.: The Polar Amplification Model Intercomparison Project (PAMIP) contribution to CMIP6: investigating the causes and consequences of polar amplification, Geosci. Model Dev., 12, 1139–1164, https://doi.org/10.5194/gmd-12-1139-2019, 2019.

Smith, D. M., Eade, R., Andrews, M. B., Ayres, H., Clark, A., Chripko, S., Deser, C., Dunstone, N. J., García-Serrano, J., Gastineau, G., Graff, L. S., Hardiman, S. C., He, B., Hermanson, L., Jung, T., Knight, J., Levine, X., Magnusdottir, G., Manzini, E., Matei, D., Mori, M., Msadek, R., Ortega, P., Peings, Y., Scaife, A. A., Screen, J. A., Seabrook, M., Semmler, T., Sigmond, M., Streffing, J., Sun, L., and Walsh, A.: Robust but weak winter atmospheric circulation response to future Arctic sea ice loss, Nat. Commun., 13, 727, https://doi.org/10.1038/s41467-022-28283-y, 2022.

Snyder, A., Prime, N., Tebaldi, C., and Dorheim, K.: Uncertainty-informed selection of CMIP6 Earth system model subsets for use in multisectoral and impact models, Earth Syst. Dyn., 15, 1301–1318, https://doi.org/10.5194/esd-15-1301-2024, 2024.

Soares, P. M. M., Careto, J. A. M., Russo, A., and Lima, D. C. A.: The future of Iberian droughts: a deeper analysis based on multi-scenario and a multi-model ensemble approach, Nat. Hazards, 117, 2001–2028, https://doi.org/10.1007/s11069-023-05938-7, 2023.

Soares, P. M. M., Johannsen, F., Lima, D. C. A., Lemos, G., Bento, V. A., and Bushenkova, A.: High-resolution downscaling of CMIP6 Earth system and global climate models using deep learning for Iberia, Geosci. Model Dev., 17, 229–259, https://doi.org/10.5194/gmd-17-229-2024, 2024.

Song, X., Wang, D.-Y., Li, F., and Zeng, X.-D.: Evaluating the performance of CMIP6 Earth system models in simulating global vegetation structure and distribution, Adv. Clim. Change Res., 12, 584–595, https://doi.org/10.1016/j.accre.2021.06.008, 2021.

Sonnewald, M. and Lguensat, R.: Revealing the Impact of Global Heating on North Atlantic Circulation Using Transparent Machine Learning, J. Adv. Model. Earth Syst., 13, e2021MS002496, https://doi.org/10.1029/2021MS002496, 2021.

Sørland, S. L., Fischer, A. M., Kotlarski, S., Künsch, H. R., Liniger, M. A., Rajczak, J., Schär, C., Spirig, C., Strassmann, K., and Knutti, R.: CH2018 – National climate scenarios for Switzerland: How to construct consistent multi-model projections from ensembles of opportunity, Clim. Serv., 20, 100196, https://doi.org/10.1016/j.cliser.2020.100196, 2020.

Steinman, B. A., Frankcombe, L. M., Mann, M. E., Miller, S. K., and England, M. H.: Response to Comment on "Atlantic and Pacific multidecadal oscillations and Northern Hemisphere temperatures," Science, 350, 1326–1326, https://doi.org/10.1126/science.aac5208, 2015.

Strobach, E. and Bel, G.: Learning algorithms allow for improved reliability and accuracy of global mean surface temperature projections, Nat. Commun., 11, 451, https://doi.org/10.1038/s41467-020-14342-9, 2020.

Su, B., Huang, J., Gemmer, M., Jian, D., Tao, H., Jiang, T., and Zhao, C.: Statistical downscaling of CMIP5 multi-model ensemble for projected changes of climate in the Indus River Basin, Atmospheric Res., 178–179, 138–149, https://doi.org/10.1016/j.atmosres.2016.03.023, 2016.

Sun, Z. and Archibald, A. T.: Multi-stage ensemble-learning-based model fusion for surface ozone simulations: A focus on CMIP6 models, Environ. Sci. Ecotechnology, 8, 100124, https://doi.org/10.1016/j.ese.2021.100124, 2021.

Takasuka, D., Satoh, M., Miyakawa, T., Kodama, C., Klocke, D., Stevens, B., Vidale, P. L., and Terai, C. R.: A protocol and analysis of year-long simulations of global storm-resolving models and beyond, Prog. Earth Planet. Sci., 11, 66, https://doi.org/10.1186/s40645-024-00668-1, 2024.

Tang, B., Hu, W., and Duan, A.: Future Projection of Extreme Precipitation Indices over the Indochina Peninsula and South China in CMIP6 Models, J. Clim., 34, 8793–8811, https://doi.org/10.1175/JCLI-D-20-0946.1, 2021.

Tang, J., Li, Q., Wang, S., Lee, D.-K., Hui, P., Niu, X., Gutowski, W. J., Dairaku, K., McGregor, J., Katzfey, J., Gao, X., Wu, J., Hong, S.-Y., Wang, Y., and Sasaki, H.: Building Asian climate change scenario by multi-regional climate models ensemble. Part I: surface air temperature: ASIAN CLIMATE CHANGE BY MULTI-MODEL ENSEMBLE, Int. J. Climatol., 36, 4241–4252, https://doi.org/10.1002/joc.4628, 2016.

Tapiador, F. J., Navarro, A., Moreno, R., Sánchez, J. L., and García-Ortega, E.: Regional climate models: 30 years of dynamical downscaling, Atmospheric Res., 235, 104785, https://doi.org/10.1016/j.atmosres.2019.104785, 2020.

Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, J. Geophys. Res. Atmospheres, 106, 7183–7192, https://doi.org/10.1029/2000JD900719, 2001.

Taylor, M., Caldwell, P. M., Bertagna, L., Clevenger, C., Donahue, A., Foucar, J., Guba, O., Hillman, B., Keen, N., Krishna, J., Norman, M., Sreepathi, S., Terai, C., White, J. B., Salinger, A. G., McCoy, R. B., Leung, L. R., Bader, D. C., and Wu, D.: The Simple Cloud-Resolving E3SM Atmosphere Model Running on the Frontier Exascale System, in: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, New York, NY, USA, 1–11, https://doi.org/10.1145/3581784.3627044, 2023.

Tebaldi, C. and Knutti, R.: The use of the multi-model ensemble in probabilistic climate projections, Philos. Trans. R. Soc. Math. Phys. Eng. Sci., 365, 2053–2075, https://doi.org/10.1098/rsta.2007.2076, 2007.

Tebaldi, C., Smith, R. L., Nychka, D., and Mearns, L. O.: Quantifying Uncertainty in Projections of Regional Climate Change: A Bayesian Approach to the Analysis of Multimodel Ensembles, https://doi.org/10.1175/JCLI3363.1, 2005.

Tebaldi, C., Dorheim, K., Wehner, M., and Leung, R.: Extreme metrics from large ensembles: investigating the effects of ensemble size on their estimates, Earth Syst. Dyn., 12, 1427–1501, https://doi.org/10.5194/esd-12-1427-2021, 2021.

Tegegne, G., Melesse, A. M., and Worqlul, A. W.: Development of multi-model ensemble approach for enhanced assessment of impacts of climate change on climate extremes, Sci. Total Environ., 704, 135357, https://doi.org/10.1016/j.scitotenv.2019.135357, 2020.

Tegegne, G., Melesse, A. M., and Alamirew, T.: Projected changes in extreme precipitation indices from CORDEX simulations over Ethiopia, East Africa, Atmospheric Res., 247, 105156, https://doi.org/10.1016/j.atmosres.2020.105156, 2021.

Teuling, A. J., de Badts, E. A. G., Jansen, F. A., Fuchs, R., Buitink, J., Hoek van Dijke, A. J., and Sterling, S. M.: Climate change, reforestation/afforestation, and urbanization impacts on evapotranspiration and streamflow in Europe, Hydrol. Earth Syst. Sci., 23, 3631–3652, https://doi.org/10.5194/hess-23-3631-2019, 2019.

Thackeray, C. W., Hall, A., Norris, J., and Chen, D.: Constraining the increased frequency of global precipitation extremes under warming, Nat. Clim. Change, 12, 441–448, https://doi.org/10.1038/s41558-022-01329-1, 2022.

Thuy, A. and Benoit, D. F.: Explainability through uncertainty: Trustworthy decision-making with neural networks, Eur. J. Oper. Res., 317, 330–340, https://doi.org/10.1016/j.ejor.2023.09.009, 2024.

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

Tibau, X.-A., Reimers, C., Gerhardus, A., Denzler, J., Eyring, V., and Runge, J.: A spatiotemporal stochastic climate model for benchmarking causal discovery methods for teleconnections, Environ. Data Sci., 1, https://doi.org/10.1017/eds.2022.11, 2022.

Toms, B. A., Barnes, E. A., and Ebert-Uphoff, I.: Physically Interpretable Neural Networks for the Geosciences: Applications to Earth System Variability, J. Adv. Model. Earth Syst., 12, e2019MS002002, https://doi.org/10.1029/2019MS002002, 2020.

von Trentini, F., Aalbers, E. E., Fischer, E. M., and Ludwig, R.: Comparing interannual variability in three regional single-model initial-condition large ensembles (SMILEs) over Europe, Earth Syst. Dyn., 11, 1013–1031, https://doi.org/10.5194/esd-11-1013-2020, 2020.

US CLIVAR: Multi-Model Large Ensemble Archive (MMLEA), 2020.

Vázquez-Patiño, A., Campozano, L., Mendoza, D., and Samaniego, E.: A causal flow approach for the evaluation of global climate models, Int. J. Climatol., 40, 4497–4517, https://doi.org/10.1002/joc.6470, 2020.

Veenadhari, S., Misra, B., and Singh, C.: Machine learning approach for forecasting crop yield based on climatic parameters, in: 2014 International Conference on Computer Communication and Informatics, 1–5, https://doi.org/10.1109/ICCCI.2014.6921718, 2014.

Vogel, M. M., Hauser, M., and Seneviratne, S. I.: Projected changes in hot, dry and wet extreme events' clusters in CMIP6 multi-model ensemble, Environ. Res. Lett., 15, 094021, https://doi.org/10.1088/1748-9326/ab90a7, 2020.

Waliser, D. E. and Gautier, C.: A Satellite-derived Climatology of the ITCZ, J. Clim., 6, 2162–2174, https://doi.org/10.1175/1520-0442(1993)006<2162:ASDCOT>2.0.CO;2, 1993.

Wang, B., Liu, D. L., Macadam, I., Alexander, L. V., Abramowitz, G., and Yu, Q.: Multi-model ensemble projections of future extreme temperature change using a statistical downscaling method in south eastern Australia, Clim. Change, 138, 85–98, https://doi.org/10.1007/s10584-016-1726-x, 2016.

Wang, B., Zheng, L., Liu, D. L., Ji, F., Clark, A., and Yu, Q.: Using multi-model ensembles of CMIP5 global climate models to reproduce observed monthly rainfall and temperature with machine learning methods in Australia, Int. J. Climatol., 38, 4891–4902, https://doi.org/10.1002/joc.5705, 2018.

Wang, D. and Yuan, F.: High-Performance Computing for Earth System Modeling, in: High Performance Computing for Geospatial Applications, edited by: Tang, W. and Wang, S., Springer International Publishing, Cham, 175–184, https://doi.org/10.1007/978-3-030-47998-5_10, 2020.

Wang, D., Liu, J., Shao, W., Mei, C., Su, X., and Wang, H.: Comparison of CMIP5 and CMIP6 Multi-Model Ensemble for Precipitation Downscaling Results and Observational Data: The Case of Hanjiang River Basin, Atmosphere, 12, 867, https://doi.org/10.3390/atmos12070867, 2021.

Wang, F. and Tian, D.: On deep learning-based bias correction and downscaling of multiple climate models simulations, Clim. Dyn., 59, 3451–3468, https://doi.org/10.1007/s00382-022-06277-2, 2022.

Wang, F. and Tian, D.: Multivariate bias correction and downscaling of climate models with trend-preserving deep learning, Clim. Dyn., 62, 9651–9672, https://doi.org/10.1007/s00382-024-07406-9, 2024.

Wang, J., Kim, H., Kim, D., Henderson, S. A., Stan, C., and Maloney, E. D.: MJO Teleconnections over the PNA Region in Climate Models. Part I: Performance- and Process-Based Skill Metrics, J. Clim., 33, 1051–1067, https://doi.org/10.1175/JCLI-D-19-0253.1, 2020.

Wang, S., Sankaran, S., and Perdikaris, P.: Respecting causality for training physics-informed neural networks, Comput. Methods Appl. Mech. Eng., 421, 116813, https://doi.org/10.1016/j.cma.2024.116813, 2024.

Weber, T., Corotan, A., Hutchinson, B., Kravitz, B., and Link, R.: Technical note: Deep learning for creating surrogate models of precipitation in Earth system models, Atmospheric Chem. Phys., 20, 2303–2317, https://doi.org/10.5194/acp-20-2303-2020, 2020.

Wehner, M. F.: Characterization of long period return values of extreme daily temperature and precipitation in the CMIP6 models: Part 2, projections of future change, Weather Clim. Extrem., 30, 100284, https://doi.org/10.1016/j.wace.2020.100284, 2020.

Weigel, A. P., Knutti, R., Liniger, M. A., and Appenzeller, C.: Risks of Model Weighting in Multimodel Climate Projections, J. Clim., 23, 4175–4191, https://doi.org/10.1175/2010JCLI3594.1, 2010.

Wenzel, S., Eyring, V., Gerber, E. P., and Karpechko, A. Yu.: Constraining Future Summer Austral Jet Stream Positions in the CMIP5 Ensemble by Process-Oriented Multiple Diagnostic Regression*, J. Clim., 29, 673–687, https://doi.org/10.1175/JCLI-D-15-0412.1, 2016.

van der Wiel, K., Lenderink, G., and de Vries, H.: Physical storylines of future European drought events like 2018 based on ensemble climate modelling, Weather Clim. Extrem., 33, 100350, https://doi.org/10.1016/j.wace.2021.100350, 2021.

Wilby, R. L. and Fowler, H. J.: Regional climate downscaling, Wiley, 85 pp., 2010.

Williams, D. N., Balaji, V., Cinquini, L., Denvil, S., Duffy, D., Evans, B., Ferraro, R., Hansen, R., Lautenschlager, M., and Trenham, C.: A Global Repository for Planet-Sized Experiments and Observations, Bull. Am. Meteorol. Soc., 97, 803–816, https://doi.org/10.1175/BAMS-D-15-00132.1, 2016.

Wing, A. A., Camargo, S. J., Sobel, A. H., Kim, D., Moon, Y., Murakami, H., Reed, K. A., Vecchi, G. A., Wehner, M. F., Zarzycki, C., and Zhao, M.: Moist Static Energy Budget Analysis of Tropical Cyclone Intensification in High-Resolution Climate Models, J. Clim., 32, 6071–6095, https://doi.org/10.1175/JCLI-D-18-0599.1, 2019.

Woldemeskel, F. M., Sharma, A., Sivakumar, B., and Mehrotra, R.: An error estimation method for precipitation and temperature projections for future climates, J. Geophys. Res. Atmospheres, 117, https://doi.org/10.1029/2012JD018062, 2012.

Wootten, A. M., Başağaoğlu, H., Bertetti, F. P., Chakraborty, D., Sharma, C., Samimi, M., and Mirchi, A.: Customized Statistically Downscaled CMIP5 and CMIP6 Projections: Application in the Edwards Aquifer Region in South-Central Texas, Earths Future, 12, e2024EF004716, https://doi.org/10.1029/2024EF004716, 2024.

Wu, H., Su, X., and Singh, V. P.: Increasing Risks of Future Compound Climate Extremes With Warming Over Global Land Masses, Earths Future, 11, e2022EF003466, https://doi.org/10.1029/2022EF003466, 2023.

Xiang, B., Zhao, M., Held, I. M., and Golaz, J.: Predicting the severity of spurious “double ITCZ” problem in CMIP5 coupled models from AMIP simulations, Geophys. Res. Lett., 44, 1520–1527, https://doi.org/10.1002/2016GL071992, 2017.

Xu, D., Ivanov, V. Y., Kim, J., and Fatichi, S.: On the use of observations in assessment of multi-model climate ensemble, Stoch. Environ. Res. Risk Assess., 33, 1923–1937, https://doi.org/10.1007/s00477-018-1621-2, 2019.

Xu, L. and Wang, A.: Application of the Bias Correction and Spatial Downscaling Algorithm on the Temperature Extremes From CMIP5 Multimodel Ensembles in China, Earth Space Sci., 6, 2508–2524, https://doi.org/10.1029/2019EA000995, 2019.

Xu, R., Chen, N., Chen, Y., and Chen, Z.: Downscaling and Projection of Multi-CMIP5 Precipitation Using Machine Learning Methods in the Upper Han River Basin, Adv. Meteorol., 2020, 8680436, https://doi.org/10.1155/2020/8680436, 2020.

Xu, Z., Han, Y., Tam, C.-Y., Yang, Z.-L., and Fu, C.: Bias-corrected CMIP6 global dataset for dynamical downscaling of the historical and future climate (1979–2100), Sci. Data, 8, 293, https://doi.org/10.1038/s41597-021-01079-3, 2021.

Yang, T., Hao, X., Shao, Q., Xu, C.-Y., Zhao, C., Chen, X., and Wang, W.: Multi-model ensemble projections in temperature and precipitation extremes of the Tibetan Plateau in the 21st century, Glob. Planet. Change, 80–81, 1–13, https://doi.org/10.1016/j.gloplacha.2011.08.006, 2012.

Yang, X., Yu, X., Wang, Y., He, X., Pan, M., Zhang, M., Liu, Y., Ren, L., and Sheffield, J.: The Optimal Multimodel Ensemble of Bias-Corrected CMIP5 Climate Models over China, J. Hydrometeorol., 21, 845–863, https://doi.org/10.1175/JHM-D-19-0141.1, 2020.

Yeganeh-Bakhtiary, A., EyvazOghli, H., Shabakhty, N., Kamranzad, B., and Abolfathi, S.: Machine Learning as a Downscaling Approach for Prediction of Wind Characteristics under Future Climate Change Scenarios, Complexity, 2022, 8451812, https://doi.org/10.1155/2022/8451812, 2022.

Yip, S., Ferro, C. A. T., Stephenson, D. B., and Hawkins, E.: A Simple, Coherent Framework for Partitioning Uncertainty in Climate Predictions, J. Clim., 24, 4634–4643, https://doi.org/10.1175/2011JCLI4085.1, 2011.

Yoon, J. and Schaar, M. van der: E-RNN : Entangled Recurrent Neural Networks for Causal Prediction, 2017.

Yu, S., Hannah, W., Peng, L., Lin, J., Bhouri, M. A., Gupta, R., Lütjens, B., Will, J. C., Behrens, G., Busecke, J., Loose, N., Stern, C., Beucler, T., Harrop, B., Hillman, B., Jenney, A., Ferretti, S. L., Liu, N., Anandkumar, A., Brenowitz, N., Eyring, V., Geneva, N., Gentine, P., Mandt, S., Pathak, J., Subramaniam, A., Vondrick, C., Yu, R., Zanna, L., Zheng, T., Abernathey, R., Ahmed, F., Bader, D., Baldi, P., Barnes, E., Bretherton, C., Caldwell, P., Chuang, W., Han, Y., Huang, Y., Iglesias-Suarez, F., Jantre, S., Kashinath, K., Khairoutdinov, M., Kurth, T., Lutsko, N., Ma, P.-L., Mooers, G., Neelin, J. D., Randall, D., Shamekh, S., Taylor, M., Urban, N., Yuval, J., Zhang, G., and Pritchard, M.: ClimSim: A large multi-scale dataset for hybrid physics-ML climate emulation, Adv. Neural Inf. Process. Syst., 36, 22070–22084, 2023.

Zappa, G. and Shepherd, T. G.: Storylines of Atmospheric Circulation Change for European Regional Climate Impact Assessment, J. Clim., 30, 6561–6577, https://doi.org/10.1175/JCLI-D-16-0807.1, 2017.

Zebarjadian, F., Dolatabadi, N., Zahraie, B., Yousefi Sohi, H., and Zandi, O.: Triple coupling random forest approach for bias correction of ensemble precipitation data derived from Earth system models for Divandareh-Bijar Basin (Western Iran), Int. J. Climatol., 44, 2363–2390, https://doi.org/10.1002/joc.8458, 2024.

Zhang, X., Zwiers, F. W., Hegerl, G. C., Lambert, F. H., Gillett, N. P., Solomon, S., Stott, P. A., and Nozawa, T.: Detection of human influence on twentieth-century precipitation trends, Nature, 448, 461–465, https://doi.org/10.1038/nature06025, 2007.

Zhang, X., Wang, X.-L., Fan, F., Cheung, Y.-M., and Bose, I.: Enhancing the Performance of Neural Networks Through Causal Discovery and Integration of Domain Knowledge, https://doi.org/10.48550/ARXIV.2311.17303, 2023.

Zhao, L., Wang, Y., Zhao, C., Dong, X., and Yung, Y. L.: Compensating Errors in Cloud Radiative and Physical Properties over the Southern Ocean in the CMIP6 Climate Models, Adv. Atmospheric Sci., 39, 2156–2171, https://doi.org/10.1007/s00376-022-2036-z, 2022.

Zhao, T. and Dai, A.: CMIP6 Model-projected Hydroclimatic and Drought Changes and Their Causes in the 21st Century, J. Clim., 1–58, https://doi.org/10.1175/JCLI-D-21-0442.1, 2021.

Zhou, W. and Xie, S.-P.: A Hierarchy of Idealized Monsoons in an Intermediate GCM, J. Clim., 31, 9021–9036, https://doi.org/10.1175/JCLI-D-18-0084.1, 2018.

Zhu, J. and Poulsen, C. J.: Last Glacial Maximum (LGM) climate forcing and ocean dynamical feedback and their implications for estimating climate sensitivity, Clim. Past, 17, 253–267, https://doi.org/10.5194/cp-17-253-2021, 2021.

Zuluaga, M., Sergent, G., Krause, A., and Püschel, M.: Active Learning for Multi-Objective Optimization, in: Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 462–470, 2013.