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] Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, 1000, Slovenia

Correspondence to: Anja Katzenberger (anja.katzenberger@gmx.de)

Abstract. Earth System Models (ESMs) are a 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

standardization 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. Here, 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, synthesizing 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 aim to support researchers working with climate MMEs, particularly in the upcoming 7th phase of CMIP.

1 Introduction

Earth system models (ESMs) are a key tool for assessing the future climate under changing conditions. Starting from the seminal work of Manabe and Hasselmann (e.g. Manabe and Strickler, 1964; Manabe and Bryan, 1969; Manabe and Wetherald, 1967; Hasselmann, 1976), who were awarded the 2021 Nobel Prize in Physics, 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. This evolution occurred in parallel with advances in Earth system observations, high-resolution numerical models giving 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 model resolution to steadily improve.

Since the beginning of large-scale atmospheric modelling, intercomparisons among models have been carried out. Initially, this intercomparison was mostly performed for numerical weather prediction as computational resources limited the intercomparison of studies in the climate context, 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 organized several intercomparison projects among climate models. 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). In the following years, 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 MME 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) coordinated by the World Climate Research Programme (WCRP), which also incorporated results from AOGCMs (Meehl et al., 2000).

While the first phase of CMIP was limited to control runs, new standardized scenarios were incorporated throughout the phases of CMIP with an increasing number of international model centres contributing simulations. Concurrently, the volume of data has been steadily increasing (Williams et al., 2016) and is stored within a standardized format at the Earth System Grid Federation (ESGF) central repository (Cinquini et al., 2012). In more 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 experiments. Finally, future climate change simulations are performed for various greenhouse gas emission scenarios such as abrupt carbon dioxide doubling or quadrupling to derive climate sensitivity (measure of how much the Earth's climate system will warm under a doubling of atmospheric CO_2 concentration), as well as for multiple plausible future emission scenarios (O'Neill et al., 2017; Riahi et al., 2017; van Vuuren et al., 2025). 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). These CMIP projections have proven essential for informing mitigation and adaptation strategies to climate change at the global and regional scales (Meehl et al., 2000). For regional analysis, the CMIP output is often downscaled to finer resolution, e.g. by using the CMIP output as boundary conditions for regional climate models (RCMs). This is done e.g. in the WCRP COordinated Regional climate Downscaling EXperiment (CORDEX), which provides a coordinated framework for producing and evaluating regional climate projections across multiple domains worldwide (Giorgi, 2019; Gutowski Jr. et al., 2016)

The main components of an ESM describe 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). In all ESMs, the continuous behavior of the atmosphere and ocean is first discretized in space and time via the so-called "model dynamical core" which encompasses known governing equations that capture resolved (grid-scale) phenomena and parameterization schemes that represent unresolved or poorly understood (subgrid-scale) processes. ESMs differ in the choice of computational grids (e.g., latitude-longitude structured grids, icosahedral grids, variable resolution cube-sphere grids), numerical methods for solving the dynamical core equations, and in parameterization schemes. 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, computational limitations restrict the accuracy with which models can represent certain relevant processes. Therefore, the decisions made at modelling centers make each ESM an imperfect attempt to represent a multitude of highly complex, nonlinear processes, and the synchronized

interplay among them. Depending on the interest of the end user, some of these necessary idealization decisions may be more suitable than others.

Combining several ESMs to multi-model ensembles (MMEs) can have numerous advantages compared to individual simulations, e.g. to account for the uncertainty arising from the differing modelling decisions (model uncertainty). Starting in the weather forecasting community, numerous studies have shown the benefits of ensemble predictions compared to predictions based on single models (Doblas-Reyes et al., 2003; Krishnamurti et al., 1999), e.g. the North American MME showed improvements in various skill metrics (correlation, RMSE, RPSS, and reliability) compared to individual models used before (Kirtman et al., 2014). Inspired by these findings, studies in the climate context also analyzed the potential benefits from working with MMEs for projections. In climate model evaluation, the MME projections have proven to outperform individual model projections 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) and variability (Zhang et al., 2007). The enhancement of the signal and cancellation of errors contribute to these advantages (Doblas-Reyes et al., 2005; Hagedorn et al., 2005; Smith et al., 2013). Becker et al. (2022) highlight the practical advantage of the continuous operation of MMEs, which can be maintained even when individual modelling centers are temporarily unable to contribute, for example due to technical or political constraints. They further provide an example where the use of a MME enabled the identification of outlier behavior in ENSO predictions, which could subsequently be traced back to previously unknown deficiencies in the underlying reanalysis dataset, thereby supporting the model improvement. Furthermore, an ensemble approach reduces the risk of selecting a model outlier with particularly large biases.

Given these benefits, MME projections have become an established tool for climate studies addressing a broad range of research questions, also being the standard method to analyze 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 synthesized 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.

It is important to recognize that CMIP constitutes an “ensemble of opportunity” (Tebaldi and Knutti, 2007; Sanderson et al., 2012; Merrifield et al., 2023), as it reflects the collection of readily available simulations rather than a systematically designed sample. Contributing institutions range from long-established, well-resourced climate modelling centers to newer groups with sufficient computational resources to run adapted versions of existing models. While this inclusivity broadens participation, such ensembles of opportunity are not designed to constitute a statistically representative sample of multi-model uncertainty (Merrifield et al., 2023). In this context, the superiority of MMEs is not universal. There are cases in which individual models can outperform the ensemble mean, for instance when the averaging inherent to MMEs suppresses relevant signals that are

well represented in only a subset of models. This can occur for specific physical processes, or extremes, where ensemble averaging may smooth physically meaningful variability or dampen circulation-driven responses. Moreover, if most models in an MME share common structural components, parameterizations, or tuning strategies, systematic biases can persist in the ensemble mean. In such cases, individual models with alternative formulations may provide more accurate representations for specific variables, regions, or applications.

The availability of standardized climate model outputs facilitated model intercomparison and has naturally inspired the use of 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. However, it is challenging to have an overview of these studies, and 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. This individuality makes it challenging or 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 climate science, but also to provide an overview of existing resources and approaches for more experienced scientists, particularly for (but not restricted to) the upcoming 7th phase of CMIP.

While the focus of this paper is on the challenges associated with combining various ESMs within a MME, it should be pointed out there are other types of climate ensembles. Besides such uninitialized simulations, there are initialized climate model ensembles that are routinely used for seasonal prediction (see e.g. Becker et al., 2020, 2022; Buontempo et al., 2022; Kirtman et al., 2014; Min et al., 2025). Initialized climate model ensembles are based on accurate initialization and thus have an emphasis on assimilation procedures to capture the atmosphere, ocean and land conditions. While their goals differ from those of CMIP, initialized prediction ensembles face similar challenges related to ensemble design, model weighting, and evaluation against observations. Further ensemble types include initial condition ensembles (ICEs) and perturbed parameter ensembles (PPEs) (IPCC, AR5). ICEs are 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 (PPEs)

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 of various ensemble types (IPCC, AR6).

In the following section, we conduct a comprehensive literature review on studies regarding model evaluation (2.1), model dependence (2.2), model selection and weighting methods (2.3) and uncertainty characterization (2.4). In this context, we also provide a summary of useful tools for MME analysis (2.5). In Section 3, we complement these guidelines with a collection of frequently occurring topics and challenges based on the experience of the WCRP Fresh Eyes on CMIP community. In Section 4, we discuss emerging trends for working with MMEs such as machine learning (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 through CMIP (https://wcrp-cmip.org/map/). When addressing any research question, the need for specific variables, scenarios, resolutions or experiments narrows the pool of available models. However, the remaining number is often still large, prompting the following questions: Should all available models be used, or only a subset? How can the models be identified that are most suitable for such a subset? The two primary objectives when selecting models are to optimize model performance and to reduce duplicated information (Herger et al., 2018). As adequate selection criteria are central to the design of MME studies, we aim to provide guidance for the choice of models in this section.

2.1 Model Evaluation

Model evaluation refers to the systematic assessment of climate model simulations against observational reference data in order to compare model performance and identify biases. For an overview of model bias see Appendix B. In practice, this involves benchmarking historical simulations with respect to observed climate statistics, such as mean states, variability, spatial patterns, and relevant physical processes.

Observation Datasets for Model Evaluation

Observational reference datasets used for model evaluation include both direct observations and reanalysis products. Reanalysis datasets are physically consistent products produced by assimilating diverse observational data into a numerical weather or climate model. They combine the broad spatial and temporal coverage of models with observational constraints and are therefore widely used as reference datasets. Direct observations include paleoclimate data, ground-based measurements over land and ocean (e.g., ships, buoys and sail drones), aircraft and balloon measurements, and satellite data. Paleoclimate

data give insight into the state of the Earth's climate hundreds to millions of years ago, offering valuable constraints for paleoclimate simulations that 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, which 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 and coverage of global reference datasets. However, satellite datasets are limited to the time after the 1970s, depending on the variable of interest.

All these datasets have distinct advantages and disadvantages: They encompass different spatial and temporal scales, cover different locations and time periods, rely on different measurement techniques, or vary in accuracy. See e.g. Sippel et al., (2024) for challenges in observational data. Associated uncertainties also differ, e.g. due to instrument uncertainty, calibration limitations, or interpolation procedures. Accounting for these uncertainties in the reference datasets can be done by combining multiple datasets (Notz et al., 2016). It also facilitates 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). Depending on the variable of interest, commonly used reanalysis datasets are ERA5 (produced by ECMWF), MERRA-2 (produced by NASA GSFC), NCEP DOE R-2 (produced by NOAA), JRA-3Q (produced by Japan Meteorological Agency).

Moreover, model evaluation using observations is not always straightforward, as observational sensors do not necessarily measure variables simulated by climate models. To ensure an "apple-to-apple comparison", observed quantities must be converted into model-output-like variables, or vice versa. For example, software has been developed which enables simulating what a satellite would observe over the model atmosphere. Moreover, it must be assured that observations and simulations have the same temporal and spatial resolution, including the horizontal grid and number of vertical levels (Simpson et al., 2025), which can be achieved by appropriate regridding methods. See Section 3.5 for details on regridding.

Generally, there are two approaches to model evaluation: (i) The performance-oriented approach focuses on identifying the models whose output is closest to observations or reanalysis data. (ii) The process-oriented approach seeks models that best capture the dynamics of interest. Regardless of the 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 referenced. But in that case it is important to make sure that they cover the variables, scales, and other factors relevant to the specific research questions.

Performance-oriented Evaluation

For shorter timescale forecasts, predictions can be verified within days as observations become available. This is typical of weather forecasting and initialized climate model simulations, in which models are started from observation-constrained initial conditions. Such near-term verifiability offers an opportunity to build confidence in models, particularly for climate services and decision-relevant applications. Although initialized and uninitialized climate projections address different time horizons, linking insights from both may help contextualize uncertainties and enhance trust in long-term projections. Climate projections addressing longer time scales cannot be directly verified in real time, as the relevant time scales (decades to centuries) preclude immediate verification. This is the case for uninitialized climate model simulations, which represent the standard approach for long-term climate projections and are the focus in this review. Accordingly, 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 fail to perform well for the past regarding some specific climate phenomena will also do so for the future. While this assumption is commonly accepted, it also is a limitation of this approach as the role of specific circulation patterns and their interactions might change throughout the 21st century. In this context, Knutti et al., 2010a found that 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 past performance does not necessarily reduce future inter-model spread. Given these pitfalls, Mendlik and Gobiet (2016) propose to only remove the severely unrealistic models. A detailed assessment on how to deal with outliers can be found in Subsection 3.3.

Because uninitialized climate model simulations are free-running and not constrained by observations, performance-oriented evaluation cannot rely on a direct comparison of individual events or temporal trajectories. Instead, model evaluation is necessarily based on climatological characteristics, such as mean states or spatial patterns. Evaluating this climatological performance 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, the model projections may be largely 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 use metrics that are relevant to the specific research question while also matching the spatial and temporal scale of the phenomenon in question.

Taylor diagrams (Taylor, 2001) have become a widely used tool to visualize performance-oriented model evaluation, helping to identify better performing models as well as outliers. They are applied across the full range of climate-related topics,

including e.g. the Indian Summer Monsoon (Roy et al., 2019) and seasonal mean temperatures (Tang et al., 2016). In a Taylor diagram, the radial distance from the origin represents the model standard deviation, the angle from the horizontal axis encodes the correlation with observations, and the geometric distance to the reference point (defined as the observed standard deviation and correlation = 1) equals the centered root mean square error, quantifying pattern mismatch after mean removal. Models closer to the observed standard deviation, along with higher correlation coefficients (and therefore lower root mean square error) are considered as better performing models for specific climate features (Taylor, 2001). The angular/azimuthal position in Taylor diagrams represents the pattern correlation coefficient between CMIP6 models and observations, while the radial distance indicates the ratio of the standard deviation of CMIP6 models to that from an observational data set. For example, the Western Pacific pattern, a prominent teleconnection pattern during the boreal winter over the North Pacific, was analyzed 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 exceed 0.6. In reproducing spatial patterns, the mean of the MME outperforms most 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. Selecting only the best performing models can improve the final MME mean.

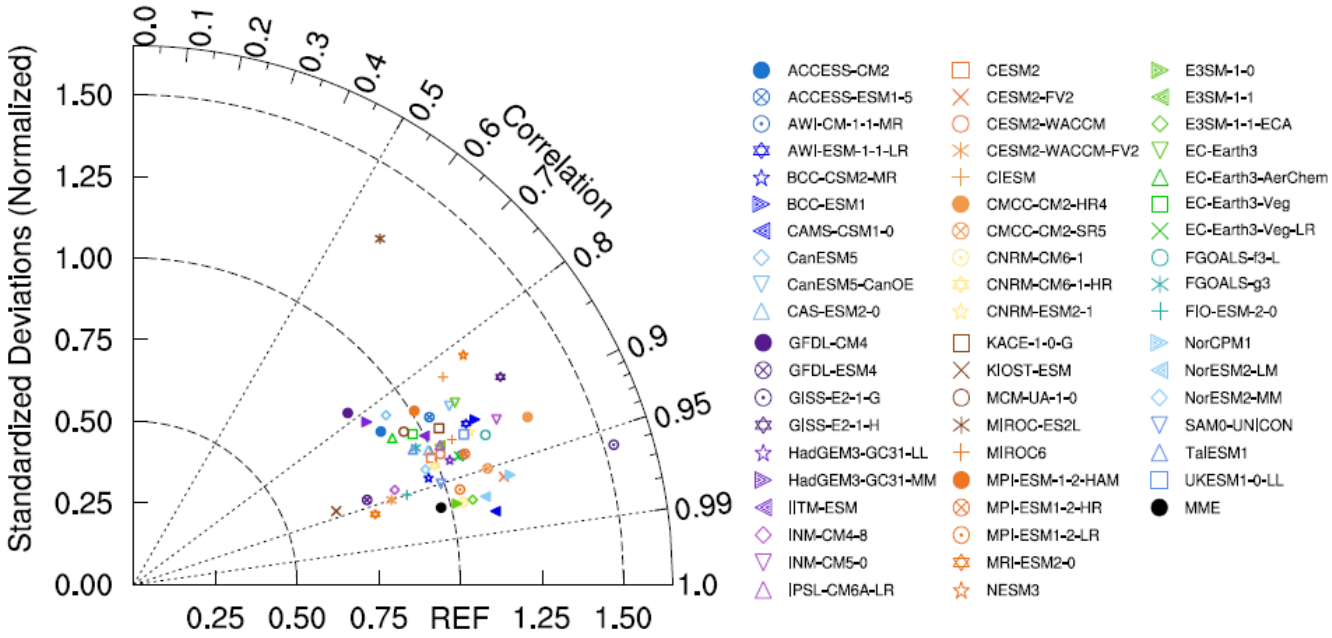


Fig. 1. Example for the use of a Taylor diagram showing the geopotential height anomalies at 500-hPa over the Western North Pacific (20°N–80°N, 120°E–120°W) in individual CMIP6 models, MME and observations, taken from Aru et al. (2023).

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 minimise error compensation.

It is important to remember that models are calibrated with the aim to reduce anomalies compared to observational data before becoming available in new CMIP generations. During this calibration (often referred to as tuning), parameters, typically associated with unresolved processes such as clouds, convection, or boundary-layer dynamics, are adjusted to improve agreement with observations. Consequently, improvements in overall model performance in new CMIP generations do not necessarily stem from enhanced capabilities in capturing relevant processes, but may instead result from optimized calibration (Knutti, 2010). A related issue is that the same observational datasets used for model calibration are often also employed for model evaluation, which is not optimal as calibration and evaluation datasets ideally should be independent. This concern is even more pronounced when using reanalysis products as reference data, since climate models are an integral part of their generation.

Additionally, observational data can influence model performance through the forcings themselves. For example, in concentration-driven CO_2 simulations, 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, since the model does not simulate atmospheric CO_2 concentrations from emissions via an interactive carbon cycle. Consequently, apparent improvements visible in the model's evaluation do not necessarily indicate a better representation of the carbon cycle itself.

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

This evaluation approach shifts from traditional performance-oriented evaluation to more detailed, process-oriented metrics, which are critical for advancing the next generation of ESMs. Eyring et al. (2005) and Gleckler et al. (2008) emphasise the need to evaluate a wide range of climate processes, since accurately simulating one aspect does not ensure accuracy in others. These authors initiated the development of a comprehensive set of model metrics to assess important processes in climate simulations. Process-oriented evaluation identifies sources and limitations of predictability, guiding model development by revealing deficiencies in the representation of physical processes and thereby enhancing the reliability of climate projections (Eyring et al., 2016). By incorporating process-oriented analysis into diagnostic packages (examples in Subsection 2.5), evaluations become reproducible, accelerating model improvements and establishing benchmarks for progress. As with any standardization effort, however, such benchmarks must be applied with care, as they have the potential to promote model similarity. Another relevant resource in the context of process-oriented evaluation is Simpson et al. (2025), who review the ability of climate models to reproduce historically observed forced trends and outline best practices for confronting modeled and observed signals. Within the MME framework, process-based approaches help identify which processes contribute most to inter-model differences and provide insights into the mechanisms behind model performance. Here, we outline some common use cases and techniques.

Process-oriented diagnostics to reduce model bias: One major focus in the development of process-oriented metrics is the investigation of phenomena with strong bias in the models, as e.g. the Madden-Julian Oscillation (MJO), the dominant mode of tropical intraseasonal variability. To better understand the origins of these biases, a number of diagnostics has been 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 process-oriented multi-model comparison study on MJO teleconnections found that biases in simulating the position of the Pacific westerly jets, together with deficiencies in MJO representation, contribute substantially to errors in MJO teleconnections (Ahn et al. 2017, Henderson et al. 2017). Similar efforts exist for the El Niño–Southern Oscillation (ENSO), for which Planton et al. (2021) provide a dedicated metrics package.

Improving projections by process-oriented multiple diagnostic ensemble regression: Karpechko et al. (2013) developed the multiple diagnostic ensemble regression (MDER) method that constrains climate projections using observed diagnostics, applying it to Antarctic ozone columns. By identifying key processes that influence ozone, MDER explains a substantial fraction of the inter-model spread in projected ozone across climate chemistry models and outperforms the unweighted multi-model mean in pseudo-realistic validation. Building on this approach, Wenzel et al. (2016) applied the MDER algorithm, implemented as a diagnostic in ESMValTool (see Subsection 2.5), to analyze projections of the austral jet position under the RCP4.5 scenario in CMIP5 simulations. They found that MDER reduces uncertainty in the ensemble-mean projection without substantially altering the long-term mean position of the jet.

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 tropical cyclones, particularly their intensity in GCMs. The findings suggest that model-specific factors, beyond large-scale environmental parameters, play a key role in shaping tropical cyclones' intensity, with differences in convection schemes contributing significantly to the inter-model spread. Wing et al. (2019) and Moon et al. (2020) further applied these methods, with Moon et al. (2020) showing that tropical cyclone wind structures are strongly influenced by model resolution. Accordingly, Dirkes et al. (2023) emphasizes the necessity of applying the developed diagnostics for tropical cyclone analysis in CMIP6 models.

Using idealization or a hierarchy of models: Another approach is to design model configurations that isolate individual processes and components, allowing to test their relevance for specific phenomena. For example, Katzenberger et al. (2024) employed an aquaplanet configuration with a circumglobal land stripe to evaluate the meridional circulation, particularly the Hadley cell, in an idealized setup. By shifting the landstripe north and southwards, and by modifying the surface albedo or aerosol concentrations, the role of these features in shaping monsoon dynamics could be systematically isolated. More generally, iteratively adding components and increasing the complexity and realism of the setup within a hierarchy of models enables the isolation of individual processes and the assessment of their contributions to the overall model performance. See also e.g. Zhou and Xie (2018) for more insights to this approach.

Using causal inference: In Section 4.1, we provide insights into how ML techniques can be applied to improve process-based evaluation by identifying causal relationships.

Another example of process-oriented assessment is provided by Fasullo et al. (2020), who present a thorough analysis of CMIP representation of the leading Earth system modes of variability. Additional applications include regime-based evaluation approaches of low-level marine clouds, where distinguishing stratocumulus from shallow cumulus regimes has helped diagnose persistent cloud-cover and radiative biases in CMIP6 and CMIP5 models and inform targeted model improvements (Črnivec et al., 2023; Cesana et al., 2023). Process-based analyses have also demonstrated that the ENSO–Indian Summer Monsoon teleconnection is robustly represented in CMIP5 and CMIP6 models, consistent with a realistic simulation of the coupled Hadley–Walker circulation and associated precipitation responses (Roy and Tedeschi, 2016; Roy et al., 2017; Fasullo et al., 2020). We provide further details and examples of process-oriented analyses in the Appendix C.

2.2 Model Dependence

ESMs are developed by multiple modelling groups worldwide. Ideally, the models in a MME would be independent, thereby providing an adequate representation of the epistemic uncertainty. Historically, climate projections are derived by calculating simple averages across the MME, based on the assumption that the ensemble mean offers the most accurate representation of

the Earth system by synthesizing the collective modelling efforts (Abramowitz et al., 2019; Knutti et al., 2010a). Assuming independence implies that the MME reflects a sufficiently broad range of uncertainties, and the averaging smooths out individual model biases. In practice, however, the development of ESMs is often not independent (Pincus et al., 2008).

Components that address modelling challenges or have demonstrated strong performance are often shared among multiple ESMs, including e.g. the dynamical core for resolving grid-scale dynamics or components addressing sub-grid-scale phenomena (e.g., parameterization schemes). For example, the McICA radiation scheme (Pincus et al., 2003) provides an efficient and flexible representation of one-dimensional radiative transfer in a cloudy atmosphere, and is thus implemented in multiple 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 modelling centers, including e.g. HadGEM3 and NorESM2, further underscoring the sharing of model components. Fig. 2 illustrates the shared model history tracing back to a few AGCMs (Kuma et al., 2023).

In addition to dependencies between modelling groups, individual centers often contribute multiple closely related model configurations, for example differing in horizontal resolution (e.g., MPI-ESM1.2-HR for high resolution versus MPI-ESM1.2-LR for low resolution) or in the inclusion of additional components, such as interactive vegetation in EC-Earth3-Veg compared to EC-Earth3. If such dependencies are not accounted for and all models are included with equal weight in a multi-model mean, modelling centers that provide several related configurations effectively receive greater weight than others. This issue is particularly relevant in multi-model studies with a limited number of included models.

Diverse analyses confirm that the number of independent models in CMIP is smaller than the total number of participating models (Jun et al., 2008; Masson and Knutti, 2011; Pennell and Reichler, 2011). Because errors across different models are often being correlated (Knutti et al., 2010a), the lack of independence can lead to amplified biases (Jun et al., 2008; Knutti, 2008; Reichler and Kim, 2008; Tebaldi and Knutti, 2007). Moreover, apparent convergence among model results and the associated reduction in ensemble uncertainty may be mistakenly interpreted as strong agreement between models, when in fact they arise from structural dependencies.

The lack of a universally accepted and unambiguous definition of model independence complicates efforts to systematically account for model dependence in MME studies. Some definitions focus on the conceptual idea of whether or not a model adds novel additional information to the MME (Masson and Knutti, 2011). Others adopt a more analytical approach to understanding model dependence, offering examples for evaluating model dependence and using their framework (e.g., Annan and Hargreaves, 2017). Despite such advances, no broadly accepted solution has yet emerged. Further approaches, such as weighting schemes (Subsection 2.3) have been proposed, but these tend to be problem-specific and struggle to capture the full complexity of model dependencies. The metadata reporting requirements introduced in CMIP6 have made comprehensive assessments of model dependence possible, thereby representing a meaningful advance in transparency. As new model

generations are developed and incorporated to CMIP, continued efforts to quantify and correct for model dependence will be essential to ensure robust ensemble projections that reflect true uncertainty.

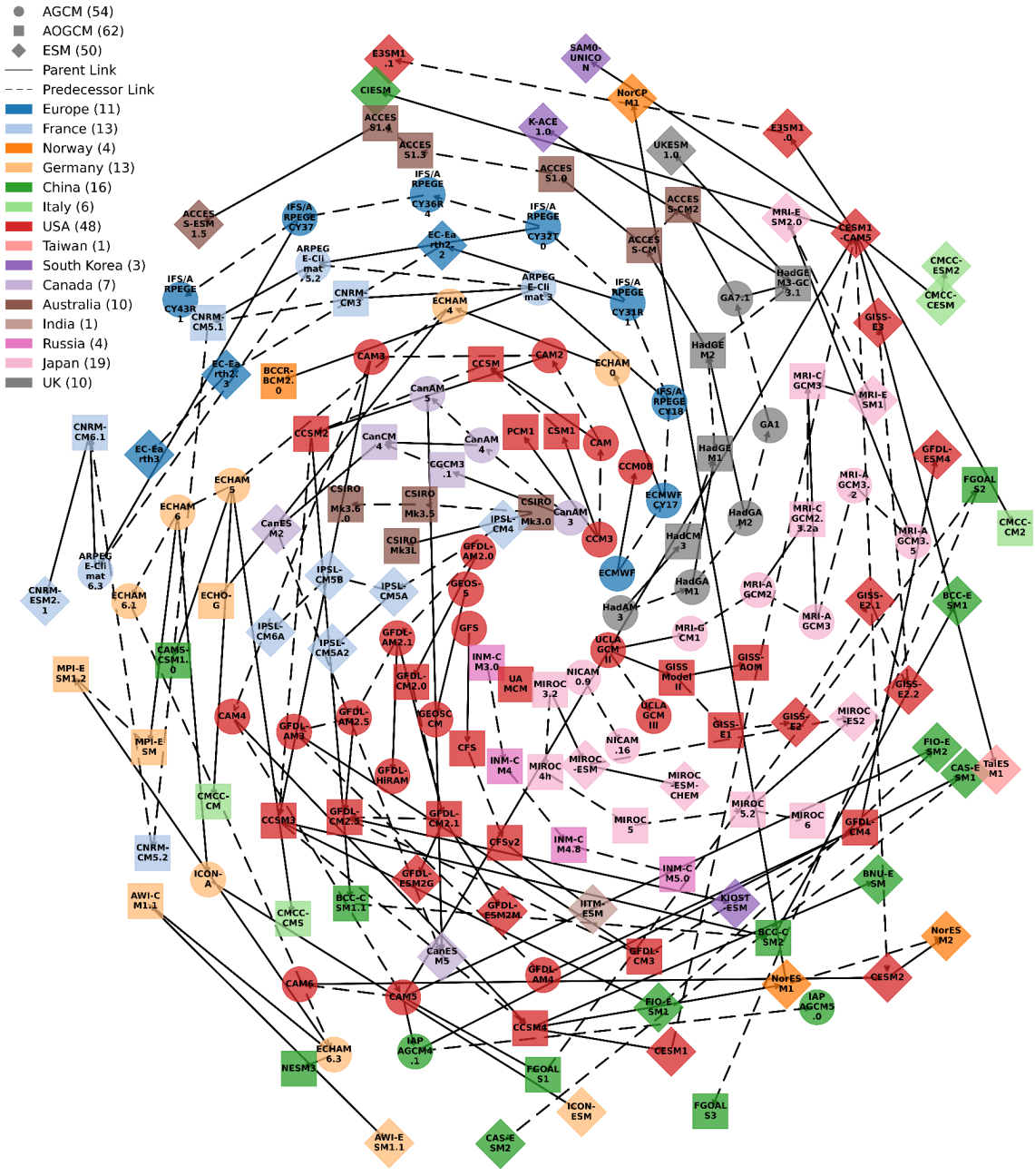


Fig. 2. Spiral plot of climate model dependencies, adapted from Kuma et al. (2023). The oldest model in any given family is in the center of the plot, spiralling out as more models are made. Model type is differentiated by shape of marker, and link type is differentiated by arrow type (solid for parent 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".

2.3 Model Selection and Weighting Methods

CMIP MME weighting and selection techniques are used to categorize the CMIP models based on historical model performance (see Subsection 2.1) and independence (see Subsection 2.2) using several metrics (Palmer et al., 2023), which is crucial for optimizing accuracy and reliability in projections (Strobach and Bel, 2020). Several performance-based and statistical approaches are used for MME weighting (Bhowmik and Sankarasubramanian, 2020; Brunner et al., 2020). Performance-based weighting assigns weights based on the ability to reproduce observed historical climate patterns, while statistical model weighting assigns weights based on properties like independence and spread (Brunner et al., 2020). Both approaches are discussed in this section, complemented by subselection approaches. Model weighting and subselecting to account for model outliers is discussed specifically in Section 3.3. It is also important to note that some studies may be primarily interested in assessing the overall performance of the full CMIP ensemble. Applying weighting or model subselection is not relevant to such analyses.

Accounting for Model Performance

Most studies in the literature use simple multi-model means, thus equally weighted MMEs to project future climate change impacts (Shuaifeng and Xiaodong, 2022). While such approaches capture the overall trends across all models, equal weighting without any model selection has been criticized for not considering model performance (Shin et al., 2020). Incorporating information on model skill —by emphasizing better-performing models and down-weighting or excluding models with poor simulation capabilities— can improve both the accuracy of projections and the assessment of uncertainty in CMIP MMEs (Merrifield et al., 2020).

Weighting, however, is a challenging process as the basis for weights must be determined and other not yet identified but equally relevant factors may be neglected. Moreover, the relevance of specific model features for a given phenomenon may change under future climate conditions, making it questionable to assign weights solely based on present-day performance, as discussed previously in the context of model evaluation. In addition, weighting schemes may inadvertently favor structurally similar models that produce "mainstream" results, while penalizing outlier models that could provide valuable insights (see also Subsection 3.3). When bias correction is applied, assessing model performance becomes particularly challenging, as

differences between models and observations are largely removed, complicating performance-based weighting. Shin et al. (2020) addressed this issue by proposing a hybrid weighting approach that preserves performance information while avoiding unrealistically extreme model weights.

Despite these challenges, several studies have demonstrated the potential benefits of performance-based model weighting for climate projections. Tang et al. (2021) found that weighted MMEs produce more robust projections of extreme precipitation over the Indo-China Peninsula and southern China than unweighted ensembles. Similarly, Shuaifeng and Xiaodong (2022) applied a rank-based weighting approach to CMIP6 MMEs for projecting and quantifying uncertainty in cold surges over northern China. 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 skills.

Another approach to account for model performance is the selection of 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 and have been shown to decrease the ensemble-mean RMSE, e.g. by roughly 10–20% for air temperature and approximately 12% for precipitation relative to the full multi-model mean (Hamed et al., 2021; Herger et al., 2018; Snyder et al., 2024). The central challenges in subselecting are the identification of performance metrics, as already discussed in Subsection 2.1, as well as the definition of selection criteria that should be made transparent. Herger et al. (2018) compared different sub-selection approaches, including random ensembles, performance-based ranking, and optimal ensemble subselection, and found improved performance over the multi-model mean in some cases. In a random ensemble, multiple models are combined randomly without an explicit optimization strategy. In performance ranking, models are ranked based on metrics such as accuracy, Q-statistics, mean square error etc. In optimal ensemble sub-selection, a subset of models is chosen that maximizes performance. Almazroui et al. (2017) similarly found that a subset of the best-performing models showed better temperature and precipitation projections over the Arabian Peninsula. Numerous further examples exist, e.g. including the ENSO teleconnection (Roy et al., 2018) and lightning over South/South-east Asia (Chandra et al., 2022).

Accounting for Model Dependence

As discussed in Subsection 2.2, climate models are not fully independent. A common approach to address this issue is by weighting models based on their independence from others. Sanderson et al. (2015) developed a mathematical formulation to quantify model uniqueness and assign corresponding weights. Boé (2018) argues that model interdependencies are more effectively assessed through code similarity instead of through result similarity. Although 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 has the potential to reveal shared model components and algorithms that may not be evident from model output comparisons alone. Recent model selection approaches also emphasize model independence (Snyder et al., 2024), with tools such as ClimSIPS explicitly accounting for model dependence (Merrifield et al., 2023).

Assessing both source-code similarity and similarity in model results enables the identification of shared methodologies that can lead to correlated predictions, thereby highlighting potential redundancies within MMEs that may bias ensemble statistics. Integrating these measures into weighting schemes can therefore improve the robustness of MMEs and contribute to more reliable and less biased projections.

Combined accounting for Model Performance and Dependence

Knutti et al. (2017) proposed a model weighting method that accounts for model performance as well as model dependency. This method 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.

2.4 Uncertainty Characterization

Model selection and weighting ideally improves uncertainty which remains inevitable when trying to predict climate (Knutti et al., 2019). Characterizing and understanding it is essential for guiding model evaluation and development, for science and risk communication, and for assessing climate 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 develop in the future, and it is accounted for by modelling different emission scenarios (O'Neill et al., 2014; van Vuuren et al., 2025). Natural or internal variability uncertainty is due to the chaotic and, thus, unpredictable evolution of the climate system (Deser et al., 2012b), having a great impact on climate projections (Lehner and Deser, 2023). The unique realization of our 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 –model uncertainty–results from our imperfect attempts to predict the aforementioned real world realization. It includes differences among models as well as the varying results that can be obtained within the same model when varying its parameters. While model uncertainty can be reduced, its interpretation and quantification depend strongly on how the ensemble is constructed (Knutti et al., 2019). An adequate understanding of uncertainty has the potential to help MMEs users with model selection and thereby 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 for reducing uncertainty. 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. 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, exploiting the increasing computational capabilities, 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 SMILEs. Instead of calculating the variance of the polynomials as in Hawkins and Sutton (2009), in this approach, the model uncertainty is calculated as the variance across ensemble means from the available SMILEs. This reduces methodological assumptions and thereby improves the results, making SMILEs currently a broadly used tool to partition uncertainty in climate projections. It is important to note that the lack of independence between models (Subsection 2.2), and the methods to account for it (Subsection 2.3) must also be considered in this context.

A question that should be considered, although it can only be partially answered, is whether the MME spread is realistic, too narrow or too broad. The uncertainty may be too broad if observations are not used correctly to tune models, or if the models have extensive and diverse structural errors. The ensemble may be too narrow, and thus overly confident if the models are structurally very similar, if they are overfitted to observations or if uncertain processes are missing. It is also important to recognize that present-day and future uncertainties arise from different sources: present-day uncertainty mainly reflects the models' ability to reproduce observations, whereas future uncertainty stems from variations in how models represent physical processes and feedbacks (Sanderson and Knutti, 2012). Care should be taken when assuming that the spread of present-day or historical simulations will be the same in the future.

As discussed in Subection 2.1, relying solely on how well a model reproduces past climate to assign confidence can be misleading: models that perform well historically may not accurately project future climate changes, while models that perform poorly may still provide useful information about future conditions (Hall et al., 2019). An evaluation and uncertainty reduction technique that avoids this bias is the development of emergent constraints (Hall et al., 2019). An emergent constraint refers to a statistically robust relationship across a MME between an observable present-day quantity (x) and a projected future change

in a quantity (Δy), typically approximated as linear (Simpson et al., 2021). When this relationship is robust, observations of x can be used to constrain the plausible range of y, thereby reducing uncertainty. This is commonly achieved by analyzing the probability distribution function of y conditioned on the observed value of x. 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 nonlinearity and feedbacks (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.

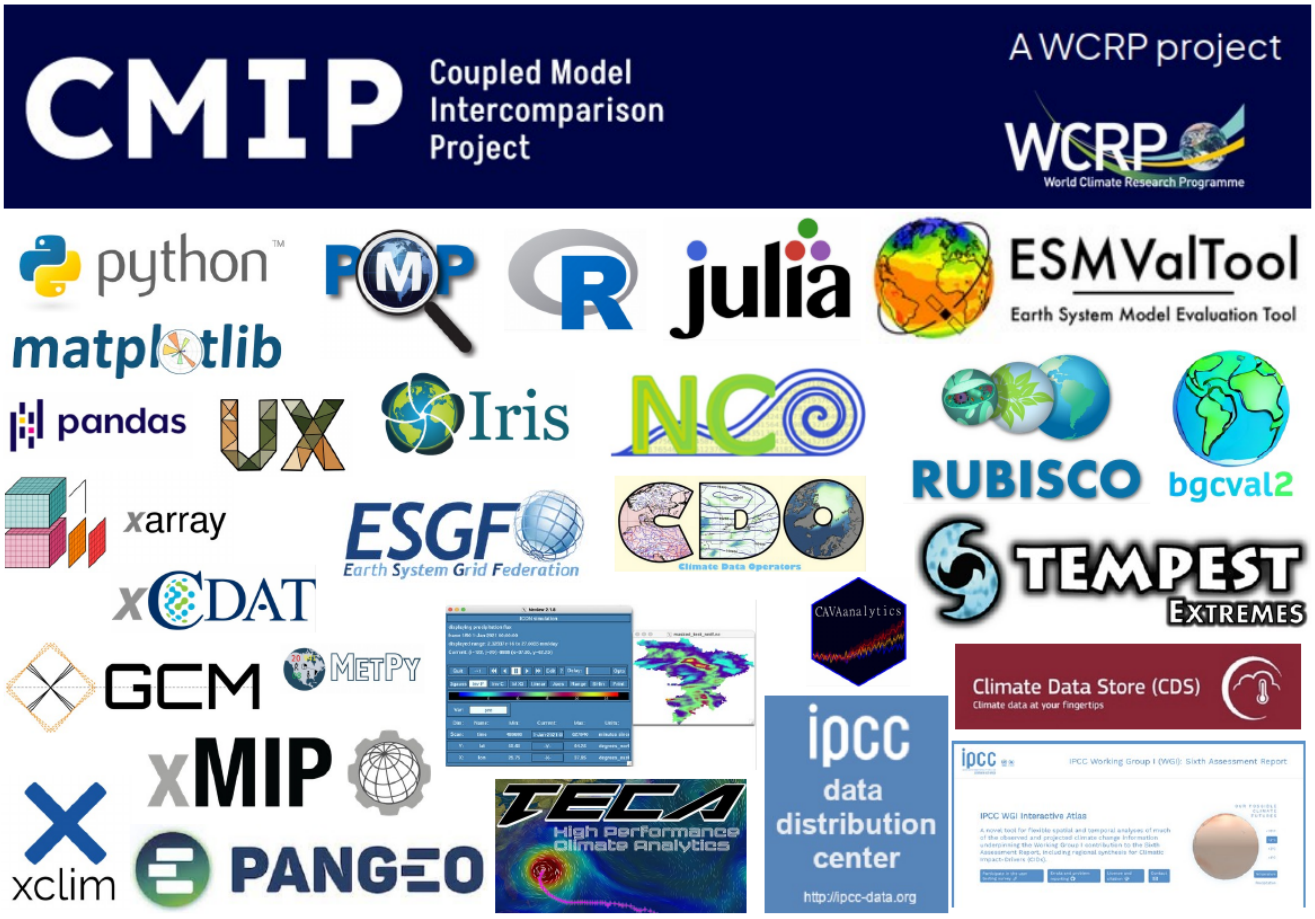
2.5 Available Tools for MME Analysis

The analysis of CMIP datasets is greatly facilitated by a variety of tools developed within the global climate science community. However, the wide range of available tools was not centrally cataloged, making it difficult to obtain a clear overview of available tools and their capabilities. To address this gap, the WCRP CMIP has undertaken an effort to compile a central repository (https://wcrp-cmip.org/tools/) that encompasses a broad range of resources.

The repository includes data access platforms (e.g., Earth System Grid Federation, Climate Data Store, IPCC data distribution centre, PANGEO, CAVA, Climate Information Portal), which facilitate accessing large and complex data volumes. It also lists widely used command line operators (e.g., ncview, NCO, CDO) and programming languages 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). In addition, the repository contains several comprehensive evaluation and benchmarking tools, such as ESMValTool, bgcval2, RUBISCO, PCMDI Metrics Package, AMBER, and the MDTF Diagnostic Package. These evaluation tools include diagnostics designed to address specific scientific questions. For example, ESMValTool incorporates the Climate Variability Diagnostics Package (CVDP, Eyring et al., 2020; Phillips et al., 2020, 2014; https://github.com/NCAR/CVDP-ncl), which facilitates the analysisof modes of climate variability and change in models and observations (Maher et al., 2024). Another important initiative in process-oriented evaluation 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 Subsection 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, and Google Earth Engine. Figure 3 highlights some of these tools, aiming to promote their use across the wider climate community. Basic information about each tool is provided by "Tools description cards" on the CMIP website, which include links to tool websites, documentation, tutorials and community support resources. Finally, it should be emphasized that the tools repository is actively maintained and continuously updated. To further enhance its utility for the broader climate science community, new contributions are highly welcome.

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



https://wcrp-cmip.org/tools/

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

3. Complementing Topics and Challenges

Building on the general workflow involved in MME studies in Section 2, we draw on the experience within the Fresh Eyes community to identify common topics and challenges that arise in this context. All of these aspects are also relevant to the subsections in Section 2; however, our aim here is to provide a dedicated overview of specific topics, allowing researchers to access the most relevant information in one place.

3.1 Number of Models

Any MME analysis has to face the question of how many models to include, which is not straightforward, as it involves the trade-offs between model diversity, computational cost, and the accuracy of the results. Increasing the number of ensemble members has the potential to enhance the robustness of the results by reducing statistical uncertainty. 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 included in MME studies. These challenges raise the question how many models are actually required to form a "good" ensemble size. Here, we focus on the number of models within a MME. A closely related question exists in the context of large ensembles where the number of perturbed simulations is discussed. Some examples for the number of simulations in large ensembles are provided in Subsection 4.2. However, many of the arguments and findings in this section apply for both contexts.

Lower threshold of ensemble size: At least 5 models

If the ensemble size is too small, the inter-model may not be fully captured. This has the potential to lead to an underestimation of uncertainties and can consequently result in an overestimation of the models' performance and thus an overconfident interpretation of the results. It is even possible that a too small ensemble size leads to qualitatively different findings, as shown by Milinski et al. (2020). In this study, the subsets of two and three models showed a warming after a volcanic eruption, while the actual known response would be cooling. 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 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 individual minimum ensemble size

If feasible, determining the appropriate minimum ensemble size on a case-by-case basis—depending on the specific research question and requirements—is preferable to adopting a general minimum. Milinski et al., 2020 proposed a procedure applicable to diverse research questions. After (1) defining the research question, (2) an error metric (e.g. RMSE) as well as a maximum acceptable error has to be decided. Then (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). This requirement is introduced to avoid resampling bias, as random subsets close to the full ensemble share many members, are no longer independent, and therefore tend to reproduce the full-ensemble signal by construction rather than providing an unbiased estimate of the required ensemble size, see Milinski et al., 2020 for details. While this method provides a straight-forward method to identify the ideal minimum number of models in an ensemble, it requires the availability and analysis of a high number of model simulations. Consequently, this method might not be feasible for all studies.

Remarks for including more models

While the considerations above address the identification of a minimum ensemble size, additional models have the potential to further improve the model performance. For some applications, larger ensemble sizes are even required, e.g. for the quantification of internal variability, as estimating higher-order moments of the distribution demands a sufficiently large ensemble (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 Subsection 2.2 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). Subsection 3.3 therefore discusses strategies for identifying and handling outliers in more detail. Another consideration becomes relevant when working with different scenarios. As the range of uncertainty increases with the number of models, using the same number of models across all scenarios is essential to ensure comparability (Knutti et al., 2010a).

3.2 Extremes

Extreme weather and climate events have significant impacts on human society and ecosystems. Understanding the drivers and producing reliable future projections of these low-frequency high-impact events is therefore essential for effective climate

change adaptation planning. When using MMEs to study extreme climate events, ensembles offer both strengths and challenges.

MMEs based on CMIP or CORDEX are widely used both in regional and global studies concerning climate extremes (Kim et al., 2020; Soares et al., 2023; Vogel et al., 2020; Yang et al., 2012). These studies typically apply statistical approaches, such as probabilistic modelling, and/or using climate extremes indices defined by the Expert Team on Climate Change Detection and Indices (ETCCDI). Extreme Value Theory (EVT) provides the theoretical foundation for analyzing extreme events by 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 modelling the tail of the distribution of rare events. For example, GEV analysis is frequently used to estimate return periods of extreme rainfall events, allowing assessment of how the frequency and intensity of such events may change under future climate scenarios (Wehner, 2020). By fitting GEV to observed and modeled data, researchers can evaluate shifts in extreme event characteristics.

A major advantage of using the mean of the MME is that averaging across models reduces noise from internal variability and thereby amplifies the climate change signal. This can also help identify trends in extreme events (IPCC, 2021). However, the MME mean might not always be the best choice, particularly when examining the intensity and frequency of extreme events (Knutti et al., 2010b). 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. This is especially the case when some models predict significantly different extreme event trends, potentially leading to an underestimation of risks. Uncertainty remains for both hot and cold extremes, with some models deviating considerably from the multi-model mean. Uncertainties are particularly large for precipitation extremes, where, despite a general tendency toward heavier precipitation and longer dry periods, several models project opposing trends in specific regions (Sillmann et al., 2013).

In studies on climate extremes, it is therefore important to evaluate how well each model performs in simulating extremes (Kim et al., 2020; Sillmann et al., 2013) and to correct for biases when appropriate. However, as discussed in Subsection 2.1, model evaluation is conducted using performance-oriented or process-oriented approaches that generally tend to focus on the models' ability to capture mean climate states or large-scale circulation patterns rather than models' extreme event representation. Dedicated evaluations tailored to extremes are therefore required, capturing relevant temporal resolution, region and variables and comparing to an appropriate reference data set. For example, Kim et al. (2020) evaluated the CMIP6 MME 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. As a step following the evaluation, employing model weighting is one possible approach to address shortcomings, enhancing the accuracy and reliability of extreme event projections (Balhane et al., 2022).

When studying extreme climate events, uncertainty is another aspect that is important to account for. As discussed in Subsection 3.1, ensemble size strongly influences uncertainty estimates, with larger ensembles allowing a more complete sampling of the range of possible outcomes. In practice, many studies of climate extremes using MMEs rely on a single ensemble member per model to ensure comparability across models (Kim et al., 2020). However, using only one ensemble member per model could miss some of the variability in extreme events that larger ensemble runs could capture, particularly as often not too extreme members are submitted for intercomparison projects as CMIP. 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 reduce uncertainties, it does not eliminate the limitations inherent to individual models. Downscaling techniques, either statistical or dynamical using 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). MMEs based on RCM projections are particularly valuable for highly vulnerable regions, offering insights into potential changes in local extreme events (Dosio, 2017; Tegegne et al., 2021) and supporting planning for challenges such as water scarcity, food security, and disaster preparedness. For more details regarding downscaling, see also Subsection 3.4.

3.3 Outliers

Outlier models have at times been disregarded in MME analyses because convergence toward the ensemble mean has been interpreted as a measure of model reliability. However, the use of convergence as a measure of model reliability has been criticized because it favors simulations that are closer to the multi-model mean, while underrepresenting uncertainty across a wider range of plausible outcomes (Tebaldi and Knutti, 2007). For example, the original version of the reliability ensemble average (REA) weighting method assigns higher weights to models that better reproduce the current climate, but also penalizes models that diverge from the ensemble mean (Giorgi and Mearns, 2002). As a result, outliers receive lower weights even when their differences may reflect physically plausible behavior rather than poor model performance. This is especially problematic because convergence toward the ensemble mean can partly arise from genealogical similarities among models, rather than independent confirmation of a result (Tebaldi and Knutti, 2007). Despite these concerns, there is a history of privileging convergence towards the MME mean within the climate science community. For example, in the third IPCC assessment report two models were discarded because of extreme warming projections associated with very high climate sensitivity (Tebaldi and Knutti, 2007). More recently, convergence-based ideas continue to influence MME subsetting approaches (Palmer et al., 2023) and are still used, at least in part, in MME evaluation frameworks (Amali et al., 2024). Whether emphasizing convergence is

appropriate depends on the purpose of a given study. For applications such as the analysis of climate extremes, averaging across models can mask the full range of possible outcomes. In these cases, outlier models may provide valuable information about plausible high-impact scenarios rather than representing spurious deviations from the ensemble mean. See also Subsection 3.2 for more details on extremes in MMEs. Building on the insights to weighting and building subsets of models in Subsection 2.3, we discuss here in more detail how and when to account for outliers.

Exclusion of Outliers

One approach to account for outliers, is exclusion, meaning the removal of models with outlier status from an ensemble. While this can help reduce unrealistic spread and improve agreement with observations, it carries the risk of omitting simulations that represent rare but physically plausible events. In some cases, however, the benefits of exclusion outweigh the 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. These models, which overestimated snow cover in areas of low snow mass, were excluded to improve the alignment between observational data and CMIP6 MME projections. Similarly, the Swiss Climate Scenarios CH2018, based on EURO-CORDEX, excluded some outlier GCMs to narrow uncertainty ranges for temperature and precipitation (Sørland et al., 2020). The consequences of outlier inclusion or exclusion have been explored in the literature. Sun and Archibald (2021) compared "aggressive" and "conservative" approaches—respectively including and excluding outliers—and found that, for their study, the differences in results were relatively minor. Similarly, Bracegirdle and Stephenson (2012) presented analyses both with and without outliers to illustrate the sensitivity of polar warming estimates to outlier inclusion and different forms of regression. Overall, while models with outlier projections may be excluded to improve MME alignment with observations or to reduce uncertainty, this should be done with caution. Exclusion is most justified when a model is known to be deeply flawed. Otherwise, removing projections of rare but plausible events may limit the assessment of adaptation strategies and risk management options (Knutti et al. 2010).

Weighting or Penalization

MME inclusion of outlier models is often accomplished through weighting. One commonly used approach is weighting based on root-mean-square error (RMSE) skill scores. For example, Tegegne et al. (2020) preserve MME spread by applying the Katsavounidis–Kuo–Zhang algorithm to select ensemble members based on their contribution to representing the full range of variability within the sampling space for extreme indices of interest, as recommended by World Meteorological Organization's ETCCDI. The IPCC characterizes this approach as suitable for detecting "moderate extremes"-events expected to occur up to 10% of the time (Seneviratne et al., 2012). In this approach, detecting extremes prior to taking the MME mean is useful for weighting members such that the full range of variability within the MME is largely preserved. To identify and characterize more extreme events, methods based on extreme value theory (EVT) are required. EVT focuses on values located in the very

ends of tails of probability distribution functions (PDFs) and is therefore better suited to representing rare, high-impact extremes (DelSole and Tippett, 2022).

Among weighting methods rooted in classical statistics is the use of outlier insensitive methods. These are methods that retain outlier models while limiting their influence on the ensemble result. Such methods include using the ensemble median instead of its mean as a measure of the MME's center which reduces sensitivity to outliers (Ge et al., 2021). Rank based tests of statistical significance provide another option. Because they rely on the rank, or position of a value within a PDF, rather than the value of a particular data point within a sample, these are largely insensitive to outliers (DelSole and Tippett, 2022). This test is recommended by the World Meteorological Organization for hydrological data analysis and is robust to outliers and non-normal data distributions (Rojpratak and Supharatid, 2022).

Weighting methods based on Bayesian statistics have been developed to sample uncertainty across a broad statistical space. 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 shape of the PDF for a given dataset (Clyde et al., 2022). Xu et al. (2019) apply Bayesian model weighting to statistically downscale precipitation data for site-specific analyses. The authors argue for the use of statistical downscaling due to its relatively low computational expense with finer spatial and temporal resolution data. At the same time, they note that dynamic downscaling can underestimate extremes and be overly sensitive to outliers. These limitations are addressed by applying a Bayesian weighted average, which reduces outlier influence while retaining information about uncertainty.

Penalization refers to methods that are explicitly designed to reduce the weight of outlier models within an MME. This can be achieved through bias correction and ridge regularization. In Shin et al. (2020) outlier models are defined as those that generate projections that are unusually close to the hydrological variable in observation data, which the authors attribute to excessive regional calibration to observations. To limit the influence of such models, they propose a hybrid method combining Bayesian weighting with bias correction. Ridge regularization, frequently used in ML context, is a form of linear regression that incorporates a penalty term to constrain variables with unusually strong linear correlations, thereby reducing the risk of overfitting. Labe and Barnes (2022) apply ridge regularization to limit the sensitivity of an artificial neural network to outlier influence.

3.4 Downscaling Techniques

Acquiring regional information on climate change is essential for impact, vulnerability, and adaptation studies. While CMIP GCMs are internationally established sources for climate projection data, their typical grid resolution of 100–250 km (Liang-Liang et al., 2022; Weigel et al., 2010) limits the ability to provide locally relevant information (Grose et al., 2023). Downscaling is therefore necessary, especially for regions with complex topography or localized climate phenomena (Wilby

and Fowler, 2010). Various downscaling techniques exist, including statistical downscaling (Gebrechorkos et al., 2023; Wootten et al., 2024), dynamical downscaling (Knutson et al., 2013; Tapiador et al., 2020), and novel machine-learning based approaches (Sachindra et al., 2018; Soares et al., 2024), each with its own strengths and limitations (Hall, 2014). These downscaling approaches differ from the regridding methods discussed in Subsection 3.5, which are purely mathematical interpolation techniques and do not introduce additional physical or statistical information.

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). Its reliability depends on the quality of observational data and on the assumption that calibrated relationships remain valid in a changing climate. Statistical downscaling is computationally efficient and can reduce the cool bias compared to the original CMIP simulations, as shown e.g. by Xu and Wang (2019) applying the Bias Correction and Spatial Downscaling (BCSD) technique for daily maximum temperature over China. However, it may struggle to represent non-linear processes or unprecedented extremes if these are not well captured in the historical record.

In dynamical downscaling, output from a GCM is used as boundary conditions for a RCM, which simulates climate on a limited-area domain and hence employs a finer resolution (Di Luca et al., 2015). Specifically, the CORDEX (Giorgi, 2019; Gutowski Jr. et al., 2016) initiative provides an international framework in which multiple institutions generate and evaluate such regional climate projections driven by GCM simulations. Dynamical downscaling can capture regional physical processes that GCMs cannot resolve (Giorgi and Gutowski, 2015). However, it is computationally demanding and depends on the availability of suitable RCMs. In addition, systematic biases in the GCM can propagate into, and potentially degrade, the regional simulations (Di Virgilio et al. 2022). One prominent example of the application of dynamic downscaling is the derivation of European Centre for Medium-Range Weather Forecasts Reanalysis 5 (ERA5) dataset (Xu et al., 2021). Another example for the Australian region is Grose et al. (2023) who used CMIP6 multimodel ensemble downscaling to provide accurate, scenario-based climate change projections. Liu et al. (2021) found that dynamical downscaling does not necessarily perform better compared to dynamic downscaling approaches.

ML-based downscaling methods have recently been applied to generate high-resolution projections from GCM simulations, leveraging their ability to capture complex and non-linear relationships. They can also efficiently process large datasets and integrate multiple variables, which has potential to improve downscaling results (Rampal et al., 2024). See Subsection 4.1 for details and examples.

3.5 Regridding Techniques

Each model produces output on its own underlying grid, often referred to as the ‘native’ grid. When combining models with different native grids into a MME, researchers must decide on whether to keep the native grids or to regrid their data to a

uniform grid. One option is to retain native grids and avoid regridding altogether. Analysing and visualization individual MME model results in the models's native grid is one way to accomplish this (Quesada et al., 2017). Another approach is to compute zonal means for each model and then average these, which allows results to be combined without regridding (Boysen, 2020). In practice, however, most MME studies regrid model output to a common grid to ensure spatial consistency prior to analysis. Regridding introduces several methodological choices related to both spatial and temporal dimensions, including the selection of a target grid resolution (coarser, intermediate, or finer) and type, the interpolation method, and the treatment of differing model calendars.

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. Regridding to a coarser grid can improve computational efficiency, and facilitate comparison across models, but may smooth spatial gradients and dampen localized extremes. Conversely, regridding to a finer grid can better preserve small-scale features and extremes, but does not add new physical information and may give a false impression of increased spatial detail. Many studies that mention regridding do not explain the direction of regridding or the rationale behind it (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 details of regridding. Where documentation on details is provided, different motivations are evident. 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. In many Python-based workflows, the default interpolation method is bilinear interpolation. This is suitable for many, but not all, variables depending on the type of analysis that is being carried out. Table 1 provides an overview of the available interpolation methods, which data types they should be applied to, and some examples of CMIP variables for each data type. Beyond the presented methods, additional interpolation methods exist (National Center for Atmospheric Research Staff 2014).

Table 1. 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

In addition to the variety of spatial resolutions present within an MME, temporal inconsistencies may also exist among members. These arise because models can use different calendar conventions when storing output in the commonly used netCDF format, which supports nearly ten calendar types (NetCDF Users Guide: NetCDF Utilities, 2025) and subsequent different encoded calendars in the modelling centers. The best choice of calendar for a given study will depend on the study particulars and researcher preference. Regardless of this choice, calendars should be brought into alignment during the regridding process to avoid inconsistencies in subsequent MME analyses.

4. Outlook

4.1 Machine Learning

With the rapid production and accumulation of climate data, automated and increasingly sophisticated analysis techniques have become essential (Rupe et al., 2017; Glymour et al., 2019). Recent advances in ML have been most pronounced at weather timescales, where large training datasets enable robust data-driven approaches. In contrast, applications to climate timescales are more challenging due to limited sample sizes, stronger non-stationarity, and a frequent occurrence of conditions outside the range of the training data, especially for extreme events. Despite these challenges, ML has emerged as a valuable tool for enhancing ensemble approaches in climate science (see Fig. 4).

Over the past 5-10 years, ML applications have demonstrated significant advantages in addressing non-linear, high-dimensional, and hierarchical problems (Li et al., 2021 and references therein). By utilizing observational data as either a reference, benchmark, or a constraint, ML offers significant potential to extract additional insights from MMEs. Approaches based on neural networks, causal inference, explainable artificial intelligence (XAI), and nonlinear multivariate emergent constraints have become increasingly competitive with traditional numerical, knowledge-based methods (see Fig. 5 and de Burgh-Day and Leeuwenburg, 2023; Eyring et al., 2024). These capabilities make ML particularly well-suited for identifying patterns and complex physical processes within climate model data, enabling a more comprehensive exploration of the valuable information embedded within the data (Reichstein et al., 2019; Wang et al., 2018). In short, ML has the potential to make climate models better and faster, while reducing their high energy consumption. Nevertheless, the application of ML algorithms in constructing MMEs for climate impact assessments remains at an early stage. Below, we provide an overview of emerging ML approaches for analyzing MMEs.

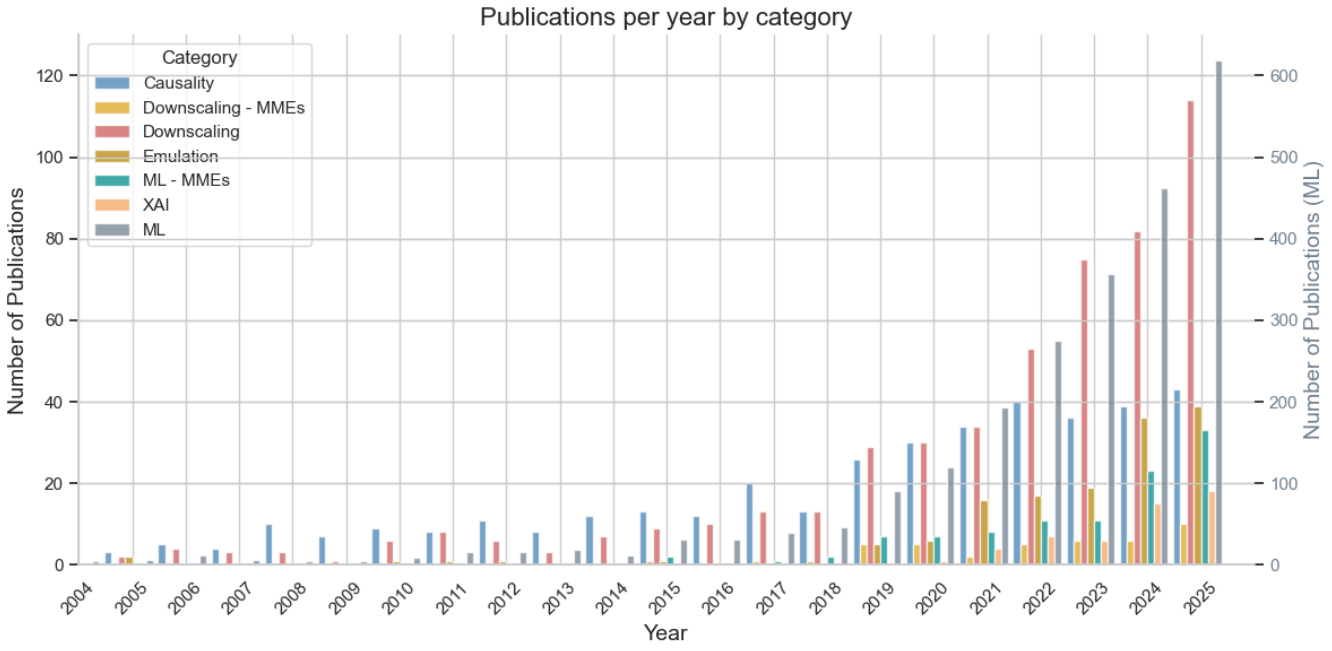


Figure 4. Number of publications per year involving different ML-related techniques in the context of climate modelling: ML (y-axis on the right), ML and MMEs, Downscaling, Downscaling and MMEs, Causality, Emulators, and Explainable AI (XAI). The data was extracted from the citation reports available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

Downscaling and Bias Correction

ESMs have horizontal resolutions often far coarser than those needed by decision makers and also exhibit substantial biases (Maraun et al., 2017). Recently, ML has been exploited to bias-correct and downscale MME's outputs–with both processes often done simultaneously. The source data are typically coarse-resolution outputs from climate models for historical periods, while the targets are high-resolution observational datasets, such as gridded products interpolated from gauge stations. It is a common practice to perform dimensionality reduction (e.g. performing principal component analysis on the raw data) before the training process. Multiple ML methods have been tested and compared, including random forests, support vector machines, relevance vector machines, and artificial neural networks (Crawford et al., 2019; Sachindra et al., 2018; Wang et al., 2018; Xu et al., 2020; Dey et al., 2022; Jose et al., 2022; Shetty et al., 2023; Zebarjadian et al., 2024; Li et al., 2021). The domain of these studies is generally limited to river basin scales (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. In many of these studies, it has been found that tree-based approaches (like random forests) commonly perform better than other algorithms, making them a strong 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. Training the algorithms with historical data can thus lead to projections becoming overly constrained. 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 of 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 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 minimise only one error metric ignoring the skill regarding other 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 neglects spatial relationships contained either within the inputs or the desired outputs. ML methods that account for spatial relationships could be of use, including convolutional neural networks (Gu et al., 2018; LeCun et al., 2015; Wang and Tian, 2022, 2024). Incorporating 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

Causal inference aims to identify the causal structure of complex systems such as the Earth and to quantify causal effects by combining domain knowledge, ML models, and data from observations and climate model simulations (Runge et al., 2023, and references therein). Because widely adopted methods based on simple descriptive statistics often fail to accurately capture the underlying physical mechanisms (Beven and Freer, 2001), structural causal models (SCMs) have become a well-established approach in statistics and ML for causal inference (Runge et al., 2019). An important resource in this area is the causality benchmark platform **causeme.net**, which aims to provide benchmarks with well-established causal structures (Runge et al., 2020). Process-oriented causal analysis has been applied across a broad range of topics, including 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), Atlantic–Pacific interactions (Karmouche et al., 2023) and subpolar gyre variability (Falkena and von der Heydt, 2024).

Building on this foundation, recent research has increasingly explored the integration of causal discovery with deep learning (DL), presenting 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 address biases and uncertainties associated with subgrid-scale processes, such as clouds and convection. Previous research has demonstrated DL's capability to effectively represent small-scale processes, 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 been criticised 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, thereby enhancing the reliability of DL algorithms. Their causally-informed approach generates

climate means and variability that closely match original simulations, while preventing spurious links typically seen in traditional DL-based parameterizations. 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 directly to multi-model analysis, their methodologies hold significant potential for future applications in this area.

Process-oriented causal model evaluation

Building on the identification of causal relationships, these frameworks also offer opportunities for systematically evaluating climate models. Detecting similar causal connections in observations and model simulations provides an opportunity to assess model performance that indicates whether models can correctly reproduce processes in the climate system. Such an evaluation framework was first introduced by Nowack et al. (2020) and was termed causal model evaluation (CME). To facilitate the comparison of causal relationships, the authors introduced a modified asymmetric F_1 score metric to classify the agreement between compared causal graphs. A similar approach was proposed by Vázquez-Patiño et al. (2020). Debeire et al. (2025) built their study upon the findings of Nowack et al. (2020) to address the practical challenges of integrating CME for CMIP6 MMEs projections. The authors adopted and adjusted the F_1 score definition and complemented it with a distance metric $1 - F_1$ with smaller distance values indicating greater similarity, both in terms of performance relative to the reference graph and in terms of dependence among the models. Based on this metrics, 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), which accounts for both model performance and interdependence of causal networks.

Ricard et al. (2024) employed a network-based approach, termed netCS, which leverages sea surface temperature (SST) variability and teleconnections to constrain Equilibrium Climate Sensitivity (ECS) and Transient Climate Response (TCR). The authors argue that the behavior of SST networks serves as a reliable proxy for how models respond to increased CO_2 concentrations. Their results show that some models capture regional SST distributions well but fail to replicate connectivity patterns, and vice versa. This distinction is crucial for evaluating model performance over historical periods, as models that realistically reproduce past SST patterns may exhibit more physically consistent behavior, even if they are not necessarily better tuned. While this does not guarantee that those models are superior for future projections (Rasp et al., 2018; Zhu and Poulsen, 2021), it provides valuable evidence. The authors further propose that causal networks, when used alongside traditional emergent constraints, offer a more reliable framework for ranking climate models in future climate projections.

Machine Learning for Climate System Emulation

Climate model emulators, including surrogate models, are simplified representations of the complex processes embedded in climate models, enabling faster computations and predictions. They mimic the behaviour of a climate model without explicitly solving the underlying equations. ML presents a unique opportunity to emulate components of the climate system through

novel and computationally efficient parameterizations. Such approaches have the potential to increase the efficiency of climate simulations while enabling higher resolution simulations (Eyring et al., 2021; Gentine et al., 2018). However, 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).

One notable initiative in this context 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 applied properly. Future studies could explore the use of MMEs as training data to develop novel ML-based emulators. Complementing the available data to train emulators, Lu and Ricciuto (2019) demonstrate an innovative approach integrating SVD, Bayesian optimization, and neural networks to create a computationally efficient surrogate model. Weber et al. (2020) provide technical notes of ML, using the example of forecasting precipitation under CO_2 forcing. The remarkable computational efficiency and ability of ML emulators to replicate complex climate processes with high precision demonstrates their potential. Nevertheless, several challenges remain, including the high computational cost of running ML models, limited diversity in training data, and the need for more robust methods to evaluate simulations.

Further promising Future ML Avenues

There are numerous promising avenues involving ML for the analysis and processing of CMIP outputs. Explainable AI (XAI), which aims to extract physical insight from otherwise black-box ML models, offers substantial potential for identifying physically meaningful changes in the Earth system simulated by CMIP models (Rader et al., 2022). For example, layer-wise relevance propagation (LRP), has been used to identify the spatial regions and input features that neural networks rely on when generating predictions (Toms et al., 2020) and has proven especially valuable for improving interpretability by visualizing relevance heatmaps (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. ML also offers effective tools for evaluating both the performance and independence of climate models within MMEs, further supporting the development of ensemble weighting metrics (Brunner and Sippel, 2023). These types of methods also support the development of process-oriented bias correction and downscaling methods for MMEs (Maraun et al., 2017). Furthermore, efforts to predict end-user-relevant variables that are not directly simulated by 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), demonstrate the potential of ML to enhance the applicability of MMEs for stakeholders and decision-makers. Given ML's growing role in improving climate projections, interpretability, and practical usability, AI-ready databases such as ClimateSet (Kaltenborn et al., 2023) represent valuable resources for the research community.

4.2 Single Model Initial Condition Large Ensembles (SMILEs)

For some activities and experiments, many models in CMIP5 and CMIP6 provide only one ensemble member (Milinski et al., 2020; Olonscheck and Notz, 2017). Consequently, modelling groups strive to provide their best performing members, carefully calibrated to the same internationally available observational datasets. This implicit incentive for modelling groups to add simulations to the CMIP MME that are less extreme can lead to a MME that underestimates the uncertainties. The outcome of this incentive is shown by findings from Sanderson et al. (2008) who found that the standard model performed comparatively to the best-performing model. To address this and other limitations of MMEs based on single-members from different modelling centers, an emerging approach (see Figure 5) is to include multiple simulations from individual models that differ only in their initial conditions (Maher et al., 2021). When there are at least 10 ensemble members, we refer to the ensembles as SMILE (Deser et al., 2020).

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), MMEs do not allow us to distinguish internal variability from the forced response. On the other hand, SMILEs allow us to quantify internal variability for any given future scenario, independent from model uncertainty (Deser et al., 2012b; Lehner et al., 2020). This capability is particularly powerful for regional detection and attribution studies and for the analysis of extreme climate events (Lehner et al., 2017; McKenna and Maycock, 2021; von Trentini et al., 2020; van der Wiel et al., 2021; Pérez-Carrasquilla et al., 2025).

While large ensemble simulations are well established as essential for studying extreme events (see also Subsection 3.2), they are equally relevant for the analysis of compound events that result from combinations of multiple weather and climate drivers (e.g. simultaneous drought and heatwave; Bevacqua et al., 2022, 2023; Wu et al., 2023). Such events are characterized by complex interactions between extreme conditions across variables, space, or time. Consequently, single variable based approaches may underestimate risks, and neglect the impact of the interplay between different variables. In this context, Bevacqua et al. (2023) showed that attributing compound events requires larger sample sizes than events based on extremes in a single variable, 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.

When multiple realizations are available, it is considered good practice to average all realizations from each model and incorporate these means into the MME. This approach is appropriate for applications focusing on long-term mean changes or the forced response, but is not appropriate for analyses of extremes or internal variability, where individual realizations carry relevant information. 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.

Some modelling centers have also applied SMILEs framework to partition the forced response into contributions from individual forcings (e.g. greenhouse gases, aerosols, biomass burning) with single-forcing large ensembles (SFLE; e.g. derived from CESM2 framework; Simpson et al., 2023). These ensembles change only one forcing at a time while holding all others fixed, thereby enabling attribution of the drivers underlying responses identified in all-forcing SMILEs.

One challenge to employing SMILEs is data accessibility. To address this, the Multi-Model Large Ensemble Archive (MMLEA) was developed (Deser et al., 2020). The newly published MMLEAv2 expands upon the original archive by including a larger number of models (18 compared to 7 previously) and more three-dimensional variables (Maher et al., 2025). Both the MMLEAv2 and a suite of corresponding observational datasets have been regridded onto a 2.5° common horizontal grid, reducing data volume and enabling straightforward model-to-model or model-to-observation comparisons. In addition, the release of MMLEAv2 is accompanied by the newest version of the CVDP (CVDPv6; Phillips et al., 2020), introduced in Subsection 2.5.

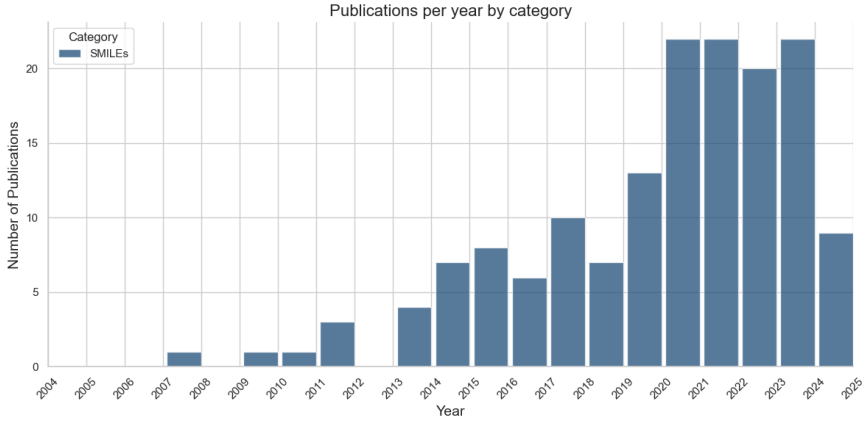


Figure 5. Number of publications per year involving SMILEs. The data was extracted from the citation report available at Web of Science (https://www.webofscience.com/wos/woscc/basic-search) for the queries provided in Appendix A.

For a variable with relatively low internal variability or high signal-to-noise ratio, 10 ensemble members can be used to sufficiently detect changes, e.g. for the global mean land temperature (Deser et al., 2012b). To robustly detect significant warming in the 2050s relative to the 2010s (at the 95% confidence level), Deser et al. (2012b) found that only a single ensemble member was required for nearly all locations. In contrast, precipitation exhibits substantially higher internal variability: detecting changes requires approximately 3–6 ensemble members in the tropics and high latitudes, while more than 15

ensemble members are needed in the mid-latitudes, with estimates reaching up to 40 members (Deser et al., 2012a, b). For sea level pressure, Deser et al. (2012b) reported that only 3-6 ensemble members are sufficient in the tropics, whereas 9-30 are required in the extratropics. The number of ensemble members required varies regionally, reflecting differences in both the strength of the forced signal and local internal variability (Bittner et al., 2016). Over the ocean, less SMILE members are required (Milinski et al., 2020). Table 2 provides a small sample of papers that have employed large ensembles for a variety of research questions.

Table 2. 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 snow trends	40-member CESM1-LE	(Mankin and Diffenbaugh, 2015)
Irreducible uncertainty	100-member MPI Grand Ensemble	(Marotzke, 2019)
Ocean ecosystem drivers	30-member GFDL Ensemble	(Rodgers et al., 2015)
Ocean carbon cycle	30-member GFDL Ensemble	(Schlunegger et al., 2019)
Weather regimes and their impact on surface extremes	100-member ensemble from CESM2-LE and 36-member ensemble from E3SM2-LE	(Pérez-Carrasquilla et al., 2025)

4.3 Computational Resources and Carbon Impact

MMEs, such as CMIP6, are powerful tools for exploring past, current, and future climate conditions, but they come with substantial computational and energy demands. MMEs typically rely on multiple simulations across different models or

multiple ensemble members of a single model, performed on high-performance computing (HPC) platforms. These execute calculations across many parallel cores and process and generate large amounts of data that require careful management and optimization. Simulating a century-scale global climate model with high spatial and temporal resolutions can take weeks, even on HPC systems. For example, the MPI-ESM1.2 model in its standard low-resolution configuration (approximately 200 km grid spacing), achieves between roughly 45 and 85 simulated years per day, representing 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 ~1 km horizontal resolution reaches a performance of only about 0.043 simulated years per day (~15.7 simulated days per day) (Fuhrer et al., 2018). Computational performance therefore represents a major constraint in the design of ESM experiments, requiring trade-offs between resolution, complexity, and the size of ensembles.

CPMIP metrics for Climate Modelling

Balaji et al. (2017) introduced a universal set of metrics to evaluate HPC and ESM performance, emphasising that traditional metrics (e.g., floating point operations per second) are no longer sufficient to characterize newer generations of computing architectures and the diverse structures of modern ESMs. Given the increasing complexity of ESM components and their heterogeneous computational characteristics, they advocated adopting these metrics (Table 3) as a standard within globally coordinated modelling initiatives and proposed their collection through the Computational Performance MIP (CPMIP). These metrics provide a consistent basis for assessing technological advances in climate models, are accessible from routine production runs, and capture efficiency across the entire modelling lifecycle.

Table 3. List of metrics introduced in CPMIP, 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 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

Addressing the performance characteristics of individual models within a MME can contribute to a more balanced and efficient use of computational resources. Building on the foundational CPMIP framework, Acosta et al. (2024) extended this work by incorporating empirical data from CMIP6, collected during long, real-time model runs across 14 institutions, encompassing 33 experiments and nearly 500,000 years of simulations. The study places particular emphasis on energy consumption, data storage demands, and operational efficiency, and provides strategic recommendations for improving the sustainability and performance of future climate modelling efforts.

The demand for computing power is steadily increasing due to several factors, including higher spatial and temporal resolution, the explicit representation of complex climate processes that were previously parameterized, increasing ensemble sizes, and the associated growth in data storage requirements for both input and output. Improving model accuracy through these processes leads to more detailed spatial and temporal outputs, but requires 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.

Kilometer-scale simulations of individual models and associated MMEs are being actively developed (Ban et al., 2021; Coppola et al., 2020; Pichelli et al., 2021; Rackow et al., 2025), as well as coordinated intercomparisons for global storm-resolving models (GSRM). To cope with the high computational and energy demands required for increase in resolution and process detail (Schär et al., 2020), such simulations are often conducted over limited regional domains (Coppola et al., 2020; Nolan and Flanagan, 2020), rely on simplified parameterizations (for processes such as radiation or soil interactions) or—when performed globally—are restricted to relatively short simulation periods of only a few weeks (Schär et al., 2020). However, this limitation is rapidly being overcome, with multi-year global simulations at such resolutions have already been 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), or the IFS model coupled to the Finite-volumE Sea ice-Ocean Model (Rackow et al., 2025).

Beyond the role of resolution and resolved processes, the internal structure of ESMs also plays a crucial role in determining computational performance. ESMs are structured with a component-based architecture, allowing scientists to update or add new components over time. While this architecture enables continuous flexibility, it also brings software engineering challenges. Modifications to individual components can alter computational demands and affect data processing, input/output (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.

Operational factors such as job scheduling also influence computational efficiency. Queue times significantly impact overall execution speed and efficiency, although they vary across different institutions (Acosta et al., 2024). Short and consistent queue times are beneficial for MMEs, as they help ensure the timely completion of simulations. Reducing queue times improves the effective use of available HPC resources, allowing more simulations to be completed within a given timeframe and thereby increasing the overall throughput of the ensemble.

Carbon Footprint of Climate Modeling: Towards "greener" Hardware

Running climate models on HPCs, particularly for large-scale MMEs, requires substantial energy and is associated with a significant carbon footprint. The climate modelling community is aware of this and is exploring ways to optimize code efficiency and transition to greener energy sources to minimise the carbon impact of their research efforts. In this context, CPMIP focuses on capturing the real energy costs of running models, aiming to help climate scientists make eco-friendly decisions in computing. Using 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), the total computational energy costs of climate experiments have been assessed, enabling estimates of the associated carbon footprint (Acosta et al., 2024). For 8 out of 49 institutions that were involved in CMIP6, the total estimated emissions amount to 1,692 t CO_2 (with total energy costs ranging from 0.41 TJ to 26.70 TJ). For context, the International Energy Agency (IEA) reports that the “global average energy-related carbon footprint” is approximately 4.7 t CO_2 per person and per year. Thus, the total emissions from this share of CMIP6 modelling centers are roughly equivalent to the annual energy-related emissions of 360 people.

As researchers recognize the environmental impact of extensive model runs, eco-friendly hardware is becoming an increasingly important consideration in HPC for climate modelling. One example of this good practice is the Energy-efficient climate simulations 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 while reducing the energy requirements of the ESM ICON (Adamidis et al., 2025). By encouraging institutions to collect the data needed to estimate their carbon footprint, to adopt eco-friendly hardware and to implement thoughtful modelling practices, the climate modelling community can reduce its carbon impact while continuing to advance its scientific mission.

HPC Facilities: petascale and beyond

As climate models continue to evolve, HPC facilities operating at the petascale and beyond are necessary. Looking ahead, exascale computing systems, capable of achieving 10^{18} floating-point operations per second, promise to greatly expand modelling capabilities, enabling longer, higher-resolution simulations, more complex process representation, and improved exploration of predictability limits in ESMs. This potential 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).

In the context of the increasing computational complexity, the findings from the CPMIP and performance metrics applied to CMIP6 experiments underline the need for better optimization of model configurations, improved coupling mechanisms, and more efficient use of HPC resources. The intercomparison between models and institutions reveals significant differences in computational costs, highlighting the need and potential for strategic advancements. Coordinated efforts are also required to integrate the latest technological advances, as e.g. outlined in Subsections 4.1 and 4.2. Standardized measurements of computational and energy costs can guide the path forward, ensuring that model performance is comparable, and thereby allowing researchers to identify areas for improvement and make informed decisions in the development.

5. Concluding Remarks

Climate modelling has been key to the understanding of past, present, and future climate change. 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, 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 SMILEs, 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 modelling 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 also want to specifically thank the two anonymous reviewers that contributed to a substantially improved version of this manuscript. We acknowledge Josh Dorrington for initial help with this project. 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: KG; Subsection 2.3: KG, PP; Subsection 2.4: JSPC, MT; Subsection 2.5: NČ, EG, MT; Subsection 3.1: AK; Subsection 3.2: MT; Subsection 3.3: CL; Subsection 3.4: PP; Subsection 3.5. CL; Subsection 4.1: EG, KG, JSPC; Section 4.2: AVC, MT, AK; Subsection 4.3: MT; Conclusion: AK; Appendix B: NČ; Appendix D: JSPC). Final details will be provided with publication.

References

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.

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.

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.: 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.

Becker, E., Kirtman, B. P., and Pegion, K.: Evolution of the North American Multi-Model Ensemble, Geophys. Res. Lett., 47, e2020GL087408, https://doi.org/10.1029/2020GL087408, 2020.

Becker, E. J., Kirtman, B. P., L'Heureux, M., Muñoz, Á. G., and Pegion, K.: A Decade of the North American Multimodel Ensemble (NMME): Research, Application, and Future Directions, Bull. Am. Meteorol. Soc., 103, E973–E995, https://doi.org/10.1175/BAMS-D-20-0327.1, 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.

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.

Buontempo, C., Burgess, S. N., Dee, D., Pinty, B., Thépaut, J.-N., Rixen, M., Almond, S., Armstrong, D., Brookshaw, A., Alos, A. L., Bell, B., Bergeron, C., Cagnazzo, C., Comyn-Platt, E., Damasio-Da-Costa, E., Guillory, A., Hersbach, H., Horányi, A., Nicolas, J., Obregon, A., Ramos, E. P., Raoult, B., Muñoz-Sabater, J., Simmons, A., Soci, C., Suttie, M., Vamborg, F., Varndell, J., Vermoote, S., Yang, X., and Garcés De Marcilla, J.: The Copernicus Climate Change Service: Climate Science in Action, Bull. Am. Meteorol. Soc., 103, E2669–E2687, https://doi.org/10.1175/BAMS-D-21-0315.1, 2022.

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.

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.

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.

Doblas-Reyes, F. J., Hagedorn, R., and Palmer, T. N.: The rationale behind the success of multi-model ensembles in seasonal forecasting – II. Calibration and combination, Tellus Dyn. Meteorol. Oceanogr., 57, 234, https://doi.org/10.3402/tellusa.v57i3.14658, 2005.

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., 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.

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%253C1962:ATAMIP%253E2.0.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.

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.

Giorgi, F. and Mearns, L. O.: Calculation of Average, Uncertainty Range, and Reliability of Regional Climate Changes from AOGCM Simulations via the "Reliability Ensemble Averaging" (REA) Method, J. Clim., 15, 1141–1158, https://doi.org/10.1175/1520-0442(2002)015%253C1141:COAURA%253E2.0.CO;2, 2002.

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.

Hagedorn, R., Doblas-Reyes, F. J., and Palmer, T. N.: The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept, Tellus Dyn. Meteorol. Oceanogr., 57, 219, https://doi.org/10.3402/tellusa.v57i3.14657, 2005.

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.

Hasselmann, K.: Stochastic climate models Part I. Theory, Tellus, 28, 473–485, https://doi.org/10.3402/tellusa.v28i6.11316, 1976.

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.

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.

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.

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.

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., 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.

Kirtman, B. P., Min, D., Infanti, J. M., Kinter, J. L., Paolino, D. A., Zhang, Q., Van Den Dool, H., Saha, S., Mendez, M. P., Becker, E., Peng, P., Tripp, P., Huang, J., DeWitt, D. G., Tippett, M. K., Barnston, A. G., Li, S., Rosati, A., Schubert, S. D., Rienecker, M., Suarez, M., Li, Z. E., Marshak, J., Lim, Y.-K., Tribbia, J., Pegion, K., Merryfield, W. J., Denis, B., and Wood, E. F.: The North American Multimodel Ensemble: Phase-1 Seasonal-to-Interannual Prediction; Phase-2 toward

Developing Intraseasonal Prediction, Bull. Am. Meteorol. Soc., 95, 585–601, https://doi.org/10.1175/BAMS-D-12-00050.1, 2014.

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.

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., 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.

Maher, N., Phillips, A. S., Deser, C., Wills, R. C. J., Lehner, F., Fasullo, J., Caron, J. M., Brunner, L., Beyerle, U., and Jeffree, J.: The updated Multi-Model Large Ensemble Archive and the Climate Variability Diagnostics Package: new tools for the study of climate variability and change, Geosci. Model Dev., 18, 6341–6365, https://doi.org/10.5194/gmd-18-6341-2025, 2025.

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.

Manabe, S. and Bryan, K.: Climate Calculations with a Combined Ocean-Atmosphere Model, J. Atmospheric Sci., 26, 786–789, https://doi.org/10.1175/1520-0469(1969)026%253C0786:CCWACO%253E2.0.CO;2, 1969.

Manabe, S. and Strickler, R. F.: Thermal Equilibrium of the Atmosphere with a Convective Adjustment, J. Atmospheric Sci., 21, 361–385, https://doi.org/10.1175/1520-0469(1964)021%253C0361:TEOTAW%253E2.0.CO;2, 1964.

Manabe, S. and Wetherald, R. T.: Thermal Equilibrium of the Atmosphere with a Given Distribution of Relative Humidity,

J. Atmospheric Sci., 24, 241–259, https://doi.org/10.1175/1520-0469(1967)024%253C0241:TEOTAW%253E2.0.CO;2, 1967.

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., 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.

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.

Min, Y., Lim, C., Yoo, J., Kim, H., Kryjov, V. N., Jeong, D., Lim, A., Ham, S., Chen, M., Xiao, Y., Gagnon, N., Muncaster, R., Liu, P., Borrelli, A., Ji, H., Lee, J., Jo, S., Kiktev, D., Tolstykh, M., Matyugin, V., McLean, P., and Molod, A. M.: A Diachronic Assessment of Advances in Seasonal Forecasting: Evolution of the APCC Multi-Model Ensemble Prediction System Over the Last Two Decades, Geophys. Res. Lett., 52, e2025GL116416, https://doi.org/10.1029/2025GL116416, 2025.

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.

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., 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.

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.

Pérez-Carrasquilla, J. S., Molina, M. J., Mayer, K. J., Dagon, K., Fasullo, J. T., & Simpson, I. R.: Observed and modeled amplification of the frequency, duration, and extreme heat impacts of the Pacific trough regime. Earth's Future, 13(12), e2025EF007140, 2025.

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.

Planton, Y. Y., Guilyardi, E., Wittenberg, A. T., Lee, J., Gleckler, P. J., Bayr, T., McGregor, S., McPhaden, M. J., Power, S., Roehrig, R., Vialard, J., and Voldoire, A.: Evaluating Climate Models with the CLIVAR 2020 ENSO Metrics Package, Bull. Am. Meteorol. Soc., 102, E193–E217, https://doi.org/10.1175/BAMS-D-19-0337.1, 2021.

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%253C3212:APADMA%253E2.0.CO;2,

2001.

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.

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.

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

Sippel, S., Kent, E. C., Meinshausen, N., Chan, D., Kadow, C., Neukom, R., Fischer, E. M., Humphrey, V., Rohde, R., De Vries, I., and Knutti, R.: Early-twentieth-century cold bias in ocean surface temperature observations, Nature, 635, 618–624, https://doi.org/10.1038/s41586-024-08230-1, 2024.

Smith, D. M., Scaife, A. A., Boer, G. J., Caian, M., Doblas-Reyes, F. J., Guemas, V., Hawkins, E., Hazeleger, W., Hermanson, L., Ho, C. K., Ishii, M., Kharin, V., Kimoto, M., Kirtman, B., Lean, J., Matei, D., Merryfield, W. J., Müller, W. A., Pohlmann, H., Rosati, A., Wouters, B., and Wyser, K.: Real-time multi-model decadal climate predictions, Clim. Dyn., 41, 2875–2888, https://doi.org/10.1007/s00382-012-1600-0, 2013.

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.

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.

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, 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., 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.

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.

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.

van Vuuren, D., O'Neill, B., Tebaldi, C., Chini, L., Friedlingstein, P., Hasegawa, T., Riahi, K., Sanderson, B., Govindasamy, B., Bauer, N., Eyring, V., Fall, C., Frieler, K., Gidden, M., Gohar, L., Jones, A., King, A., Knutti, R., Kriegler, E., Lawrence, P., Lennard, C., Lowe, J., Mathison, C., Mehmood, S., Prado, L., Zhang, Q., Rose, S., Ruane, A., Schleussner, C.-F., Seferian, R., Sillmann, J., Smith, C., Sörensson, A., Panickal, S., Tachiiri, K., Vaughan, N., Vishwanathan, S., Yokohata, T., and Ziehn, T.: The Scenario Model Intercomparison Project for CMIP7 (ScenarioMIP-CMIP7), https://doi.org/10.5194/egusphere-2024-3765, 30 January 2025.

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, 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.

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.

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.

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, 462–470, 2013.

Appendix

A 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")

B Systematic Model Biases

Some systematic biases are present in the vast majority of CMIP models at the global and regional scale, and some 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 their origins and consequences of these systematic model biases. Those are as follows: a) General evaluation, b) Sea surface temperature (SST) and ocean model biases, c) The Intertropical Convergence Zone (ITCZ) bias, d) biases in extratropical cyclones, e) Marine tropical/subtropical low cloud biases, f*)* biases in the cryosphere, *g)* biases in extremes. 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 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 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 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.

C Further Examples of process-based Evaluation

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).

The teleconnection between the Indian Summer Monsoon (ISM) and the El Niño–Southern Oscillation (ENSO) serves as a further example, which is well captured by MMEs of CMIP5 and CMIP6 models (Roy and Tedeschi, 2016; Roy et al., 2017). The teleconnection is strongest over central northeast India (Roy et al., 2017), where El Niño events are associated with a significant rainfall deficit, while La Niña events lead to a significant rainfall excess. In these studies, the MME is constructed using a simple mean ("one-model-one-vote") approach. Similar results are obtained when the MME is restricted to a subset of well-performing models, as identified for the ISM by Jourdain et al. (2013). Precipitation anomalies associated with different ENSO phases are reproduced well by most individual models and by the MME, in agreement with observations (see Roy et al., 2017 for details). Furthermore, the model ensemble analysis of ISM precipitation and Pacific SSTs reveals a clear linkage between the Walker circulation and the ISM over central northeast India, consistent with observations. This region represents the meeting point of the Hadley and Walker circulations during the ISM season, and the associated coupling and teleconnection processes appear to be well represented in most CMIP models and in the MME. This process-based understanding helps explain why the ENSO–ISM teleconnection is robustly captured in climate models.

Ahmed and Neelin (2021) utilised the observed relationship between tropical precipitation and buoyancy as the basis for a process-oriented analysis of CMIP6 models. They quantitatively assessed the thermodynamic sensitivity of convection across models using regime-oriented diagnostics. Their results showed that several models exhibit excessive moisture sensitivity, potentially arising from underactive convective schemes or tuning assumptions. Consequently, these models tend to produce mean precipitation states that are biased towards grid-scale saturation.

D Storylines for understanding Uncertainty

Considering this, uncertainty in climate projections can be communicated through climate storylines (Shepherd et al., 2018), which emphasizes exploring and understanding physically plausible events or pathways. The storyline approach differs from traditional methods of uncertainty evaluation in climate models in that it does not assume that model spread adequately represents uncertainty. This assumption may not hold for dynamically driven climate phenomena, where MME means may obscure 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). Instead of quantifying the likelihood of events, storylines focus on the physical drivers and interactions that make an event possible (Shepherd et al., 2018), constructing a causal network and conditioning on specific 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).

References

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.

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. Hydrometeor, 4, 1147–1167, https://doi.org/10.1175/1525-7541(2003)004%253C1147:TVGPCP%253E2.0.CO;2, 2003.

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

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

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.

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

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, JGR Atmospheres, 125, e2019JD032321, https://doi.org/10.1029/2019JD032321, 2020.

Cesana, G. V. and Del Genio, A. D.: Observational constraint on cloud feedbacks suggests moderate climate sensitivity, Nat. Clim. Chang., 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.

Chang, E. K. M., Guo, Y., and Xia, X.: CMIP5 multimodel ensemble projection of storm track change under global warming, J. Geophys. Res., 117, 2012JD018578, https://doi.org/10.1029/2012JD018578, 2012.

Clark, P. A. and Gray, S. L.: Sting jets in extratropical cyclones: a review, Quart J Royal Meteoro Soc, 146, 1065–1065, https://doi.org/10.1002/qj.3761, 2020.

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, JGR Atmospheres, 128, e2022JD038437, https://doi.org/10.1029/2022JD038437, 2023.

Dacre, H. F.: A review of extratropical cyclones: observations and conceptual models over the past 100 years, Weather, 75, 4–7, https://doi.org/10.1002/wea.3653, 2020.

Dutheil, C., Lengaigne, M., Bador, M., Vialard, J., Lefèvre, J., Jourdain, N. C., Jullien, S., Peltier, A., Sultan, B., and Menkès, C.: Impact of projected sea surface temperature biases on tropical cyclones projections in the South Pacific, Sci Rep, 10, 4838, https://doi.org/10.1038/s41598-020-61570-6, 2020.

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.

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, Journal of Climate, 37, 4257–4274, https://doi.org/10.1175/JCLI-D-23-0452.1, 2024.

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, Journal of Hydrometeorology, 8, 38–55, https://doi.org/10.1175/JHM560.1, 2007.

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.

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.

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, Climate Risk Management, 40, 100496, https://doi.org/10.1016/j.crm.2023.100496, 2023.

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

Luo, F., Ying, J., Liu, T., and Chen, D.: Origins of Southern Ocean warm sea surface temperature bias in CMIP6 models, npj Clim Atmos Sci, 6, 1–8, https://doi.org/10.1038/s41612-023-00456-6, 2023.

Nam, C., Bony, S., Dufresne, J. -L., and Chepfer, H.: The 'too few, too bright' tropical low-cloud problem in CMIP5 models, Geophysical Research Letters, 39, 2012GL053421, https://doi.org/10.1029/2012GL053421, 2012.

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

Priestley, M. D. K., Ackerley, D., Catto, J. L., and Hodges, K. I.: Drivers of Biases in the CMIP6 Extratropical Storm Tracks. Part I: Northern Hemisphere, Journal of Climate, 36, 1451–1467, https://doi.org/10.1175/JCLI-D-20-0976.1, 2023a.

Priestley, M. D. K., Ackerley, D., Catto, J. L., and Hodges, K. I.: Drivers of Biases in the CMIP6 Extratropical Storm Tracks. Part II: Southern Hemisphere, Journal of Climate, 36, 1469–1486, https://doi.org/10.1175/JCLI-D-20-0977.1, 2023b.

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., Dörr, J., Holmes, C. R., Massonnet, F., Blockley, E. W., Notz, D., Rackow, T., Raphael, M. N., O'Farrell, S. P., Bailey, D. A., and Bitz, C. M.: Antarctic Sea Ice Area in CMIP6, Geophysical Research Letters, 47, e2019GL086729, https://doi.org/10.1029/2019GL086729, 2020.

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, Intl Journal of Climatology, 37, 1794–1813, https://doi.org/10.1002/joc.4811, 2017.

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.

Schultz, D. M., Bosart, L. F., Colle, B. A., Davies, H. C., Dearden, C., Keyser, D., Martius, O., Roebber, P. J., Steenburgh, W. J., Volkert, H., and Winters, A. C.: Extratropical Cyclones: A Century of Research on Meteorology's Centerpiece, Meteorological Monographs, 59, 16.1-16.56, https://doi.org/10.1175/AMSMONOGRAPHS-D-18-0015.1, 2019.

Sen Gupta, A., Santoso, A., Taschetto, A. S., Ummenhofer, C. C., Trevena, J., and England, M. H.: Projected Changes to the Southern Hemisphere Ocean and Sea Ice in the IPCC AR4 Climate Models, Journal of Climate, 22, 3047–3078, https://doi.org/10.1175/2008JCLI2827.1, 2009.

Shepherd, T. G.: Storyline approach to the construction of regional climate change information, Proc. R. Soc. A., 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, Climatic Change, 151, 555–571, https://doi.org/10.1007/s10584-018-2317-9, 2018.

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.

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

Van Westen, R. M., Kliphuis, M., and Dijkstra, H. A.: Physics-based early warning signal shows that AMOC is on tipping course, Sci. Adv., 10, eadk1189, https://doi.org/10.1126/sciadv.adk1189, 2024.

Waliser, D. E. and Gautier, C.: A Satellite-derived Climatology of the ITCZ, Journal of Climate, 6, 2162–2174, https://doi.org/10.1175/1520-0442(1993)006%253C2162:ASDCOT%253E2.0.CO;2, 1993.

Wang, C., Zhang, L., Lee, S.-K., Wu, L., and Mechoso, C. R.: A global perspective on CMIP5 climate model biases, Nature Clim Change, 4, 201–205, https://doi.org/10.1038/nclimate2118, 2014.

Wang, S., Sankaran, S., and Perdikaris, P.: Respecting causality for training physics-informed neural networks, Computer Methods in Applied Mechanics and Engineering, 421, 116813, https://doi.org/10.1016/j.cma.2024.116813, 2024.

van Westen, R. M. and Dijkstra, H. A.: Persistent climate model biases in the Atlantic Ocean's freshwater transport, Ocean Science, 20, 549–567, https://doi.org/10.5194/os-20-549-2024, 2024.

Wills, R. C. J., Dong, Y., Proistosecu, C., Armour, K. C., and Battisti, D. S.: Systematic Climate Model Biases in the Large-Scale Patterns of Recent Sea-Surface Temperature and Sea-Level Pressure Change, Geophysical Research Letters, 49, e2022GL100011, https://doi.org/10.1029/2022GL100011, 2022.

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, Geophysical Research Letters, 44, 1520–1527, https://doi.org/10.1002/2016GL071992, 2017.

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