Constraining uncertainty in projected precipitation over land with causal discovery

Kevin Debeire[1], Lisa Bock[1], Peer Nowack[2,3], Jakob Runge[4,5], and Veronika Eyring[1,6]

[1]Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen,Germany
[2]Institute of Theoretical Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany
[3]Institute of Meteorology and Climate Research, Karlsruhe Institute of Technology, Karlsruhe, Germany
[4]Technische Universität Berlin, Faculty of Electrical Engineering and Computer Science, Berlin, Germany
[5]Technische Universität Dresden, Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Dresden/Leipzig, Germany
[6]University of Bremen, Institute of Environmental Physics (IUP), Bremen, Germany

Correspondence: Kevin Debeire (kevin.debeire@dlr.de)

Abstract. Accurately projecting future precipitation patterns over land is crucial for understanding climate change and developing effective mitigation and adaptation strategies. However, projections of precipitation changes in state-of-the-art climate models still exhibit considerable uncertainty, in particular over vulnerable and populated land areas. This study aims to address this challenge by introducing a novel methodology for constraining climate model precipitation projections with causal discovery. Our approach involves a multistep procedure that integrates dimension reduction, causal network estimation, causal network evaluation, and a causal weighting scheme which is based on the historical performance (the distance of the causal network of a model to the causal network of a reanalysis dataset) and the interdependence of Coupled Model Intercomparison Project Phase 6 (CMIP6) models (the distance of the causal network of a model to the causal network of other climate models). To uncover the significant causal pathways crucial for understanding dynamical interactions in the climate models and reanalysis datasets, we estimate the time-lagged causal relationships using the PCMCI causal discovery algorithm. In the last step, a novel causal weighting scheme is introduced, assigning weights based on the performance and interdependence of the CMIP6 models' causal networks. For the end-of-century period 2081-2100, our method reduces the very likely ranges (5-95 percentile) of projected precipitation changes over land between 10 and 16 % relative to the unweighted ranges across three global warming scenarios (SSP2-4.5, SSP3-7.0 and SSP5-8.5). The sizes of the likely ranges (17-83 percentile) are further reduced between 16 and 41 %. This methodology is not limited to precipitation over land and can be applied to other climate variables, supporting better mitigation and adaptation strategies to tackle climate change.

1 Introduction

Global mean precipitation and evaporation are expected to rise with warming by approximately 2-3 % per °C, driven by increased atmospheric water vapor according to thermodynamics (Allan et al., 2020). Although recent observations have struggled to detect a response of global precipitation to the current warming level, new research has demonstrated that precipitation variability has already increased globally over the past century (Zhang et al., 2024). The Coupled Model Intercomparison

Project Phase 6 (CMIP6) models represent the latest generation of climate models used to simulate past, present, and future climate conditions, providing vital projections to inform policy and adaptation strategies (Eyring et al., 2016). However, a significant challenge associated with these models is the large uncertainty range in land precipitation projections, reflecting the complex nature of precipitation processes and their representation in climate models (Tebaldi et al., 2021). Studies have shown that uncertainty in climate projections can be attributed to multiple factors, including, e.g., model structure, parameterization, and internal variability (Hawkins and Sutton, 2009). Model uncertainty is commonly assessed as the range of values projected by different climate models for a given future scenario (also known as intermodel spread). According to the Intergovernmental Panel on Climate Change (IPCC) Sixth Assessment Report (Douville et al., 2021), the average projected precipitation rates over land increases by 2.4 % in the low-emission scenario by 2081-2100 (with the 17-83 percentile range varying from -0.2 % to +4.7 %) relative to the period 1995-2014. In comparison, the very high-emission scenario shows a more substantial increase of 8.3 % (with the 17-83 percentile range varying from 0.9 % to 12.9 %) by 2081-2100. Reducing intermodel spread in precipitation projections is crucial for enhancing the reliability of climate projections.

These changes in future precipitation patterns have profound implications for various sectors, including natural and human systems (IPCC, Seneviratne et al., 2021). Kotz et al. (2022) discuss the extensive economic impacts associated with precipitation shifts, emphasizing the need for precise and reliable projections. The economic consequences of climate change, particularly in regions vulnerable to precipitation changes, underscore the urgency of reducing the uncertainty in these projections. Accurate projections are therefore critical for developing effective adaptation and mitigation strategies to minimize these negative impacts and enhance resilience.

To reduce the intermodel spread of future climate projections, a common method is the usage of an emergent constraint (Hall and Qu, 2006; Eyring et al., 2019). An emergent constraint identifies a statistically significant relationship between a constrained observable and a future climate variable. This observable can be a trend or variation observed during the historical period and includes metrics such as temperature variability (Cox et al., 2018) and shortwave low cloud feedback (Qu et al., 2018). The future climate variable often relates to key climate sensitivity metrics such as Equilibrium Climate Sensitivity (ECS) and Transient Climate Response (TCR) (Nijsse et al., 2020; Schlund et al., 2020b). By establishing a robust statistical relationship and combining it with observed data, the probability distribution of ECS and TCR can be constrained, leading to a narrower range in future climate projections.

However, when it comes to global precipitation and precipitation over land, the emergent constraints tend to be weaker compared to those for TCR or ECS (Ferguglia et al., 2023). Previous studies attribute this to the complexity of precipitation processes, model parameterizations, and observational constraints (Ferguglia et al., 2023). For example, complex atmospheric processes affecting precipitation, including aerosol impacts on cloud microphysics (Allen and Ingram, 2002; Beydoun and Hoose, 2019), convection, and large-scale circulation, are challenging to model accurately, leading to larger uncertainties. Furthermore, climate models use different parameterizations for subgrid-scale processes such as cloud formation, contributing to the spread. In addition, precipitation observations are often limited and carry substantial uncertainties (Trenberth and Zhang, 2018), weakening the relationships between historical predictors and future projections. The robustness of emergent constraints also depends on the specific ensemble of models used. For example, the constraints identified in the previous generation of models, CMIP5, may not hold in CMIP6 (Pendergrass, 2020; Schlund et al., 2020b). Building on this understanding, Shiogama et al. (2022) investigated emergent constraints related to future global precipitation changes using past temperature and precipitation trends. They revised the upper bound (95th percentile) of global precipitation change for 2051–2100 under a medium greenhouse gas concentration scenario from 6.2 % change to a range of 5.2–5.7 %. Additionally, other studies have also explored constraining future precipitation projections using observational data and past warming trends. Thackeray et al. (2022) developed an emergent constraint to reduce the uncertainty in projections of future heavy rainfall occurrence. Dai et al. (2024) propose an emergent constraint that utilizes past observational warming trends to constrain future projections of mean and extreme precipitation on both global and regional scales. They constrained the projected globally averaged mean precipitation fractional changes under the high-emission scenario for the 2081–2100 period relative to 1981–2014, reducing the average estimate from 6.9 % to 5.2 % and narrowing the 5–95 % range from $[3.0-10.9]$ % to $[1.9-8.5]$ %.

Other methods have been developed to constrain future climate projections. For instance, Schlund et al. (2020a) employed a machine learning regression approach known as Gradient Boosted Regression Tree (GBRT) on historical climate data to reduce the uncertainty range of future projections of Gross Primary Production (GPP). Another key method to address the intermodel spread is multimodel weighting based on model performance and interdependence. This addresses the issues in the commonly used "model democracy" approach, used in the IPCC Sixth Assessment Report (IPCC, Lee et al., 2021), where each climate model is given equal weight regardless of its performance and interdependence with other models. Equal weighting can lead to significant issues, such as overrepresenting similar models and ignoring differences in model performance (IPCC, Doblas-Reyes et al., 2021). The methodology introduced by Knutti et al. (2017) and further explored by Brunner et al. (2020) evaluates the historical model performance and interdependence based on several diagnostics and applies weights to combine the model outputs. This technique refines ensemble projections by prioritizing models that more accurately simulate historical climate conditions and accounts for redundancy among models.

These previous studies highlight the importance of using advanced statistical techniques and observational data. Notably, Nowack et al. (2020) introduced a causal model evaluation framework which assesses models based on their ability to capture cause-and-effect relationships within the system. In particular, Nowack et al. (2020) applied a causal discovery algorithm to sea level pressure (SLP) data from CMIP5 simulations and meteorological reanalyses. They constructed causal networks, referred to as *fingerprints*, to conduct a process-oriented evaluation of the models. Interestingly, models with fingerprints closer to observations better reproduced precipitation patterns over various regions, including South Asia, Africa, East Asia, Europe, and North America. These findings highlight the potential of causal model evaluation to address uncertainties in climate projections but have not yet been applied as such.

Nowack et al. (2020) also underscores the role of using SLP components as proxies for modes of variability to better understand precipitation patterns. Furthermore, we emphasize the strong connection between dynamical interactions imprinted in SLP fields and precipitation patterns. Numerous studies (e.g., Lavers et al., 2013; Thompson and Green, 2004; Müller-Plath et al., 2022) revealed how large-scale pressure variations, such as the North Atlantic Oscillation (NAO), the Azores high, the Arctic oscillation and the North Sea Caspian pattern, can influence precipitation variability across Europe and the Mediterranean basin. Dia-Diop et al. (2021) have shown that SLP anomalies over specific areas such as the Azores and St. Helena

High are interconnected with monthly mean precipitation in West Africa, indicating a relationship between SLP and rainfall. Furthermore, Benestad et al. (2007) demonstrated the importance of statistical models that use SLP to predict interannual variations in rainfall, revealing the connection between these variables. Costa-Cabral et al. (2016) confirmed the importance of large-scale climate indices, particularly the North Pacific High (NPH) wintertime anomaly, in predicting precipitation variability in Northern California. These studies support the broader applicability of SLP indices to understand precipitation patterns.

A research gap lies in the need to explore new methods, such as causal model evaluation, to more accurately assess the performance of climate models. Combining these advanced evaluation techniques with multimodel weighting schemes promises to reduce the uncertainty of climate projections. The goal of this study is to explore causal discovery for evaluating climate models and reducing the uncertainty of their projections, particularly for precipitation over land. We further demonstrate the application of causal approaches in capturing complex climate dynamics. Additionally, we address the practical challenges of integrating causal model evaluation with multimodel weighting. As such, this research will ultimately help to improve projections of precipitation change over land, enhancing our ability to anticipate and respond to the consequences of climate change in populated and vulnerable areas. This is essential for water resource management, agriculture, infrastructure planning, and overall climate resilience efforts (IPCC, 2021).

To constrain precipitation change projections over land, our methodology involves a multistep process that integrates data preprocessing, dimension reduction, causal relationship estimation, causal network evaluation, and model weighting. Our study utilizes CMIP6 historical simulations of Sea Level Pressure (SLP) complemented by reanalysis datasets which serve as references. Future projections based on Shared Socioeconomic Pathways (SSP, O'Neill et al., 2014) are employed to calibrate the weighting scheme and project precipitation changes. To address the high dimensionality of the data, Principal Component Analysis (PCA, Shaffer, 2002; Ramsay and Silverman, 2005) with Varimax rotation (Rohe and Zeng, 2023; Kaiser, 1958) is utilized, extracting 60 components that capture the essential modes of variability. We estimate the time-lagged causal relationships using the PCMCI (Peter-Clark Momentary Conditional Independence, Runge et al., 2019b) causal discovery algorithm to uncover the significant causal pathways crucial for understanding dynamical interactions in the climate models. The identified causal networks are evaluated against the reference networks derived from the reanalysis data using the F_1 score and its complement $1 - F_1$. The causal networks of the climate models are also compared with one another with the F_1 score to measure their similarities. This quantitative approach provides insights into the relative performance and uniqueness of each model's representation of dynamical processes. In the last step, a novel *causal weighting* scheme is introduced, assigning weights based on the performance and interdependence metrics of the causal networks. This scheme prioritizes models closely matching the reference causal network and exhibiting distinctive causal structures. The resulting weights inform the computation of the multimodel weighted means and ranges of precipitation changes over land.

Section 2 provides an overview of the materials and methods used in this study. The results are detailed in Sect. 3. We summarize and discuss our findings in Sect. 4.

2 Materials and methods

Here we introduce the data and methodology used in this study. Section 2.1 describes the CMIP6 and reanalysis data that we integrate. Section 2.2 explains the pre-processing of the data and its relevance in this study. Sections 2.3, 2.4, 2.5, 2.6, respectively, introduce our multistep methodology consisting of dimension reduction, causal network estimation, causal network evaluation, and causal weighting of climate models. **Fig.1** presents the different steps of our framework.

2.1 Data

This study utilizes CMIP6 historical simulations of SLP spanning from 1979 to 2014, with daily time resolution. These simulations, derived from 23 different climate models, each with 2 to 10 ensemble members, are used to estimate the historical causal networks. This results in a total of 154 ensemble members. The daily resolution of the SLP data provides a robust foundation for analyzing the climate models and evaluating their performance in simulating SLP patterns. In addition to the historical simulations of CMIP6, the study employs ERA5 reanalysis data sets (Hersbach et al., 2020), and NCEP/NCAR (Kalnay et al., 1996), covering the period 1979-2014 with daily resolution. These reanalysis datasets serve as a reference for estimating the causal networks of SLP, providing a benchmark against which the performance of the climate models could be assessed. Additionally to the SLP datasets, future climate projections of precipitation of the same climate models are incorporated using simulations based on three Shared Socioeconomic Pathways (SSP, O'Neill et al., 2014): the medium-emission SSP2-4.5 scenario (101 total members), the high-emission SSP3-7.0 scenario (96 total members), and the very high-emission SSP5-8.5 scenario (107 total members), for the period 2015-2100 focusing on precipitation over land with yearly time resolution. These simulations are used to calibrate the model performance parameter σ_D of the weighting scheme. They are also used for the projections of precipitation changes over land. A complete list of included models and members for the SLP historical simulations, and precipitation SSP simulations is available in supplementary Table S1.

2.2 Data preprocessing

The preprocessing of the SLP data involved several crucial steps to ensure consistency and enhance the quality of the analysis. Firstly, all ~~the~~ datasets (including the ERA5 dataset) are linearly interpolated to the 2.5° latitude × 2.5° longitude grid of NCEP/NCAR. Subsequently, the daily data ~~is detrended and anomalized by subtracting the climatological monthly~~ are detrended on a grid-cell basis to remove small trends and ensure robust causal discovery. Anomalies are then calculated using a long-term daily climatology by subtracting each day's mean and dividing by ~~the monthly variance, which reduces seasonality in the data. In addition~~its standard deviation. While SLP data is largely stationary even under historical forcing (Nowack et al., 2020), these steps improve the stationarity of the time series which is essential for the effective application of the PCMCI causal discovery algorithm (Runge et al., 2023). Additionally, the data are separated to isolate winter (DJF: December, January, February), spring (MAM: March, April, May), summer (JJA: June, July, August), and autumn (SON: September, October, November), as different causal dependencies are expected for each meteorological season. ~~These preprocessing steps~~

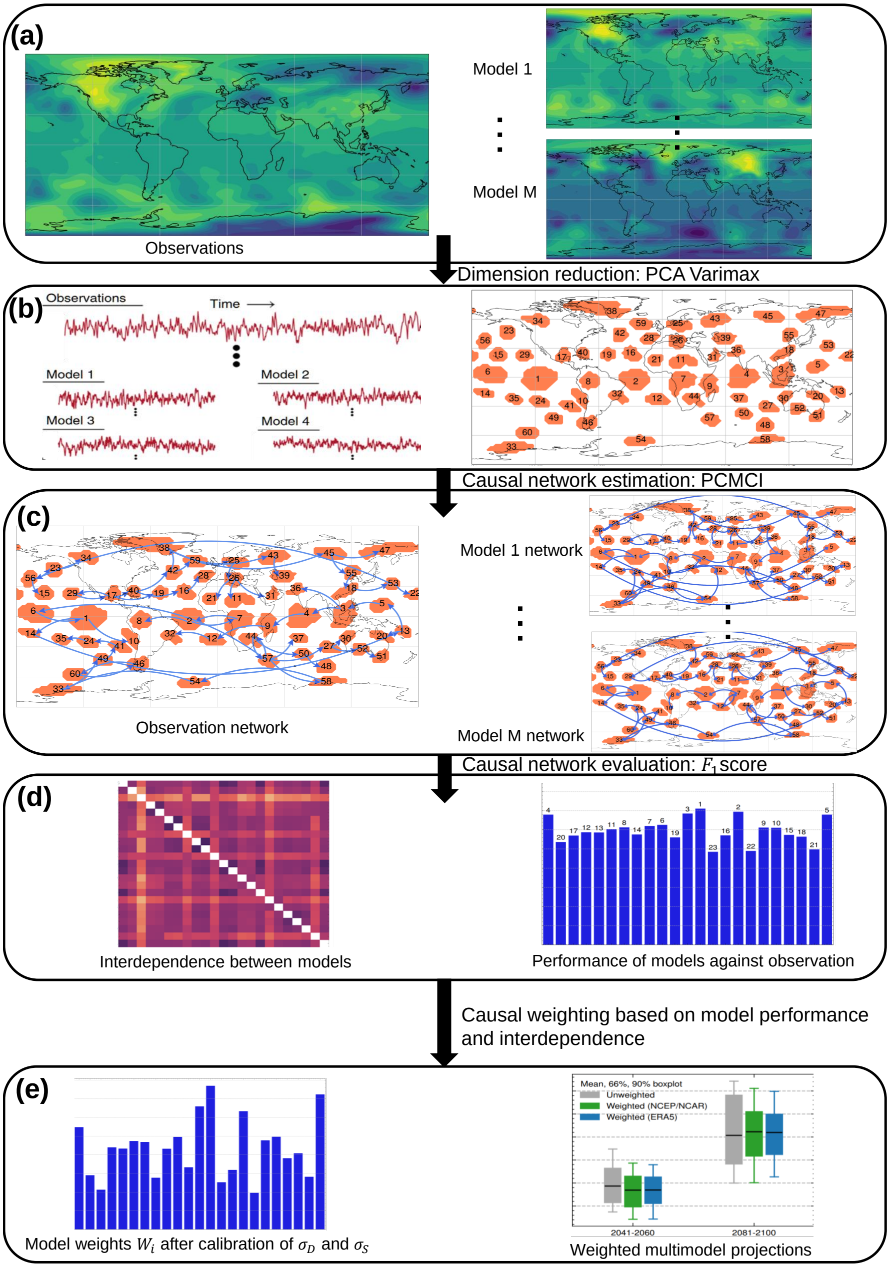

(a)
Model 1
Model M
Observations
Dimension reduction: PCA Varimax
(b)
Observations
Time ⟶
Model 1
Model 2
Model 3
Model 4
Causal network estimation: PCMCI
(c)
Model 1 network
Model M network
Observation network
Causal network evaluation: F_1 score
(d)
Interdependence between models
Performance of models against observation
Causal weighting based on model performance and interdependence
(e)
Model weights W_i after calibration of σ_D and σ_S
Mean, 66%, 90% boxplot
Unweighted
Weighted (NCEP/NCAR)
Weighted (ERA5)
2041-2060
2081-2100
Weighted multimodel projections

Figure 1. Overview of the causal weighting framework. (a) Daily SLP data from NCEP/NCAR and ERA5 reanalyses is reduced using PCA-Varimax to yield (b) regionally confined climate modes for each meteorological season and climate model. PCMCI estimates lagged causal relationships, resulting in (c) dataset-specific causal networks for reanalysis and climate models. These networks enable (d) ~~casual~~ causal model evaluation via network similarity and (e) causal model weighting, which informs the weighted multi-model precipitation projections over land.

~~are designed to improve the stationarity of the time series, which is essential for the effective application of the PCMCI causal discovery algorithm (Runge et al., 2023).~~

2.3 Dimension reduction (Step 1)

As in Nowack et al. (2020), a PCA with Varimax rotation is used to extract the main modes of variability and to manage the high dimensionality of the SLP dataset. This dimension reduction step is crucial to represent the processes of interest.

PCA serves as a dimension reduction technique, preserving as much information as possible while reducing the number of dimensions (Shaffer, 2002; Ramsay and Silverman, 2005). This is accomplished by identifying orthogonal linear combinations (known as principal components) from the original spatial data. These initial components often lack straightforward interpretation. Varimax rotation's role is to enhance interpretability by transforming the principal components. Varimax rotation achieves this by maximizing the variance of loadings on each component (Rohe and Zeng, 2023; Kaiser, 1958). The loadings become more localized on specific variables, making them distinct and easier to interpret.

The PCA-Varimax transformation is derived from the reference reanalysis datasets individually for each season and subsequently applied to the datasets of all climate models. For the final analysis, the first 60 components are selected that capture the essential characteristics of the variability of SLP data. In the remainder of the study, we use the terms components and modes interchangeably to refer to the PCA-Varimax components.

2.4 Causal network estimation (Step 2)

The next step is the estimation of time-lagged causal relationships within the reduced datasets. As correlation alone does not establish causation, we choose to apply causal discovery methods which come with certain assumptions. Here, the assumptions are that all the relevant variables are included in the analysis (causal sufficiency), the causal relationships and the distributions of the variables remain consistent in the sample data (stationarity), and the statistical dependencies and independencies are a true reflection of the underlying causal structure (faithfulness and Markov condition) (Runge et al., 2023). We underscore that not all of these assumptions are strictly verified in this study. For instance, causal sufficiency is not fully met, as our analysis is restricted to SLP causal networks. However, these assumptions are less critical in our case because our main goal is to derive a metric of the data, rather than to determine the exact causal relevance of each link. The rationale behind using causal discovery is that it offers a more precise estimation of dynamical interactions compared to correlation networks, thanks to its ability to filter out spurious relationships.

Given this last assumption, we choose to implement the PCMCI causal discovery algorithm, which is well-suited for time series data with no contemporaneous effects (Runge et al., 2019b). PCMCI aims to uncover causal relationships among variables by assessing conditional dependencies over different time lags. PCMCI builds on the PC algorithm — a constraint-based causal discovery method — by incorporating momentary conditional independence (MCI) tests. These tests help identify causal links even when variables exhibit high autocorrelations, which is common with climate time series (Runge et al., 2019a). During the MCI step, the PCMCI algorithm tests for conditional independence among variables. A causal link is only considered significant if the p-value of the test is less than or equal to a significance level α_{MCI} set by the user.

In this study, the PCMCI algorithm is applied to the principal component time series of each dataset (one dataset per member and season), which are derived from the previous dimension reduction step. The PCMCI algorithm outputs a causal network, enabling the identification of causal pathways between the SLP modes of individual climate datasets or reanalysis datasets. This step identifies significant causal relationships, which is crucial for understanding the dynamical interactions between the SLP modes.

2.5 Causal network evaluation (Step 3)

Following the identification of causal relationships, the resulting causal networks are evaluated using similarity and distance metrics. Given the relatively large size, consisting of 60 variables and a maximum time lag of 20 days, it is challenging to discern patterns. This complexity underscores the necessity of employing a similarity metric to facilitate the comparison of causal networks. The similarity is quantified using the F_1 score introduced in Nowack et al. (2020), while its complement $1 - F_1$ score serves as a measure of distance. The F_1 score is defined as the harmonic mean between precision and recall where: Precision $= \frac{TP}{TP+FP}$, Recall $= \frac{TP}{TP+FN}$ and $F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

Compared to a reference network, FP represents the count of falsely identified links, FN is the count of undetected links, and TP denotes the number of correctly identified links. Like Nowack et al. (2020), we adjusted the traditional F_1 score definition to account for the sign of the dependencies and integrated a relaxation of the time-lags of identified links. Specifically, if a link exists in reference network A and corresponds to a link in network B with the same causal direction within a time range of $\pm\tau_{Diff}$ time lags, we consider it a correctly identified link (TP).

The performance of each climate model's causal network is assessed against a reference causal network derived from observational data, with the distance to this reference network serving as the performance metric. Furthermore, the interdependence among the causal networks of the climate models is quantified, reflecting the degree of similarity or divergence among the networks. Smaller distance values indicate greater similarity, both in terms of performance relative to the reference and in terms of dependence among the models. These measures are averaged over separate causal networks obtained for the four meteorological seasons for each model and reanalysis dataset. The results provide insights into the relative performance and distinctiveness of each model's representation of atmospheric dynamical processes.

2.6 Causal weighting scheme based on performance and interdependence (Step 4)

In this study, we develop a new weighting scheme called causal weighting, which is based on the performance and interdependence of the model causal networks. Specifically, we measure performance and assess interdependence between the networks using the complement of the F_1 scores, calculated as $1 - F_1$ score. These scores are then normalized by the median score across all models. The causal weighting scheme aims to assign higher weights to models that closely match the reference causal network (indicating high performance) and exhibit unique causal structures (indicating high independence). The scheme is formulated as:

$$w_i \propto \frac{e^{\frac{-(1-F_1^i)^2}{\sigma_D^2}}}{1+\sum_{j\neq i}^{M} e^{\frac{-(1-F_1^{ij})^2}{\sigma_S^2}}} \quad (1)$$

In Eq. 1, M indicates the number of models in the ensemble, $1 - F_1^i$ is the normalized "distance" of model i relative to observations or reanalyses, and $1 - F_1^{ij}$ is the normalized "distance" of model i relative to model j. Weights are normalized to sum to 1. The causal weighting is inspired by the scheme introduced in Knutti et al. (2017) and further explored in several follow-up studies (Brunner et al., 2020). In the original scheme, the performance and interdependence are measured with root-mean-square differences (RMSD).

The parameters σ_D and σ_S determine the balance between model performance and interdependence. The calibration of the interdependence shape parameter σ_S is performed first. In the original weighting scheme, different options are available to calibrate σ_S as reported in Merrifield et al. (2020). We choose one of the more robust options. Namely, we identify an interdependence shape parameter larger than the typical distances between members of the same model but smaller than the typical intermodel distances. More independent models are given smaller denominators resulting in larger weights.

The other weighting parameter is the performance shape parameter σ_D. Large σ_D values result in equal weighting across models, whereas small σ_D values cause aggressive weighting, with high-performance models receiving the majority of the weights. After calibrating σ_S, a perfect model test is used to estimate the performance shape parameter σ_D by evaluating climate models based on their historical performance without being overconfident (Karpechko et al., 2013; Abramowitz and Bishop, 2015; Wenzel et al., 2016; Sanderson et al., 2017; Knutti et al., 2017; Brunner et al., 2020). In the perfect model test approach, each model is sequentially treated as the “truth”, while the other models are weighted to project the future target response of the perfect model. After testing σ_D values between 0.1 and 2.0, the calibration selects the smallest σ_D value for which the projection is not overconfident, i.e., when 80% of these "perfect models" fall within the 10–90 percentile range of the weighted distribution in the target period. To prioritize performance over interdependence in the weighting trade-off, we reduce this proportion to 70%. In this study, the target to predict is the precipitation over land for different SSPs and periods (2041-2060 and 2081-2100), resulting in different calibrated values.

Once the two shape parameters have been calibrated, the weights are computed to obtain weighted multimodel means and ranges of future climate projections. The weighting scheme and associated figures were developed using the Earth System Model Evaluation Tool (ESMValTool) version 2 (Eyring et al., 2020; Righi et al., 2020; Lauer et al., 2020; Brunner et al., 2020; Schlund et al., 2023).

2.7 Technical details

Both the observational data and the climate model simulations contain internal variability, which can introduce noise and potentially bias the comparison between models and observations. To mitigate its influence, multiple ensemble members for each model were processed, with causal networks derived independently for each. The final F1-scores represent an ensemble average, which reduces the variability effects by smoothing out member-specific results. Recognizing that reanalysis datasets themselves are subject to internal variability and measurement uncertainties, we have analyzed multiple reanalysis products (ERA5 and NCEP/NCAR).

In the dimension reduction step, we keep the first 60 components of the 100 obtained from the PCA-Varimax analysis. Some tests are also performed with only the first 50 components. Components with unresolved frequency spectra or dipolar patterns are discarded similar to the methodology used in Nowack et al. (2020). This selection ensures that only the most significant and stable modes of variability are considered, enhancing the quality of the following steps in our methodology.

Time-lagged dependencies within the data are estimated using the PCMCI algorithm, with a minimum time lag $\tau_{\min}$ of 1 day and with a maximum time lag $\tau_{\max}$ set to 20 days, though trials with a maximum time lag of 10 days are also tested. PCMCI outputs a time series Directed Acyclic Graph (DAG, Runge et al., 2023), where the nodes represent variables, the directed edges indicate lagged causal relationships, and there are no cycles in the graph. We assume that the causal dependencies are linear and with additive Gaussian noise. Under such assumptions, we employ the partial correlation conditional independence tests within PCMCI to detect these dependencies. The hyperparameter α_{MCI}, which controls the significance threshold for the PCMCI algorithm's MCI step, is set to 10^{-5} in the results presented in the main text. We briefly investigate the sensitivity of the causal model evaluation to larger values of this parameter in the appendix D1.

The causal network evaluation employs the F_1 score, which is "relaxed" by counting links as true positives even if they occur at slightly different time lags than the reference. We set this window at 2 days ($\tau_{Diff} = 2$ days).

3 Results

In this section, we present the findings for each step of our methodology as applied to the CMIP6 model datasets.

3.1 Dimension reduction and causal network estimation (Step 1 & 2)

Results of the dimension reduction step are shown in **Fig. A1** and **A2** in the appendix. In ~~**Fig. A1** and **A2** in the appendix, we show the centers of the 60 first PCA-Varimax components. By comparing the spatial patterns for each season between ERA5 and NCEP/NCAR, we can observe similarities and differences in the distribution of components. Generally, we see similar large-scale patterns since both datasets are reanalyses of atmospheric variables. However, differences arise due to variations in data assimilation methods, and model physics. PCA-Varimax identifies major modes of variability for all seasons and datasets as reported in Vejmelka et al. (2015) and Nowack et al. (2020). The components explaining the most variance are located in the tropics (for example El Niño region), influencing atmospheric circulation globally. In~~ our analysis, we chose to retain the first 60 components from the 1979-2014 data to better cover the Northern Hemisphere, particularly during the JJA season. Using SLP data from 1948 to 2017, Nowack et al. (2020) truncated and kept a selection of 50 components, discarding additional components due to unphysical time series, such as sudden jumps observed in 1979 when entering the satellite era. We do not encounter these jumps in the time series that start in 1979. We also perform tests with 50 components to investigate the stability of the methodology. Components retrieved from NCEP/NCAR were used across all climate models to obtain reduced datasets. Additionally, components derived from ERA5 were used as an alternative reference for all models.

3.2 ~~Causal network estimation (Step 2)~~

~~Although a maximum time lag of 20 days was set for PCMCI, 99.9 % of the dependencies were found within the first 10 days. The causal networks are too complex to visualize, with an average of 18 causal dependencies per mode for NCEP/NCAR and 20 for ERA5. For this reason, we choose to inspect only the most significant causal dependencies of each mode. **Fig. ??** and **??** in the appendix display the most significant causal dependencies for each mode in the two reanalysis datasets during the winter months (DJF). Despite the lack of spatial information provided to the PCMCI causal discovery algorithm, the most significant dependencies predominantly originate from neighboring modes. This finding indicates~~ Discussed in more detail in appendix B, our findings indicate that the causal network estimation step identifies physically meaningful dependencies between the SLP modes for both reanalysis datasets.

3.2 Causal network evaluation (Step 3)

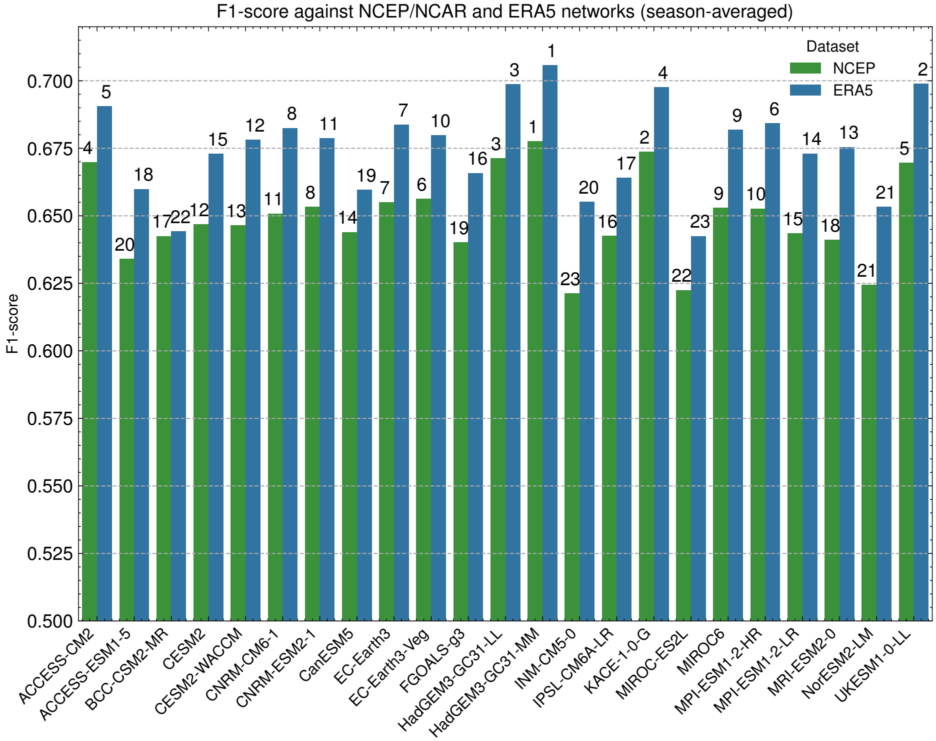


Figure 2. Comparison of the climate models' causal networks F_1 scores with NCEP-NCAR (~~a~~green) ~~NCEP-NCAR~~ and ERA5 (~~b~~blue) ~~ERA5~~ as reference. This figure illustrates the similarity between climate models' causal networks and those of the reference reanalysis datasets, averaged across all available members and seasons, using the F_1 score. Higher F_1 scores indicate greater similarity. The rank of each model's similarity is denoted on top of each bar.

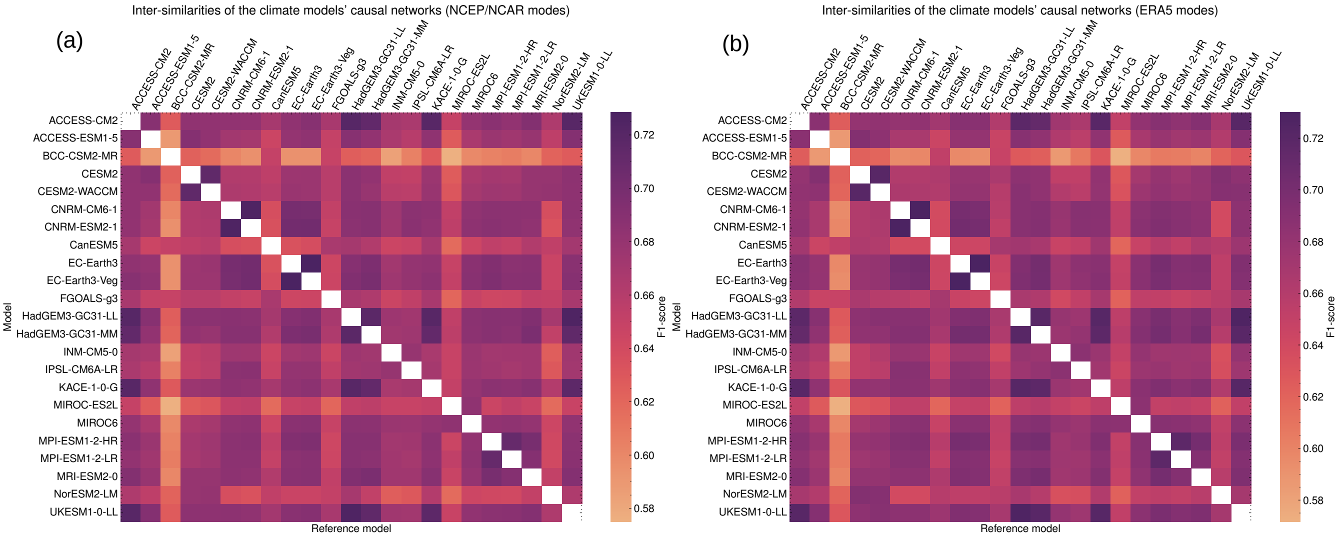


Figure 3. Similarities of the climate models' causal networks when the modes are obtained from (a) NCEP/NCAR and (b) ERA5. This figure illustrates the similarity between the causal networks of different climate models. Similarity is quantified using the F_1 scores between two models. Higher values denote greater similarity or lower independence. The causal networks of one climate model (row) are compared against the causal networks of other climate models used as reference (columns). The values are averaged across the members of each climate model and over all seasons.

Fig. 2 compares climate models' causal networks' F_1 scores relative to ~~(a)~~ NCEP-NCAR and ~~(b)~~ ERA5 reference datasets. Higher F_1 scores indicate greater similarity between model networks and the references, averaged across all members and seasons. Interestingly, the F_1 scores are consistently higher when compared to ERA5 than to NCEP-NCAR, indicating that the models' dynamical SLP patterns generally match ERA5 more closely. ~~The Spearman's rank correlation coefficient is calculated between the rankings of climate models to assess variation in rankings across the different reference datasets (NCEP/NCAR and ERA5), yielding a coefficient of 0.91. This confirms a strong~~ A more detailed analysis of the consistency between the ~~climate model rankings with the NCEP-NCAR and ERA5 references. The obtained p-value from the Student's t-test is 1.1×10^{-9}, rejecting the null hypothesis of no ordinal correlation between the rankings of models with NCEP/NCAR or ERA5 taken as reference~~model performance across reanalysis datasets and meteorological seasons is given in appendix C.

~~In **Fig. D1**, we varied~~ Sensitivity tests for the significance level α_{MCI} ~~of PCMCI from 10^{-5} to 10^{-4} and 10^{-3}. **Fig. D2** demonstrates the effects of reducing the number of modes in the networks from 60 to 50 and decreasing the~~, the maximum time lag ~~in the PCMCIalgorithm from 20 to 10 days. While these variations affected the F_1 score values moderately, they had a minimal influence on the rankings of the climate models. This was evaluated by calculating the Spearman's rank correlation coefficient for the modified experiments against the baseline experiment presented in the main text (Figure 2a), which used the NCEP/NCAR reference with 60 modes and $\alpha_{MCI} = 10^{-5}$. The correlation coefficients were close to 1, ranging from 0.95 to 0.98, confirming a strong ordinal correlation between the rankings of models in the different experiments. The p-values, all~~

~~smaller than 10^{-11}, rejected the null hypothesis of no ordinal correlation between the alternative experiments and the baseline experiment.~~ parameter in PCMCI, and the number of components retained during the dimension reduction step are provided in Appendix D.

In **Fig. 3**, models with ~~similar causal networks have high similarities. Climate models sharing development in their atmospheric models exhibit the highest F_1 scores~~shared developmental features exhibit higher causal network similarity, likely due to their comparable dynamical representations. For example, the ACCESS, UKESM, HadGEM and K-ACE models share more similar causal networks as measured by the F_1 scores. As reported in the genealogy tree of CMIP6 models in Kuma et al. (2023), the HadGEM2 model was an ancestor of the aforementioned models. Additionally, climate models developed by the same institute (such as the CNRM-CM6-1 and CNRM-ESM2-1) exhibit more similar causal networks, as indicated by the F_1 scores. This finding confirms that the evaluation of the SLP causal networks can identify models with similar physical cores and, consequently, similar dynamical sea-level pressure processes. This result is consistent with previous literature, as Nowack et al. (2020) showed that CMIP5 models with shared development and atmospheric models also exhibited more similar causal networks.

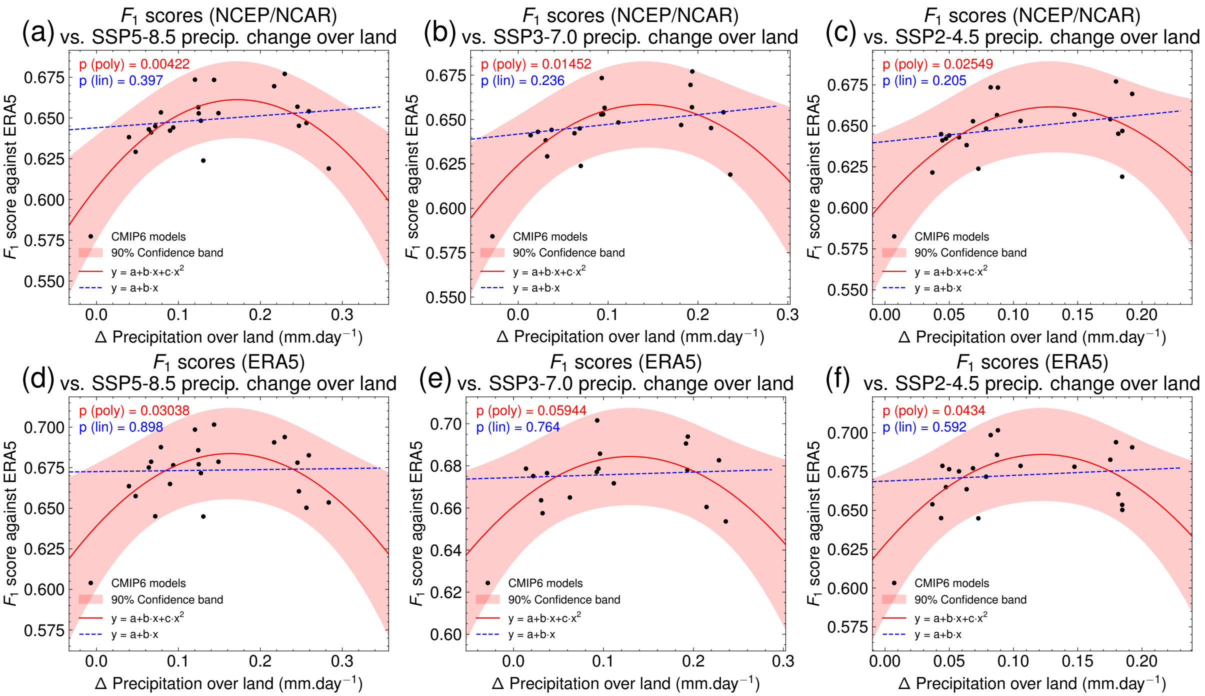


Figure 4. Relationship between precipitation change over land under SSP2-4.5 (c,f), SSP3-7.0 (b,e) and SSP5-8.5 (a,d) scenarios and F_1 scores, using NCEP/NCAR (a,b,c) and ERA5 (d,e,f) causal networks as references. The x-axis shows precipitation changes between 1850-1900 and 2050-2099, while the y-axis represents F_1 scores of climate model causal networks relative to the reference. F_1 ~~score values~~ scores are averaged across all seasons and available members of a model. The red solid line shows a polynomial fit and the red filled area depicts the 90% confidence band based on a two-tailed t-test. The blue dashed line corresponds to linear fit.

Fig. 4 shows the relationship between the F_1 scores of the CMIP6 climate models' causal network and the changes in precipitation over land for the SSP2-4.5, SSP3-7.0 and SSP5-8.5 scenarios. The shape indicates ~~a parabolic relationship~~ an approximately parabolic relationship over the space of opportunities covered by CMIP6 models between the F_1 scores and the precipitation changes. ~~A statistical analysis confirms a parabolic~~ Statistically significant parabolic relationships (polynomial of degree 2) ~~relationship with a p-value~~ with p-values of less than 0.05 ~~, indicating a strong and statistically significant relationship~~ (except for a p-value of 0.06 for SSP3-7.0 and ERA5 reference) are found. Significant parabolic relationships are found for all SSPs and the two different references, underscoring the robustness of this relationship for different global warming scenarios and reference reanalysis datasets. Notably, climate models with higher F_1 scores, indicating better representations of observed dynamical sea level pressure patterns, tend to cluster around the center of the parabola. These models project precipitation changes in the mid-range compared to other CMIP6 models. On the contrary, climate models with lower F_1 scores, indicating lower representations of observed dynamical sea level pressure patterns, tend to either overestimate or underestimate precipitation changes over land. Nowack et al. (2020) previously reported a significant parabolic relationship between precipitation changes under the RCP8.5 scenario and F_1 scores of CMIP5 models. Our findings extend this relationship to CMIP6 models using daily data, compared to the three-day resolution in Nowack et al. (2020), suggesting that F_1 scores may serve as a robust constraint for projecting precipitation changes over land.

Unlike emergent constraints, which typically display linear relationships, we present a different approach. On the x-axis ~~, we have~~ in Fig. 4, we consider a metric which is an observable (the causal network) relative to the observed values (the reference causal network), rather than the observable itself. As a result, the relationship is not linear but rather a concave function with a ~~distinct~~ peak, here ~~a~~ an approximately parabolic relationship between the F_1 scores and the precipitation changes.

3.3 Causal weighting scheme based on performance and interdependence (Step 4)

Table 1. Calibrated performance shape parameters σ_D for different target periods (columns), SSPs (rows) and reference reanalysis dataset (sub-tables).

NCEP/NCAR	2041-2060	2081-2100	ERA5	2041-2060	2081-2100
SSP2-4.5	0.4	0.3	SSP2-4.5	0.36	0.26
SSP3-7.0	0.28	0.3	SSP3-7.0	0.25	0.29
SSP5-8.5	0.53	0.29	SSP5-8.5	0.48	0.28

~~The~~ Our previous findings suggest that leveraging the climate models' causal networks' similarity to reference reanalysis causal networks and the intermodel similarities can be promising to constrain precipitation changes over land. We found that models sharing atmospheric characteristics exhibit higher causal network similarity, highlighting the ability of the methodology in capturing sea level pressure (SLP) dynamics accurately. Furthermore, the parabolic relationship between F_1 scores — measuring a model's ability to replicate observed SLP dynamics — and its projection of precipitation changes over land

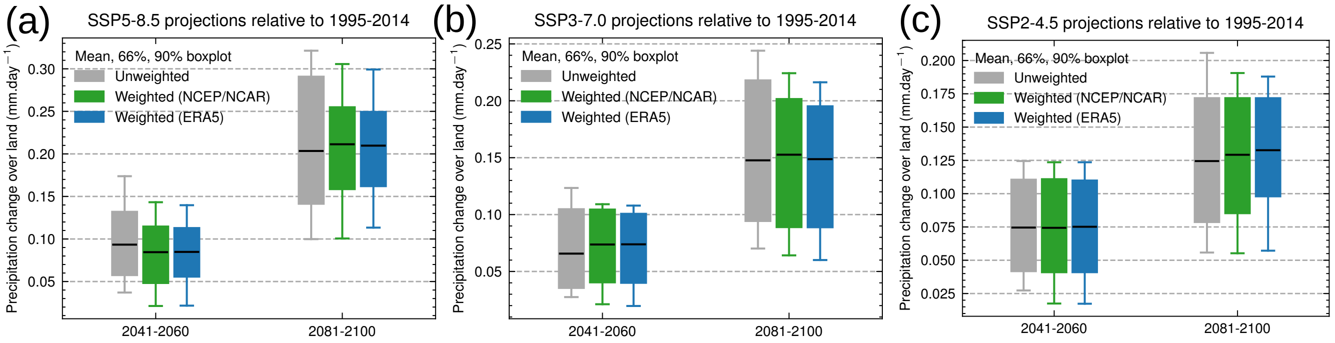


Figure 5. Boxplots of weighted and unweighted projections of precipitation over land relative to 1995-2014 for (a) the SSP5-8.5, (b) SSP3-7.0, (c) and SSP2-4.5 scenarios. The grey boxplots represent unweighted projections, the green boxplots represent projections weighted using NCEP/NCAR as a reference, and the blue boxplots represent projections weighted using ERA5 as a reference. Each boxplot displays the mean (black solid line), likely ranges (17–83 percentile), and very likely ranges (5–95 percentile). The y-axis indicates the precipitation change over land, while the x-axis indicates the target period.

support the use of the F_1 scores as a diagnostic to weight climate projections of precipitation based on their SLP representation skill.

Using the notation of Eq. 1, the complement of the F_1 score serves to measure the distance of models to a chosen reference reanalysis dataset and to evaluate the interdependence among the different models. These distances are separately normalized by the median over all models. After this normalization, the distances can range from 0 to values greater than 1.

The interdependence shape parameter σ_S is calibrated first. We calculate the average distance between members of the same model and the average distance between members of different models. A robust choice for σ_S should lie between these typical distances. These distances are presented in **Fig. E1**, leading to a calibrated σ_S value of 0.9.

The model performance parameter σ_D is then calibrated using the perfect model test described in Sect. 2.6. A specific σ_D value was calibrated for each SSP (SSP5-8.5, SSP3-7.0, SSP2-4.5), target period (2041-2060 and 2081-2100), and reference dataset (NCEP/NCAR and ERA5). The calibration results, reported in Table 1, range from 0.25 to 0.53.

3.4 Weighted projections of land precipitation changes

Using the calibrated shape parameters, the weights for each combination of SSP, target period, and reference dataset are derived by applying Eq. 1. These weights are used to calculate the weighted projections for a medium (SSP2-4.5), high (SSP3-7.0) and very high (SSP5-8.5) emission scenario.

The weighted and unweighted projections are shown in **Fig. 5**. The boxplot indicates the mean, likely (17–83 percentile) and very likely (5–95 percentile) ranges of projected precipitation changes over land relative to 1995–2014. In general, we observed narrower ranges for the weighted projections. Across all scenarios and reference datasets, the weighted means of precipitation over land do not significantly differ from the unweighted mean. However, the likely and very likely weighted ranges are generally reduced compared to the unweighted ranges, except for those based on the SSP2-4.5 scenario in the 2041-2060 period. The reduction in uncertainty is consistently higher when ERA5 is used in the dimension reduction and causal model evaluation steps compared to NCEP/NCAR. In particular, the upper bounds of the weighted ranges (83th percentiles for the likely range, 95th percentiles for the very likely range) are consistently shifted downward, indicating that ensembles with larger projected precipitation changes over land are less probable. The most substantial reductions in uncertainty ranges occur for the SSP5-8.5 scenario during the 2081-2100 period. This reduction in the weighted upper bound aligns with previous studies that constrained global (not only land) mean precipitation, which also reported lower upper bounds of projections for various SSPs and target periods (Shiogama et al., 2022; Dai et al., 2024). In contrast, no consistent trend is observed for the lower bounds of the weighted ranges across the SSPs and target periods. For the period 2081–2100, the very likely range in the ERA5 weighted projections is narrowed compared to raw CMIP6 projections. Under SSP5-8.5, the range is reduced from $[0.099 - 0.321]$ mm day^{-1} to $[0.113 - 0.299]$ mm day^{-1}. Similarly, under SSP3-7.0, the range decreases from $[0.070 - 0.244]$ mm day^{-1} to $[0.060 - 0.216]$ mm day^{-1}, and under SSP2-4.5, it is reduced from $[0.055 - 0.205]$ mm day^{-1} to $[0.057 - 0.188]$ mm day^{-1}. This represents a decrease from 10 to 16 % in range sizes relative to the unweighted ranges and across the different SSP scenarios. The reduction is even more pronounced for the likely ranges, decreasing substantially by 16 to 41 % relative to the unweighted ranges and across the different SSP scenarios. These findings highlight the effectiveness of the weighting method in narrowing the projection uncertainty of precipitation over land.

Given that the causal weighting accounts for models that better represent the dynamical pattern of SLP globally, we also examine the spatial pattern of precipitation change over land under global warming. **Fig. 6(a-c)** shows the spatial distribution of the causally weighted projections of mean precipitation changes for three SSP scenarios (SSP2-4.5, SSP3-7.0, and SSP5-8.5) for the period 2081–2100 relative to 1995–2014. ERA5 was used as a reference for the causally weighting. The projections indicate substantial regional variability across all scenarios. Significant increases in mean precipitation are projected in Northern Europe, Northern Asia, parts of North America, as well as East and South Asia, and Central and Eastern Africa. These regions could see increases of up to 1.2 mm day^{-1} under the SSP5-8.5 scenario. Conversely, decreases in precipitation are projected for the Mediterranean basin, Central America, Northern South America with reductions reaching up to -1.2 mm day^{-1}. These trends are consistent across all three SSP scenarios, though the intensity varies, with the most pronounced changes observed under the SSP5-8.5 scenario.

Fig. 6(d-f) presents the difference between the absolute changes of the causally weighted and unweighted mean for the period 2081–2100 relative to 1995–2014, while Fig. 6(g-i) depicts the difference between the relative changes of the causally weighted and unweighted mean. Despite the spatially averaged weighted projections of precipitation change over land showing no significant deviation from the unweighted averages (refer to Figure 5), Fig. 6(d-i) highlights that the weighted patterns exhibit notable spatial variations compared to the unweighted mean precipitation absolute change. Regions with positive absolute differences indicate areas where the weighted projections forecast greater increases in precipitation relative to the unweighted mean. Conversely, negative absolute differences denote areas where the weighted projections give smaller increases or larger decreases in precipitation than the unweighted mean. In particular, South America demonstrates the most significant variations in the weighted projections, with absolute differences reaching up to ±0.4 mm day^{-1}. However, the map of differences between

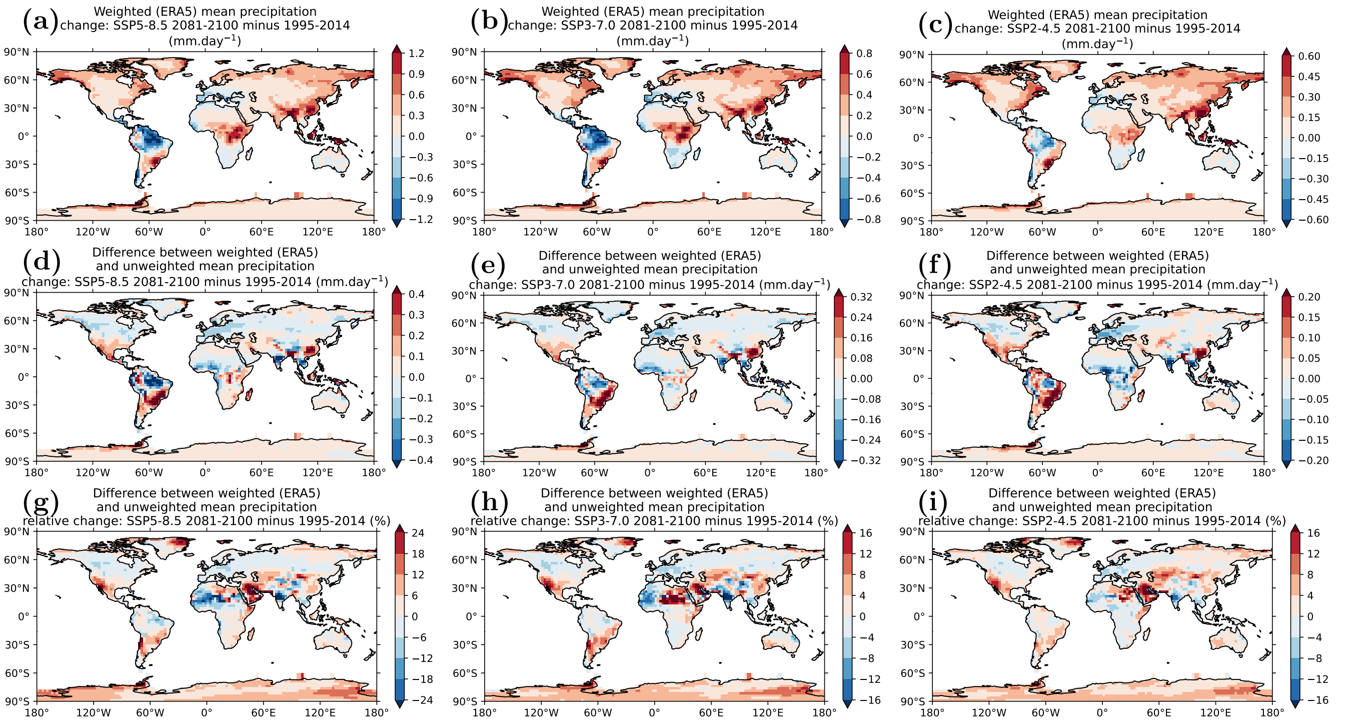

(a)
Weighted (ERA5) mean precipitation change: SSP5-8.5 2081-2100 minus 1995-2014 (mm.day^{-1})
1.2
0.9
0.6
0.3
0.0
−0.3
−0.6
−0.9
−1.2
(b)
Weighted (ERA5) mean precipitation change: SSP3-7.0 2081-2100 minus 1995-2014 (mm.day^{-1})
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
(c)
Weighted (ERA5) mean precipitation change: SSP2-4.5 2081-2100 minus 1995-2014 (mm.day^{-1})
0.60
0.45
0.30
0.15
0.00
−0.15
−0.30
−0.45
−0.60
(d)
Difference between weighted (ERA5) and unweighted mean precipitation change: SSP5-8.5 2081-2100 minus 1995-2014 (mm.day^{-1})
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
(e)
Difference between weighted (ERA5) and unweighted mean precipitation change: SSP3-7.0 2081-2100 minus 1995-2014 (mm.day^{-1})
0.32
0.24
0.16
0.08
0.00
−0.08
−0.16
−0.24
−0.32
(f)
Difference between weighted (ERA5) and unweighted mean precipitation change: SSP2-4.5 2081-2100 minus 1995-2014 (mm.day^{-1})
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
(g)
Difference between weighted (ERA5) and unweighted mean precipitation relative change: SSP5-8.5 2081-2100 minus 1995-2014 (%)
24
18
12
6
0
−6
−12
−18
−24
(h)
Difference between weighted (ERA5) and unweighted mean precipitation relative change: SSP3-7.0 2081-2100 minus 1995-2014 (%)
16
12
8
4
0
−4
−8
−12
−16
(i)
Difference between weighted (ERA5) and unweighted mean precipitation relative change: SSP2-4.5 2081-2100 minus 1995-2014 (%)
16
12
8
4
0
−4
−8
−12
−16
90°N
60°N
30°N
0°
30°S
60°S
90°S
180°
120°W
60°W
0°
60°E
120°E
180°

Figure 6. Patterns of causally weighted projections of mean precipitation change over land in 2081–2100 relative to 1995–2014 for (a) SSP5-8.5, (b) SSP3-7.0, (c) and SSP2-4.5 scenarios. The differences between the weighted and unweighted mean precipitation change are shown in (d) for SSP5-8.5, (e) for SSP3-7.0, and (f) for SSP2-4.5 scenarios. The differences between the weighted and unweighted mean precipitation relative change are shown in (d) for SSP5-8.5, (e) for SSP3-7.0, and (f) for SSP2-4.5 scenarios. ERA5 was used as a reference for the causal weighting.

the relative changes of the weighted and unweighted mean precipitation suggests that these absolute changes are not the largest relative changes globally. The region of the Sahara, the Arabian Peninsula, Southwest and North America, and Northeastern Greenland exhibit more pronounced relative changes, with values reaching up to 20 %.

A figure comparable to Fig. 6 is presented in **Fig. F1** of the appendix, illustrating the projected changes for the period 2041–2060. The observed trends for 2081–2100 remain consistent for this earlier period.

4 Summary and discussion

Climate projections derived from an ensemble of multiple climate models participating in the Coupled Model Intercomparison Project Phase 6 (CMIP6, Eyring et al., 2016) continue to have large uncertainties for precipitation (Tebaldi et al., 2021). This hinders accurate information to be delivered for mitigation and adaptation. Eyring et al. (2019) argue that advanced methods for model weighting are needed to distil more credible information on regional climate changes, pointing out the importance of considering both model performances and interdependencies in model weighting studies as for example presented by Knutti et al. (2017) and Brunner et al. (2020). Machine learning can play an important role in pushing the frontiers of climate model analysis Eyring et al. (2024a), including approaches to weight multimodel projections (Schlund et al., 2020a). Here we build on a previous study that evaluates the performance of a CMIP ensemble with causal networks (Nowack et al., 2020) and expand this concept to a weighting scheme for precipitation projections with causal discovery.

We first demonstrate that causal model evaluation of CMIP6 models can effectively identify specific causal fingerprints of sea level pressure (SLP) that influence precipitation patterns and their projections. Notably, we identify a parabolic relationship between the ability of climate models to represent observed dynamical SLP patterns in causal networks, quantified by the networks' F_1 scores, and the projected precipitation changes over land by the end of the century. CMIP6 models that better represent reference dynamical interactions in their causal networks produce projections within the middle range of the CMIP6 ensemble, while models with lower skill either overestimate or underestimate the mean projections. This pattern is consistent across various global warming scenarios (SSP2-4.5, SSP3-7.0, and SSP5-8.5) and reference reanalysis datasets (NCEP/NCAR and ERA5). Similar findings were reported by Nowack et al. (2020) for the RCP8.5 simulations of CMIP5 models.

Additionally, our study reveals that CMIP6 models with shared development, such as those with a common ancestor model or the same atmospheric model, exhibit more similar causal pathways. This result underscores the ability of the causal model evaluation to effectively identify interdependencies of the CMIP6 models.

Building on these findings, the study introduces a causal weighting scheme for climate projections based on the performance and interdependence of their causal networks. By combining causal model evaluation with multimodel weighting, this approach offers a convincing alternative to traditional weighting based on metrics such as root-mean-square error or trend analysis (Knutti et al., 2017; Brunner et al., 2020; Liang et al., 2020; Tokarska et al., 2020).

The implementation of this causal weighting scheme for projecting precipitation over land significantly reduces the uncertainty range of the climate projections. While the weighted mean projections are closely aligned with the unweighted means, the likely (17–83 percentile) and very likely (5–95 percentile) weighted ranges were notably narrower, and the spatial patterns revealed regional differences in precipitation. For the end-of-century period 2081–2100, the sizes of the very likely weighted ranges under SSP2-4.5, SSP3-7.0, and SSP5-8.5 are reduced by 10 to 16 %, while the likely ranges show an even greater reduction, ranging from 16 to 41 %, when ERA5 was used as a reference.

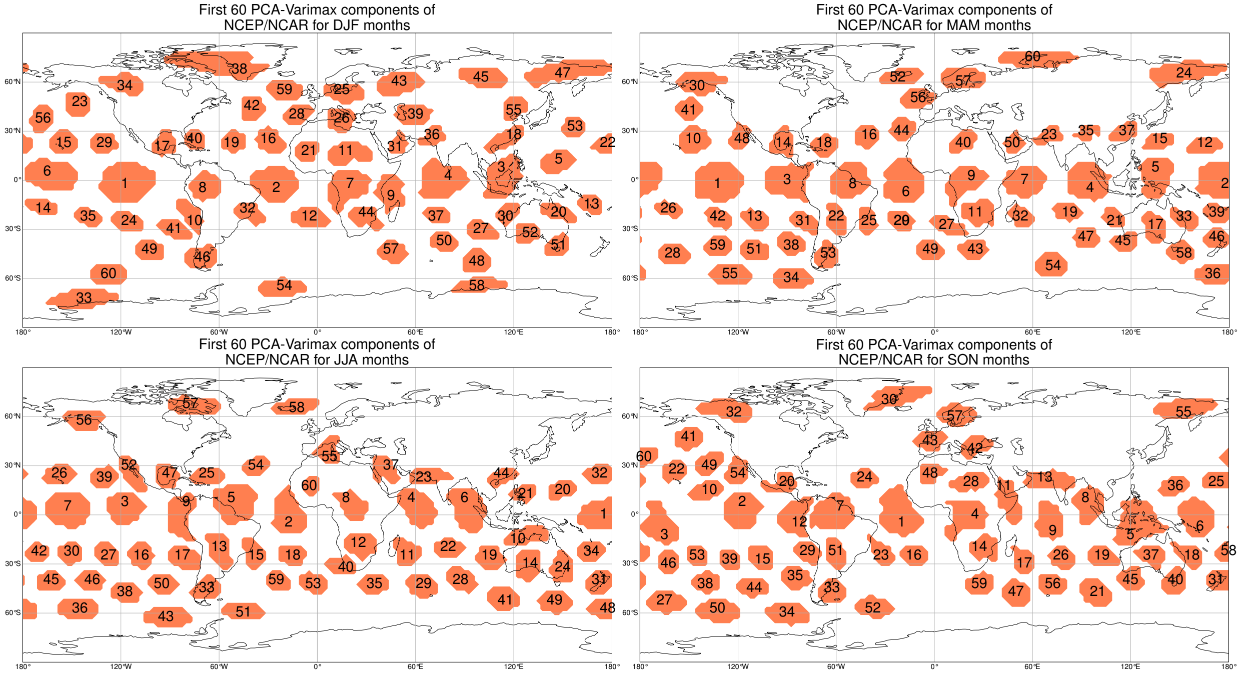
For future research, we consider several areas to be particularly promising. One potential direction is the development of multi-diagnostic weighting (Schlund et al., 2020a), which involves integrating multiple metrics alongside the SLP causal network distance metric into the weighting process. This multi-diagnostic approach could improve precipitation projections further by addressing model differences more comprehensively. By considering additional diagnostics, such as temperature trends, weighted projections may further reduce the uncertainty of projected precipitation over land. Another promising direction is the regional weighting of precipitation change. This approach would focus the weighting scheme specifically on regional precipitation projections, incorporating both global and region-specific diagnostics. Tailoring multimodel weighting to specific regions could prove especially effective. Exploring alternative similarity measures is also a key area for future investigation. Currently, F_1 scores are used to measure the similarity between causal networks, but alternative measures that better discriminate between causal networks or that consider causal effects could provide new insights.

Finally, we want to emphasize that our methodology is not limited to projecting precipitation changes over land. Its applicability could extend to any target variable, provided that pertinent variables and diagnostics exhibiting a robust and consistent relationship (e.g., a parabolic relationship) with the target variable are selected. Our results highlight the importance of integrating advanced evaluation methods and weighting schemes to reduce the uncertainty ranges of climate projections (Nowack and Watson-Parris, 2024). Alongside the development of improved hybrid Earth system models with machine learning with demonstrated reduction of long-standing systematic errors (Eyring et al., 2024a, b), this research provides a novel methodology to constrain uncertainties in multimodel climate projections towards more robust climate change information and more effective mitigation and adaptation strategies.

Appendix A: Maps of sea level pressure components

In Fig. A1 and A2, we show the centers of the 60 first PCA-Varimax components. By comparing the spatial patterns for each season between ERA5 and NCEP/NCAR, we can observe similarities and differences in the distribution of components. Generally, we see similar large-scale patterns since both datasets are reanalyses of atmospheric variables. However, differences arise due to variations in data assimilation methods, and model physics. PCA-Varimax identifies major modes of variability for all seasons and datasets as reported in Vejmelka et al. (2015) and Nowack et al. (2020). The components explaining the most variance are located in the tropics (for example El Niño region), influencing atmospheric circulation globally.

A1 NCEP/NCAR components


First 60 PCA-Varimax components of NCEP/NCAR for DJF months

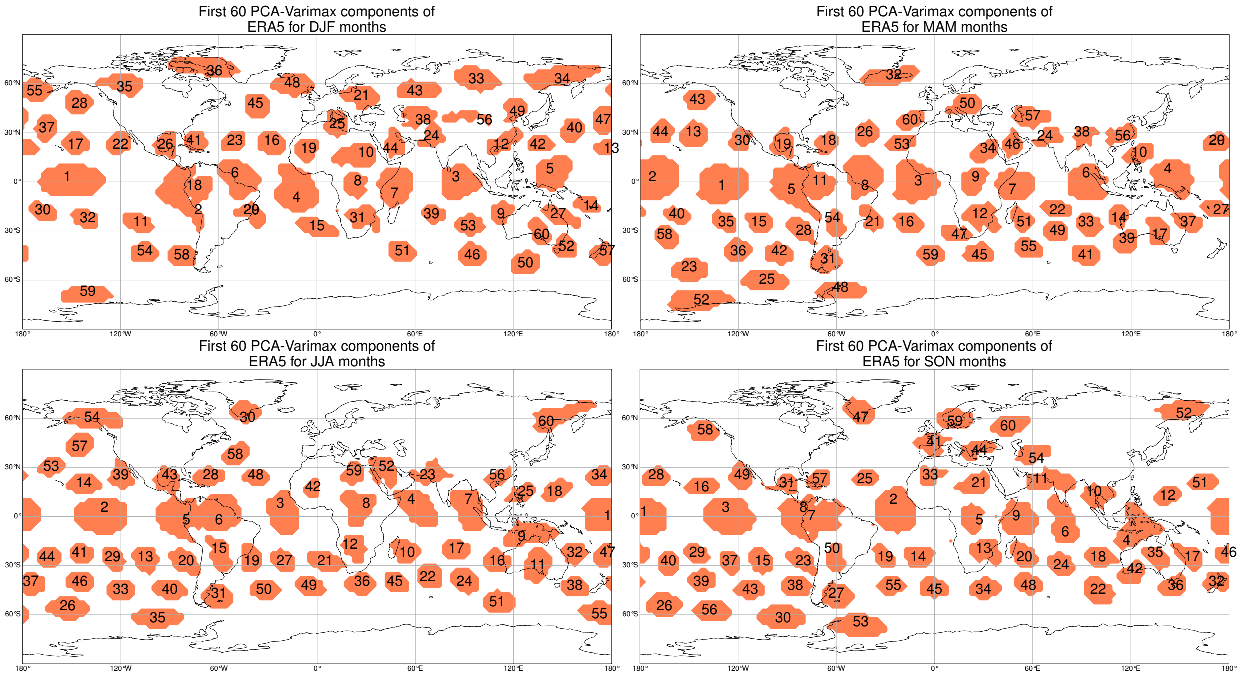
First 60 PCA-Varimax components of NCEP/NCAR for MAM months

First 60 PCA-Varimax components of NCEP/NCAR for JJA months

First 60 PCA-Varimax components of NCEP/NCAR for SON months

Figure A1. Principal Component Analysis (PCA) with Varimax rotation for the NCEP/NCAR dataset during DJF (December, January, February), MAM (March, April, May), JJA (June, July, August) and SON (September, October, December). Here, each component is represented by its core, which consists of loadings greater than 80% of the maximum loading.

A2 ERA5 components


First 60 PCA-Varimax components of ERA5 for DJF months

First 60 PCA-Varimax components of ERA5 for MAM months

First 60 PCA-Varimax components of ERA5 for JJA months

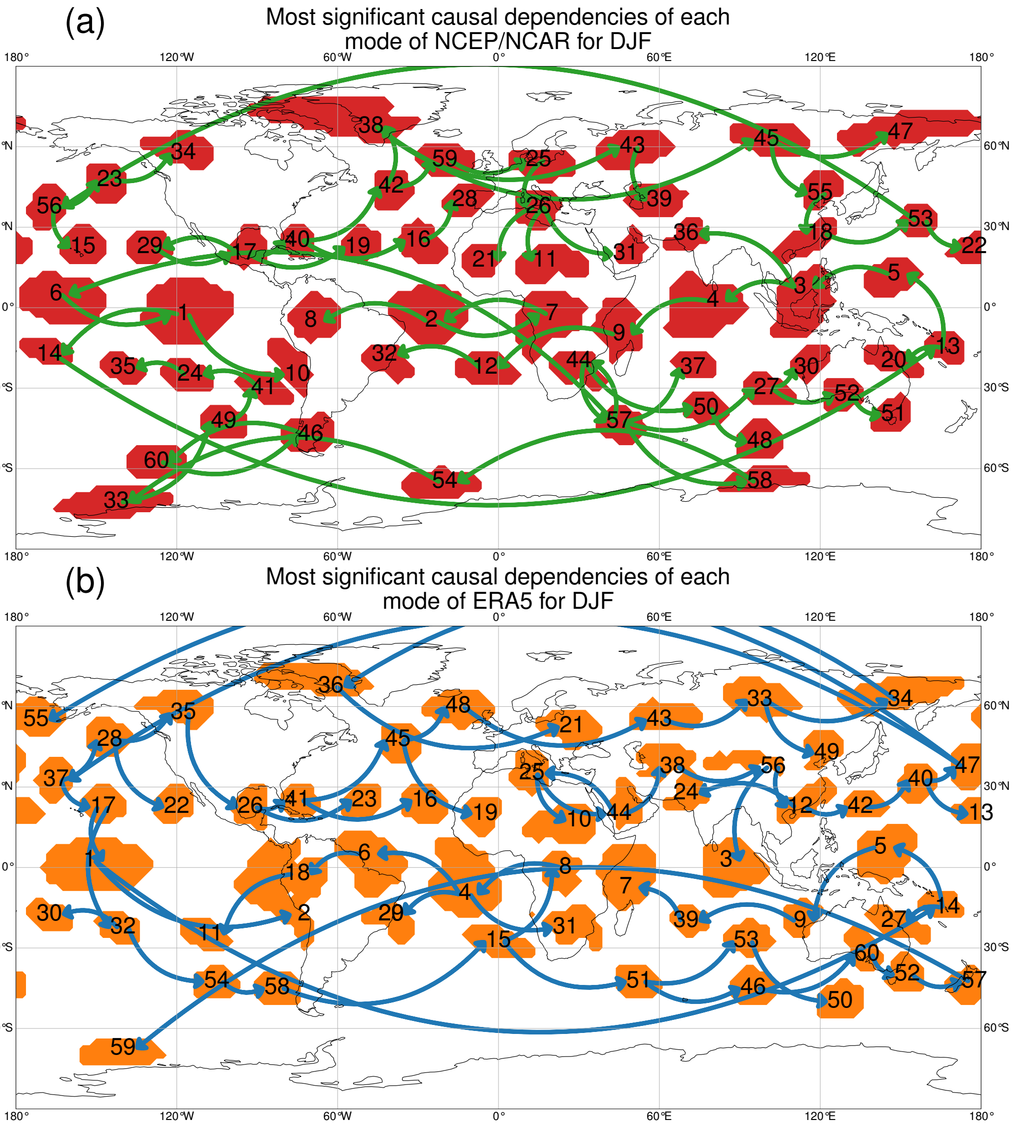
First 60 PCA-Varimax components of ERA5 for SON months

Figure A2. Principal Component Analysis (PCA) with Varimax Rotation for the ERA5 dataset during DJF (December, January, February), MAM (March, April, May), JJA (June, July, August) and SON (September, October, December). Here, each component is represented by its core, which consists of loadings greater than 80% of the maximum loading.

Appendix B: Causal network estimation results

B1 ~~Simplified NCEP/NCAR causal network~~

Although a maximum time lag of 20 days was set for PCMCI, 99.9 % of the dependencies were found within the first 10 days. The causal networks are too complex to visualize, with an average of 18 causal dependencies per mode for NCEP/NCAR and 20 for ERA5. For this reason, we choose to inspect only the most significant causal dependencies of each mode. **Fig. B1** in the appendix displays the most significant causal dependencies for each mode in the two reanalysis datasets during the winter months (DJF). Despite the lack of spatial information provided to the PCMCI causal discovery algorithm, the most significant dependencies predominantly originate from neighboring modes, indicating that the causal network estimation step identifies physically meaningful dependencies between the SLP modes for both reanalysis datasets.


(a) Most significant causal dependencies of each mode of NCEP/NCAR for DJF

(b) Most significant causal dependencies of each mode of ERA5 for DJF

Figure B1. Most significant causal dependencies of each ~~ERA5~~ mode in DJF (December, January, February) for the (a) NCEP/NCAR or (b) ERA5 dataset. The PCMCI causal discovery algorithm identifies physically meaningful links. Despite the lack of spatial information provided to the algorithm, the most significant dependency for a mode generally originates from a neighboring mode. Each mode has, on average, 18 or 20 causal dependencies for NCEP/NCAR or ERA5 respectively, with time lags ranging from 1 to 20 days. Notably, 99.9 % of these dependencies are found with a time lag of less than 10 days.

B1 ~~Simplified ERA5 causal network~~

Appendix C: Causal network evaluation results

For Fig. 2, the Spearman's rank correlation coefficient is calculated between the rankings of climate models to assess variation in rankings across the different reference datasets (NCEP/NCAR and ERA5), yielding a coefficient of 0.91. This confirms a strong consistency between the climate model rankings with the NCEP-NCAR and ERA5 references. The obtained p-value from the Student's t-test is 1.1×10^{-9}, rejecting the null hypothesis of no ordinal correlation between the rankings of models with NCEP/NCAR or ERA5 taken as reference. In **Fig. C1**, we compare the climate models' causal networks' F_1 scores relative to the NCEP-NCAR and ERA5 reference datasets across different seasons (DJF, MAM, JJA, and SON). Although the structure of the causal networks exhibits substantial seasonal variation, the comparison of F_1 scores consistently highlights similar performance patterns across seasons. This consistency reinforces the validity of using season-averaged F_1 scores in the rest of this study.

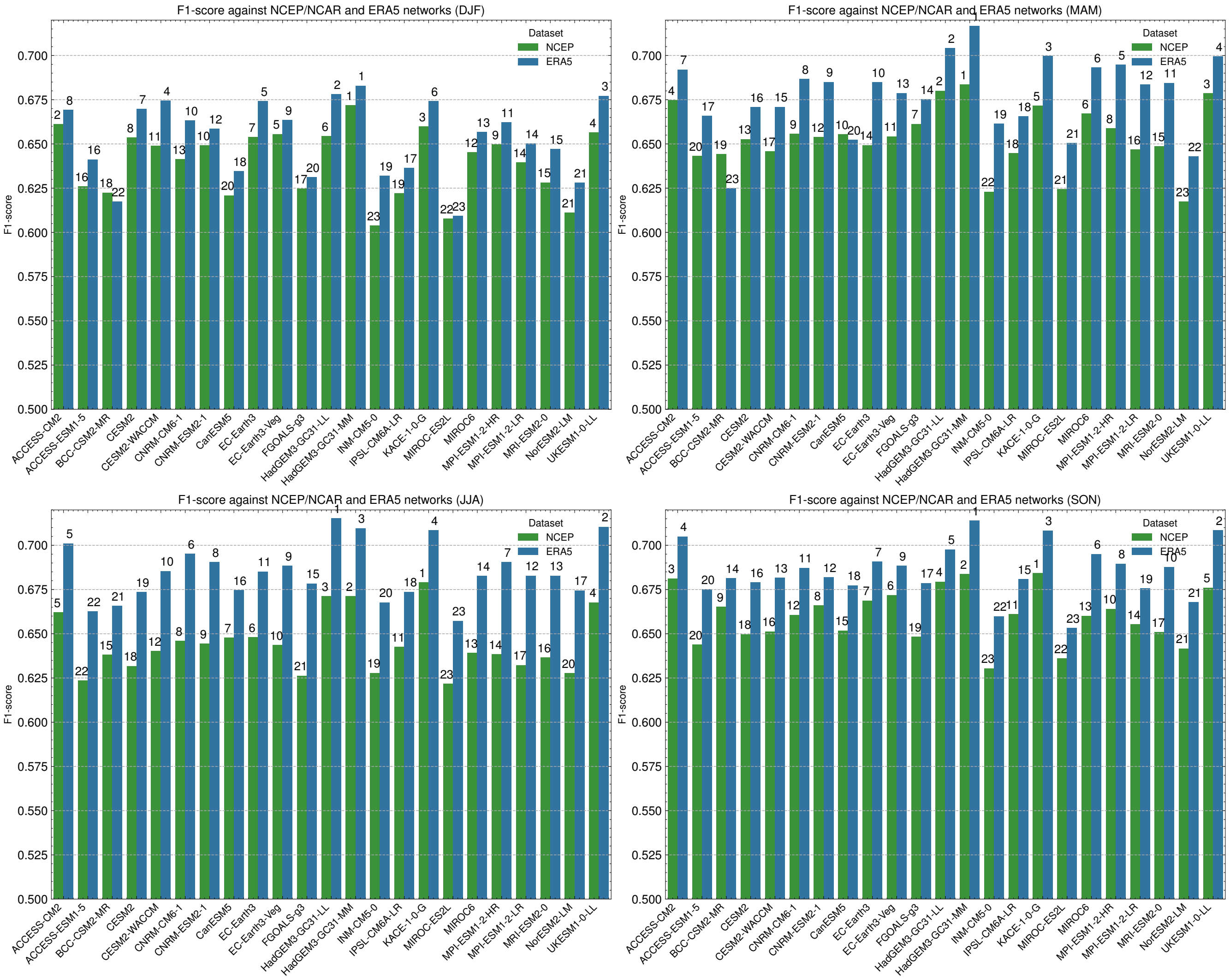


Figure C1. ~~Most significant causal dependencies~~ Comparison of ~~each~~ the climate models' causal networks F_1 scores with NCEP-NCAR (green) and ERA5 ~~mode in DJF~~ (~~December, January, February~~blue) as reference for the four meteorological seasons. ~~The PCMCI causal discovery algorithm identifies physically meaningful links. Despite~~ This figure illustrates the ~~lack~~ similarity between climate models' causal networks and those of ~~spatial information provided to~~ the ~~algorithm~~reference reanalysis datasets, averaged across all available members, using the ~~most significant dependency for a mode generally originates from a neighboring mode~~F_1 score. ~~Each mode has, on average, 20 causal dependencies, with time lags ranging from 1 to 20 days~~Higher F_1 scores indicate greater similarity. ~~Notably, 99.9 %~~ The rank of ~~these dependencies are found with a time lag~~ each model's similarity is denoted on top of ~~less than 10 days~~each bar.

Appendix D: Impact of the number of modes, maximum time lag of PCMCI, and α_{MCI}

In **Fig. D1**, we varied the significance level α_{MCI} of PCMCI from 10^{-5} to 10^{-4} and 10^{-3}. **Fig. D2** demonstrates the effects of reducing the number of modes in the networks from 60 to 50 and decreasing the maximum time lag in the PCMCI algorithm from 20 to 10 days. While these variations affected the F_1 score values moderately, they had a minimal influence on the rankings of the climate models. This was evaluated by calculating the Spearman's rank correlation coefficient for the modified experiments against the baseline experiment presented in the main text (Figure 2a), which used the NCEP/NCAR reference with 60 modes and $\alpha_{MCI} = 10^{-5}$. The correlation coefficients were close to 1, ranging from 0.95 to 0.98, confirming a strong ordinal correlation between the rankings of models in the different experiments. The p-values, all smaller than 10^{-11}, rejected the null hypothesis of no ordinal correlation between the alternative experiments and the baseline experiment.

D1 Impact of α_{MCI} on the causal model evaluation step

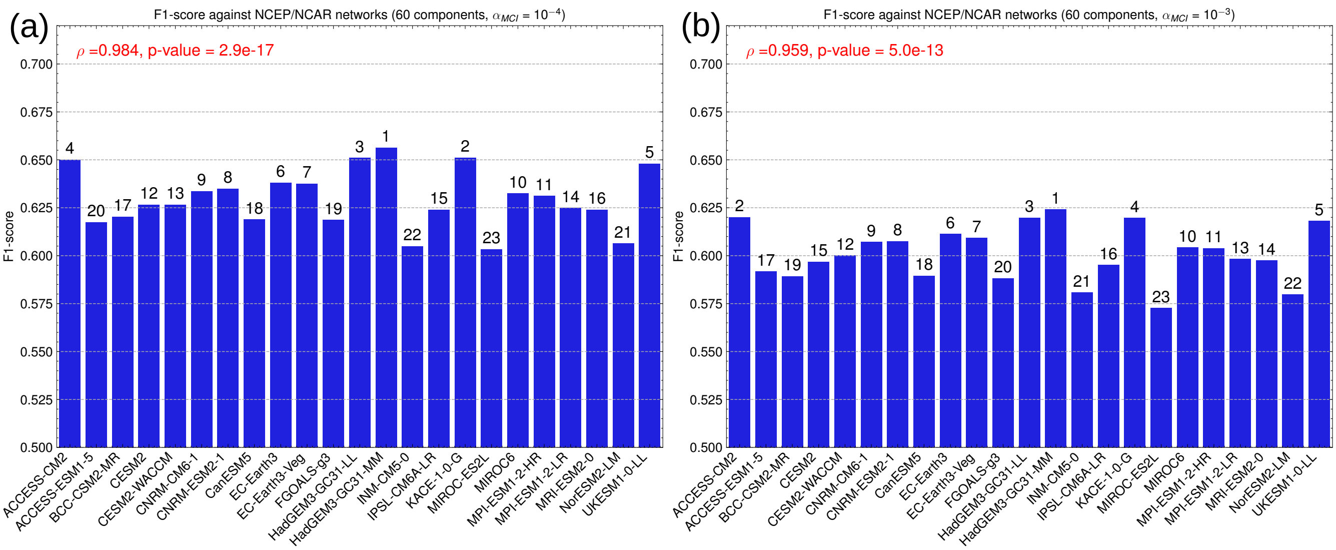


Figure D1. Impact of α_{MCI} on climate models' causal networks F_1 scores with NCEP-NCAR as reference. The causal networks, composed of 60 modes, were constructed using varying levels of α_{MCI}. Specifically, α_{MCI} was varied from (a) 10^{-4} to (b) 10^{-3}, whereas $\alpha_{MCI} = 10^{-5}$ was used in the main text. α_{MCI} represents the significance level for the MCI step in PCMCI. A causal link is established if the MCI test value is equal to or smaller than α_{MCI}. The Spearman's rank correlation coefficient was calculated to compare the variation in model rankings relative to the main text results in Fig. 2a. The resulting Spearman's rank correlation coefficient and the associated p-value from a Student's t-test, testing the null hypothesis of no ordinal correlation between the rankings, are displayed in red.

D2 Impact of the number of modes and maximum time lag on the causal model evaluation step

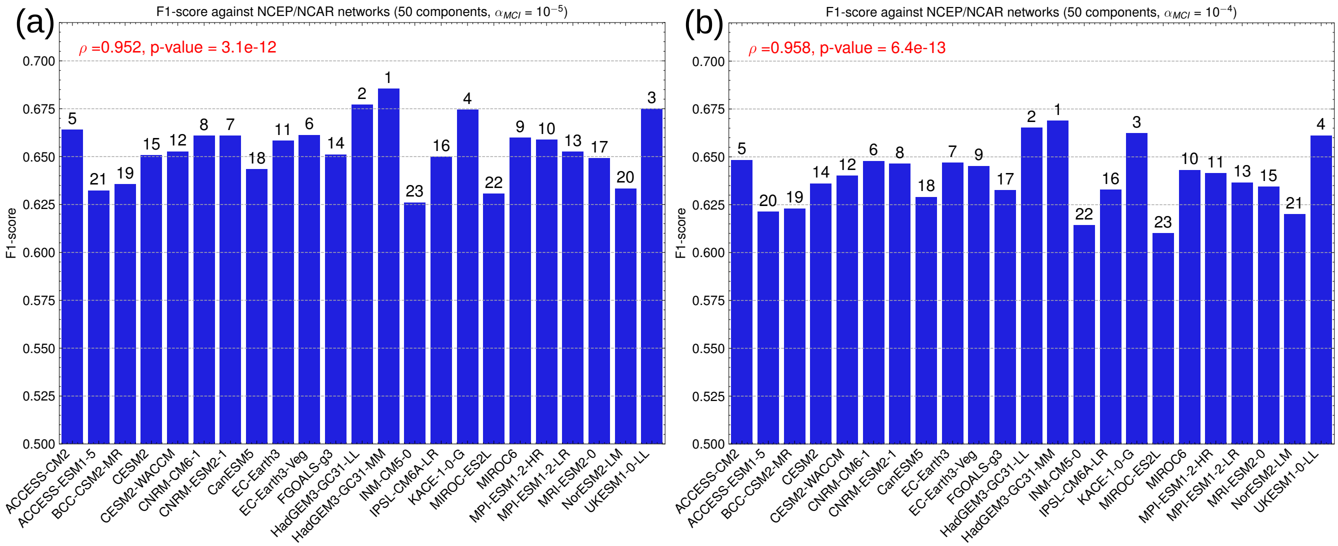


Figure D2. Impact of the number of PCA-Varimax modes on climate models' causal networks F_1 scores with NCEP-NCAR as reference. The causal networks are composed of 50 modes, in contrast to the 60 modes used in the main text. Additionally, the maximum time lag in PCMCI is set to 10 days instead of 20 days. The parameter α_{MCI} of PCMCI is also varied from (a) 10^{-5} to (b) 10^{-4}. The Spearman's rank correlation coefficient was calculated to compare the variation in model rankings relative to the main text results in Fig. 2a. The resulting Spearman's rank correlation coefficient and the associated p-value from a Student's t-test, which tests the null hypothesis of no ordinal correlation between the rankings, are displayed in red at the top of each subfigure.

Appendix E: Calibration of interdependence shape parameter σ_S

In Fig. E1, we note that internal variability itself offers opportunities to learn about the robustness of our method. Specifically, we have found differences between the causal networks of the models, which were shown to be larger than the differences between the causal networks across ensemble members of individual models. This supports the idea that the differences we capture are meaningful and not purely due to internal variability. This finding aligns with results from previous work (Nowack et al., 2020), where this was demonstrated clearly.

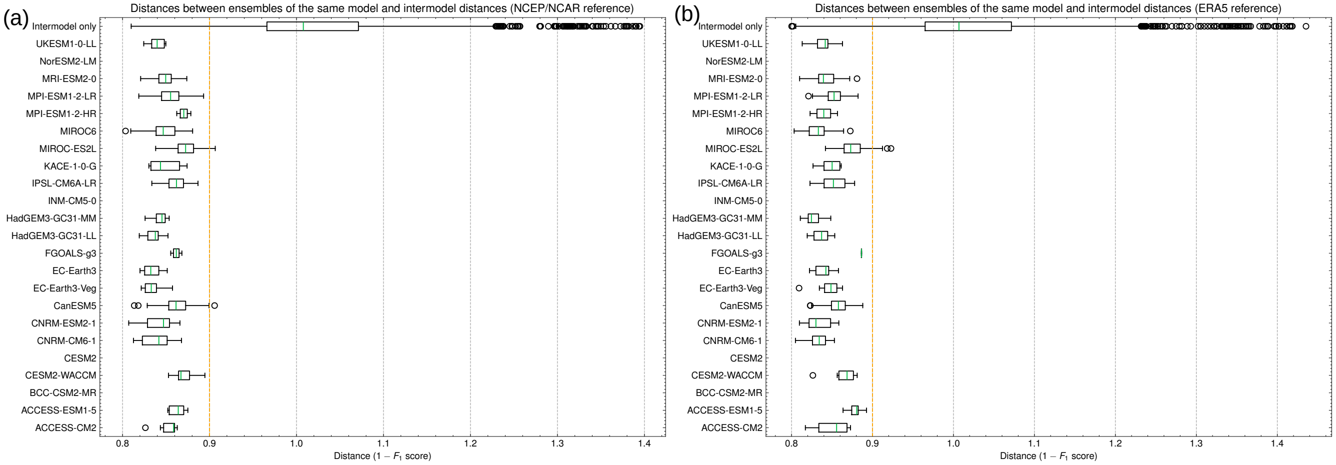


Figure E1. Distances between ensembles of the same model and intermodel distances for (a) NCEP/NCAR and (b) ERA5 taken as reference. The distances are calculated using the complement of the F_1 scores normalized by the median across all models. σ_S is set to 0.9 (orange dashed line) which separates most of the intermodel distances and the intramodel distances.

Appendix F: Weighted projections of land precipitation changes

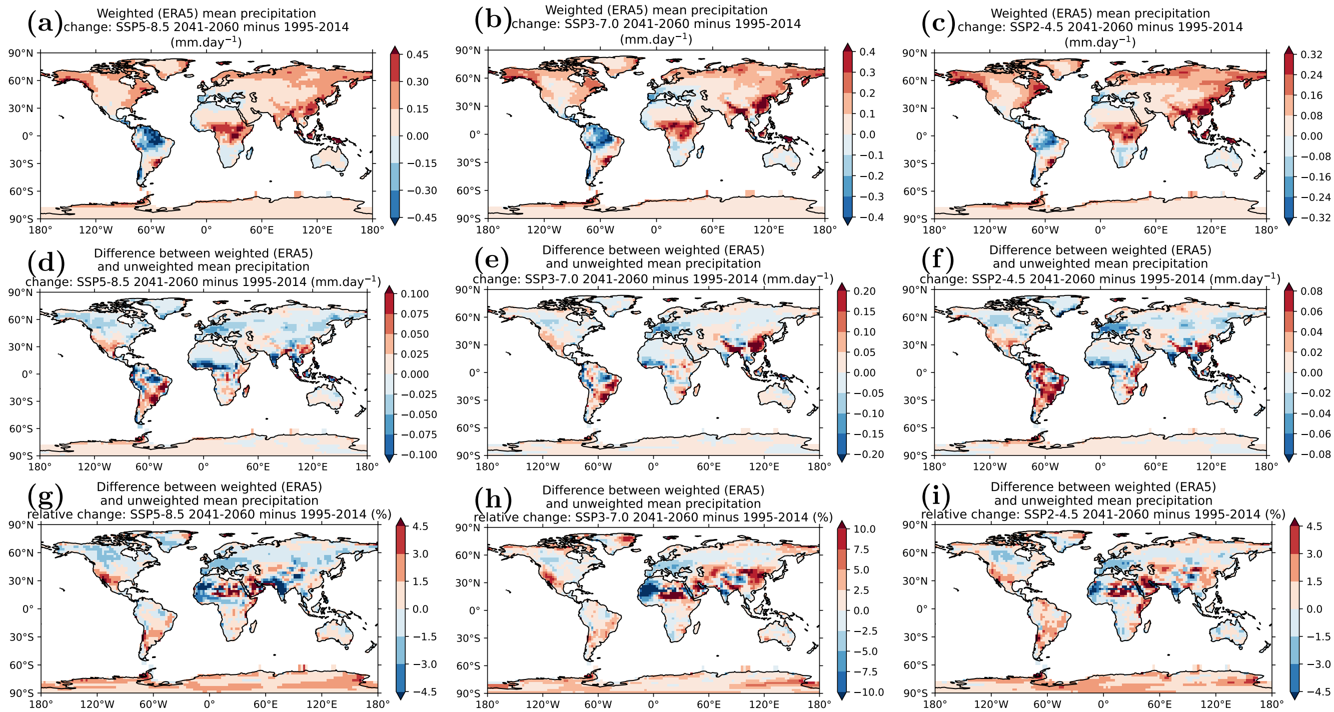

(a)
Weighted (ERA5) mean precipitation change: SSP5-8.5 2041-2060 minus 1995-2014 (mm.day^{-1})
0.45
0.30
0.15
0.00
−0.15
−0.30
−0.45
(b)
Weighted (ERA5) mean precipitation change: SSP3-7.0 2041-2060 minus 1995-2014 (mm.day^{-1})
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
(c)
Weighted (ERA5) mean precipitation change: SSP2-4.5 2041-2060 minus 1995-2014 (mm.day^{-1})
0.32
0.24
0.16
0.08
0.00
−0.08
−0.16
−0.24
−0.32
(d)
Difference between weighted (ERA5) and unweighted mean precipitation change: SSP5-8.5 2041-2060 minus 1995-2014 (mm.day^{-1})
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100
(e)
Difference between weighted (ERA5) and unweighted mean precipitation change: SSP3-7.0 2041-2060 minus 1995-2014 (mm.day^{-1})
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
(f)
Difference between weighted (ERA5) and unweighted mean precipitation change: SSP2-4.5 2041-2060 minus 1995-2014 (mm.day^{-1})
0.08
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
−0.08
(g)
Difference between weighted (ERA5) and unweighted mean precipitation relative change: SSP5-8.5 2041-2060 minus 1995-2014 (%)
4.5
3.0
1.5
0.0
−1.5
−3.0
−4.5
(h)
Difference between weighted (ERA5) and unweighted mean precipitation relative change: SSP3-7.0 2041-2060 minus 1995-2014 (%)
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
(i)
Difference between weighted (ERA5) and unweighted mean precipitation relative change: SSP2-4.5 2041-2060 minus 1995-2014 (%)
4.5
3.0
1.5
0.0
−1.5
−3.0
−4.5
90°N
60°N
30°N
0°
30°S
60°S
90°S
180°
120°W
60°W
0°
60°E
120°E
180°

Figure F1. Patterns of causally weighted projections of mean precipitation change over land in 2041–2060 relative to 1995–2014 for (a) SSP5-8.5, (b) SSP3-7.0, (c) and SSP2-4.5 scenarios. The differences between the weighted and unweighted mean precipitation change are shown in (d) for SSP5-8.5, (e) for SSP3-7.0, and (f) for SSP2-4.5 scenarios. The differences between the weighted and unweighted mean precipitation relative change are shown in (d) for SSP5-8.5, (e) for SSP3-7.0, and (f) for SSP2-4.5 scenarios. ERA5 was used as a reference for the causal weighting.

Code and data availability. The code is written in Python and will be published under https://github.com/EyringMLClimateGroup/debeire24esd_Causalweighting. CMIP6 data were accessed from https://esgf-metagrid.cloud.dkrz.de/search. ERA5 data are provided by the ECMWF from https://cds.climate.copernicus.eu (Hersbach et al., 2020). NCEP/NCAR data are provided by the NOAA Physical Sciences Laboratory from https://psl.noaa.gov/data/reanalysis/reanalysis.shtml (Kalnay et al., 1996).

Author contributions. KD conceptualized the study with the help of all co-authors, he developed the coding scripts, conducted the analysis and generated all plots. KD wrote the manuscript with help of LB and VE, with contributions and feedback from all other co-authors.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. KD, LB, and VE have received funding for this study by the European Research Council (ERC) Synergy Grant "Understanding and Modelling the Earth System with Machine Learning (USMILE)" under the Horizon 2020 research and innovation programme (Grant agreement No. 855187). KD and LB received additional funding from the DLR-internal project CausalAnomalies. VE was additionally supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through the Gottfried Wilhelm Leibniz Prize awarded to Veronika Eyring (Reference No. EY 22/2-1). PN was supported by the UK Natural Environment Research Council (grant no. NE/V012045/1). JR received funding from the ERC Starting Grant CausalEarth under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 948112). We acknowledge the computational resources of the Deutsches Klimarechenzentrum (DKRZ, Hamburg, Germany) used in this study (grant no. 1083). We acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modeling, coordinated and promoted CMIP, and thank the climate modeling groups for producing and sharing their model outputs, as well as the Earth System Grid Federation (ESGF) for archiving and providing access to the data. ChatGPT was used on an earlier version of the draft of this manuscript to improve the clarity of certain paragraphs.

References

Abramowitz, G. and Bishop, C. H.: Climate Model Dependence and the Ensemble Dependence Transformation of CMIP Projections, Journal of Climate, 28, 2332–2348, https://doi.org/10.1175/JCLI-D-14-00364.1, 2015.

Allan, R. P., Barlow, M., Byrne, M. P., Cherchi, A., Douville, H., Fowler, H. J., Gan, T. Y., Pendergrass, A. G., Rosenfeld, D., Swann, A. L. S., Wilcox, L. J., and Zolina, O.: Advances in understanding large-scale responses of the water cycle to climate change, Annals of the New York Academy of Sciences, 1472, 49–75, https://doi.org/https://doi.org/10.1111/nyas.14337, 2020.

Allen, M. R. and Ingram, W. J.: Constraints on future changes in climate and the hydrologic cycle, Nature, 419, 224–232, 2002.

Benestad, R. E., Hanssen-Bauer, I., and Førland, E. J.: An Evaluation of Statistical Models for Downscaling Precipitation and Their Ability to Capture Long-Term Trends, International Journal of Climatology, 27, 649–665, https://doi.org/10.1002/joc.1421, 2007.

Beydoun, H. and Hoose, C.: Aerosol-Cloud-Precipitation Interactions in the Context of Convective Self-Aggregation, Journal of Advances in Modeling Earth Systems, 11, 1066–1087, https://doi.org/10.1029/2018MS001523, 2019.

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 System Dynamics, 11, 995–1012, https://doi.org/10.5194/esd-11-995-2020, 2020.

Costa-Cabral, M., Rath, J. S., Mills, W. B., Roy, S. B., Bromirski, P. D., and Milesi, C.: Projecting and Forecasting Winter Precipitation Extremes and Meteorological Drought in California Using the North Pacific High Sea Level Pressure Anomaly, Journal of Climate, 29, 5009–5026, https://doi.org/10.1175/JCLI-D-15-0525.1, 2016.

Cox, P. M., Huntingford, C., and Williamson, M. S.: Emergent Constraint on Equilibrium Climate Sensitivity from Global Temperature Variability, Nature, 553, 319–322, https://doi.org/10.1038/nature25450, 2018.

Dai, P., Nie, J., Yu, Y., and Wu, R.: Constraints on Regional Projections of Mean and Extreme Precipitation under Warming, Proceedings of the National Academy of Sciences, 121, e2312400 121, https://doi.org/10.1073/pnas.2312400121, 2024.

Dia-Diop, A., Wade, M., Zebaze, S., Diop, A. B., Efon, E., Lenouo, A., and Diop, B.: Influence of Sea Level Pressure on Inter-Annual Rainfall Variability in Northern Senegal in the Context of Climate Change, Atmospheric and Climate Sciences, 12, 113–131, https://doi.org/10.4236/acs.2022.121009, 2021.

Doblas-Reyes, F., Sörensson, A., Almazroui, M., Dosio, A., Gutowski, W., Haarsma, R., Hamdi, R., Hewitson, B., Kwon, W.-T., Lamptey, B., Maraun, D., Stephenson, T., Takayabu, I., Terray, L., Turner, A., and Zuo, Z.: Linking Global to Regional Climate Change, in: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, edited by Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J. B. R., Maycock, T. K., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., book section 10, pp. 1363–1512, Cambridge University Press, Cambridge, UK and New York, NY, USA, https://doi.org/10.1017/9781009157896.012, 2021.

Douville, H., Raghavan, K., Renwick, J., Allan, R., Arias, P., Barlow, M., Cerezo-Mota, R., Cherchi, A., Gan, T., Gergis, J., Jiang, D., Khan, A., Pokam Mba, W., Rosenfeld, D., Tierney, J., and Zolina, O.: Water Cycle Changes, in: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, edited by Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J. B. R., Maycock, T. K., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., book section 8, pp. 1055–1210, Cambridge University Press, Cambridge, UK and New York, NY, USA, https://doi.org/10.1017/9781009157896.010, 2021.

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

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

Eyring, V., Bock, L., Lauer, A., Righi, M., Schlund, M., Andela, B., Arnone, E., Bellprat, O., Brötz, B., Caron, L.-P., Carvalhais, N., Cionni, I., Cortesi, N., Crezee, B., Davin, E. L., Davini, P., Debeire, K., de Mora, L., Deser, C., Docquier, D., Earnshaw, P., Ehbrecht, C., Gier, B. K., Gonzalez-Reviriego, N., Goodman, P., Hagemann, S., Hardiman, S., Hassler, B., Hunter, A., Kadow, C., Kindermann, S., Koirala, S., Koldunov, N., Lejeune, Q., Lembo, V., Lovato, T., Lucarini, V., Massonnet, F., Müller, B., Pandde, A., Pérez-Zanón, N., Phillips, A., Predoi, V., Russell, J., Sellar, A., Serva, F., Stacke, T., Swaminathan, R., Torralba, V., Vegas-Regidor, J., von Hardenberg, J., Weigel, K., and Zimmermann, K.: Earth System Model Evaluation Tool (ESMValTool) v2.0 – an Extended Set of Large-Scale Diagnostics for Quasi-Operational and Comprehensive Evaluation of Earth System Models in CMIP, Geoscientific Model Development, 13, 3383–3438, https://doi.org/10.5194/gmd-13-3383-2020, 2020.

Eyring, V., Collins, W. D., Gentine, P., Barnes, E. A., Barreiro, M., Beucler, T., Bocquet, M., Bretherton, C. S., Christensen, H. M., 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 modeling and analysis with machine learning, Nature Climate Change, https://doi.org/10.1038/s41558-024-02095-y, 2024a.

Eyring, V., Gentine, P., Camps-Valls, G., Lawrence, D. M., and Reichstein, M.: AI-empowered Next-generation Multiscale Climate Modeling for Mitigation and Adaptation, Nature Geoscience, https://doi.org/10.1038/s41561-024-01527-w, 2024b.

Ferguglia, O., von Hardenberg, J., and Palazzi, E.: Robustness of Precipitation Emergent Constraints in CMIP6 Models, Climate Dynamics, 61, 1439–1450, https://doi.org/10.1007/s00382-022-06634-1, 2023.

Hall, A. and Qu, X.: Using the Current Seasonal Cycle to Constrain Snow Albedo Feedback in Future Climate Change, Geophysical Research Letters, 33, https://doi.org/10.1029/2005GL025127, 2006.

Hawkins, E. and Sutton, R.: The Potential to Narrow Uncertainty in Regional Climate Predictions, Bulletin of the American Meteorological Society, 90, 1095–1108, https://doi.org/10.1175/2009BAMS2607.1, 2009.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, last accessed: July 2024, 2020.

IPCC: Summary for Policymakers, in: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, edited by Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J. B. R., Maycock, T. K., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., pp. 1–31, Cambridge University Press, Cambridge, UK and New York, NY, USA, https://doi.org/10.1017/9781009157896.001, 2021.

Kaiser, H. F.: The Varimax Criterion for Analytic Rotation in Factor Analysis, Psychometrika, 23, 187–200, https://doi.org/10.1007/BF02289233, 1958.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R., and Joseph, D.: The NCEP/NCAR 40-Year Reanalysis Project, Bulletin of the American Meteorological Society, 77, 437–472, last accessed: July 2024, 1996.

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

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, Geophysical Research Letters, 44, 1909–1918, https://doi.org/10.1002/2016GL072012, 2017.

Kotz, M., Levermann, A., and Wenz, L.: The Effect of Rainfall Changes on Economic Production, Nature, 601, 223–227, https://doi.org/10.1038/s41586-021-04283-8, 2022.

Kuma, P., Bender, F. A.-M., and Jönsson, A. R.: Climate Model Code Genealogy and Its Relation to Climate Feedbacks and Sensitivity, Journal of Advances in Modeling Earth Systems, 15, e2022MS003 588, https://doi.org/10.1029/2022MS003588, 2023.

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

Lavers, D., Prudhomme, C., and Hannah, D. M.: European Precipitation Connections with Large-Scale Mean Sea-Level Pressure (MSLP) Fields, Hydrological Sciences Journal, 58, 310–327, https://doi.org/10.1080/02626667.2012.754545, 2013.

Lee, J.-Y., Marotzke, J., Bala, G., Cao, L., Corti, S., Dunne, J., Engelbrecht, F., Fischer, E., Fyfe, J., Jones, C., Maycock, A., Mutemi, J., Ndiaye, O., Panickal, S., and Zhou, T.: Future Global Climate: Scenario-Based Projections and Near-Term Information, in: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, edited by Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J. B. R., Maycock, T. K., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., book section 4, pp. 553–672, Cambridge University Press, Cambridge, UK and New York, NY, USA, https://doi.org/10.1017/9781009157896.006, 2021.

Liang, Y., Gillett, N. P., and Monahan, A. H.: Climate Model Projections of 21st Century Global Warming Constrained Using the Observed Warming Trend, Geophysical Research Letters, 47, e2019GL086 757, https://doi.org/10.1029/2019GL086757, 2020.

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 System Dynamics, 11, 807–834, https://doi.org/10.5194/esd-11-807-2020, 2020.

Müller-Plath, G., Lüdecke, H.-J., and Lüning, S.: Long-Distance Air Pressure Differences Correlate with European Rain, Scientific Reports, 12, 10 191, https://doi.org/10.1038/s41598-022-14028-w, 2022.

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 System Dynamics, 11, 737–750, https://doi.org/10.5194/esd-11-737-2020, 2020.

Nowack, P. and Watson-Parris, D.: Opinion: Why All Emergent Constraints Are Wrong but Some Are Useful &ndash; a Machine Learning Perspective, EGUsphere, pp. 1–28, https://doi.org/10.5194/egusphere-2024-1636, 2024.

Nowack, P., Runge, J., Eyring, V., and Haigh, J. D.: Causal Networks for Climate Model Evaluation and Constrained Projections, Nature Communications, 11, 1415, https://doi.org/10.1038/s41467-020-15195-y, 2020.

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, Climatic Change, 122, 387–400, https://doi.org/10.1007/s10584-013-0905-2, 2014.

Pendergrass, A. G.: The Global-Mean Precipitation Response to CO2-Induced Warming in CMIP6 Models, Geophysical Research Letters, 47, e2020GL089 964, https://doi.org/10.1029/2020GL089964, 2020.

Qu, X., Hall, A., DeAngelis, A. M., Zelinka, M. D., Klein, S. A., Su, H., Tian, B., and Zhai, C.: On the Emergent Constraints of Climate Sensitivity, Journal of Climate, 31, 863–875, https://doi.org/10.1175/JCLI-D-17-0482.1, 2018.

Ramsay, J. O. and Silverman, B. W.: Principal Components Analysis for Functional Data, in: Functional Data Analysis, pp. 147–172, Springer, New York, NY, ISBN 978-0-387-22751-1, https://doi.org/10.1007/0-387-22751-2_8, 2005.

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

Rohe, K. and Zeng, M.: Vintage Factor Analysis with Varimax Performs Statistical Inference, Journal of the Royal Statistical Society Series B: Statistical Methodology, 85, 1037–1060, https://doi.org/10.1093/jrsssb/qkad029, 2023.

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, Nature Communications, 10, 2553, https://doi.org/10.1038/s41467-019-10105-3, 2019a.

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D.: Detecting and Quantifying Causal Associations in Large Nonlinear Time Series Datasets, Science Advances, 5, eaau4996, https://doi.org/10.1126/sciadv.aau4996, 2019b.

Runge, J., Gerhardus, A., Varando, G., Eyring, V., and Camps-Valls, G.: Causal Inference for Time Series, Nature Reviews Earth & Environment, 4, 487–505, https://doi.org/10.1038/s43017-023-00431-y, 2023.

Sanderson, B. M., Wehner, M., and Knutti, R.: Skill and Independence Weighting for Multi-Model Assessments, Geoscientific Model Development, 10, 2379–2395, https://doi.org/10.5194/gmd-10-2379-2017, 2017.

Schlund, M., Eyring, V., Camps-Valls, G., Friedlingstein, P., Gentine, P., and Reichstein, M.: Constraining Uncertainty in Projected Gross Primary Production With Machine Learning, Journal of Geophysical Research: Biogeosciences, 125, e2019JG005 619, https://doi.org/10.1029/2019JG005619, 2020a.

Schlund, M., Lauer, A., Gentine, P., Sherwood, S. C., and Eyring, V.: Emergent Constraints on Equilibrium Climate Sensitivity in CMIP5: Do They Hold for CMIP6?, Earth System Dynamics, 11, 1233–1258, https://doi.org/10.5194/esd-11-1233-2020, 2020b.

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

Seneviratne, S., Zhang, X., Adnan, M., Badi, W., Dereczynski, C., Di Luca, A., Ghosh, S., Iskandar, I., Kossin, J., Lewis, S., Otto, F., Pinto, I., Satoh, M., Vicente-Serrano, S., Wehner, M., and Zhou, B.: Weather and Climate Extreme Events in a Changing Climate, in: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, edited by Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J. B. R., Maycock, T. K., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., book section 11, pp. 1513–1765, Cambridge University Press, Cambridge, UK and New York, NY, USA, https://doi.org/10.1017/9781009157896.013, 2021.

Shaffer, R. E.: Multi- and Megavariate Data Analysis. Principles and Applications, I. Eriksson, E. Johansson, N. Kettaneh-Wold and S. Wold, Umetrics Academy, Umeå, 2001, ISBN 91-973730-1-X, 533pp., Journal of Chemometrics, 16, 261–262, https://doi.org/10.1002/cem.713, 2002.

Shiogama, H., Watanabe, M., Kim, H., and Hirota, N.: Emergent Constraints on Future Precipitation Changes, Nature, 602, 612–616, https://doi.org/10.1038/s41586-021-04310-8, 2022.

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

Thackeray, C. W., Hall, A., Norris, J., and Chen, D.: Constraining the Increased Frequency of Global Precipitation Extremes under Warming, Nature Climate Change, 12, 441–448, https://doi.org/10.1038/s41558-022-01329-1, 2022.

Thompson, R. and Green, D. N.: Mediterranean Precipitation and Its Relationship with Sea Level Pressure Patterns, Annals of Geophysics, 47, https://doi.org/10.4401/ag-3364, 2004.

Tokarska, K. B., Hegerl, G. C., Schurer, A. P., Forster, P. M., and Marvel, K.: Observational Constraints on the Effective Climate Sensitivity from the Historical Period, Environmental Research Letters, 15, 034 043, https://doi.org/10.1088/1748-9326/ab738f, 2020.

Trenberth, K. E. and Zhang, Y.: How Often Does It Really Rain?, Bulletin of the American Meteorological Society, 99, 289–298, https://doi.org/10.1175/BAMS-D-17-0107.1, 2018.

Vejmelka, M., Pokorná, L., Hlinka, J., Hartman, D., Jajcay, N., and Paluš, M.: Non-Random Correlation Structures and Dimensionality Reduction in Multivariate Climate Data, Climate Dynamics, 44, 2663–2682, https://doi.org/10.1007/s00382-014-2244-z, 2015.

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

Zhang, W., Zhou, T., and Wu, P.: Anthropogenic Amplification of Precipitation Variability over the Past Century, Science, 385, 427–432, https://doi.org/10.1126/science.adp0212, 2024.