Geophysical Research Letters

RESEARCH LETTER
10.1029/2022GL102466

How Credibly Do CMIP6 Simulations Capture Historical Mean and Extreme Precipitation Changes?

Markus G. Donat[1,2], Carlos Delgado-Torres[1], Paolo De Luca[1], Rashed Mahmood[1,3], Pablo Ortega[1], and Francisco J. Doblas-Reyes[1,2]

[1]Barcelona Supercomputing Center, Barcelona, Spain, [2]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain, [3]Department of Geography, University of Montreal, Montreal, QC, Canada

Abstract Future precipitation changes are typically estimated from climate model simulations, while the credibility of such projections needs to be assessed by their ability to capture observed precipitation changes. Here we evaluate how skillfully historical climate simulations contributing to the Coupled Model Intercomparison Project Phase 6 (CMIP6) capture observed changes in mean and extreme precipitation. We find that CMIP6 historical simulations skillfully represent observed precipitation changes over large parts of Europe, Asia, northeastern North America, parts of South America and western Australia, whereas a lack of skill is apparent in western North America and parts of Africa. In particular in regions with moderate skill the availability of very large ensembles can be beneficial to improve the simulation accuracy. CMIP6 simulations are regionally skillful where they capture observed (positive or negative) trends, whereas a lack of skill is found in regions characterized by negative observed precipitation trends where CMIP6 simulates increases.

Plain Language Summary Climate models are the primary tools to predict future changes in precipitation related to global warming. These predictions can however only usefully inform adaptation measures if they can be trusted. Here we evaluate the trustworthiness of climate model-simulated precipitation changes based on their capability to correctly capture observed precipitation changes. We apply skill measures commonly used for the evaluation of seasonal to decadal climate predictions to historical climate simulations. We perform this analysis for total precipitation accumulations and indicators of precipitation extremes. The level of skill differs between regions and can be sensitive to the number of available simulations, with some regions benefitting from very large simulation ensembles. Mean and extreme precipitation are skillfully predicted in similar regions, including large parts of Europe and Asia. Lack of skill typically occurs in regions where observed precipitation is characterized by downward trends but Coupled Model Intercomparison Project Phase 6 models simulate increases. This study helps understand the trustworthiness of climate simulations to realistically capture precipitation changes, identifying regions where current models are more or less capable.

Key Points:

- Coupled Model Intercomparison Project Phase 6 (CMIP6) realistically simulates observed changes in mean and extreme precipitation in large parts of Europe and Asia and other land regions
- In regions with moderate skill and observed precipitation subject to multi-decadal variations the availability of very large ensembles is beneficial
- Lack of skill occurs primarily in regions where negative precipitation trends are observed but CMIP6 simulates increases

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

M. G. Donat,
markus.donat@bsc.es

Citation:

Donat, M. G., Delgado-Torres, C., De Luca, P., Mahmood, R., Ortega, P., & Doblas-Reyes, F. J. (2023). How credibly do CMIP6 simulations capture historical mean and extreme precipitation changes? *Geophysical Research Letters*, *50*, e2022GL102466. https://doi.org/10.1029/2022GL102466

Received 8 DEC 2022
Accepted 15 JUN 2023

Author Contributions:

Conceptualization: Markus G. Donat
Formal analysis: Carlos Delgado-Torres, Paolo De Luca, Rashed Mahmood
Funding acquisition: Markus G. Donat
Investigation: Markus G. Donat, Carlos Delgado-Torres, Paolo De Luca, Rashed Mahmood

1. Introduction

The local characteristics and spatial distribution of precipitation are changing as a consequence of global warming (Allen & Ingram, 2002; Trenberth, 2011). These changes include a wettening of high latitude regions and a precipitation decrease in some sub-tropical regions, and an intensification of precipitation events in large parts of the world (Contractor et al., 2021). This intensification affects in particular the stronger precipitation events, which manifests in an increasing frequency and intensity of heavy precipitation extremes (Donat et al., 2016, 2019; Westra et al., 2013).

Precipitation changes are typically analyzed from observational data sets or climate model simulations, and both types of data have their specific advantages and shortcomings. While observational data sets represent the climate trajectory of the real world, the network of high-quality long-term observations is sparse (Donat et al., 2013; Dunn et al., 2020), and the uncertainties across different observational precipitation data sets are substantial (Alexander et al., 2020; Herold et al., 2016). Climate models, on the other hand, simulate their own specific attractors, and the representation of climate variations may not be in phase with the observed climate. Furthermore, state-of-the-art climate models (such as those contributing to Coupled Model Intercomparison Project Phase 6, CMIP6; Eyring et al., 2016) have well-known shortcomings in the simulation of precipitation and relevant processes. These shortcomings include that key processes related to cloud formation and the formation of water droplets, or atmospheric convection, happen on spatial scales that are orders of magnitude smaller than the typical grid resolution

Methodology: Markus G. Donat, Carlos Delgado-Torres, Paolo De Luca, Rashed Mahmood, Pablo Ortega, Francisco J. Doblas-Reyes
Project Administration: Markus G. Donat
Supervision: Markus G. Donat
Validation: Carlos Delgado-Torres, Paolo De Luca, Rashed Mahmood
Visualization: Carlos Delgado-Torres, Paolo De Luca, Rashed Mahmood
Writing – original draft: Markus G. Donat
Writing – review & editing: Markus G. Donat, Carlos Delgado-Torres, Paolo De Luca, Rashed Mahmood, Pablo Ortega, Francisco J. Doblas-Reyes

of global climate models, and are therefore parameterized. The choice of micro-physics or convection schemes represents major uncertainties in the simulation of precipitation (Morrison et al., 2020; O'Gorman, 2012).

In light of these shortcomings and inconsistencies, it is important to evaluate the model-simulated precipitation in comparison to observations. Such evaluations have in the past often focused on regional mean values or spatial patterns (e.g., Bador et al., 2020; Lorenz et al., 2014; Sillmann et al., 2013), but these are not informative about the credibility of simulated changes. Some studies have compared linear trend estimates of precipitation between model simulations and observations, and then discuss areas of agreement or disagreement of these trends (Knutson & Zeng, 2018; Kumar et al., 2013; Vicente-Serrano et al., 2022). To some extent also interpretable as evaluation of simulated against observed changes, detection and attribution studies quantified to what extent simulations with specific forcings can explain observed changes (Min et al., 2011; Paik et al., 2020; Zhang et al., 2007). These studies usually focus on larger regional averages to remove noise effects, and detected effects from increasing atmospheric greenhouse gas concentrations (and other forcings) in global and regional changes in mean and extreme precipitation.

Here we assess how well forced simulations of the historical climate can capture observed precipitation changes, using skill metrics typically applied to evaluate climate predictions (e.g., Goddard et al., 2012). While these metrics are widely used to evaluate (initialized) seasonal to decadal climate predictions (Delgado-Torres et al., 2022; Meehl et al., 2021; Smith et al., 2019), we demonstrate here their use for the evaluation of (uninitialized) simulations of the historical climate over the past several decades since 1950. Compared to previous evaluations of mean values or trend slopes mentioned above, this allows us to obtain quantifiable measures of agreement between observed and simulated precipitation changes.

We focus the evaluation on averages over 20-year windows, as this is the temporal aggregation often used when analyzing climate change projections (Lee et al., 2021). This is often based on the assumption that the effect of internal climate variability is small on these timescales, and the ensemble mean highlights a forced signal (Risbey et al., 2022). It is however largely unclear how representative such simulations can be for changes in the climate of the real world. In particular for observations and individual ensemble members the impact of internal variability on regional precipitation can still be substantial for 20-year averages, which (besides incorrect response to forcing) is why the simulated precipitation may not well represent the climate evolution of the real world. We consider different aspects of precipitation, such as annual precipitation totals and indices representing the intensity and accumulated amounts of precipitation extremes. We further investigate the effects of ensemble size on the skill of simulated precipitation, and discuss the spatial patterns of skill in comparison to global trend patterns.

2. Data and Methods

We use precipitation data from a total of 32 global climate models that contributed to CMIP6. Most of these models provide several ensemble members, and we included all available simulations at the time of analyses. When performing the analyses we were able to include 211 different ensemble members providing monthly precipitation rates (used for the analysis of annual totals). For the analysis of precipitation extremes indices based on daily precipitation rates we had access to 149 members from 19 different climate models (Table S1 in Supporting Information S1).

The CMIP6 models provide daily or monthly average precipitation fluxes, which are converted into daily (for the calculation of extremes indices) and monthly (for the calculation of annual precipitation totals) precipitation amounts, respectively. We use data from the historical simulations until 2014, and concatenate these with data from the SSP2-4.5 scenario (O'Neill et al., 2016) from 2015 onwards to match the temporal coverage of the observational data (until 2019 for mean precipitation measures and until 2016 for extreme precipitation measures, defined by the availability of observational data to compare against; note that differences between the scenarios are very small in these early years).

We use a range of different gridded observational data sets as references for the evaluations. These are monthly precipitation provided by the Global Precipitation Climatology Centre (GPCC) Full Data Monthly Product Version 2020 (Becker et al., 2013) and the Climate Research Unit CRU TS (Mitchell & Jones, 2005) version 4.04. We use these monthly precipitation data sets to evaluate annual precipitation accumulations (also referred to as mean precipitation in this study). We use the Rainfall Estimates on a Gridded Network (REGEN; Contractor et al.; 2020) data set of gridded daily precipitation fields to calculate extreme precipitation indices. We also

use extreme precipitation indices from the HadEX3 data set (Dunn et al., 2020), which was generated by gridding extremes indices calculated from station time series. These two data sets best cover observed precipitation extremes over the investigation period from the mid-20th century to the recent past.

We analyze annual totals of precipitation (PR), annual precipitation from very wet days (R95p), and annual maximum 5-day precipitation amount (Rx5day) (Zhang et al., 2011). PR is simply the annual sum of monthly precipitation totals (which is proportional to the annual mean of average precipitation rates). R95p is defined as the annual sum of precipitation on days that exceed the 95th percentile of wet-day precipitation amounts (the percentile was calculated over the base period 1981–2010). Rx5day is the annual maximum of 5-day running accumulated precipitation. The CMIP6 precipitation indices were further masked to remove ocean grid cells and all latitudes south of 60S, focusing the study on land regions excluding Antarctica. These precipitation indices are calculated at the native grid resolution of each model, and then regridded (using first-order conservative remapping) to a common $2.8° \times 2.8°$ grid (the resolution of the model with the coarsest resolution in this study, CanESM5) for multi-model analysis and evaluation against the observational data. The remapped fields are than further masked to match the coverage of the observational products, removing those grid cells with more than 50% missing values in the observed data sets. Our analysis focuses on the data period 1950 to 2019 (for PR) and 1950 to 2016 (for R95p and Rx5day), defined by common temporal coverage and times with reasonable spatial coverage across observational data sets.

Observational data sets of precipitation are subject to substantial uncertainties with regard to the estimated precipitation intensities and amounts (Alexander et al., 2020) and may be affected by scaling issues related to the representativeness of a grid box value (Avila et al., 2015). To avoid potential biases related to such uncertainties to affect those skill measures accounting for bias, we transformed the data of precipitation amounts (PR) and intensities (Rx5day) by dividing each annual value by the climatology for that index during 1981–2010. This results in time series of relative anomalies where the value of 1 corresponds to average conditions, values larger than 1 indicate above-average precipitation in a certain year, and values below 1 indicate below-average precipitation.

Within CMIP6, modeling centers are requested to provide several ensemble members, which sample different phases of internal climate variability. In this study we use several ensemble approaches to see how such choices may affect the skill of the ensemble. One approach is to use all available simulations and pool them into a large multi-model ensemble (211 members for total precipitation and 149 members for extremes indices), which results in an ensemble mean where models providing many ensemble members contribute with a higher weight. As an alternative approach we construct a multi-model ensemble with equal weight for each model, by first averaging all members of any one model before combining these model-specific averages into a multi-model mean or distribution. A third (also equally weighted) approach is to construct a multi-model ensemble using only one run per model (typically member r1i1p1f1 when available), resulting in 32 ensemble members for annual precipitation totals and 19 ensemble members for precipitation extremes indices.

For a more systematic analysis of how the skill measures depend on the ensemble size, we also sub-sampled ensembles for all possible ensemble sizes that can be constructed out of the available members. For each ensemble size between one and the total available simulations, we randomly sampled 1,000 possible ensembles that could be constructed based on the available simulations. We used a resampling without replacement, to ensure that each simulation cannot be included more than once in any resampled ensemble. Repetition is, however, allowed between the different random samples of a given ensemble size (and occurs in particular for large ensemble sizes).

For evaluating regional precipitation averages, we mask the data to the IPCC reference regions as defined for the Working Group I 6th Assessment Report (Iturbide et al., 2020, see also Figure S1 in Supporting Information S1). Regional averages are calculated as area-weighted averages given the different grid cell sizes.

We consider three different measures typically used for forecast evaluation to assess different aspects of the degree of agreement between the CMIP6 simulated precipitation and observations. These are the Spearman rank correlation (Spearman, 1904), the Root-Mean-Squared Skill Score (RMSSS; Murphy, 1988), and the Ranked Probability Skill Score (RPSS; Wilks, 2011). Please refer to Text S1 in Supporting Information S1 for a more detailed description of which aspects of forecast quality these measure and how they are calculated.

We quantify temporal trends by fitting a linear least-squares regression to the annual data points. The trend significance is estimated using a modified Mann-Kendall test for serially correlated data following the variance correction approach by Hamed and Ramachandra Rao (1998).

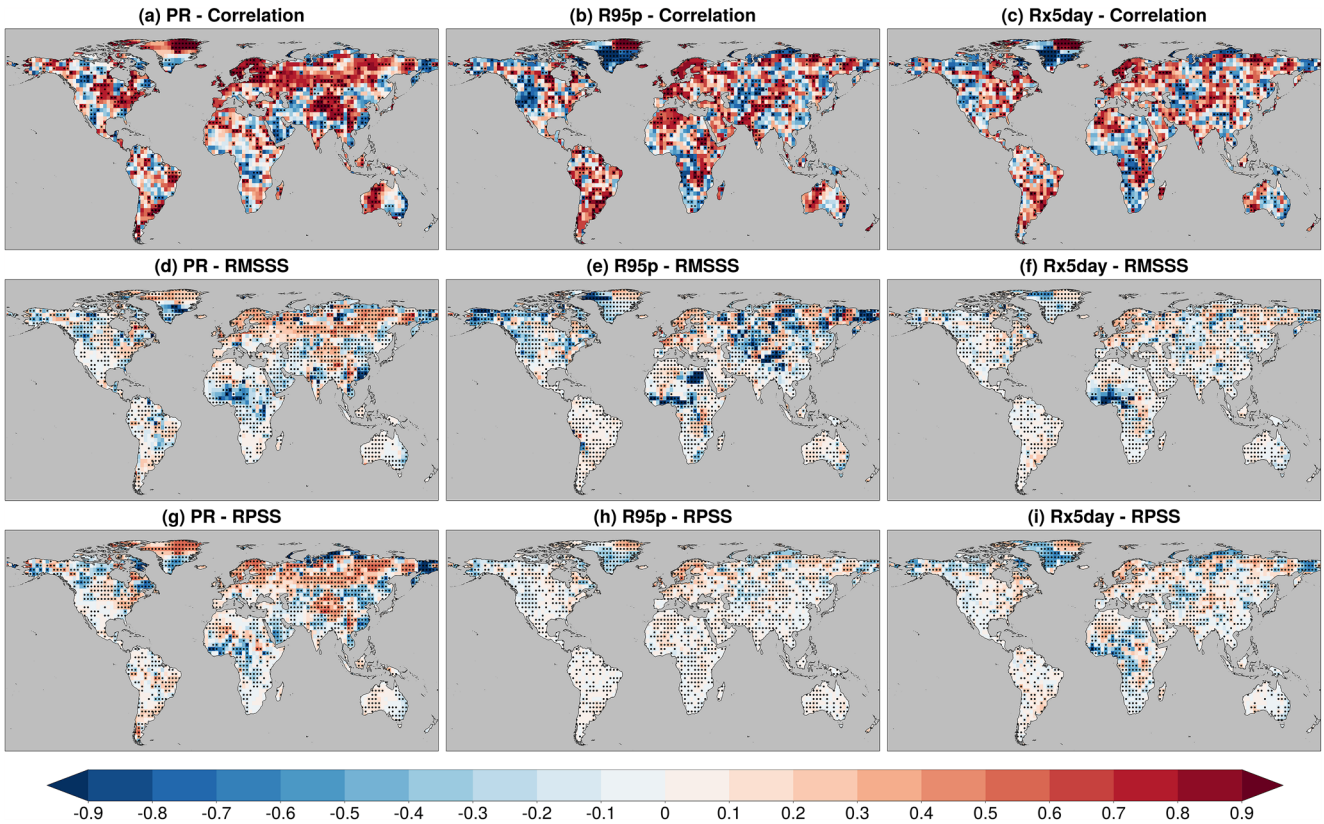

Figure 1. Forecast quality measures for different aspects of precipitation using all available ensemble members: (a, d, g) total annual precipitation, (b, e, h) R95p, (c, f, i) Rx5day. First row (a, b, c) shows the correlation, second row (d, e, f) Root-Mean-Squared Skill Score (RMSSS) and third row (g, h, i) Probability Skill Score (RPSS). RMSSS and RPSS are calculated against climatology as reference. Dots indicate grid cells where the skill measures are statistically significant ($p \leq 0.05$; see Section 2). Observational reference data sets: Global Precipitation Climatology Centre (PR) and Rainfall Estimates on a Gridded Network (R95p, Rx5day).

3. Results

The historical simulations show skill in reproducing the observed PR evolution, indicated by positive correlations, in large parts of western and Northern Europe (NEU), northern and central Asia, northeastern North America, southern and eastern South America and western part of Australia (Figure 1a), indicating that variations of precipitation are realistically simulated in these regions. Positive significant RMSSS values are found in similar regions (Figure 1d), indicating that also the magnitude of the relative precipitation measures is simulated with some skill. The RPSS also shows significant positive values in similar regions where correlations and RMSSS values are positive and significant (Figure 1g), indicating that specific parts of the distribution can be predicted based on external forcing, in particular in northern parts of Eurasia, northeastern North America, southern South America, the Tibetan Plateau and western Australia. In turn, the CMIP6 simulations show a lack of skill (apparent in most of the skill measures) for large areas in western North America, northern and central parts of South America, parts of Africa and Southeast Asia.

Also the indices of precipitation extremes show positive correlations in large parts of central and NEU, parts of Asia, northeastern North America, South America, western Australia, and northwestern parts of Africa (Figures 1b and 1c). Overall these correlation maps appear more heterogeneous for R95p and Rx5day than for PR while indicating skill in similar regions, however in particular R95p shows larger areas with positive correlation than PR, in particular in northwestern Africa (the area of Algeria and Morocco), southern central Africa and also parts of eastern Australia. RMSSS and RPSS indicate, respectively, more deterministic and probabilistic skill than the climatology for R95p and Rx5day most prominently in NEU, but also in parts of Asia, central Europe, and northwestern Africa.

The other approaches to construct an equally-weighted multi-model result in largely similar patterns of skill, showing agreement with observations in similar regions (Figures S2 and S4 in Supporting Information S1,

compare to Figure 1). Some regional differences are found in the southern USA or the Middle East, where calculating model-specific averages first results in slightly higher correlations (Figure S3 in Supporting Information S1); and eastern parts of Africa or western Australia, where pooling all members results in higher skill. A direct comparison of the skill differences between the all-member and one-member ensembles highlights in particular western Australia, central Europe, parts of Asia and North America as regions where including all ensemble members is beneficial, for example, in terms of improved correlations (Figure S5 in Supporting Information S1). In turn, the one-member ensemble shows slightly higher skill than the all-member ensemble in a few regions such as the Middle East, southern Africa and southeastern Australia. Some of these differences may be due to the different weighting of models in the all-member (models are weighted according to the number of members they provide) compared to the one-member ensembles or averaging all members of each model first (all models are equally weighted)—in particular those common differences between Figures S3 and S5 in Supporting Information S1 suggest these would be due to model weighting, for example, in the southern USA, the Middle East and western Australia. We will investigate the sensitivity of the skill measures to ensemble size more systematically further down.

There is substantial uncertainty across different precipitation data sets which can affect the skill estimates. We therefore perform the skill evaluations also using alternative precipitation data sets (Figure S6 in Supporting Information S1). This analysis confirms that the CMIP6 ensemble captures the observed PR, R95p and Rx5day in large parts of Europe (in particular NEU), Asia, eastern North America, southern and eastern South America and western Australia. Please note that the spatial coverage of HadEX3 (providing R95p and Rx5day) is limited, with large gaps in particular over Africa, the Amazon and high northern latitudes.

We next investigate the sensitivity of the skill metrics to the ensemble size at regional scale, by resampling 1,000 ensembles (based on all available simulations) for ensemble sizes between 1 and the total available number of simulations for each index (see Section 2).

In Northern Europe (NEU) as a region of generally high skill, all three skill metrics converge relatively quickly to the respective maximum value (see Figures 2a–2c; Figures S7a–S7c and S8a–S8c in Supporting Information S1). The median of the 1,000 different ensembles reaches the asymptotic value already for ensemble sizes below 20 members and does not increase further for larger ensemble sizes. The spread of skill values (e.g., inter-quartile range across all 1,000 ensembles for each size) shows only relatively small deviations from the median value. Due to the high correlation threshold to reach significance in this region, still relatively large ensembles are needed for 95% of the possible ensembles to have significant skill: for example, 69 members for PR, 136 for Rx5day and 70 for R95p.

In regions with lower (but in most cases still positive) skill score values, their dependence on ensemble size can be larger. For example, for precipitation averages over the Mediterranean region (MED; Figures 2d–2f), the median correlation value requires an ensemble size of approximately 60 members for PR and Rx5day to reach the quasi-asymptotic values (for PR reaching the threshold for significance). For PR, 95% of possible ensembles reach significant correlation for ensemble sizes of 123 and larger, whereas the significance threshold is not reached for the extremes indicators. Similar results are found for Eastern Asia (EAS), although in that case relatively small ensemble sizes of below 15 (for Rx5day) and 26 (for PR) are sufficient to reach significant correlations in 95% of the ensembles—associated with a relatively low threshold for correlations to be significant (Figures 2g–2i).

The dependence of skill on ensemble size is particularly strong in Central North America (CNA, Figures 2j–2l) and Central Australia (CAU, Figures 2m–2o)—both regions where the ensemble using all available members showed higher skill than the ensemble based on only one simulation per model (Figure S2 in Supporting Information S1). For these regions the median correlation value still increases for ensemble sizes well above 100, and in particular for PR and R95p does not clearly converge to a maximum value based on the available ensemble size. In CAU significant correlation values for 95% of the resampled ensembles are reached for ensemble sizes of above 201 members (PR) and ensembles larger than 114 members (R95p). Also the range of correlation values is substantial for these regions, and there is a non-negligible probability for negative correlation values still for 100-member ensembles, while the correlation for PR reaches 0.6 (in CNA) and 0.7 (in CAU) for ensemble sizes of more than 200 members. This large range reflects a substantial uncertainty of skill across the different ensemble members, which is likely indicative of an important contribution of internal climate variability to the local precipitation changes.

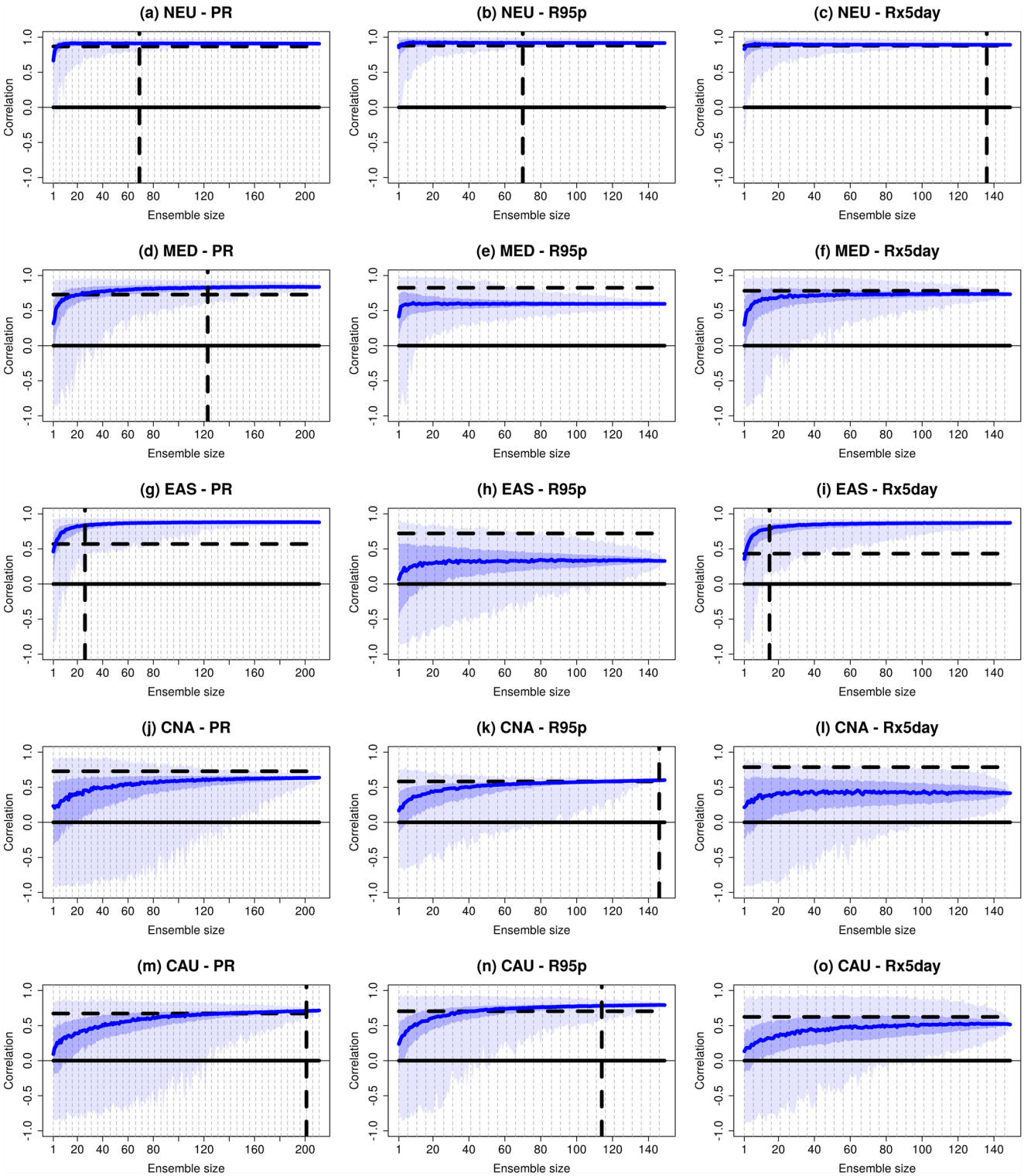

Figure 2. Correlation as function of ensemble size for precipitation averaged over selected IPCC AR6 regions. (a, b, c) Northern Europe, (d, e, f) Mediterranean region, (g, h, i) Eastern Asia, (j, k, l) Central North America, (m, n, o) Central Australia. Correlation is shown for (left column) PR, (middle) R95p and (right) Rx5day. For each ensemble size the range of correlation values is estimated from randomly resampling 1,000 sub-ensembles (see Section 2). The blue line represents the median across the 1,000 realizations, dark blue shading indicates the interquartile range (25th to 75th percentile) and light blue the full range across all iterations. The horizontal dashed black line indicates the value above which correlations are statistically significant ($p \leq 0.05$). The vertical dashed black line indicates the ensemble size for which at least 95% of correlation coefficients are statistically significant.

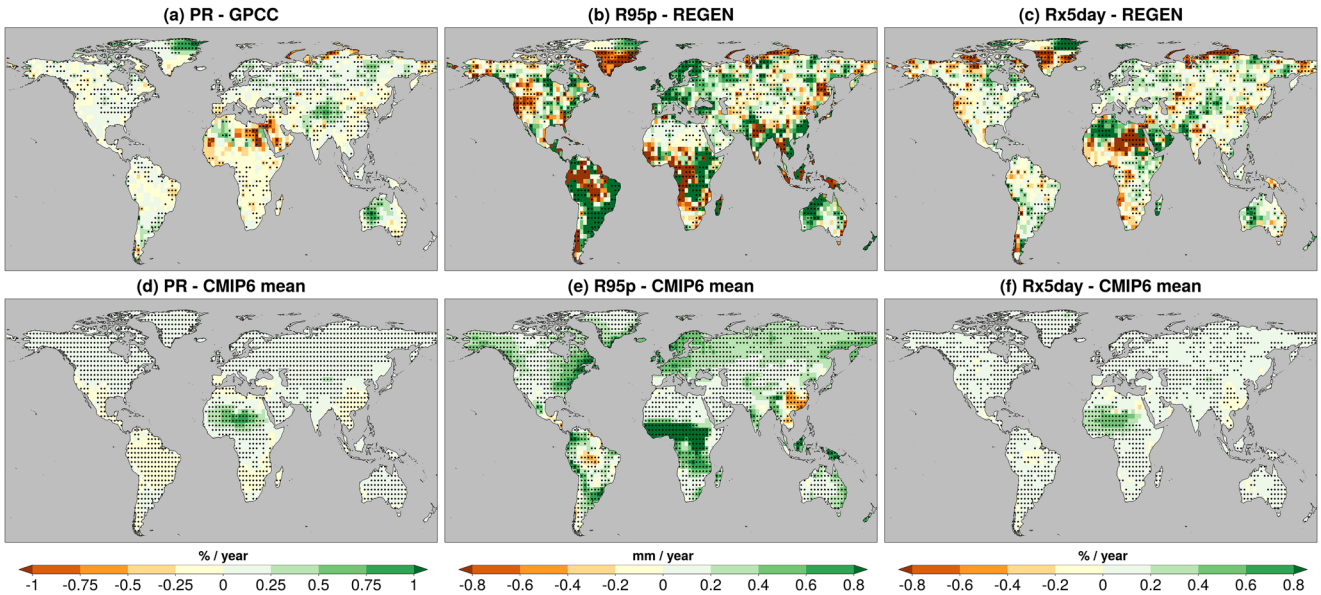

Figure 3. Trend maps of 20-year averages of (a, d) PR, (b, e) R95p, (c, f) Rx5day. (Top row) observations (Global Precipitation Climatology Centre for PR, Rainfall Estimates on a Gridded Network for R95p and Rx5day), (bottom row) Coupled Model Intercomparison Project Phase 6 ensemble mean based on all available ensemble members. Dots indicate grid cells where trends are statistically significant ($p \leq 0.05$).

In NEU, where high correlation values are reached already for small ensemble sizes, the observed increase of precipitation is relatively monotonic and exhibits little variability on decadal to multi-decadal scales (Figure S9 in Supporting Information S1). In contrast, in CNA and CAU, where correlations continue to increase for large ensemble sizes, the observed time series exhibit more pronounced multi-decadal variations. These variations are not captured by the CMIP6 ensemble mean, suggesting they are not a response to forcing and may be related to internal climate variability. This illustrates how large ensembles can be particularly beneficial in cases where a predictable signal (in this case the response to external forcing) occurs in combination with non-predictable "noise" (in this case the multi-decadal variations cannot be predicted by the forcing).

RMSSS and RPSS generally converge faster to their respective maximum value compared to correlations. For all examples shown the asymptotic values are reached for ensemble sizes below 40 (Figures S7 and S8 in Supporting Information S1). This indicates that the benefit of larger ensembles is seen in particular when correlation is the skill measure of interest.

To understand the characteristics of precipitation changes that are or are not coherent between simulations and observations, we compare the spatial patterns of long-term trends between the CMIP6 ensemble mean and observations (Figure 3) and the patterns of the skill metrics (Figure 1). This comparison shows that significant skill (in particular correlations and RPSS) is typically found in regions where trends toward more or stronger precipitation are observed (e.g., NEU, midlatitude parts of Asia and the Himalaya region, eastern North America, southern South America, western Australia). The CMIP6 ensemble simulates upward trends in larger regions globally compared to the observations, and this includes most regions with observed increases. Only in a few cases positive skill values are associated with decreasing precipitation, for example, PR in the Mediterranean region (in particular the Iberian Peninsula) and eastern South America or R95p in the central Amazon region of South America. A lack of skill is generally found in regions where observed precipitation is characterized by negative trends but the CMIP6 ensemble simulates positive trends (e.g., western North America, western sub-Saharan and western tropical Africa, southeastern Australia).

4. Summary, Discussion and Conclusions

This study evaluates to what extent the simulations with historical forcings (and scenario SSP2-4.5 for the last part of the evaluation period) agree with observed changes in precipitation since 1950. We find that measures of both mean and extreme precipitation are realistically simulated in large parts of Europe and Asia, as well as some areas in (eastern) North America, (southern and eastern) South America and (western) Australia. The level

of agreement is assessed with skill measures typically used to verify forecast accuracy. The skill can show some dependence on the ensemble size. In particular for some regions with moderate agreement with observations the availability of very large ensembles (>100 members) can be beneficial to achieve statistically significant skill metrics, which are indicative of a good agreement between the forced signal in CMIP6 and observed precipitation changes. Whereas in other regions (e.g., those with highest skill) quasi-asymptotic skill values are already reached for relatively smaller ensembles.

Comparing the spatial patterns of skill and long-term trends highlights that high skill is typically found in regions where both observations and simulations show trends towards wetter (i.e., more intense and more accumulated rain), and in a few regions where both observations and simulations show trends towards drier conditions (i.e., less intense and smaller precipitation amounts, such as the Mediterranean or eastern South America). We find lack of skill primarily in regions where observations show trends towards drier conditions but CMIP6 simulates precipitation increases. The exact regions of trend discrepancies differ between the different precipitation indices, but most of these regions are located in low-to-mid latitudes of the tropics and sub-tropics (Figure 3), similar also to earlier trend comparisons based on CMIP5 for annual mean precipitation (Knutson & Zeng, 2018). These regions include (based on the different precipitation indices) parts of tropical and sub-tropical Africa, tropical South America, and the western United States. Most of these regions are affected by seasonal dry or wet regimes, and one possible avenue to better understand these discrepancies could focus in the representation of these seasonal regimes. In general, as the ensemble of CMIP6 historical simulations represents the precipitation response to (historical) forcing, this regional lack of skill might be due to dominating effects of internal climate variability on the observed precipitation time series, or model errors in simulating precipitation responses to the forcing, or erroneous observations. Also the quality of the observational data needs to be taken into account. In particular for the extreme precipitation indices, while we find relatively similar results in the regions where both data sets have coverage, some of the regions with low skill (e.g., parts of Africa) are only covered by one data set—and even here are not part of the regions with high-quality data (Contractor et al., 2020). Overall the CMIP6 ensemble mean shows more widespread increases in precipitation compared to observations, and does not capture several of the regional decreases found in the observations.

The evaluation presented in this study corresponds to the "forcing skill" discussed by Risbey et al. (2022), and does not take other drivers (e.g., related to internal variability) into account. Based on the evaluations presented, the CMIP6 ensemble is relatively trustworthy in simulating (mean and extreme) precipitation changes in large parts of Europe and Asia, northeastern North America, parts of South America and western Australia. However, in the regions where a lack of agreement has been highlighted (e.g., western North America and substantial parts of Africa), the CMIP6 ensemble may not be a credible prediction system (due to missing either regional variability effects or forcing response)—and also future projections may need to be considered with particular caution.

We suggest that evaluations similar to the ones presented here can also be useful to test potential improvements related to model developments, for example, to test whether models at higher resolution simulate more realistic precipitation changes (Bador et al., 2020; Moreno-Chamarro et al., 2021). Similar assessments can also be useful to test the efficacy of post-processing methods aiming to obtain more realistic information about climate changes in the coming decades, for example, by constraining projections based on specific characteristics (e.g., Brunner et al., 2020; Hegerl et al., 2021; Mahmood et al., 2021, 2022), or by applying calibration techniques aiming to correct model biases and improve the reliability of the simulations (Bellprat et al., 2019; Doblas-Reyes et al., 2005).

In conclusion, current climate models provide credible simulations of observed changes in both mean and extreme precipitation for some land regions in the northern hemisphere extra-tropics, including large parts of Europe and Asia and some parts of other continents. However, substantial regions of the globe also show a lack of skill, in particular regions where precipitation measures have been observed to decrease but CMIP6 simulates increases. This lack of skill in some regions highlights the need to better understand the relative roles of forcing and internal variability in the observed precipitation changes. Improving the models to more realistically capture observed precipitation changes over the past decades will be essential to increase their credibility to inform about precipitation changes in the decades to come.

Data Availability Statement

CMIP6 data is available through the Earth System Grid Federation system (ESGF; https://esgf-node.llnl.gov/search/cmip6/), the models and ensemble members used in this study are listed in Table S1 in Supporting Information S1. GPCC at https://www.dwd.de/EN/ourservices/gpcc/gpcc.html, CRU precipitation:

https://crudata.uea.ac.uk/cru/data/hrg/, REGEN at https://doi.org/10.25914/5ca4c380b0d44, HadEX3 at https://www.metoffice.gov.uk/hadobs/hadex3/. Extreme precipitation indices have been calculated with the climpact package: https://github.com/ARCCSS-extremes/climpact/.

Acknowledgments

We are grateful for support by the Departament de Recerca i Universitats de la Generalitat de Catalunya for the Climate Variability and Change (CVC) Research Group (Reference: 2021 SGR 00786), and research funding by the Horizon 2020 LANDMARC project (grant agreement no. 869367), the Horizon Europe ASPECT project (Grant 101081460), and the AXA Research Fund. CDT acknowledges financial support from the Spanish Ministry for Science and Innovation (FPI PRE2019–509 08864 financed by MCIN/AEI/ http://doi.org/10.13039/501100011033). PDL received funding from the Horizon Europe Research and Innovation Programme, Grant 101059659. We thank the climate modeling groups contributing to CMIP6 for producing and making available their model output. We are grateful to Margarida Samsó and Pierre–Antoine Bretonnière for downloading, formatting and managing the large data sets of climate simulations and observations used in this study.

References

Alexander, L. V., Bador, M., Roca, R., Contractor, S., Donat, M. G., & Nguyen, P. L. (2020). Intercomparison of annual precipitation indices and extremes over global land areas from in situ, space-based and reanalysis products. *Environmental Research Letters*, *15*(5), 055002. https://doi.org/10.1088/1748-9326/ab79e2

Allen, M. R., & Ingram, W. J. (2002). Constraints on future changes in climate and the hydrologic cycle. *Nature*, *419*(6903), 224–232. https://doi.org/10.1038/nature01092

Avila, F. B., Dong, S., Menang, K. P., Rajczak, J., Renom, M., Donat, M. G., & Alexander, L. V. (2015). Systematic investigation of gridding-related scaling effects on annual statistics of daily temperature and precipitation maxima: A case study for south-east Australia. *Weather and Climate Extremes*, *9*, 6–16. https://doi.org/10.1016/j.wace.2015.06.003

Bador, M., Boé, J., Terray, L., Alexander, L. V., Baker, A., Bellucci, A., et al. (2020). Impact of higher spatial atmospheric resolution on precipitation extremes over land in global climate models. *Journal of Geophysical Research: Atmospheres*, *125*(13), e2019JD032184. https://doi.org/10.1029/2019JD032184

Becker, A., Finger, P., Meyer-Christoffer, A., Rudolf, B., Schamm, K., Schneider, U., & Ziese, M. (2013). A description of the global land-surface precipitation data products of the Global Precipitation Climatology Centre with sample applications including centennial (trend) analysis from 1901-present. *Earth System Science Data*, *5*(1), 71–99. https://doi.org/10.5194/essd-5-71-2013

Bellprat, O., Guemas, V., Doblas-Reyes, F., & Donat, M. G. (2019). Towards reliable extreme weather and climate event attribution. *Nature Communications*, *10*(1), 1732. https://doi.org/10.1038/s41467-019-09729-2

Brunner, L., McSweeney, C., Ballinger, A. P., Befort, D. J., Benassi, M., Booth, B., et al. (2020). Comparing methods to constrain future European climate projections using a consistent framework. *Journal of Climate*, *33*(20), 8671–8692. https://doi.org/10.1175/JCLI-D-19-0953.1

Contractor, S., Donat, M. G., & Alexander, L. V. (2021). Changes in observed daily precipitation over global land areas since 1950. *Journal of Climate*, *34*(1), 3–19. https://doi.org/10.1175/JCLI-D-19-0965.1

Contractor, S., Donat, M. G., V Alexander, L., Ziese, M., Meyer-Christoffer, A., Schneider, U., et al. (2020). Rainfall estimates on a gridded network (REGEN)—A global land-based gridded dataset of daily precipitation from 1950 to 2016. *Hydrology and Earth System Sciences*, *24*(2), 919–943. https://doi.org/10.5194/hess-24-919-2020

Delgado-Torres, C., Donat, M. G., Gonzalez-Reviriego, N., Caron, L. P., Athanasiadis, P. J., Bretonnière, P. A., et al. (2022). Multi-model forecast quality assessment of CMIP6 decadal predictions. *Journal of Climate*, *35*(13), 4363–4382. https://doi.org/10.1175/JCLI-D-21-0811.1

Doblas-Reyes, F. J., Hagedorn, R., & Palmer, T. N. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting—II. Calibration and combination. *Tellus A: Dynamic Meteorology and Oceanography*, *57*(3), 234. https://doi.org/10.3402/tellusa.v57i3.14658

Donat, M. G., Alexander, L. V., Yang, H., Durre, I., Vose, R., Dunn, R. J. H., et al. (2013). Updated analyses of temperature and precipitation extreme indices since the beginning of the twentieth century: The HadEX2 dataset. *Journal of Geophysical Research: Atmospheres*, *118*(5), 2098–2118. https://doi.org/10.1002/jgrd.50150

Donat, M. G., Angélil, O., & Ukkola, A. M. (2019). Intensification of precipitation extremes in the world's humid and water-limited regions. *Environmental Research Letters*, *14*(6), 065003. https://doi.org/10.1088/1748-9326/ab1c8e

Donat, M. G., Lowry, A. L., Alexander, L. V., O'Gorman, P. A., & Maher, N. (2016). More extreme precipitation in the world's dry and wet regions. *Nature Climate Change*, *6*(5), 508–513. https://doi.org/10.1038/nclimate2941

Dunn, R. J. H., Alexander, L. V., Donat, M. G., Zhang, X., Bador, M., Herold, N., et al. (2020). Development of an updated global land in situ-based data set of temperature and precipitation extremes: HadEX3. *Journal of Geophysical Research: Atmospheres*, *125*(16), e2019JD032263. https://doi.org/10.1029/2019JD032263

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the coupled model Intercomparison project phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Goddard, L., Kumar, A., Solomon, A., Smith, D., Boer, G., Gonzalez, P., et al. (2012). A verification framework for interannual-to-decadal predictions experiments. *Climate Dynamics*, *40*(1–2), 245–272. https://doi.org/10.1007/s00382-012-1481-2

Hamed, K. H., & Ramachandra Rao, A. (1998). A modified Mann-Kendall trend test for autocorrelated data. *Journal of Hydrology*, *204*(1–4), 182–196. https://doi.org/10.1016/S0022-1694(97)00125-X

Hegerl, G. C., Ballinger, A. P., Booth, B. B. B., Borchert, L. F., Brunner, L., Donat, M. G., et al. (2021). Toward consistent observational constraints in climate predictions and projections. *Frontiers in Climate*, *3*, 43. https://doi.org/10.3389/fclim.2021.678109

Herold, N., Alexander, L. V., Donat, M. G., Contractor, S., & Becker, A. (2016). How much does it rain over land? *Geophysical Research Letters*, *43*(1), 341–348. https://doi.org/10.1002/2015GL066615

Iturbide, M., Gutiérrez, J. M., Alves, L. M., Bedia, J., Cerezo-Mota, R., Cimadevilla, E., et al. (2020). An update of IPCC climate reference regions for subcontinental analysis of climate model data: Definition and aggregated datasets. *Earth System Science Data*, *12*(4), 2959–2970. https://doi.org/10.5194/essd-12-2959-2020

Knutson, T. R., & Zeng, F. (2018). Model assessment of observed precipitation trends over land regions: Detectable human influences and possible low bias in model trends. *Journal of Climate*, *31*(12), 4617–4637. https://doi.org/10.1175/JCLI-D-17-0672.1

Kumar, S., Merwade, V., Kinter, J. L., & Niyogi, D. (2013). Evaluation of temperature and precipitation trends and long-term persistence in CMIP5 twentieth-century climate simulations. *Journal of Climate*, *26*(12), 4168–4185. https://doi.org/10.1175/JCLI-D-12-00259.1

Lee, J.-Y., Marotzke, J., Bala, G., Cao, L., Corti, S., Dunne, J. P., et al. (2021). Chapter 4: Future global climate: Scenario-based projections and near-term information (pp. 553–672). https://doi.org/10.1017/9781009157896.006

Lorenz, R., Pitman, A. J., Donat, M. G., Hirsch, A. L., Kala, J., Kowalczyk, E. A., et al. (2014). Representation of climate extreme indices in the ACCESS1.3b coupled atmosphere-land surface model. *Geoscientific Model Development*, *7*(2), 545–567. https://doi.org/10.5194/gmd-7-545-2014

Mahmood, R., Donat, M. G., Ortega, P., Doblas-Reyes, F. J., Delgado-Torres, C., Samsó, M., & Bretonnière, P.-A. (2022). Constraining low-frequency variability in climate projections to predict climate on decadal to multi-decadal timescales—A poor man's initialized prediction system. *Earth System Dynamics*, *13*(4), 1437–1450. https://doi.org/10.5194/esd-13-1437-2022

Mahmood, R., Donat, M. G., Ortega, P., Doblas-Reyes, F. J., & Ruprich-Robert, Y. (2021). Constraining decadal variability yields skillful projections of near-term climate change. *Geophysical Research Letters*, *48*(24), e2021GL094915. https://doi.org/10.1029/2021GL094915

Meehl, G. A., Richter, J. H., Teng, H., Capotondi, A., Cobb, K., Doblas-Reyes, F., et al. (2021). Initialized Earth System prediction from subseasonal to decadal timescales. *Nature Reviews Earth & Environment*, *17*(5), 1–18. https://doi.org/10.1038/s43017-021-00155-x

Min, S.-K., Zhang, X., Zwiers, F. W., & Hegerl, G. C. (2011). Human contribution to more-intense precipitation extremes. *Nature*, *470*(7334), 378–381. https://doi.org/10.1038/nature09763

Mitchell, T. D., & Jones, P. D. (2005). An improved method of constructing a database of monthly climate observations and associated high-resolution grids. *International Journal of Climatology*, *25*(6), 693–712. https://doi.org/10.1002/joc.1181

Moreno-Chamarro, E., Caron, L.-P., Ortega, P., Loosveldt Tomas, S., & Roberts, M. J. (2021). Can we trust CMIP5/6 future projections of European winter precipitation? *Environmental Research Letters*, *16*(5), 054063. https://doi.org/10.1088/1748-9326/abf28a

Morrison, H., van Lier-Walqui, M., Fridlind, A. M., Grabowski, W. W., Harrington, J. Y., Hoose, C., et al. (2020). Confronting the challenge of modeling cloud and precipitation microphysics. *Journal of Advances in Modeling Earth Systems*, *12*(8), e2019MS001689. https://doi.org/10.1029/2019MS001689

Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*, *116*(12), 2417–2424. https://doi.org/10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2

O'Gorman, P. A. (2012). Sensitivity of tropical precipitation extremes to climate change. *Nature Geoscience*, *5*(10), 697–700. https://doi.org/10.1038/ngeo1568

O'Neill, B. C., Tebaldi, C., van Vuuren, D. P., Eyring, V., Friedlingstein, P., Hurtt, G., et al. (2016). The scenario model Intercomparison project (ScenarioMIP) for CMIP6. *Geoscientific Model Development*, *9*(9), 3461–3482. https://doi.org/10.5194/gmd-9-3461-2016

Paik, S., Min, S., Zhang, X., Donat, M. G., King, A. D., & Sun, Q. (2020). Determining the anthropogenic greenhouse gas contribution to the observed intensification of extreme precipitation. *Geophysical Research Letters*, *47*(12), e2019GL086875. https://doi.org/10.1029/2019GL086875

Risbey, J. S., Squire, D. T., Baldissera Pacchetti, M., Black, A. S., Chapman, C. C., Dessai, S., et al. (2022). Common issues in verification of climate forecasts and projections. *Climate*, *10*(6), 83. https://doi.org/10.3390/cli10060083

Sillmann, J., Kharin, V. V., Zhang, X., Zwiers, F. W., & Bronaugh, D. (2013). Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate. *Journal of Geophysical Research: Atmospheres*, *118*(4), 1716–1733. https://doi.org/10.1002/jgrd.50203

Smith, D. M., Eade, R., Scaife, A. A., Caron, L. P., Danabasoglu, G., DelSole, T. M., et al. (2019). Robust skill of decadal climate predictions. *Npj Climate and Atmospheric Science*, *2*(1), 13. https://doi.org/10.1038/s41612-019-0071-y

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*(1), 72. https://doi.org/10.2307/1412159

Trenberth, K. E. (2011). Changes in precipitation with climate change. *Climate Research*, *47*(1), 123–138. https://doi.org/10.3354/cr00953

Vicente-Serrano, S. M., García-Herrera, R., Peña-Angulo, D., Tomas-Burguera, M., Domínguez-Castro, F., Noguera, I., et al. (2022). Do CMIP models capture long-term observed annual precipitation trends? *Climate Dynamics*, *58*(9–10), 2825–2842. https://doi.org/10.1007/s00382-021-06034-x

Westra, S., Alexander, L. V., & Zwiers, F. W. (2013). Global increasing trends in annual maximum daily precipitation. *Journal of Climate*, *26*(11), 3904–3918. https://doi.org/10.1175/JCLI-D-12-00502.1

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences*. Elsevier/Academic Press.

Zhang, X., Alexander, L., Hegerl, G. C., Jones, P., Tank, A. K., Peterson, T. C., et al. (2011). Indices for monitoring changes in extremes based on daily temperature and precipitation data. *Wiley Interdisciplinary Reviews: Climate Change*, *2*(6), 851–870. https://doi.org/10.1002/wcc.147

Zhang, X., Zwiers, F. W., Hegerl, G. C., Lambert, F. H., Gillett, N. P., Solomon, S., et al. (2007). Detection of human influence on twentieth-century precipitation trends. *Nature*, *448*(7152), 461-5–465. https://doi.org/10.1038/nature06025

References From the Supporting Information

Andrews, M. B., Ridley, J. K., Wood, R. A., Andrews, T., Blockley, E. W., Booth, B., et al. (2020). Historical simulations with HadGEM3-GC3.1 for CMIP6. *Journal of Advances in Modeling Earth Systems*, *12*(6), e2019MS001995. https://doi.org/10.1029/2019MS001995

Bao, Y., Song, Z., & Qiao, F. (2020). FIO-ESM version 2.0: Model description and evaluation. *Journal of Geophysical Research: Oceans*, *125*(6), e2019JC016036. https://doi.org/10.1029/2019JC016036

Bi, D., Dix, M., Marsland, S., O'Farrell, S., Sullivan, A., Bodman, R., et al. (2020). Configuration and spin-up of ACCESS-CM2, the new generation Australian community climate and Earth system simulator coupled model. *Journal of Southern Hemisphere Earth Systems Science*, *70*(1), 225–251. https://doi.org/10.1071/ES19040

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., et al. (2020). Presentation and evaluation of the IPSL-CM6A-LR climate model. *Journal of Advances in Modeling Earth Systems*, *12*(7), e2019MS002010. https://doi.org/10.1029/2019MS002010

Cao, J., Wang, B., Yang, Y.-M., Ma, L., Li, J., Sun, B., et al. (2018). The NUIST Earth System Model (NESM) version 3: Description and preliminary evaluation. *Geoscientific Model Development*, *11*(7), 2975–2993. https://doi.org/10.5194/gmd-11-2975-2018

Cherchi, A., Fogli, P. G., Lovato, T., Peano, D., Iovino, D., Gualdi, S., et al. (2019). Global mean climate and main patterns of variability in the CMCC-CM2 coupled model. *Journal of Advances in Modeling Earth Systems*, 2018MS001369. https://doi.org/10.1029/2018MS001369

Christian, J. R., Denman, K. L., Hayashida, H., Holdsworth, A. M., Lee, W. G., Riche, O. G. J., et al. (2022). Ocean biogeochemistry in the Canadian Earth system model version 5.0.3: CanESM5 and CanESM5-CanOE. *Geoscientific Model Development*, *15*(11), 4393–4424. https://doi.org/10.5194/gmd-15-4393-2022

Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D. A., DuVivier, A. K., Edwards, J., et al. (2020). The community Earth system model version 2 (CESM2). *Journal of Advances in Modeling Earth Systems*, *12*, e2019MS001916. https://doi.org/10.1029/2019MS001916

DelSole, T., & Tippett, M. K. (2016). Forecast comparison based on random walks. *Monthly Weather Review*, *144*(2), 615–626. https://doi.org/10.1175/MWR-D-15-0218.1

Döscher, R., Acosta, M., Alessandri, A., Anthoni, P., Arsouze, T., Bergman, T., et al. (2022). The EC-Earth3 Earth system model for the coupled model Intercomparison project 6. *Geoscientific Model Development*, *15*(7), 2973–3020. https://doi.org/10.5194/gmd-15-2973-2022

Gettelman, A., Mills, M. J., Kinnison, D. E., Garcia, R. R., Smith, A. K., Marsh, D. R., et al. (2019). The whole atmosphere community climate model version 6 (WACCM6). *Journal of Geophysical Research: Atmospheres*, *124*(23), 12380–12403. https://doi.org/10.1029/2019JD030943

Guangqing, Z., Yunquan, Z., Jinrong, J., He, Z., Baodong, W., Hang, C., et al. (2020). Earth system model: CAS-ESM. *Frontiers of Data and Computing*, *2*(1), 38–54. https://doi.org/10.11871/jfdc.issn.2096-742X.2020.01.004

Hajima, T., Watanabe, M., Yamamoto, A., Tatebe, H., Noguchi, M. A., Abe, M., et al. (2020). Development of the MIROC-ES2L Earth system model and the evaluation of biogeochemical processes and feedbacks. *Geoscientific Model Development*, *13*(5), 2197–2244. https://doi.org/10.5194/gmd-13-2197-2020

He, B., Yu, Y., Bao, Q., Lin, P., Liu, H., Li, J., et al. (2020). CAS FGOALS-f3-L model dataset descriptions for CMIP6 DECK experiments. *Atmospheric and Oceanic Science Letters*, *13*(6), 582–588. https://doi.org/10.1080/16742834.2020.1778419

Kelley, M., Schmidt, G. A., Nazarenko, L. S., Bauer, S. E., Ruedy, R., Russell, G. L., et al. (2020). GISS-E2.1: Configurations and climatology. *Journal of Advances in Modeling Earth Systems*, *12*(8), e2019MS002025. https://doi.org/10.1029/2019MS002025

Krishnan, R., Swapna, P., Vellore, R., Narayanasetti, S., Prajeesh, A. G., Choudhury, A. D., et al. (2019). The IITM Earth system model (ESM): Development and future roadmap. In D. A. Randall, J. Srinivasan, R. S. Nanjundiah, & P. Mukhopadhyay (Eds.), *Current trends in the representation of physical processes in weather and climate models* (pp. 183–195). Springer. https://doi.org/10.1007/978-981-13-3396-5_9

Lovato, T., Peano, D., Butenschön, M., Materia, S., Iovino, D., Scoccimarro, E., et al. (2022). CMIP6 simulations with the CMCC Earth system model (CMCC-ESM2). *Journal of Advances in Modeling Earth Systems*, *14*(3), e2021MS002814. https://doi.org/10.1029/2021MS002814

Mauritsen, T., Bader, J., Becker, T., Behrens, J., Bittner, M., Brokopf, R., et al. (2019). Developments in the MPI-M Earth system model version 1.2 (MPI-ESM1.2) and its response to increasing CO_2. *Journal of Advances in Modeling Earth Systems*, *11*(4), 998–1038. https://doi.org/10.1029/2018MS001400

Pak, G., Noh, Y., Lee, M.-I., Yeh, S.-W., Kim, D., Kim, S.-Y., et al. (2021). Korea Institute of ocean science and technology Earth system model and its simulation characteristics. *Ocean Science Journal*, *56*(1), 18–45. https://doi.org/10.1007/s12601-021-00001-7

Pu, Y., Liu, H., Yan, R., Yang, H., Xia, K., Li, Y., et al. (2020). CAS FGOALS-g3 model datasets for the CMIP6 scenario model Intercomparison project (ScenarioMIP). *Advances in Atmospheric Sciences*, *37*(10), 1081–1092. https://doi.org/10.1007/s00376-020-2032-0

Rong, X. Y., Li, J., Chen, H. M., Xin, Y. F., Su, J. Z., Hua, L. J., & Zhang, Z. Q. (2019). Introduction of CAMS-CSM model and its participation in CMIP6. *Climate Change Research*, *15*(5), 540–544. https://doi.org/10.12006/j.issn.1673-1719.2019.186

Séférian, R., Nabat, P., Michou, M., Saint-Martin, D., Voldoire, A., Colin, J., et al. (2019). Evaluation of CNRM Earth system model, CNRM-ESM2-1: Role of Earth system processes in present-day and future climate. *Journal of Advances in Modeling Earth Systems*, *11*(12), 4182–4227. https://doi.org/10.1029/2019MS001791

Sellar, A. A., Jones, C. G., Mulcahy, J. P., Tang, Y., Yool, A., Wiltshire, A., et al. (2019). UKESM1: Description and evaluation of the U.K. Earth system model. *Journal of Advances in Modeling Earth Systems*, *11*(12), 4513–4558. https://doi.org/10.1029/2019MS001739

Seland, Ø., Bentsen, M., Olivié, D., Toniazzo, T., Gjermundsen, A., Graff, L. S., et al. (2020). Overview of the Norwegian Earth System Model (NorESM2) and key climate response of CMIP6 DECK, historical, and scenario simulations. *Geoscientific Model Development*, *13*(12), 6165–6200. https://doi.org/10.5194/gmd-13-6165-2020

Siegert, S., Bellprat, O., Ménégoz, M., Stephenson, D. B., & Doblas-Reyes, F. J. (2017). Detecting improvements in forecast correlation skill: Statistical testing and power analysis. *Monthly Weather Review*, *145*(2), 437–450. https://doi.org/10.1175/MWR-D-16-0037.1

Swart, N. C., Cole, J. N. S., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., et al. (2019). The Canadian Earth system model version 5 (CanESM5.0.3). *Geoscientific Model Development*, *12*(11), 4823–4873. https://doi.org/10.5194/gmd-12-4823-2019

Tatebe, H., Ogura, T., Nitta, T., Komuro, Y., Ogochi, K., Takemura, T., et al. (2019). Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in MIROC6. *Geoscientific Model Development*, *12*(7), 2727–2765. https://doi.org/10.5194/gmd-12-2727-2019

Volodin, E. M., Mortikov, E. V., Kostrykin, S. V., Galin, V. Y., Lykossov, V. N., Gritsun, A. S., et al. (2017). Simulation of the present-day climate with the climate model INMCM5. *Climate Dynamics*, *49*(11–12), 3715–3734. https://doi.org/10.1007/s00382-017-3539-7

Voldoire, A., Saint-Martin, D., Sénési, S., Decharme, B., Alias, A., Chevallier, M., et al. (2019). Evaluation of CMIP6 DECK experiments with CNRM-CM6-1. *Journal of Advances in Modeling Earth Systems*, *11*(7), 2177–2213. https://doi.org/10.1029/2019MS001683

Wu, T., Lu, Y., Fang, Y., Xin, X., Li, L., Li, W., et al. (2019). The Beijing climate center climate system model (BCC-CSM): The main progress from CMIP5 to CMIP6. *Geoscientific Model Development*, *12*(4), 1573–1600. https://doi.org/10.5194/gmd-12-1573-2019

Yukimoto, S., Kawai, H., Koshiro, T., Oshima, N., Yoshida, K., Urakawa, S., et al. (2019). The meteorological Research institute Earth system model version 2.0, MRI-ESM2.0: Description and basic evaluation of the physical component. *Journal of the Meteorological Society of Japanese Series II*, *97*(5), 931–965. https://doi.org/10.2151/jmsj.2019-051

Ziehn, T., Chamberlain, M. A., Law, R. M., Lenton, A., Bodman, R. W., Dix, M., et al. (2020). The Australian Earth system model: ACCESS-ESM1.5. *Journal of Southern Hemisphere Earth Systems Science*, *70*(1), 193–214. https://doi.org/10.1071/ES19035

Zwiers, F. W., & von Storch, H. (1995). Taking serial correlation into account in tests of the mean. *Journal of Climate*, *8*(2), 336–351. https://doi.org/10.1175/1520-0442(1995)008<0336:TSCIAI>2.0.CO;2