AGU ADVANCING EARTH AND SPACE SCIENCE

Geophysical Research Letters

RESEARCH LETTER
10.1029/2020GL092026

Using Large Ensembles to Identify Regions of Systematic Biases in Moderate-to-Heavy Daily Precipitation

N. Goldenson[1], C. W. Thackeray[1], A. D. Hall[1,2], D. L. Swain[2,3,4], and N. Berg[1]

[1]Atmospheric and Oceanic Sciences, University of California, Los Angeles, CA, USA, [2]Institute of the Environment and Sustainability, University of California, Los Angeles, CA, USA, [3]Capacity Center for Climate and Weather Extremes, National Center for Atmospheric Research, Boulder, CO, USA, [4]The Nature Conservancy of California, San Francisco, CA, USA

Key Points:

- The short time series of the satellite era often inadequately sample precipitation distributions, posing challenges for model evaluation
- Large ensembles can be used to identify coherent areas where no ensemble member simulates daily precipitation consistent with a reference
- Most models simulate the extratropical winter hemisphere moderate-to-heavy precipitation less inconsistently than that of the tropics

Supporting Information:

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

Correspondence to:

N. Goldenson,
ngoldenson@ucla.edu

Citation:

Goldenson, N., Thackeray, C. W., Hall, A. D., Swain, D. L., & Berg, N. (2021). Using large ensembles to identify regions of systematic biases in moderate to heavy daily precipitation. *Geophysical Research Letters*, *48*, e2020GL092026. https://doi.org/10.1029/2020GL092026

Received 7 DEC 2020
Accepted 10 APR 2021

Abstract Because of internal variability in both the real-world and global climate models, it is unclear whether disagreement between models and observations reflects true systematic differences, or different phasing of internal variability in the short observational period. Here, we address this issue through an examination of moderate-to-heavy precipitation in large ensembles of global climate models. We find that model inconsistency with a global observational product is lowest for extratropical precipitation in northern hemisphere winter. The inconsistency is systematically greater for the southern hemisphere winter, but the difference between hemispheres could be due to observational quality. Moderate-to-heavy extratropical winter precipitation is less inconsistent than moderate-to-heavy tropical precipitation in most models. Within the tropics, moderate-to-heavy precipitation is particularly inconsistent with the reference in regions including the Caribbean (especially during JJA), the northern and southern flanks of the Pacific and Atlantic ITCZ, and the Indian Ocean.

Plain Language Summary Among the challenges of evaluating precipitation in climate simulations are relatively short periods of observations in the satellite era. We introduce a method using large ensembles of global climate model (GCM) simulations that are identical but for perturbed initial conditions, which can help to circumvent this challenge. By examining large ensembles from five different GCMs, we identify common areas in which there are deficiencies either in observational assimilation and/or GCM process representation. GCMs generally agree better with observations of moderate-to-heavy precipitation in the northern hemisphere winter, where they can resolve the processes leading to precipitation, than in the tropics where convection dominates or the southern hemisphere where the quality of the observations is lower.

1. Introduction and Background

Precipitation is a complex field to simulate in global climate models (GCMs), as even the most sophisticated models must rely on multiple sub-grid scale parameterizations to represent it. The choice of convective parameterization is a key source of model disagreement, relating to microphysics (e.g., Zhao et al., 2016), closure schemes (e.g., Qiao & Liang, 2016), or nonlinear interaction among parameterization choices (e.g., Liu et al., 2018). These structural differences lead to large differences among GCM projections of future local precipitation change (Fläschner et al., 2016; Kharin et al. 2007), despite the changes being controlled at the global scale by atmospheric energy balance (DeAngelis et al., 2015; Pendergrass & Hartmann, 2014). Sometimes the choice of parameter value within a scheme (parametric uncertainty) can be as large an uncertainty source as the choice of method (e.g., Morales et al., 2019).

Despite these issues, there are important aspects of precipitation projections about which models agree. GCMs universally agree that global mean precipitation increases with planetary warming, a result that has a simple thermodynamic explanation (Chou et al., 2009; Held & Soden, 2006). Extreme precipitation is projected to increase at an even greater rate than mean precipitation as the planet warms (Kharin et al., 2007, 2013; Sun et al., 2007), a feature that has already been observed (Fischer & Knutti, 2016).

Studies that seek to improve process understanding of precipitation should be complemented with insights into model quality that uncover systematic errors across an ensemble (Eyring et al., 2019). However, if we wish to compare a GCM simulation to an observational product for precipitation we confront several challenges: short time series, the known systematic excess of light precipitation in GCMs, and the quality of the observations.

The first challenge is the relatively short observational time series. Satellites can be used to fill spatial gaps left by sparse station data, but have only been in operation since approximately the early 1980s or later. Our understanding of internal variability in temperature (Deser et al., 2016) and modes of variability in long GCM control runs (e.g., Deser et al., 2012) suggests that the period since the beginning of the satellite era may be too short. Further evidence comes from studies showing major decadal fluctuations in mean precipitation (Gu & Adler, 2018). The highly skewed nature of this distribution also dictates caution in evaluating precipitation extremes. The main contribution of this work is to show how large ensembles can be used to circumvent this problem of short time series. Other work that uses large ensembles to compare with observations focuses on better-sampled fields like temperature (e.g., Maher et al., 2019; McKinnon et al., 2017; Suarez-Gutierrez et al., 2018), and/or smaller regions with more complete datasets (e.g., von Trentini et al., 2020), for which longer time-series are available.

A second issue is a well-known excess of light precipitation in GCMs (Stephens et al., 2010). There are promising hypotheses to explain this model behavior and its connection to projected changes, in terms of energy balance, and a trade-off between extreme and non-extreme precipitation in simulations. For example, increases in extreme precipitation are typically offset by decreases at less intense precipitation rates in a regionally (and temporally) dependent manner (Thackeray et al., 2018). The excess of simulated light precipitation is an issue for approaches that compare the entire distribution, as it can influence the total precipitation and shape of the frequency distribution. Here, we apply a threshold meant to exclude these lower rain rates and examine only the rest of the precipitation distribution instead.

The third challenge is the quality of the observational record. Precipitation varies greatly on small spatial scales. Stations are scattered unevenly and are often absent altogether at topographic extremes (Touma et al., 2018). As mentioned, the satellite record is short and incomplete, and is often used to fill data gaps in observationally sparse regions, especially over ocean basins. Despite these issues, attempts have been made to synthesize the available historical precipitation data in a global gridded fashion (e.g., Sun, Alexander, et al., 2018). These gridded products can differ substantially in their characterization of extremes both globally (Donat et al., 2014; Gehne et al., 2016) and in specific regions where more data are available (e.g., Timmermans et al., 2019). As this question of observational uncertainty has been the focus of several recent efforts (e.g., Roca et al., 2019), here we choose to focus primarily on the outstanding challenge posed by the inadequate sampling over the short time period of satellite observations.

2. Data and Methods

2.1. GCM Simulations

Single model large ensembles, with perturbed initial conditions, form the central tool of this analysis. Each ensemble member over a short time period can be seen as sampling from the true larger distribution of model daily precipitation under a given forcing. The bulk of this analysis relies on five single-model large ensembles: the 40-member Community Earth System Model version 1.1 Large ensemble (CESM1-LENS, Kay et al., 2015), the 50-member Canadian Earth System Model (CanESM2, Kirchmeier-Young et al., 2017), the 30-member Commonwealth Scientific and Industrial Research Organisation Mk3.6 model (CSIRO-Mk3.6.0, Jeffrey et al., 2013), the 20-member Geophysical Fluid Dynamics Laboratory CM3 model (GFDL-CM3, Sun, Miao, et al., 2018), and the 16-member SMHI/KNMI EC-Earth model (EC-Earth, Hazeleger et al., 2010). These five ensembles were assembled under the auspices of the US CLIVAR Large Ensembles Working Group (Deser et al., 2020).

Daily precipitation data also come from GCMs participating in the Coupled Model Intercomparison Project Phase 5 (CMIP5; Taylor et al., 2012) and Phase 6 (CMIP6; Eyring et al., 2016). When examining this larger set of CMIP models, we only use the first realization from each GCM as internal variability is already accounted for by our use of the single model large ensembles. A full list of CMIP GCMs used is presented in Tables S1 and S2.

We evaluate December-February (DJF) and June-August (JJA) precipitation separately as there is often a strong seasonal variation in both the frequency and magnitude of precipitation extremes, and in the physical processes driving them. To match the availability of observational data (1996–2017), we use the historical simulation from 1996 to 2005 and append the first 12 years of the Representative Concentration Pathway (RCP, van Vuuren et al., 2011) for the end-of-century 8.5 W m^{-2} forcing (RCP8.5 for CMIP5, Riahi et al., 2011; SSP5-8.5 for CMIP6, O'Neill et al., 2016; Tebaldi et al., 2021). (Emissions scenarios do not substantially diverge by 2017.) Lastly, we regrid the GCM output to a 2° × 2° grid using bilinear interpolation, so that no model is evaluated at a resolution much finer than its native grid.

2.2. Observational Estimates

The most comprehensive effort to synthesize precipitation data globally and produce a homogeneous daily product based on satellite data and ground-based measurements is the Global Precipitation Climatology Project (GPCP; Adler et al., 2003, 2018; Huffman et al., 1997, 2009), which offers climatologically corrected complete global coverage. In addition, it is desirable to use a product that has a similar native resolution to the GCMs being evaluated, as downsampling from much higher resolution could result in substantial differences in frequency and intensity (Chen & Dai, 2018). We use the one-degree daily (1DD, Huffman et al., 2001) GPCP product version 1.3 for all comparisons between observed and simulated precipitation distributions. This dataset is then regridded to the aforementioned common 2° grid. It should be noted that the quality of the available satellite observations is not uniform. Poleward of 50° latitude, the GPCP 1DD product uses only Television and Infrared Observation Satellite Operational Vertical Sounder (TOVS) data (Susskind et al., 1997), which may explain the apparent discontinuity in the dataset around 50° north and south. (This discontinuity is visible in the plots of daily intensity at several percentiles; Figure S1.)

2.3. Ruling Out Sameness of Distributions

To focus on the moderate-to-heavy part of the precipitation distribution, we limit our sample to those days above the 85th percentile for wet days (i.e., those days with precipitation rate >0.1 mm/day), calculated on a grid-cell by grid-cell basis (see Figure S1). We evaluate the sensitivity to the choice of threshold (see supporting information). With a higher percentile cutoff, there are fewer demonstrable areas of disagreement because there are often too few samples to reject the null hypothesis, whereas with a lower percentile cutoff larger areas of disagreement emerge as lighter rain rates are included. Examining the variation with percentile for a given location would provide an alternative perspective. Thus, we find that the 85th percentile provides a reasonable balance between sample size and sufficient exclusion of low rain rates.

For each grid cell, we take the subset of days exceeding the GPCP 85th percentile threshold and compare the GCM distribution to that of GPCP using a two-sample Kolmogorov–Smirnov test (KS test, Massey, 1951). The null hypothesis for this test is that the two datasets are drawn from the same distribution; rejecting the null hypothesis ($P < 0.1$) suggests the simulated and observed precipitation distributions are inconsistent with one another. Failing to reject the null hypothesis does not indicate that the distributions are the same, but rather that we cannot rule out their consistency on the basis of the limited data available.

Figure 1 compares the precipitation in Community Earth System Model version 1 Large Ensemble (CESM1-LENS) with GPCP over an example grid cell. Figure 1a illustrates the paucity of samples in the tail of the distribution, as well as the great differences at very low rain rates (note the log frequency scale). Figure 1b shows the two empirical cumulative distribution functions (CDFs). In this case, the two CDFs fail the KS test for sameness of distribution. It fails because the KS statistic, which is simply the maximum distance between the CDFs, is large. Figure 1c shows the CDFs calculated only above the 85th percentile threshold from GPCP. With the two distributions shown, the null hypothesis of sameness of distribution cannot be rejected at the $P < 0.1$ level. This is the test that we adopt. It is most sensitive to the differences in the rain rates near the center of this remaining moderate-to-heavy portion of the distribution. Finally, the other 39 ensemble members of CESM1-LENS are shown in a cloud of CDFs (Figure 1c). Some of these can be statistically distinguished from GPCP while others cannot.

All of the large ensemble members are sampled from the same population defined by that model's behavior. However, for the real-world observational product we have only one realization. If the null hypothesis is

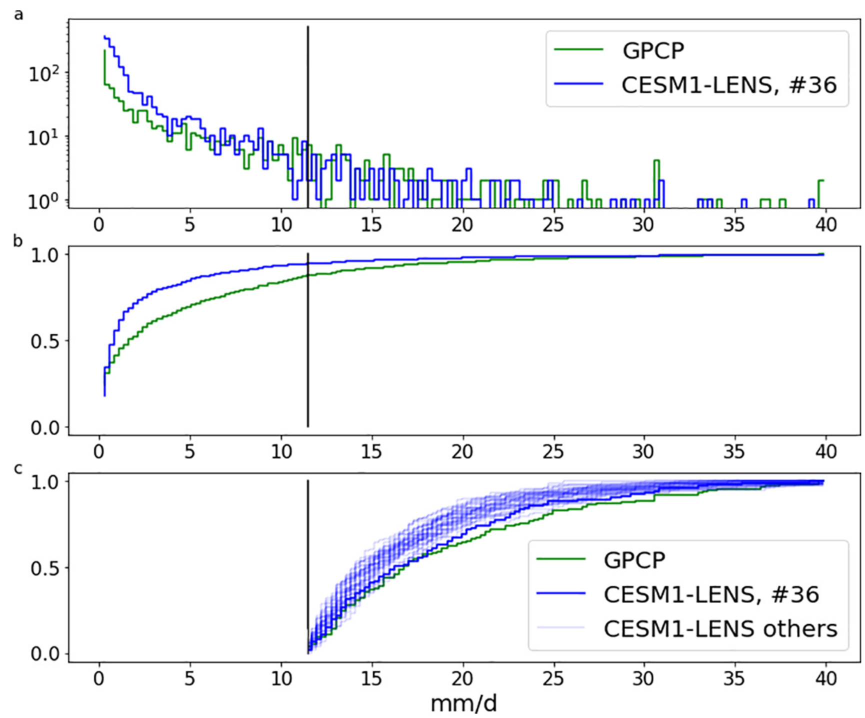

Figure 1. Precipitation distributions for a sample grid cell (33°N, 326°E), comparing one ensemble member of CESM1-LENS with GPCP in (a) frequency and (b) an empirical CDF. In (c), the empirical CDF of the same ensemble member is emphasized (dark blue), and compared with the other 39 ensemble members above the GPCP 85th percentile threshold for the grid cell (vertical black line). CDFs, cumulative distribution functions; GPCP, Global Precipitation Climatology Project.

rejected for every ensemble member, then we rule out that the model and observational product are drawing from the same true larger distribution. In some grid cells with very few samples, we may erroneously fail to reject the null hypothesis, which is why we only make statements about those locations that are demonstrably inconsistent between datasets (see Figure S3 for the geographic distribution of sample size). Further, we apply the false discovery rate method (FDR; see, e.g., Wilks, 2016) to account for field significance.

Those locations which are demonstrably inconsistent between datasets are those for which even the ensemble member with the smallest distance rejects the null hypothesis. We show the minimum distance found among any of the ensemble members at a particular location to illustrate this. We must consider the possibility that the observations are unrepresentative of the real climate's full statistics, a likely outcome in locations subject to significant low frequency (e.g., ~decadal or longer) variability. In such a case, we might hope to find at least one large ensemble member that is in a similar phase of variability to the observations, and which gives a low value for the KS statistic. Conversely, if the test gives a large value for the KS statistic and fails in all 40 ensemble members, then it is highly unlikely that any phasing of variability can make them appear consistent. It is these locations, for which we have the highest confidence of systematic inconsistency, that we highlight when we examine the distance between each ensemble member's CDF and that of GPCP and then find the minimum.

3. Results and Discussion

In Section 3.1, we first present the results for the CESM1-LENS and then extend the analysis to all five large ensembles. In Section 3.2, we aggregate across large areas to compare the large ensembles with the CMIP5 and CMIP6 ensembles of opportunity.

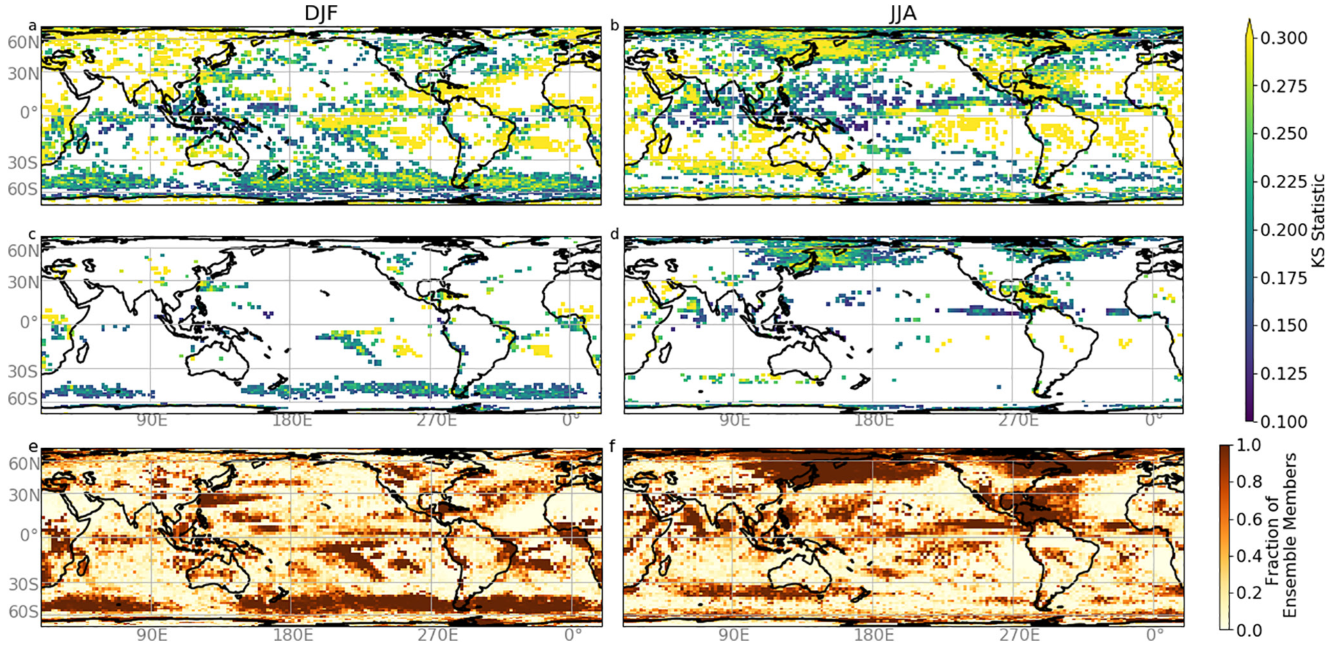

Figure 2. The distance between the CESM1-LENS and GPCP CDFs above the GPCP 85th percentile threshold is shown for ensemble member #1 in (a) DJF and (b) JJA. Grid cells where the distance is not significant at $P < 0.1$ are masked as white. The minimum such distance across all 40 ensemble members at each grid cell independently is shown for (c) DJF and (d) JJA. The fraction of ensemble members for which $P < 0.1$ at each grid cell is shown for (e) DJF and (f) JJA. CDFs, cumulative distribution functions; GPCP, Global Precipitation Climatology Project.

3.1. Identifying Regions of Consistent Disagreement

If we evaluate only the first member of CESM1-LENS against GPCP, we find that most locations fail the KS test at the $P < 0.1$ level (Figures 2a and 2b), meaning that the simulated and observed precipitation distributions are different. The KS test statistic above the 85th percentile (comparing CESM member 1 and GPCP) is used to illustrate this point (yellow = larger distance). Large areas of the subtropics and mid-latitudes may fail to match GPCP in one ensemble member or another, in part because of internal variability. Examining where the null hypothesis is rejected even for the ensemble member for which the distance is smallest (Figures 2c and 2d), we rule out internal variability as the explanation (to the extent that the ensemble has sufficient members—see Figures S2). Generally, the areas of inconsistency that emerge are spatially coherent, lending credence to the idea that the model-observation differences are systematic in these areas. While these locations of systematic differences are robust, they are not necessarily comprehensive: Grid cells with few events above the 85th percentile compared to GPCP may not have enough samples (see Figure S3), yielding an erroneous failure to identify significant differences in these locations.

Coherent areas of difference between the simulated and observed CDFs remain, notably across the Caribbean and North Pacific in JJA and the South Pacific in DJF (Figures 2c–2f). Those grid cells that are inconsistent when using a single ensemble member (as in Figures 2a and 2b), or only some ensemble members, are likely to be inconsistent because of low-frequency internal variability rather than true model bias. Furthermore, a sensitivity study treating one ensemble member of CESM1-LENS as the reference in place of an observational product approaches uniform success across all grid cells after only 10 ensemble members are included (Figure S2). Thus, it is not possible to claim that 40 ensemble members are insufficient to characterize the low-frequency variability for this purpose.

Where the two datasets are inconsistent, it is not necessarily because the model is faulty. For example, in the Southern Ocean, discrepancies are especially clear in December-February (DJF). But GPCP has relatively few observations to assimilate and the daily product uses different satellite inputs poleward of 50°. It is unclear why some Northern high-latitude locations display more inconsistency with GPCP than others. On the other hand, it seems likely that lower latitude locations are inconsistent with GPCP because the model fails to resolve key processes like tropical cyclones (see, e.g., the Caribbean in JJA). However, there is an

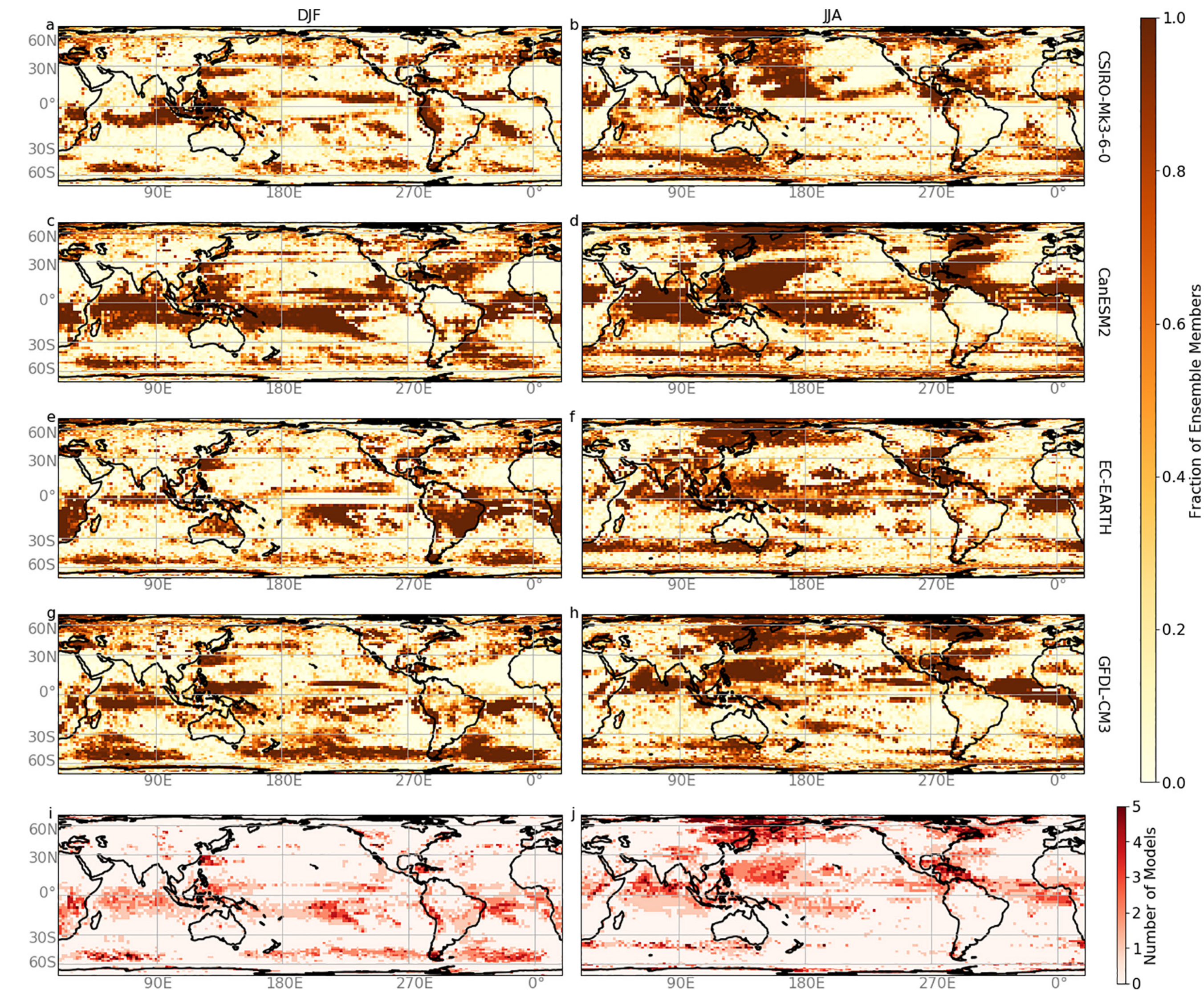

Figure 3. Figures 2e and 2f are duplicated for the four other large ensembles, as radiolabeled. A summary metric is also presented counting the number of the models (maximum of five) to have a significant difference in all ensemble members at each grid cell in (i) DJF and (j) JJA.

absence of similarly poor across-the-board performance in the western Pacific. Other possibilities include a misplaced ITCZ and unrealistic sea surface temperatures.

Next, we conduct the same exercise for four other large ensembles, varying in size from 16 to 50 realizations. Figures 3a–3h show the equivalent to Figures 2c and 2d for each of those other models. Commonalities across all five models are illustrated by a heat map (Figures 3i and 3j). The same regions noted above for CESM1-LENS are common areas of disagreement. The other models typically show disagreement in more of the tropics, where it is easy to imagine that biases in the position and shape of the inter-tropical convergence zone (ITCZ), or the parameterization of convection could lead to inconsistencies with observations. As observational products do vary considerably (e.g., Alexander et al., 2020; Bador et al., 2020) some of these regions of common disagreement may be dependent on the observational product (see Figure S4). Rather than address this issue here, we have focused on this method of visualizing differences between datasets using large ensembles to solve the problem of short time series.

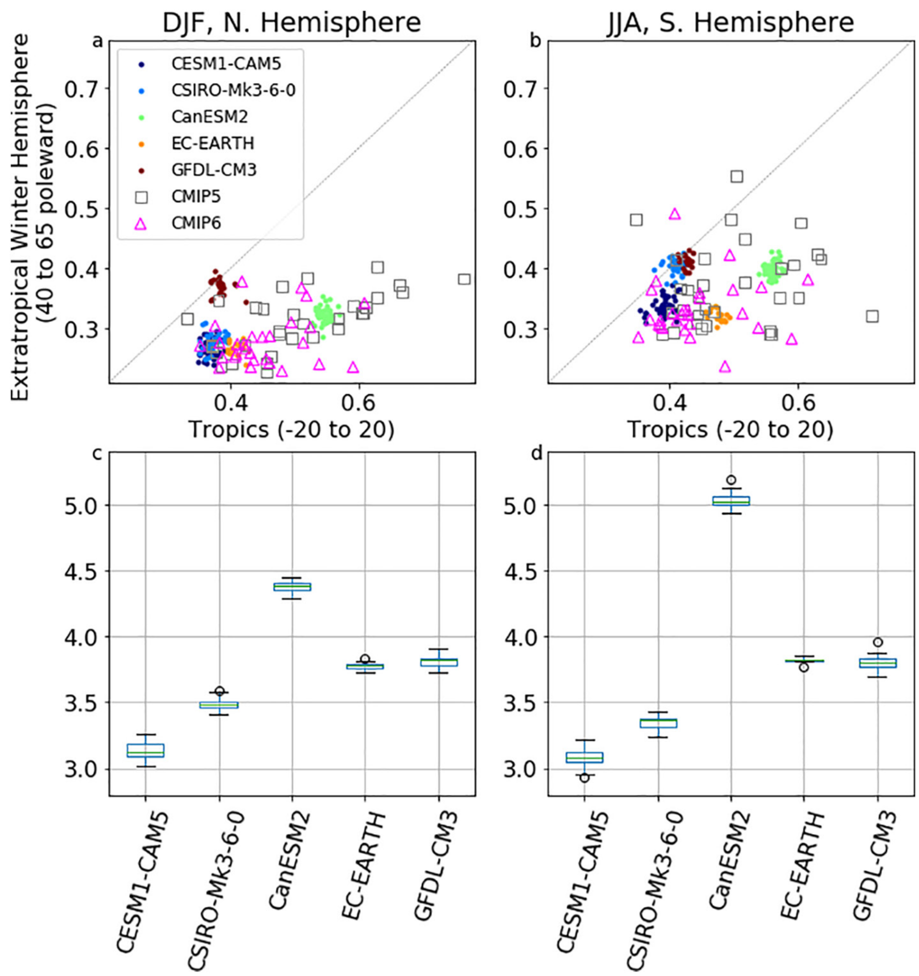

Figure 4. The fraction of grid cells that are inconsistent with GPCP is shown for each ensemble member separately across the five large ensembles and a set of CMIP5 and CMIP6 models for the tropics (equatorward of 20°) and the extratropical winter hemisphere (40°–65° poleward) for (a) DJF/N. hemisphere and (b) JJA/S. hemisphere. The RMSE for mean tropical precipitation above the GPCP 85th percentile is shown for (c) DJF and (d) JJA. The boxplot represents the spread across ensemble members. The models plotted in Figures 4a and 4b are listed in Tables S1 and S2. CDFs, cumulative distribution functions; GPCP, Global Precipitation Climatology Project.

3.2. Aggregating over Tropics and Extratropics

We next look at internal variability another way, by comparing the overall fraction of grid cells in disagreement with GPCP across two broad regions where different precipitation processes might be expected to dominate. The regions are the tropics, equatorward of 20° latitude, where convection is expected to dominate, and the extratropical winter hemisphere (from 40 to 65° poleward) in each season, where large-scale processes dominate. The colored dots of Figures 4a and 4b show the fraction of grid cells in disagreement for each of the individual realizations of the five large ensembles. This allows us to quantify the differences across precipitation regimes that are suggested by inspection of the maps in Figures 2 and 3.

While most of the large ensembles perform similarly in the tropics (~50% agreement), CanESM2 shows substantially larger areas in disagreement with GPCP (>60% of grid cells are in disagreement across all its ensemble members). For the other four models, the clusters of ensemble members all overlap. Of the five model large ensembles, CanESM2 has the lowest native resolution (~2.8 square degrees of latitude/longitude), which might explain its systematically lower performance. Although GFDL-CM3 also has relatively low native resolution, it has a lower RMSE for the pattern of mean tropical precipitation on in-sample days (Figures 4c and 4d) and clusters with the other models in the tropics. It should also be noted that in the sensitivity test with CESM1-LENS (where one ensemble member was treated as "observations"—see supporting information), grid cells in the tropics are the last locations to achieve complete statistical agreement with the reference ensemble member, as ensemble members are added one by one (not shown). We hypothesize that this is due to low-frequency internal variability—in particular ENSO, which might require more realizations before the phasing aligns with that in a historical reference period.

In the extratropics, variability in model performance across the large ensembles is substantially reduced. CanESM2 does not stand out as having a poorer agreement with GPCP compared to the other four models. In the Northern winter extratropics, it is GFDL-CM3 that has slightly larger areas of demonstrable inconsistency with GPCP compared with the others. Generally, CESM1-LENS and EC-EARTH have the smallest areas of demonstrable inconsistency with GPCP in the extratropical winter in both hemispheres.

For JJA extratropical winter, the fractions of inconsistent grid cells are typically greater than the corresponding fractions for DJF extratropical winter. Models appear to better capture the Northern Hemisphere winter precipitation regime, but this difference may not stem solely from factors shaping model quality. Observational products are lower in quality in the Southern hemisphere because it is so much more poorly instrumented. This deficit may make it more difficult to detect high model quality there or may lead to an erroneous impression of low model quality, regardless of the season.

By aggregating across large areas for this portion of the analysis, we increase our effective sample size and can include the CMIP ensembles of opportunity for comparison. We perform the same analysis across the CMIP5 and CMIP6 ensembles (squares and triangles in Figures 4a and 4b). In general, the five models with large ensembles are among those that agree better with GPCP, compared to the full range of CMIP models. The CMIP6 models do not appear to show improved agreement with GPCP on the whole, although some of the worst outliers are absent. The spread within the large ensembles is not negligible compared to the inter-model spread in the CMIP ensembles, further underscoring the need for large single-model ensembles to perform a rigorous model evaluation of precipitation extremes.

If these large ensembles are any guide, then the full internal variability within each CMIP model would likely produce variations in the fraction of inconsistent grid cells similar to those seen in the five models shown. These variations have a maximum magnitude of 0.05 for all five ensembles, while the CMIP ensembles of opportunity in Figures 4a and 4b span a range of roughly 0.3. One model can span 17% of the range spanned by the entire CMIP ensemble, further highlighting the importance of large ensembles for detailed inter-model comparisons.

At the same time, the positioning of CMIP models in Figures 4a and 4b should allow us to make useful statements about the *overall* agreement of the CMIP5 and CMIP6 models when it comes to moderate-to-heavy precipitation aggregated across broad regions. In general, we find that the models are in better agreement with GPCP in the extratropical winter hemisphere than in the tropics. This is probably a reflection of the fact that the processes associated with large precipitation events in the extratropical winter are better resolved compared to their tropical counterparts. We also see from Figure 4 that the models performing better in one regime are not necessarily those that perform better in the other. This probably stems from the fact that different processes produce precipitation in the extratropical and tropical regimes. Although there are greater demonstrable differences with GPCP in the tropics than the extratropics in general, there are some GCMs for which the moderate-to-heavy portion of the precipitation distribution is potentially consistent with GPCP in as much as half the convective-dominated region, despite the greater uncertainties associated with convective parameterizations in the GCMs.

While we have noted and interpreted some of the geographic patterns of inconsistency between models and GPCP, we caution that conclusions about model quality in particular regions are somewhat dependent on the observational reference product, as observational uncertainty is known to be quite large. In this portion of the analysis, we have shown that aggregated across large regions, the CMIP ensembles of opportunity tell a similar story to the large ensembles about how the tropics compare with the extratropics in terms of demonstrable inconsistency with GPCP. The prior spatial disaggregation in the maps provides further detail for those models for which a large ensemble is available.

4. Conclusions

We have addressed one of the central challenges of evaluating moderate-to-heavy daily precipitation simulated by GCMs in the face of short observational time series and large decadal variability. To distinguish whether disagreement between models and observations reflects true systematic differences, or different phasing of internal variability, single model large ensembles are a useful tool. We use five single-model large ensembles to identify coherent regions where each GCM is systematically unable to capture moderate-to-heavy precipitation consistently with GPCP. The CMIP5 and CMIP6 ensembles are consistent in aggregate in their relative performance across distinct precipitation regimes.

Most of the areas for which we cannot rule out the consistency with GPCP that emerge across the models are in the extratropical northern winter. This is both a well-instrumented geographic area as well as a precipitation regime that is dominated by large-scale drivers that are well resolved in GCMs. Areas of demonstrable inconsistency do not necessarily reflect model inadequacies, as either the GCM or the observational product could be lacking. One such area of systematically lower agreement that may be explained by lower observational quality is the southern hemisphere winter.

By contrast, there are other locations where a more likely explanation is the failure of the GCM, either to resolve essential processes or to properly place large-scale circulation features. Moderate-to-heavy extratropical winter precipitation is less inconsistent with GPCP than moderate-to-heavy tropical precipitation in almost every model. Within the tropics, moderate-to-heavy precipitation is particularly inconsistent in certain regions, including the Caribbean (especially during JJA), where resolution may be insufficient to capture processes like tropical cyclones. In the northern and southern flanks of the Pacific and Atlantic ITCZ, and the Indian Ocean, biases in large-scale circulation could be an additional source of greater distance between two CDFs. Thus, the method we have introduced makes it easy to visualize those locations where caution is required in drawing conclusions based on simulated moderate-to-heavy precipitation, be it due to model or observational inadequacy. There are different sources that could explain the inconsistencies that we detect in CDFs, including biases in the number of days over a precipitation threshold, and differences in trends that are not uniform across precipitation intensities. An initial diagnostic of locations of disagreement could lead to further investigation in a region of interest. We encourage others to apply this method with other datasets and other metrics of interest—all that is required of the metric is a measure of whether two values differ at some significance level.

The methods described in this study reveal that the limitations imposed by internal variability and the short observational record can be at least partly circumvented with additional samples from large ensembles of simulations. Our work assumes that the internal variability itself is not biased. If one chose a field that is better sampled observationally, less subject to low-frequency internal variability, or less inherently skewed, one could go farther and evaluate the model's representation of the variability itself, as in McKinnon et al. (2017). Doing so might rule out consistent moderate-to-heavy precipitation in even more locations.

We also can identify differentials in summary metrics over very large latitude zones and seasons (e.g., tropics vs. extratropics). This work has implications for model evaluation. As we go to smaller scales, complications arising from internal variability become increasingly difficult to navigate. For example, it is common to compare observed and simulated precipitation over a small area to select GCMs that perform better for regional applications. But we have shown that the relatively short observational period of the satellite era is usually insufficient to properly sample the daily precipitation distribution in individual grid cells (e.g., Figures 1 and 2). So it is inappropriate to rule out a model because it does not match the particular phasing of the internal variability in daily precipitation that occurred historically. Thus, our analysis also highlights the importance of avoiding regional-scale model evaluations of the precipitation distribution based on a single ensemble member.

Data Availability Statement

All data used in this study are publicly available. The authors acknowledge the World Climate Research Program's Working Group on Coupled Modeling and the individual modeling groups who contributed data for CMIP5 and CMIP6. CMIP data are downloaded from https://esgf-node.llnl.gov. The authors also acknowledge the US CLIVAR Working Group on Large Ensembles (data available from http://www.cesm.ucar.edu/projects/community-projects/MMLEA/). The GPCP 1DD v1.3 product description and download instructions can be found at http://eagle1.umd.edu/GPCP_ICDR/GPCP_Monthly.html. The TRMM (TMPA) Rainfall Estimate L3 3 h 0.25 × 0.25° V7 data (Huffman et al., 2007) can be found at https://disc.gsfc.nasa.gov/datasets/TRMM_3B42_7/summary.

Acknowledgments

Support for this research is provided by the US Department of Energy (DE-SC0016605, "An Integrated Evaluation of the Simulated Hydroclimate of the Continental US" and DE-SC0016438, "A Hierarchical Evaluation Framework for Assessing Climate Simulations Relevant to the Energy-Water-Land Nexus"), as well as by the California State Water Resources Control Board (Agreement No. 16-024-140). Contributions by D. L. Swain were made possible through a joint collaboration between the Institute of the Environment and Sustainability at the University of California, Los Angeles; the Center for Climate and Weather Extremes at the National Center for Atmospheric Research; and the Nature Conservancy of California. Finally, the authors acknowledge the reviewers of this as well as an earlier related manuscript for helpful comments that substantially improved the quality of this work.

References

Adler, R., Sapiano, M., Huffman, G., Wang, J.-J., Gu, G., Bolvin, D., et al. (2018). The global precipitation climatology Project (GPCP) monthly analysis (new version 2.3) and a review of 2017 global precipitation. *Atmosphere*, *9*(4), 138. https://doi.org/10.3390/atmos9040138

Adler, R. F., Huffman, G. J., Chang, A., Ferraro, R., Xie, P.-P., Janowiak, J., et al. (2003). The version-2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979–present). *Journal of Hydrometeor*, *4*(6), 1147–1167. https://doi.org/10.1175/1525-7541(2003)004<1147:tvgpcp>2.0.co;2

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

Bador, M., Alexander, L. V., Contractor, S., & Roca, R. (2020). Diverse estimates of annual maxima daily precipitation in 22 state-of-the-art quasi-global land observation datasets. *Environmental Research Letters*, *15*(3), 035005. https://doi.org/10.1088/1748-9326/ab6a22

Chen, D., & Dai, A. (2018). Dependence of estimated precipitation frequency and intensity on data resolution. *Climate Dynamics*, *50*(9), 3625–3647.

Chou, C., Neelin, J. D., Chen, C.-A., & Tu, J.-Y. (2009). Evaluating the "Rich-Get-Richer" mechanism in tropical precipitation change under Global Warming. *Journal of Climate*, *22*(8), 1982–2005. https://doi.org/10.1175/2008jcli2471.1

DeAngelis, A. M., Qu, X., Zelinka, M. D., & Hall, A. (2015). An observational radiative constraint on hydrologic cycle intensification. *Nature*, *528*, 249. https://doi.org/10.1038/nature15770

Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., et al. (2020). Insights from Earth system model initial-condition large ensembles and future prospects. *Nature Climate Change*, *10*, 277–286. https://doi.org/10.1038/s41558-020-0731-2

Deser, C., Phillips, A. S., Tomas, R. A., Okumura, Y. M., Alexander, M. A., Capotondi, A., et al. (2012). Enso and pacific decadal variability in the community climate system model version 4. *Journal of Climate*, *25*(8), 2622–2651. https://doi.org/10.1175/jcli-d-11-00301.1

Deser, C., Terray, L., & Phillips, A. S. (2016). Forced and internal components of winter air temperature trends over North America during the past 50 years: Mechanisms and implications. *Journal of Climate*, *29*(6), 2237–2258. https://doi.org/10.1175/jcli-d-15-0304.1

Donat, M. G., Sillmann, J., Wild, S., Alexander, L. V., Lippmann, T., & Zwiers, F. W. (2014). Consistency of temperature and precipitation extremes across various global gridded in situ and reanalysis datasets. *Journal of Climate*, *27*(13), 5019–5035. https://doi.org/10.1175/jcli-d-13-00405.1

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* (p. 9). Geoscientific Model Development (Online).

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., et al. (2019). Taking climate model evaluation to the next level. *Nature Climate Change*, *9*, 102–110. https://doi.org/10.1038/s41558-018-0355-y

Fischer, E. M., & Knutti, R. (2016). Observed heavy precipitation increase confirms theory and early models. *Nature Climate Change*, *6*(11), 986–991. https://doi.org/10.1038/nclimate3110

Fläschner, D., Mauritsen, T., & Stevens, B. (2016). Understanding the intermodel spread in global-mean hydrological sensitivity. *Journal of Climate*, *29*(2), 801–817. https://doi.org/10.1175/jcli-d-15-0351.1

Gehne, M., Hamill, T. M., Kiladis, G. N., & Trenberth, K. E. (2016). Comparison of global precipitation estimates across a range of temporal and spatial scales. *Journal of Climate*, *29*(21), 7773–7795. https://doi.org/10.1175/jcli-d-15-0618.1

Gu, G., & Adler, R. F. (2018). Precipitation intensity changes in the tropics from observations and models. *Journal of Climate*, *31*(12), 4775–4790. https://doi.org/10.1175/jcli-d-17-0550.1

Hazeleger, W., Severijns, C., Semmler, T., Ştefănescu, S., Yang, S., Wang, X., et al. (2010). EC-Earth. *Bulletin of the American Meteorological Society*, *91*(10), 1357–1364. https://doi.org/10.1175/2010bams2877.1

Held, I. M., & Soden, B. J. (2006). Robust responses of the hydrological cycle to global warming. *Journal of Climate*, *19*(21), 5686–5699. https://doi.org/10.1175/jcli3990.1

Huffman, G. J., Adler, R. F., Arkin, P., Chang, A., Ferraro, R., Gruber, A., et al. (1997). The global precipitation climatology project (GPCP) combined precipitation dataset. *Bulletin of the American Meteorological Society*, *78*(1), 5–20. https://doi.org/10.1175/1520-0477(1997)078<0005:tgpcpg>2.0.co;2

Huffman, G. J., Adler, R. F., Bolvin, D. T., & Gu, G. (2009). Improving the global precipitation record: GPCP version 2.1. *Geophysical Research Letters*, *36*(17). https://doi.org/10.1029/2009gl040000

Huffman, G. J., Adler, R. F., Morrissey, M. M., Bolvin, D. T., Curtis, S., Joyce, R., et al. (2001). Global precipitation at one-degree daily resolution from multisatellite observations. *Journal of Hydrometeorology*, *2*(1), 36–50. https://doi.org/10.1175/1525-7541(2001)002<0036:gpaodd>2.0.co;2

Huffman, G. J., Bolvin, D. T., Nelkin, E. J., Wolff, D. B., Adler, R. F., Gu, G., et al. (2007). The TRMM multisatellite precipitation analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *Journal of Hydrometeorology*, *8*(1), 38–55. https://doi.org/10.1175/jhm560.1

Jeffrey, S., Rotstayn, L., Collier, M., Dravitzki, S., Hamalainen, C., Moeseneder, C., et al. (2013). Australia's CMIP5 submission using the CSIRO-Mk3.6 model. *Australian Journal of Meterology*. *63*(1), 1–13. https://doi.org/10.22499/2.6301.001

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., et al. (2015). The community earth system model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bulletin of the American Meteorological Society*, *96*(8), 1333–1349. https://doi.org/10.1175/bams-d-13-00255.1

Kharin, V. V., Zwiers, F. W., Zhang, X., & Hegerl, G. C. (2007). Changes in temperature and precipitation extremes in the IPCC ensemble of global coupled model simulations. *Journal of Climate*, *20*(8), 1419–1444. https://doi.org/10.1175/jcli4066.1

Kharin, V. V., Zwiers, F. W., Zhang, X., & Wehner, M. (2013). Changes in temperature and precipitation extremes in the CMIP5 ensemble. *Climatic Change*, *119*(2), 345–357. https://doi.org/10.1007/s10584-013-0705-8

Kirchmeier-Young, M. C., Zwiers, F. W., & Gillett, N. P. (2016). Attribution of extreme events in arctic sea ice extent. *Journal of Climate*, *30*(2), 553–571.

Liu, D., Yang, B., Zhang, Y., Qian, Y., Huang, A., Zhou, Y., & Zhang, L. (2018). Combined impacts of convection and microphysics parameterizations on the simulations of precipitation and cloud properties over Asia. *Atmospheric Research*, *212*, 172–185. https://doi.org/10.1016/j.atmosres.2018.05.017

Maher, N., Milinski, S., Suarez-Gutierrez, L., Botzet, M., Dobrynin, M., Kornblueh, L., et al. (2019). The Max Planck institute grand ensemble: Enabling the exploration of climate system variability. *Journal of Advances in Modeling Earth Systems*, *11*(7), 2050–2069. https://doi.org/10.1029/2019ms001639

Massey, F. J., Jr (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, *46*(253), 68–78. https://doi.org/10.1080/01621459.1951.10500769

McKinnon, K. A., Poppick, A., Dunn-Sigouin, E., & Deser, C. (2017). An "observational large ensemble" to compare observed and modeled temperature trend uncertainty due to internal variability. *Journal of Climate*, *30*(19), 7585–7598. https://doi.org/10.1175/jcli-d-16-0905.1

Morales, A., Posselt, D. J., Morrison, H., & He, F. (2019). Assessing the influence of microphysical and environmental parameter perturbations on orographic precipitation. *Journal of the Atmospheric Sciences*, *76*(5), 1373–1395. https://doi.org/10.1175/jas-d-18-0301.1

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.

Pendergrass, A. G., & Hartmann, D. L. (2014). The atmospheric energy constraint on global-mean precipitation change. *Journal of Climate*, *27*(2), 757–768. https://doi.org/10.1175/jcli-d-13-00163.1

Qiao, F., & Liang, X.-Z. (2016). Effects of cumulus parameterization closures on simulations of summer precipitation over the United States coastal oceans. *Journal of Advances in Modeling Earth Systems*, *8*(2), 764–785. https://doi.org/10.1002/2015ms000621

Riahi, K., Rao, S., Krey, V., Cho, C., Chirkov, V., Fischer, G., et al. (2011). RCP 8.5—A scenario of comparatively high greenhouse gas emissions. *Climatic Change*, *109*(1), 33. https://doi.org/10.1007/s10584-011-0149-y

Roca, R., Alexander, L. V., Potter, G., Bador, M., Jucá, R., Contractor, S., et al. (2019). FROGS: a daily 1° × 1° gridded precipitation database of rain gauge, satellite and reanalysis products. *Earth System Science Data*, *11*(3), 1017–1035. https://doi.org/10.5194/essd-11-1017-2019

Stephens, G. L., L'Ecuyer, T., Forbes, R., Gettelmen, A., Golaz, J.-C., Bodas- Salcedo, A., et al. (2010). Dreary state of precipitation in global models. *Journal of Geophysical Research*, *115*(D24). https://doi.org/10.1029/2010jd014532

Suarez-Gutierrez, L., Li, C., Müller, W. A., & Marotzke, J. (2018). Internal variability in European summer temperatures at 1.5°C and 2°C of global warming. *Environmental Research Letters*, *13*(6), 064026. https://doi.org/10.1088/1748-9326/aaba58

Sun, L., Alexander, M., & Deser, C. (2018). Evolution of the global coupled climate response to Arctic Sea ice loss during 1990–2090 and its contribution to climate change. *Journal of Climate*, *31*(19), 7823–7843. https://doi.org/10.1175/jcli-d-18-0134.1

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., & Hsu, K. L. (2018). A review of global precipitation data sets: Data sources, estimation, and intercomparisons. *Reviews of Geophysics*, *56*(1), 79–107. https://doi.org/10.1002/2017rg000574

Sun, Y., Solomon, S., Dai, A., & Portmann, R. W. (2007). How often will it rain?. *Journal of Climate*, *20*(19), 4801–4818. https://doi.org/10.1175/jcli4263.1

Susskind, J., Piraino, P., Rokke, L., Iredell, L., & Mehta, A. (1997). Characteristics of the TOVS Pathfinder Path A dataset. *Bulletin of the American Meteorological Society*, *78*(7), 1449–1472. https://doi.org/10.1175/1520-0477(1997)078<1449:cottpp>2.0.co;2

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, *93*(4), 485–498. https://doi.org/10.1175/bams-d-11-00094.1

Tebaldi, C., Debeire, K., Eyring, V., Fischer, E., Fyfe, J., Friedlingstein, P., et al. (2021). Climate model projections from the scenario model intercomparison project (ScenarioMIP) of CMIP6. *Earth System Dynamics*, *12*(1), 253–293. https://doi.org/10.5194/esd-12-253-2021

Thackeray, C. W., DeAngelis, A. M., Hall, A., Swain, D. L., & Qu, X. (2018). On the connection between global hydrologic sensitivity and regional wet extremes. *Geophysical Research Letters*, *45*(20), 11–343. https://doi.org/10.1029/2018gl079698

Timmermans, B., Wehner, M., Cooley, D., O'Brien, T., & Krishnan, H. (2019). An evaluation of the consistency of extremes in gridded precipitation data sets. *Climate Dynamics*, *52*(11), 6651–6670. https://doi.org/10.1007/s00382-018-4537-0

Touma, D., Michalak, A. M., Swain, D. L., & Diffenbaugh, N. S. (2018). Characterizing the spatial scales of extreme daily precipitation in the United States. *Journal of Climate*, *31*(19), 8023–8037. https://doi.org/10.1175/jcli-d-18-0019.1

van Vuuren, D. P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., et al. (2011). The representative concentration pathways: An overview. *Climatic Change*, *109*(1–2), 5. https://doi.org/10.1007/s10584-011-0148-z

von Trentini, F., Aalbers, E. E., Fischer, E. M., & Ludwig, R. (2020). Comparing interannual variability in three regional single-model initial-condition large ensembles (SMILEs) over Europe. *Earth System Dynamics*, *11*(4), 1013–1031. https://doi.org/10.5194/esd-11-1013-2020

Wilks, D. S. (2016). "The stippling shows statistically significant grid points": How research results are routinely overstated and overinterpreted, and what to do about it. *Bulletin of the American Meteorological Society*, *97*(12), 2263–2273. https://doi.org/10.1175/bams-d-15-00267.1

Zhao, M., Golaz, J.-C., Held, I. M., Ramaswamy, V., Lin, S.-J., Ming, Y., et al. (2016). Uncertainty in model climate sensitivity traced to representations of cumulus precipitation microphysics. *Journal of Climate*, *29*(2), 543–560. https://doi.org/10.1175/jcli-d-15-0191.1