JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2019MS001639

Special Section:
The Max Planck Institute for Meteorology Earth System Model version 1.2

The Max Planck Institute Grand Ensemble: Enabling the Exploration of Climate System Variability

Nicola Maher[1], Sebastian Milinski[1,2], Laura Suarez-Gutierrez[1,2], Michael Botzet[1], Mikhail Dobrynin[3,4], Luis Kornblueh[1], Jürgen Kröger[1], Yohei Takano[1], Rohit Ghosh[1], Christopher Hedemann[1], Chao Li[1], Hongmei Li[1], Elisa Manzini[1], Dirk Notz[1], Dian Putrasahan[1], Lena Boysen[1], Martin Claussen[1,3], Tatiana Ilyina[1], Dirk Olonscheck[1], Thomas Raddatz[1], Bjorn Stevens[1], and Jochem Marotzke[1]

[1]Max Planck Institute for Meteorology, Hamburg, Germany, [2]International Max Planck Research School on Earth System Modelling, Hamburg, Germany, [3]Centrum für Erdsystemforschung und Nachhaltigkeit (CEN, Universität Hamburg, Hamburg, Germany, [4]Deutscher Wetterdienst (DWD), Hamburg, Germany

Abstract The Max Planck Institute Grand Ensemble (MPI-GE) is the largest ensemble of a single comprehensive climate model currently available, with 100 members for the historical simulations (1850–2005) and four forcing scenarios. It is currently the only large ensemble available that includes scenario representative concentration pathway (RCP) 2.6 and a 1% CO_2 scenario. These advantages make MPI-GE a powerful tool. We present an overview of MPI-GE, its components, and detail the experiments completed. We demonstrate how to separate the forced response from internal variability in a large ensemble. This separation allows the quantification of both the forced signal under climate change and the internal variability to unprecedented precision. We then demonstrate multiple ways to evaluate MPI-GE and put observations in the context of a large ensemble, including a novel approach for comparing model internal variability with estimated observed variability. Finally, we present four novel analyses, which can only be completed using a large ensemble. First, we address whether temperature and precipitation have a pathway dependence using the forcing scenarios. Second, the forced signal of the highly noisy atmospheric circulation is computed, and different drivers are identified to be important for the North Pacific and North Atlantic regions. Third, we use the ensemble dimension to investigate the time dependency of Atlantic Meridional Overturning Circulation variability changes under global warming. Last, sea level pressure is used as an example to demonstrate how MPI-GE can be utilized to estimate the ensemble size needed for a given scientific problem and provide insights for future ensemble projects.

Key Points:
- The 100-member MPI-GE is currently the largest publicly available ensemble of a comprehensive climate model
- MPI-GE currently has the most forcing scenarios of all large ensemble projects: RCP2.6, RCP4.5, RCP8.5, and 1% CO_2
- The power of MPI-GE is to estimate the forced response and internal variability, including changing variability, to unprecedented precision

Correspondence to:
N. Maher,
nicola.maher@mpimet.mpg.de

Citation:
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*, 2050–2069. https://doi.org/10.1029/2019MS001639

Received 29 JAN 2019
Accepted 28 MAY 2019
Accepted article online 4 JUN 2019
Published online 5 JUL 2019

©2019. The Authors.
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

1. Introduction

Internal variability and uncertainties in model physics and the future forcing all contribute to uncertainties in climate projections (Hawkins & Sutton, 2009). While multimodel ensembles such as the Coupled Model Intercomparison Project (CMIP; Taylor et al., 2012) can be used to effectively investigate the combined effect of all three in climate projections, it is difficult to separate internal variability from the forced response with a limited number of ensemble members of each single model. For single realizations, linear detrending is often used with the intention of removing the forced response and isolating the internal variability (e.g., Frankcombe et al., 2018). However, this then introduces biases in the amplitude and phase of internal variability, with more complicated scaling methods needed to better separate the two quantities (Frankcombe et al., 2015, 2018). Internal variability can be quantified using a long control simulation in the absence of external forcing (e.g., Thompson et al., 2015; Wittenberg et al., 2014). However, internal variability may itself be influenced by external forcing (e.g., Maher et al., 2015) in ways that are difficult to account for a priori. This means that a long control run cannot be used to address projections where the variability itself might change. A large ensemble of a single model can be used to estimate changes in variability in this model, uncertainties due to future forcing, and together with other model ensembles can be used to address uncertainties in model physics. The Max Planck Institute Grand Ensemble (MPI-GE) is currently the largest such ensemble and will be introduced in this paper.

Large-ensemble projects of comprehensive coupled climate models are gaining traction as methods to robustly estimate internal variability in transient simulations and to quantify the forced signal (e.g., Kay

et al., 2015). The first large ensemble project was a 62-member simulation of Community Climate System Model 1.4 run for the period 1940–2080 (e.g., Branstator & Selten, 2009; Zelle et al., 2005). Three other large ensembles are currently publicly available. One is the Community Earth System Modelling Large Ensemble Project (LENS), which was run by National Center for Atmospheric Research (NCAR) for the period 1920–2100 and has 42 members of the historical simulation and representative concentration pathway (RCP) 8.5 scenario (CESM-LE) and 15 members of the RCP4.5 scenario (CESM-ME) (Kay et al., 2015; Sanderson et al., 2018). Another is the Geophysical Fluid Dynamics Laboratory large ensemble, which consists of 30 members from the RCP8.5 scenario (2006–2100; e.g., von Känel et al., 2017). The third is the Canadian Earth System Model Large Ensembles, which has 50 members of three single-forcing experiments run from 1950–2020 and 50 members of the historical simulation run from 1950–2005 and continued using the RCP8.5 scenario run from 2006–2100 (Kirchmeier-Young et al., 2017). Other modeling groups have also recently completed large ensembles; however, they are not yet publicly available (e.g., Frankignoul et al., 2017; Stolpe et al., 2018).

Studies that utilize large ensembles have been extensively used to investigate the internal variability of the climate system (e.g., Dai & Bloecker, 2019; Fasullo & Nerem, 2016; Frankignoul et al., 2017; Smith & Jahn, 2019) and extreme events (e.g., Diffenbaugh et al., 2015; Gibson et al., 2017; Kirchmeier-Young et al., 2017; Tebaldi & Wehner, 2018; Wang et al., 2018). They have also been used as a test bed for new methodologies such as creating an observational large ensemble (McKinnon et al., 2017; McKinnon & Deser, 2018), built by combining the forced response from CESM-LE and the estimated internal variability from observations. These ensembles have also been used as a test bed for dynamical adjustment, which can be used to remove the internal dynamical signal and consequently bring observations closer to the forced response (Deser et al., 2016; Lehner et al., 2017). Additionally, large ensembles have been used to inform observing systems, for example, for marine ecosystem drivers such as ocean acidification and to provide information to optimize the observing system (Rodgers et al., 2015). The previously available ensembles and the work associated with them have provided a treasure trove of information, but more large ensembles are still useful, particularly to investigate the robustness of simulated internal variability and the forced response. Additional information can also be gained from having a very large ensemble, such as investigating extreme events and computing the forced signal for highly variable quantities.

MPI-GE is the largest ensemble of simulations for any given scenario currently available, with 100 members for the historical simulation and each of four forcing scenarios. It is currently the only large ensemble available where three future scenarios (RCP2.6, RCP4.5, and RCP8.5), each consisting of 100 members, can be compared. It additionally enables studies of the targets set by the Paris agreement using the RCP2.6 scenario. MPI-GE has the advantage that it is initialized by sampling the preindustrial control state, which effectively samples the full phase space of both the ocean and atmosphere states. This method has been shown to produce a larger spread than the atmospheric perturbation method and hence a better sample of the full range of variability from the beginning of the simulation (Hawkins et al., 2016). This allows investigation of the late nineteenth century and the early twentieth century warming, given that the ensemble is initialized in 1850, something that is not possible with the other existent ensembles due to their later start dates. Being able to contrast different warming states provides useful constraints on the behavior of the climate system and the magnitude of different forcings (Stevens, 2015). The initialization method is particularly important for investigating the variability of quantities that have been shown to have longer times of divergence when initialized with atmospheric perturbations, such as regional temperature and precipitation trends (Hawkins et al., 2016), the Atlantic Meridional Overturning Circulation (AMOC), and 2,000-m ocean heat content (Marotzke, 2019). Overall, these advantages make MPI-GE very powerful.

The utility of MPI-GE itself has also been demonstrated in previous studies (Bittner et al., 2016; Bengtsson & Hodges, 2018; Dessler et al., 2018; Hedemann et al., 2017; Li & Ilyina, 2018; Manzini et al., 2018; Maher et al., 2018; Marotzke, 2019; Niederdrenk & Notz, 2018; Plesca et al., 2018; Rädel et al., 2016; Stevens, 2015; Suárez-Gutiérrez et al., 2017, 2018; Zhang et al., 2018). Some high-profile examples include the investigation into the 1998–2012 hiatus and extreme events; for example, Hedemann et al. (2017) used MPI-GE to investigate the recent surface warming hiatus. Whereas most studies suggested that ocean heat uptake caused the hiatus, Hedemann et al. (2017) found that energy radiated upward from the surface could have caused the hiatus as well and that observational uncertainty is too large for us to know which explanation is correct. Studies investigating extreme events also benefit from the large ensemble size of MPI-GE. Suarez-Gutierrez et al. (2018) investigated European summer temperature extremes using MPI-GE. They found that in a climate that warms globally by 2 °C, the European summer extremes are 1 °C warmer than in a climate that

warms by 1.5 °C. They also found that the 2003 heatwave has a 1 in 2,000 chance of happening under preindustrial conditions, while it occurs under 1.5 °C warming every other year. An ensemble of 100 members allows the sampling of 1 in 100-year extreme events in the ensemble simulated every year on average and further allows the simulation and characterization of large samples of extreme events with return periods over hundreds of years. The use of MPI-GE allowed Suarez-Gutierrez et al. (2018) to investigate events with return periods of up to 500 years, without explicitly parametrizing the tails of the distributions using extreme value statistics.

Internal variability and the forced response can be separated with high precision using MPI-GE. MPI-GE has previously been used to show that the observed negative decadal trend in the ocean carbon sink in the 1990s can be attributed to internal variability (Li & Ilyina, 2018). Li and Ilyina (2018) also indicate that, in the presence of large internal variability, the emergence time of the forced response in the ocean carbon sink is beyond a decade. Forced temperature trends in the upper tropical troposphere are larger in most models that in observations. Suárez-Gutiérrez et al. (2017) used MPI-GE to show that most of these differences can be explained by internal variability alone. This indicates that differences between models and observations may also be misinterpreted in the absence of large ensembles, such as MPI-GE.

Given its large size, MPI-GE can also be used to address the question of how many ensemble members of a single model are needed to address a given problem. While Daron and Stainforth (2013) suggested that ensembles of several hundred members may be required to characterize a model's climate, Drótos et al. (2017) suggest that 100 members may be sufficient for analyzing the forced response.. Maher et al. (2018) found that 30–40 members are needed to robustly estimate ENSO variability in MPI-GE. Olonscheck and Notz (2017) used CMIP5 and MPI-GE to suggest that when investigating sea ice variability, multiple small ensembles of coupled climate models are of more use than either a large ensemble of a single model or a multimodel ensemble of single realizations. Other studies have looked at how many members are needed to detect forced changes in specific variables. Li and Ilyina (2018) found that up to 79 ensemble members are needed to detect the forced decadal trends in the carbon sink under the RCP4.5 forcing regime, with the largest number of members needed in the Southern Ocean. Bittner et al. (2016) investigated how many ensemble members are needed to identify a forced change in the northern hemisphere polar vortex in the winter after the Pinatubo eruption. They found that 7 to 40 members are needed, depending on the latitude considered. Another example based on sea level pressure (SLP) trends is used in this paper to demonstrate how MPI-GE can be used to determine the ensemble size needed.

The purpose of this paper is twofold. The first is to present MPI-GE to the wider community, the second to further demonstrate the usefulness of this 100-member ensemble by presenting a variety of examples and some novel analyses. In section 2, MPI-GE is presented, and the ensemble simulations are described. In section 3, we use MPI-GE to investigate specific quantities and how they evolve in time in different scenarios. In section 4, we demonstrate how to compare MPI-GE to observations and show a novel approach for evaluating the model internal variability. In section 5, we show four examples of scientific problems that can be best investigated with a large ensemble of a single climate model. The first investigates whether temperature and precipitation behave similarly at the same warming levels in the different forcing scenarios. The second demonstrates the quantification of the forced signals in the northern hemisphere atmospheric circulation when there is strong variability. The third investigates changes in variability itself, using the AMOC as an example. The fourth demonstrates how MPI-GE can be used to determine the ensemble size needed for a specific quantity, in this case for projected trends in SLP. Finally, we discuss the use of the MPI-GE in the context of current climate modeling in the scientific community.

2. MPI-GE

2.1. The Model

MPI-ESM is described by Giorgetta et al. (2013). MPI-GE uses MPI-ESM1.1 (version MPI-ESM 1.1.00p2), is run in low-resolution configuration, and consists of the following components. The ocean component is MPIOM (version mpiom-1.6.1p1; Marsland et al., 2003), run on the GR15L40 grid. The ocean biogeochemistry model is HAMOCC5.2 and is run as described by Ilyina et al. (2013). ECHAM (version echam-6.3.01p3; Stevens et al., 2013) provides the atmosphere component, run in a T63L47 configuration. The land component is the JSBACH model (version jsbach-3.00) including dynamic vegetation and land use transitions with the standard fire module (Reick et al., 2013). This model configuration has an atmosphere of approximately

Table 1
Initialization Branching Times From the Preindustrial Control Run

Ensemble member	Branch time	Ensemble member	Branch time
1	1898	51	3164
2	1946	52	3188
3	1994	53	3212
4	2042	54	3236
5	2090	55	3260
6	2138	56	3284
7	2186	57	3308
8	2234	58	3332
9	2282	59	3356
10	2330	60	3380
11	2378	61	3404
12	2426	62	3428
13	2474	63	3452
14	2522	64	3476
15	2570	65	3500
16	2618	66	3524
17	2666	67	2906
18	2714	68	2930
19	2762	69	2954
20	2810	70	2978
21	1874	71	2822
22	1922	72	2846
23	1970	73	2870
24	2018	74	2894
25	2066	75	2918
26	2114	76	2942
27	2162	77	2966
28	2210	78	2990
29	2258	79	3014
30	2306	80	3038
31	2354	81	3062
32	2402	82	3086
33	2450	83	3110
34	2498	84	3134
35	2546	85	3158
36	2594	86	3182
37	2642	87	3206
38	2690	88	3230
39	2738	89	3254
40	2786	90	3278
41	2834	91	3302
42	2882	92	3326
43	2858	93	3350
44	3006	94	3374
45	3020	95	3398
46	3044	96	3422
47	3068	97	3446

Table 1 *(continued)*

Ensemble member	Branch time	Ensemble member	Branch time
48	3092	98	3470
49	3116	99	3494
50	3140	100	3518

1.8° and an ocean of approximately 1.5° resolution, although the ocean resolution increases closer to the poles in the grid.

MPI-ESM1.1 has some similarities to MPI-ESM used in CMIP5 (Giorgetta et al., 2013), but overall behaves closer to MPI-ESM1.2 (Mauritsen et al., 2019), which is used in CMIP6. The ocean component is very similar to the CMIP5 version, with some minor differences. The atmosphere is based on ECHAM6.3, rather than ECHAM6.1 as was used in CMIP5. MPI-ESM1.1 and MPI-ESM1.2 have specifically tuned cloud feedbacks to better match the historical warming. The equilibrium climate sensitivity is hence lowered from 3.4 K in CMIP5 to 2.8 K in MPI-GE, as calculated using linear extrapolation from 150 years of an abrupt 4 xCO_2 experiment (Andrews et al., 2012). HAMOCC is run in the same configuration as CMIP5, and JSBACH is the CMIP5 version of the model component, however now including the soil carbon model YASSO (Goll et al., 2015) and a five-layer soil hydrology scheme (Hagemann & Stacke, 2014).

2.2. Initialization and Forcing

MPI-GE follows the protocol of the CMIP5 simulations (Taylor et al., 2012). The historical and idealized forcing simulations are branched from different years of the preindustrial control simulation after it has reached a state of quasi-stationarity. Both historical and 1% CO_2 simulations are initialized from the state on the first of January in different years of the control simulation (Table 1) and thus sample differences in the possible state of the atmosphere, land, and ocean assuming a stationary and volcano-free 1850 climate.

2.3. Simulations and Data Availability

Monthly mean data are available for all components except the ocean biogeochemistry. Ocean biogeochemical data are available as monthly mean surface variables and annual mean three-dimensional variables, except for the RCP8.5 scenario where the three-dimensional variables are also available as monthly means. The deep ocean biogeochemistry variables in the first 500 years of the preindustrial control simulation are prone to model drift. Hence, in ensemble members branched from an early state of the control simulation (Table 1), this model drift could introduce spurious trends on top of the internal variability and forced response. However, the magnitude of the model drift is much smaller than both the internal variability and forced response, at least for the CO_2 flux and subsurface (upper few hundred meters) dissolved oxygen concentrations. Additionally, we note that the carbon cycle in the land component is not fully equilibrated early in the preindustrial control. As such, for analysis of the carbon cycle on land and any variables affected by it, ensemble members 1–18 and 21–39 in both the historical and 1% CO_2 scenario should not be used without drift removal. As with many other climate simulations, deep ocean temperature drift also occurs. Details on drift removal and the best current methods to perform these calculations can be found in Gupta Sen et al. (2013). We emphasize that for drift removal calculations using this method, smoothing must be applied to the control simulation. This is because subtracting the control directly from the forced simulations would initially artificially dampen the anomalies due to internal variability, making all members more similar and deflating the variability across the ensemble dimension. Later in the simulation, a lack of coherence between the preindustrial control and the ensemble members would artificially inflate the variability.

MPI-GE is an evolving ensemble, which will include additional experiments in the future. Additional simulations to add more monthly mean biogeochemistry data and high-frequency atmosphere output are currently being undertaken. Currently, the following simulations are available for the ensemble, with 100 ensemble members completed for every simulation besides the preindustrial control:

- Preindustrial control simulation (2,000 years);
- Historical (1850–2005);
- RCP2.6 (2006–2099);
- RCP4.5 (2006–2099);
- RCP8.5 (2006–2099); and
- 1% CO_2 (150 years).

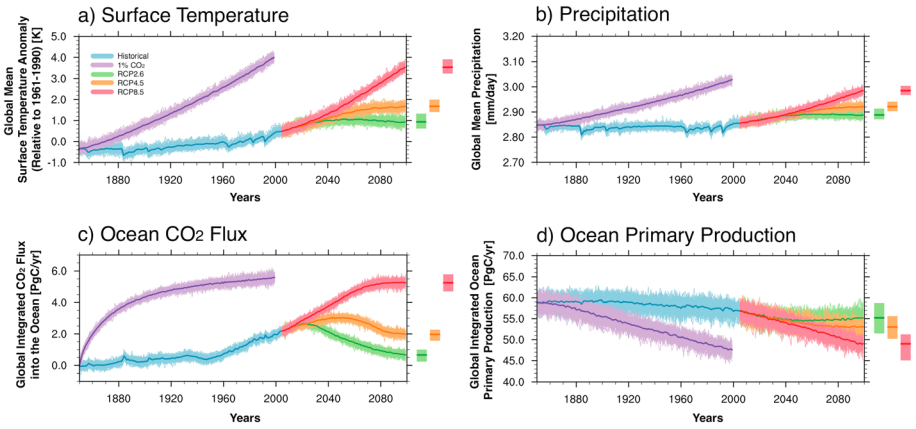

Figure 1. Ensemble spread (thin lines), ensemble mean (thick lines), and projected range in 2099 (bars) for the historical simulation (blue), 1% CO_2 scenario (purple), RCP2.6 (green), RCP4.5 (orange), and RCP8.5 (red). Shown for global mean (a) surface temperature anomalies (relative to 1961–1990) and (b) precipitation and globally integrated (c) CO_2 flux into the ocean and (d) primary production in the ocean. RCP = representative concentration pathway.

Details of how to download the data can be found on the website (https://www.mpimet.mpg.de/en/grand-ensemble).

3. Quantifying the Transient Forced Response and Evolving Internal Variability

MPI-GE and other large ensembles can be used to quantify the forced response at each time step as well as the internal variability of the climate system and how it evolves in time. We estimate the transient forced response (F_t) by taking the ensemble mean at each time step:

$$F_t = \frac{\sum_{e=1}^{e=100} f_{et}}{100}, \tag{1}$$

where f_{et} is a single ensemble member with ensemble number e at time step t. This estimation of the forced response can be seen in the solid lines in Figure 1. To robustly estimate the internal variability (IV_t) at time t, we then remove the ensemble mean (F_t) from each ensemble member (f_{et}) and calculate the variability (here as the standard deviation) across the ensemble:

$$IV_t = std(f_{et} - F_t). \tag{2}$$

This method allows us to estimate transient internal variability. The spread of the realizations shown in Figure 1 demonstrates the variability of the ensemble around the forced response (F_t). While the parametric method of using the ensemble standard deviation is used in this paper, a non-parametric method or alternate parametric method to estimate variability can also be used in place of the standard deviation.

The superposition of internal variability and externally forced changes in the climate due to external drivers such as anthropogenic emissions and volcanic eruptions is illustrated in Figure 1. In the historical simulation, the long-term trend of each quantity over time can be attributed to anthropogenic forcing (Bindoff et al., 2013), with an increase in global mean surface temperature (GMST) and net carbon dioxide (CO_2) flux into the ocean, a small decrease in primary production in the ocean, and little change in precipitation. The forced response to external volcanic forcing is also clear, with decreases in GMST and precipitation, an increase in the CO_2 flux, and little response in the ocean primary production seen just after large tropical eruptions (e.g., Segschneider et al., 2013).

In all four quantities, the 1% CO_2 scenario quickly distinguishes itself from the historical simulation, with the strong forcing causing large changes in all quantities. Different quantities evolve in different ways. Precipitation demonstrates an almost linear increase over time and primary production an almost linear decrease, while GMST shows a slightly stronger increase at the end of the time series compared to the beginning. The ocean CO_2 flux, however, shows the strongest increase right at the beginning of the 1% CO_2 forcing scenario and begins to plateau near the end of the scenario as a consequence of ocean warming, increased thermal

stratification, and a slower AMOC. There is also more internal variability shown in the CO_2 flux at the end (ensemble mean variability ≈ 0.21 PgC/year in the last year), compared to the beginning of the 1% CO_2 scenario (ensemble mean variability ≈ 0.15 PgC/year in the last year). When considering the strongest forcing, from the 1% CO_2 scenario, we also find a decrease of variability of the primary production (from ≈1.35 to ≈1.1 PgC/year) and an increase in global mean precipitation variability (from ≈0.0085 to ≈0.0095 mm/day).

The role of the forced response and internal variability in determining how the future may look is demonstrated in the scenarios. For GMST, all forcing scenarios are distinct at the end of the century, with the highest emission scenario (RCP8.5) showing the most warming. This distinct response between the three scenarios is also the case for the CO_2 flux into the ocean. However, we also see that a plateau in the increase in CO_2 flux occurs in the strongest scenario (RCP8.5), and a decrease in the flux compared to the beginning of the forcing scenario occurs in both lower emission scenarios (RCP2.6 and RCP4.5).

Global mean precipitation increases with warming, but the spread of realizations overlaps in the two weaker scenarios until the end of the 21st century. Here the distinction between the forced response in RCP2.6 compared to RCP4.5 is smaller than the internal variability. This means that single realizations from both scenarios could show similar precipitation in any given year. This overlap is even larger in the ocean primary production, where single realizations from all three scenarios could exhibit the same primary production at the end of the century. We note that time averages may also be used to distinguish between the scenarios, but this strong overlap tells us that the primary production will take longer to distinguish itself between scenarios, than other variables such as GMST. It is important for future projections to quantify both the forced response and the role of internal variability, because even though mitigation might happen in the future For example, RCP2.6 could still look similar to a high warming future (depending on the strength of the internal variability; Marotzke, 2019; Suarez-Gutierrez et al., 2018).

We note that for all quantities presented in this paper, we rely on the ability of MPI-GE to adequately represent the real world, but multiple models are needed to assess uncertainties due to model differences. The superposition of internal variability and the forced response in a single climate realization can cause confusion as to whether trends seen in one realization or an observational change are due to internal variability or can be characterized as a forced response (Hasselmann, 1976; Hawkins & Sutton, 2009). The use of MPI-GE allows us to disentangle these quantities in the model. In the following section, we demonstrate how to evaluate the performance of MPI-GE with observations for the current climate.

4. Comparison to Observations

In this section we use observations and MPI-GE to both evaluate the model and to interpret observations in the context of the model's internal variability. It is not appropriate to expect observations to match either a single realization or the ensemble mean; however, observations can be put in the context of the model's ensemble mean and variability.

Three methods of evaluating the ensemble are demonstrated. The first method is ideal for quantities that have good observational coverage over a long time span. This method puts the observations in the context of the transient model spread in time (e.g., Bengtsson & Hodges, 2018; Marotzke & Forster, 2015; Risbey et al., 2014) and additionally uses a rank histogram to evaluate the spread in the ensemble dimension (e.g., Marotzke & Forster, 2015). The second method can be used for observations where there is high quality data available but only for a short period of time. Here we compare a single observational estimate with a histogram of the ensemble spread (e.g., Bengtsson & Hodges, 2018; Flato et al., 2013). The third method provides a novel way to assess the agreement of the model's internal variability and forced response with observations on a global map.

To demonstrate the first method, we use GMST. Observed GMST largely falls within the MPI-GE spread, with some observational values sitting on the edges of the model spread (Figure 2a). Due to internal variability, we expect observational GMST to occur everywhere across the ensemble with uniform frequency. We also expect observational GMST to occasionally sit outside of the ensemble by chance. To test this, we perform a rank histogram to evaluate the model (Figure 2b). The rank histogram indicates with which frequency observations occur across the ensemble. For each year, the rank represents the place that the observations would take in a list of ensemble members ordered by ascending GMST values. If the observed value is smaller than all ensemble members, the rank is 1. If the observed value is higher than all ensemble members, the rank

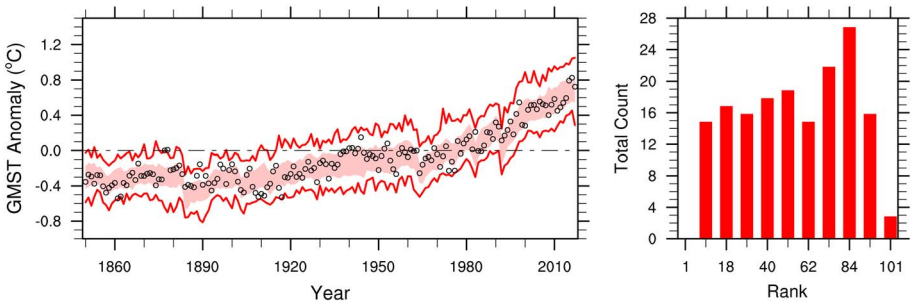

Figure 2. (a) Time series of annual GMST anomalies (relative to 1961–1990) simulated by the Max Planck Institute Grand Ensemble in the historical and representative concentration pathway 4.5 simulations (red) and GMST HadCRUT4 observations (Morice et al., 2012; black circles), shown for the period 1850–2017. The red shading represents the ensemble spread within the 12.5th to 87.5th percentile range, while the red lines represent the ensemble maximum and minimum. (b) Rank histogram of the total count of each rank (binned for 11 members centered on the rank indicated, note that bins 1 and 101 are single bins for this purpose) for HadCRUT4 GMST observations shown as a member of the Max Planck Institute Grand Ensemble. GMST = global mean surface temperature.

is $n + 1$, with n the number of members (here 100). For a large enough record, one would expect that, if variability is perfectly simulated, observations take all ranks with no preferred frequency. That would lead to a "flat" rank histogram. The relatively flat rank histogram (Figure 2b) again demonstrates that MPI-GE is performing well in simulating GMST variability. We find some bias toward higher ranks, suggesting that the observations occur more often in the upper part of the model spread. Additionally, we would expect observations to occasionally sit at ranks 1 and 101 by chance. For a model that underestimates variability, there will be many occurrences at these ranks. For MPI-GE GMST, we find no occurrences at rank 1 but a few occurrences at rank 101, which might be related to a too-strong cooling associated with volcanic eruptions. Overall, MPI-GE performs well for GMST.

To demonstrate the second method, we investigate the monthly variability of the net outgoing longwave, absorbed shortwave, and net top of the atmosphere irradiances, which have far fewer observed data points than GMST. Here the most robust estimate is a single observational estimate of the variability, based on the Clouds and the Earth's Radiant Energy System Energy Balanced and Filled (Loeb et al., 2018) top of the atmosphere irradiance product, which only covers the period 2000–2015. While a previous study claims that climate models do not represent the variability of these fluxes well (Stephens et al., 2015), Figure 3 shows that the observations are within the model spread, and hence, the model is consistent with observations.

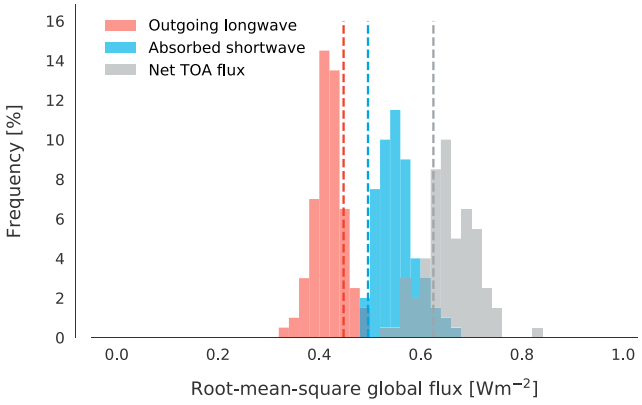

Figure 3. Monthly variability in TOA fluxes: net TOA flux (gray), outgoing longwave radiation (red), and absorbed shortwave radiation (blue), shown as a histogram of root-mean-square monthly means for the Max Planck Institute Grand Ensemble and as dotted lines for observations (Clouds and the Earth's Radiant Energy System Energy Balanced and Filled-TOA Ed2.8, these data were obtained from the NASA Langley Research Center Atmospheric Science Data Center). Both the Max Planck Institute Grand Ensemble and the observations are shown for the period 2000–2015 (RCP4.5 is used for the extension of the historical simulation from 2006 to 2015). This figure is redrafted following Hedemann et al. (2017) with minor modifications.. TOA = top of the atmosphere.

We demonstrate the third method of comparison in Figure 4. This is a novel model evaluation method, where we transfer the methodology used on European temperature (Suarez-Gutierrez et al., 2018, Supplementary Figure 3) to the globe. While previous methods to investigate this have used standard deviations and often detrended quantities (Bengtsson & Hodges, 2018; Lehner et al., 2017; McKinnon et al., 2017), this new method allows quantification of whether the whole distribution, including the extremes, agree well with observations. Additionally, by combining the map in Figure 4 with the method from Figure 2, we can investigate exactly why the model does not agree with observations in specific regions identified on the map and can differentiate between discrepancies in internal variability and the forced response.

We use surface temperature for this evaluation due to its long observational record and near global coverage (Figure 4). We first identify where observations lie outside the ensemble, using blue shading to show where observations lie below the ensemble minimum and red where they lie above the ensemble maximum. We then determine where the

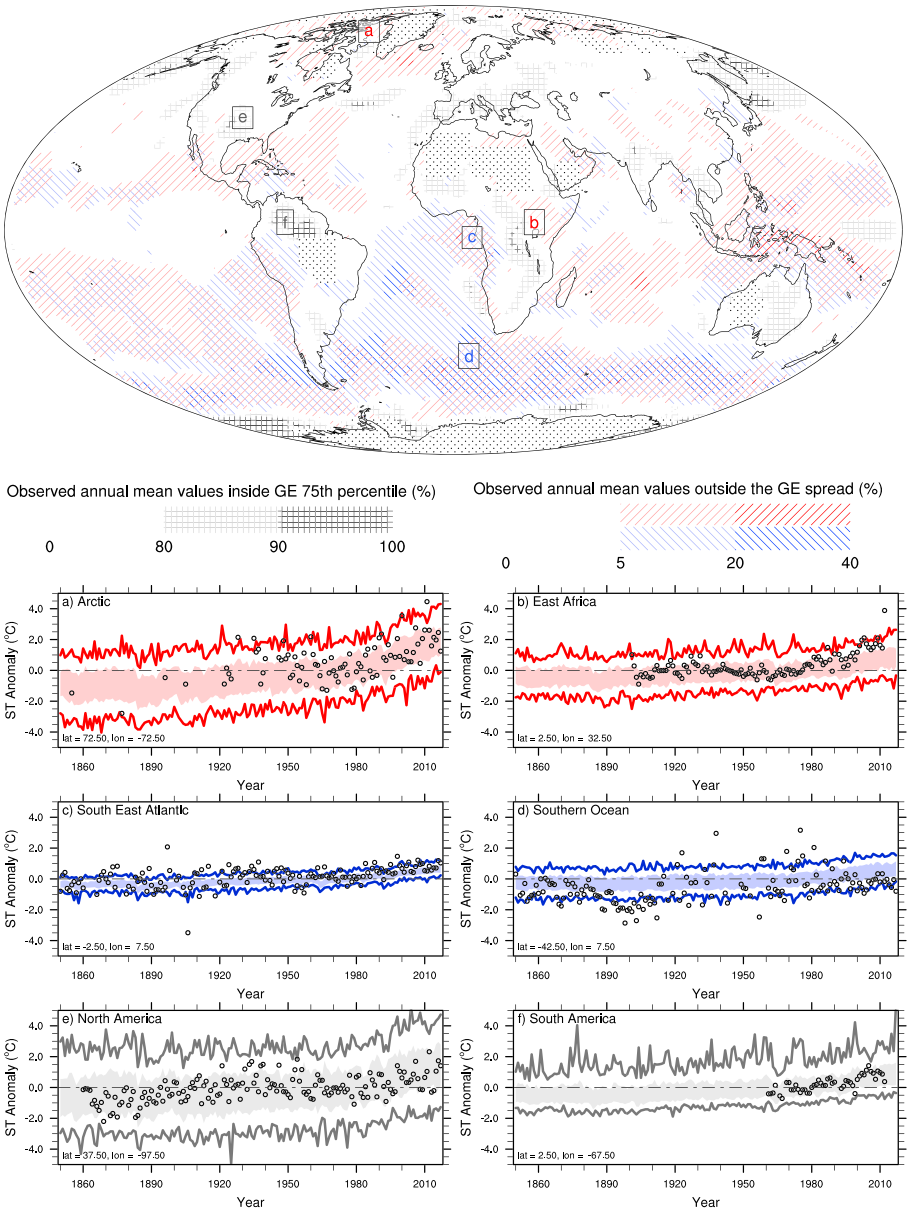

Figure 4. Surface temperature variability (1850–2017) in MPI-GE (historical and representative concentration pathway 4.5) versus observations. Global map representing the frequency of HadCRUT4 (Morice et al., 2012) annually averaged surface temperature observations occurring outside the limits of the MPI-GE spread. Red shading represents regions where observed anomalies are larger than the ensemble maximum, while blue shading represents regions where observed anomalies are smaller than the ensemble minimum. Hatching indicates the percentage of time that the observed annual mean values fall inside the 12.5 to 87.5 percentile range of MPI-GE. Dotted regions represent regions where no observations are available or where observed estimates of surface temperatures are available for less than 10 years. Surface temperature is represented by 2-m air temperature over land grid cells and by sea surface temperature over the oceans. Anomalies are calculated with respect to the climatology baseline, defined by the period 1961–1990. Time series are shown for (a) Arctic, (b) East Africa, (c) South East Atlantic, (d) Southern Ocean, (e) North America, and (f) South America. The shading represents the ensemble spread within the 12.5th to 87.5th percentile range, while the solid lines represent the ensemble maximum and minimum, and the black circles are the observations. The time series are specific points on the observed grid as stated in the bottom left of each panel. For this analysis, model data were regridded to match the observational grid (approximately 5°). MPI-GE = Max Planck Institute Grand Ensemble.

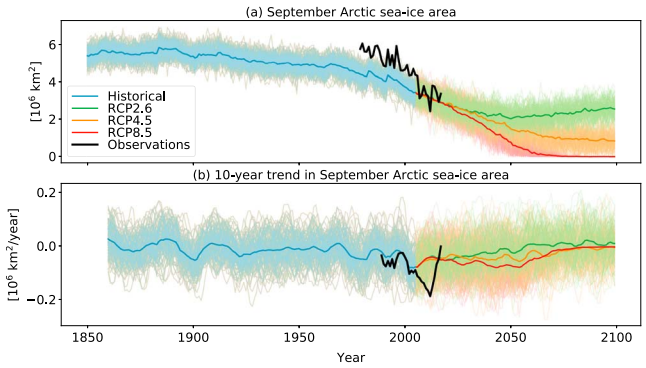

Figure 5. a) Evolution of September Arctic sea-ice area and (b) evolution of 10-year linear trends of September Arctic sea ice area, plotted at the end point of the 10-year period. Shown for the historical simulation (blue), RCP2.6 (green), RCP4.5 (orange), RCP8.5 (red), and observations (black). Observations are based on the Arctic sea ice index from Fetterer et al. (2017). RCP = representative concentration pathway.

observations crowd too much in the center of the ensemble by using hatching to show where the observations sit inside the 75th percentile range (12.5 to 87.5 percentile) more than 75% of the time. Crowding too much in the center of the ensemble indicates that internal variability is overestimated in the model. White regions with no hatching or shading are regions where the variability is similar in the observations and the model. The Northern Hemisphere Oceans and the European continents observed variability are well represented in the model. However, we find that in general, the Southern Ocean has too low variability in the model, and parts of the land surface and the ice edge show too high variability.

To delve into specific regions identified as having biases, we can use time series plots similar to Figure 2. Six gridpoints are highlighted. We find that the Arctic point that shows observations lying above but not below the ensemble spread does so because MPI-GE captures cold extremes but not warm extremes accurately (Figure 4a). The East African land point also exhibits a similar bias, but does so because the observed trend over the period at the end of the time series is not completely captured by MPI-GE (Figure 4b). The South American land point shows too high variability in MPI-GE, presenting a distribution with too many warm extremes (Figure 4f), whereas the North American land point shows too strong variability in both warm and cold extremes (Figure 4e). The ocean off the west coast of Africa exhibits too small variability in MPI-GE (Figure 4c). The Southern Ocean also shows too small variability (Figure 4d). This is likely due to the low model resolution and the lack of eddies (e.g., Screen et al., 2009). Poor observational coverage may also contribute to differences between the observed GMST and the ensemble in this region. Overall, this method gives us a means to evaluate model variability and consequently hypothesize why it is overestimated or underestimated in various regions.

As well as using observations to evaluate the models performance, the combination of a large ensemble and observations can be used to put observations into the context of the model's internal variability. As previously mentioned, decadal trends can be misinterpreted due to lack of understanding about internal variability (Marotzke & Forster, 2015). Large-scale sea ice loss, a prominent indication of climate change (e.g., Notz & Marotzke, 2012), is on short time scales strongly influenced by internal variability, (see Figure 5a) with a large effect even on decadal trends (Notz, 2015; Swart et al., 2015). To predict sea ice loss on interannual to decadal time scales, internal variability as well as the forced response of sea ice to greenhouse gas increases must be understood.

To exemplify this point, we examine decadal trends in September Arctic sea-ice area (Figure 5b) as computed from the annual data (Figure 5a). The decadal trend of the ensemble mean is to a substantial degree a direct reflection of changes in the external forcing, primarily increasing greenhouse gas concentration and a few

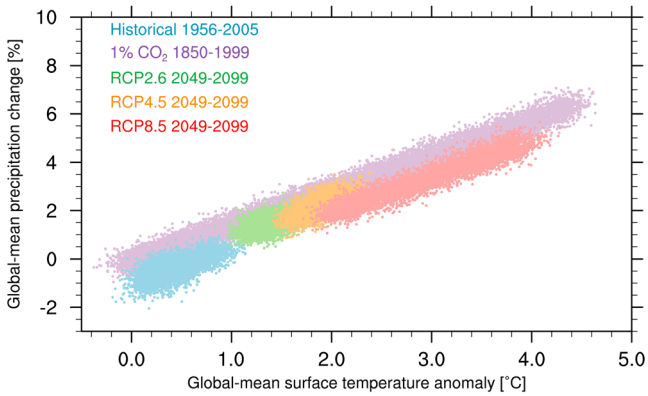

Figure 6. Global-mean annual surface temperature anomaly (relative to the preindustrial control; °C) plotted against global-mean precipitation change (%) for the 1% CO_2 simulation (purple) and the last 50 years of the historical simulation (blue), RCP2.6 scenario (green), RCP4.5 scenario (orange), and RCP8.5 scenario (red). RCP = representative concentration pathway.

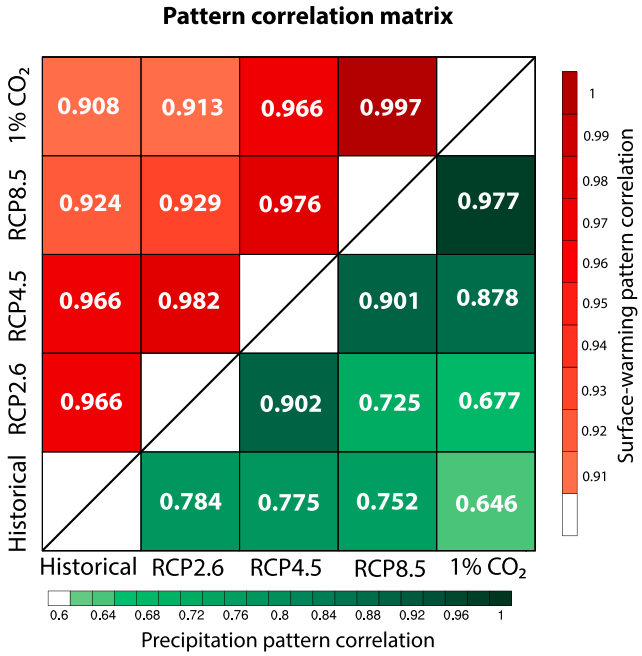

Figure 7. Correlation between the ensemble mean scaled temperature change (red) and ensemble mean scaled precipitation change (green) patterns for the different model simulations for the last 10 years of each simulation. For each grid point and each year, the temperature and precipitation change for each ensemble member is scaled by the global mean temperature. A global ensemble mean of the 10 years is then taken and used for the correlation. RCP = representative concentration pathway.

volcanic eruptions. The response of individual ensemble members, in contrast, shows a very clear impact of internal variability, which overshadows the impact of the external forcing on these short time scales. The same is true for the observational record, whose decadal trends fluctuate initially around the ensemble mean, but then deviate because internal variability causes rapid ice loss primarily in the summer of 2007 and the summer of 2012. More recently, the observed trend again recovered to the ensemble mean trend. For more information, see Notz (2017) and Olonscheck et al. (2019).

5. The Power of MPI-GE

In this section we present four novel analyses, which utilize the power of MPI-GE. The first application demonstrates the pathway dependence of future changes, a question that can only currently be answered using MPI-GE. The second analysis shows how a very large ensemble can be used to identify forced changes in the atmospheric circulation that are difficult to observe due to high internal variability. The third analysis utilizes the ensemble dimension to determine whether projected changes in AMOC variability are linear in time. Finally, because MPI-GE is currently the largest ensemble available, we demonstrate how it can be used to determine the ensemble size needed when investigating SLP trends.

5.1. The Value of Multiple Scenarios

MPI-GE is a unique large ensemble in that it can be used to compare multiple scenarios. Previously, Giorgetta et al. (2013) showed how to compare projected warming under different scenarios in MPI-ESM (low-resolution configuration), by scaling the warming (taken in comparison to the preindustrial control) by the global mean warming value. By doing this, Giorgetta et al. (2013) concluded that the warming pattern was generally consistent between the historical, RCP2.6, RCP4.5, and RCP8.5 scenarios, with the absolute magnitude of the warming dependent on the scenario. While they were able to conclude that the patterns were similar, they were unable to quantify the role of internal variability in the comparison. We extend this analysis to precipitation as well as temperature and use MPI-GE to both quantify the differences between the scenarios and compare this to the magnitude of the internal variability.

We plot the global mean precipitation change versus the global mean temperature change for the end (last 50 years) of the historical simulation and three future scenarios for different global annual mean surface temperature anomalies, compared to the change for every year of the 1% CO_2 simulation (Figure 6). We find that the relationship between temperature and precipitation is not completely consistent between scenarios, indicating a pathway dependence in the precipitation response.

Additionally, we correlate the scaled ensemble mean temperature and precipitation change patterns from the last 10 years of each scenario with the temperature and precipitation change patterns from the last 10 years of each other scenario (Figure 7). There is a high correlation between the temperature patterns for all simulations, demonstrating that similar information can be found for temperature without running all scenarios, adding strength to the qualitative description by Giorgetta et al. (2013). There is lower correlation between the simulations for precipitation, showing that in this case we get different information by running different scenarios, again pointing to a pathway dependence of the precipitation response. This is likely due to the differing aerosol forcing between scenarios, as projected precipitation changes have previously been linked to aerosol forcing (e.g., Lin et al., 2016, 2018; Pendergrass et al., 2015).

To determine where the pathway dependence of precipitation is most important, we compare the scaled precipitation patterns from the last 10 years of the strong warming (RCP8.5) and weak warming (RCP2.6) scenarios (Figure 8). We find that the differences are largest in the Pacific Ocean, Eastern South America, and the North Atlantic. The standard deviation of the pattern difference (Figure 8d) is used to investigate whether the pattern differences are greater than the noise from internal variability. By subtracting the internal variability from the absolute value of the ensemble mean, we can specifically identify regions where this

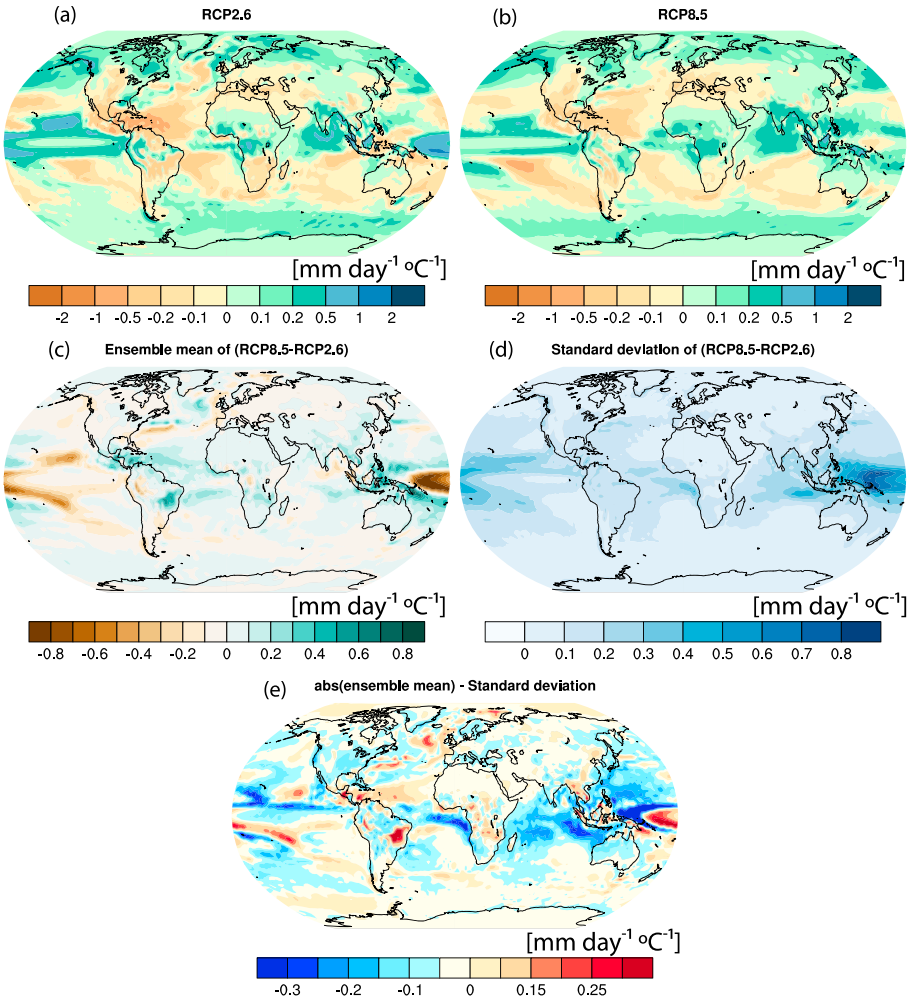

Figure 8. Ensemble mean precipitation change in the last 10 years of the (a) RCP2.6 and (b) RCP8.5 scenarios scaled by the warming over the same time period as compared to the preindustrial control simulation (mm·day^{-1} · °C^{-1}). Here each grid point, for each year, for each ensemble member is scaled by the mean warming for this year, and then the ensemble time mean is computed. (c) The difference between the scaled RCP2.6 and RCP8.5 patterns (a,b). (d) Standard deviation of the difference pattern shown in (c) across the ensemble. Here the difference between each corresponding ensemble member is taken, and then a standard deviation across the ensemble is computed. (e) The difference between the absolute value of the difference between the scenarios and the standard deviation (abs(c), d). RCP = representative concentration pathway.

signal is larger than the inter-member variability (Figure 8e; red regions). We find that the western Pacific Ocean, eastern South America, and parts of the African and southern Asian land masses show differences between the scenarios that are larger than the inter-member variability, indicating that the scenario differences matter in these regions. This shows that multiple scenarios are beneficial, but emphasizes that in some regions, they are not necessary.

5.2. Identifying the Forced Response Under High Variability

Due to the high variability of the atmospheric circulation, projected changes of the tropospheric eddy-driven jets and of the stratospheric polar vortex are highly uncertain and have traditionally been made over longer periods (30- to 50-year averages; Barnes & Polvani, 2013; Manzini et al., 2014; Simpson et al., 2018). Future projections on these time scales have shown a weakening of the stratospheric vortex under anthropogenic warming (Manzini et al., 2014; Simpson et al., 2018). However, this weakening, along with a strengthening of the Brewer-Dobson circulation, may already be occurring, but may be masked by internal variability (Fu et al., 2015).

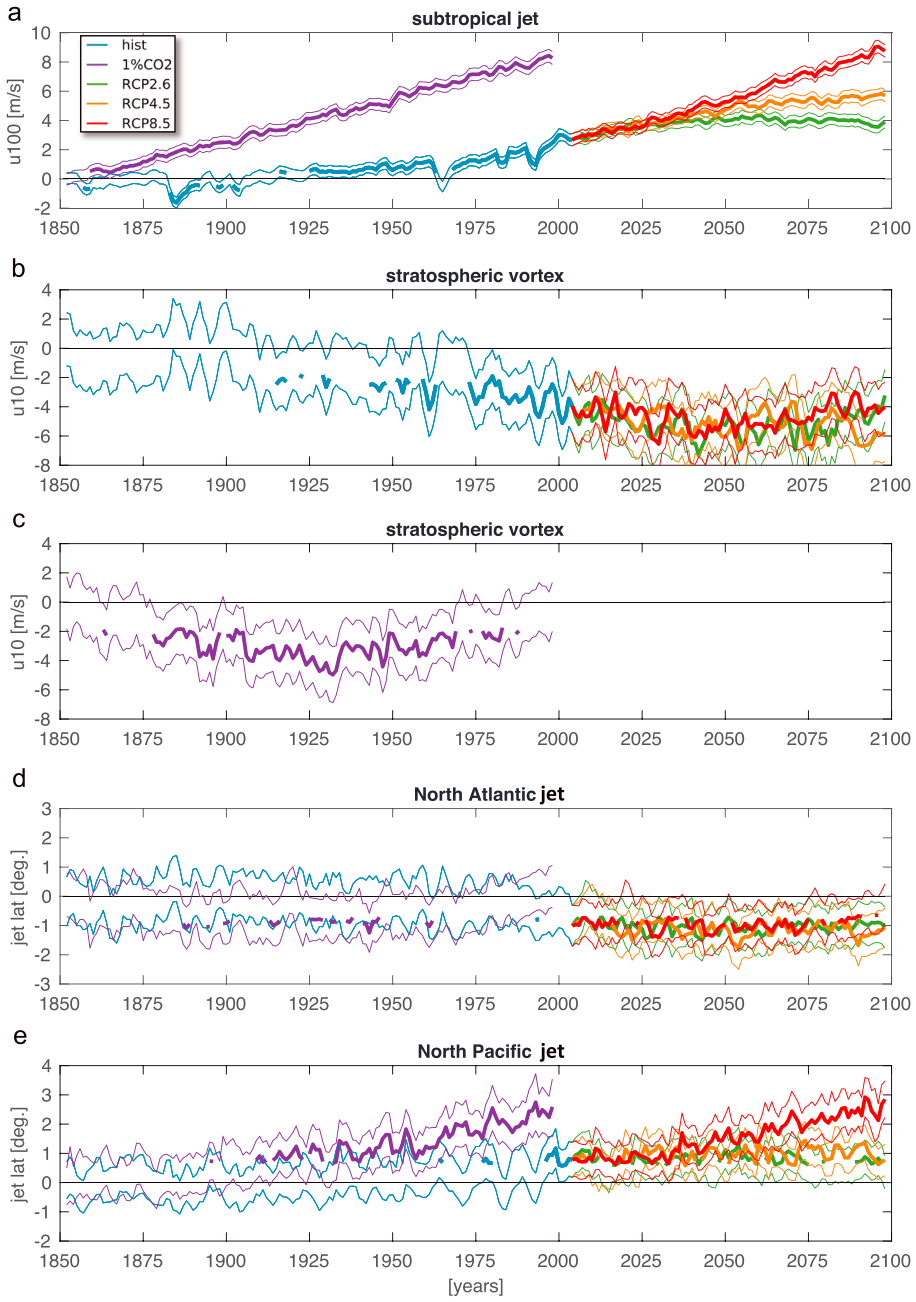

Figure 9. Atmospheric circulation changes in January of the (a) subtropical tropospheric jet (m/s), 20–40°N zonal average, zonal wind at 100 hPa; (b,c) stratospheric polar vortex (m/s), 70–80°N zonal average, zonal wind at 10 hPa; and (e,f) latitude of the maximum in tropospheric eddy-driven jet (degrees), from the 850- to 700-hPa zonal wind for the North Atlantic (15–70°N, 60°W–0°) and North Pacific (15–70°N, 135°E–125°W) sectors, respectively. Prior to any calculation, interannual variability is reduced by applying a 3-year running mean to each member time series. The ensemble mean change is shown in bold where it is significant with respect to the ensemble mean at 1850. The 95% confidence intervals (shown as thin solid lines) are calculated using the two-sample *t* test. Results are shown for the historical simulation (blue), RCP2.6 (green), RCP4.5 (orange), RCP8.5 (red), and 1% CO_2 simulation (purple). RCP = representative concentration pathway.

We take advantage of MPI-GE to estimate the historical and scenario evolution in time of changes in circulation indices representing key aspects of the subtropical jet, the stratospheric vortex, and the eddy-driven jet position (Figure 9). To this end, we calculate time series of yearly differences in the ensemble means of the considered circulation indices minus their ensemble means at 1850. In so doing, we can identify future circulation changes with respect to the preindustrial state. To quantify the forced response, a 3-year running mean is applied to each ensemble member (to reduce the high internal variability) before estimating the forced response. These questions are different from asking if a change has occurred during the time period for which we have reanalysis (≈1955 onward). Indeed, we cannot directly compare the modeled changes to observations because we are computing yearly changes as deviations of ensemble mean quantities from a mean reference state at each point in time (the ensemble mean at a specific year). This procedure cannot be reproduced with a single observational time series due to the high interannual variability of the circulation indices.

The changes in the subtropical jet (zonal-mean zonal wind change at 100 hPa averaged over 20–40°N) are significant at the 95% level for all scenarios and after 1925 in the historical simulation (Figure 9a), with distinct divergence of the three projection scenarios occurring by 2075. The changes in the stratospheric vortex (zonal-mean zonal wind change at 10 hPa averaged over 70–80°N) are more complex (Figures 9b and 9c). Although all scenarios indicate a weakening of the vortex, the behavior is nonlinear, as can be particularly seen in the 1% CO_2 scenario, where the vortex is first decreasing and then increasing in the second half of the scenario, despite the monotonous increase in GMST/atmospheric CO_2. (Figure 9c). This nonlinear behavior is described in more detail in Manzini et al. (2018). This behavior means that the mean state is very similar under both low and high radiative forcing (e.g., beginning and end of the 1% CO_2 scenario). The tendency for a weaker stratospheric vortex during present day indicates that the strengthening of the Brewer-Dobson circulation could indeed be already underway, with respect to preindustrial conditions. Although within the period for which we can compare to observations (1980–2016), MPI-GE shows a trend in the forced change, this trend is weak compared to the confidence intervals, consistent with other large ensemble estimates (Seviour, 2017).

In contrast to the projected changes of the subtropical jet, the projected changes of the stratospheric vortex (Figure 9b) and the tropospheric eddy-driven jets (Figures 9d and 9e) cannot be distinguished between the three scenarios, due to high internal variability. In addition, contrasting influences in different regions lead to different shifts of the latitude of the tropospheric eddy-driven jets. In the North Atlantic, the projected equatorward shifts (Figure 9d) suggest that the stratospheric influence dominates, given that the stratospheric vortex weakens (Kidston et al., 2015). These projected changes are significant with respect to preindustrial conditions. While the changes are significant in this case, it is likely that if compared to present day, they would no longer be significant, as suggested by other large ensemble estimates (Kwon et al., 2018). In the North Pacific, the poleward shift (Figure 9e) instead indicates that the tropospheric response (Figure 9a) dominates in this region with little influence of the stratosphere. Similarly, the lack of change in the North Atlantic jet for the 1% CO_2 simulation is likely due to opposing stratospheric and tropospheric changes, while the North Pacific projection for this simulation clearly branches off and evolves similar to the RCP8.5 projection scenario. MPI-GE therefore clearly illustrates that different processes dominate the forced response of these jets to climate change in different regions.

5.3. Are Changes in Variability Time Dependent?

With projections of a decreasing AMOC strength under anthropogenic warming (e.g., Collins et al., 2013), the North American and European climates are expected to significantly alter (e.g., Sutton & Hodson, 2005). While a weakening of the AMOC is associated with stronger global warming, internal variability also plays an important role in driving the climate response (Maroon et al., 2018). In a large ensemble, ensemble members with a stronger AMOC due to internal variability are likely to show increased surface warming, compared to ensemble members with a weaker AMOC (Maroon et al., 2018). By utilizing the ensemble dimension (similar to Herein et al., 2017; Maher et al., 2018) and computing the standard deviation of any given quantity across the ensemble, we can assess whether projected changes in AMOC variability are time-dependent.

MPI-GE captures the observed AMOC reasonably well, with the short observed time series largely fitting within the MPI-GE ensemble spread (Figure 10a). Similar to previous studies (e.g., Collins et al., 2013), MPI-GE projects a decrease (from 19 to 14 Sv) in the AMOC strength under strong future warming scenarios

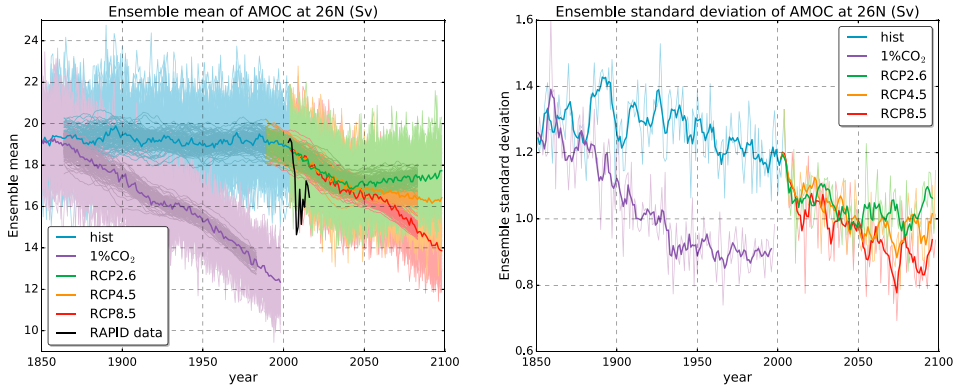

Figure 10. (a) The AMOC ensemble mean (solid line), ensemble spread (pale shaded region), and 30-year running mean (darker individual lines) calculated at 26°N for each year (Sv). Observations from RAPID are shown in black (https://www.rapid.ac.uk/data.php). (b) Standard deviation of AMOC (pale line, 5-year running mean is shown in the solid line) across the ensemble for each year (Sv). Shown for the historical simulation (blue), RCP2.6 (green), RCP4.5 (orange), RCP8.5 (red), and 1% CO_2 simulation (purple). AMOC = Atlantic Meridional Overturning Circulation; RCP = representative concentration pathway.

(Figure 10a). The CMIP5 models show a weakening of the AMOC in RCP8.5 and a weakening in RCP4.5 in the first half of the 21st century with a recovery thereafter (Cheng et al., 2013), similar to the forced response found in MPI-GE. The forced response of the AMOC in CESM-LE at 26.5°N has also been investigated and is shown in Figure 1 of Maroon et al. (2018). There are some differences between the two large ensembles. MPI-GE has a more realistic AMOC strength in the historical period, whereas in CESM-LE, it is somewhat too strong. The forced response in the historical period shows changes in CESM-LE that do not exist in MPI-GE, and the recovery in the AMOC in RCP4.5 only occurs in MPI-GE. However, both models show overall similar trends in the AMOC forced response, with a weakening in both RCP4.5 and RCP8.5 scenarios.

CMIP5 projections suggest that under global warming, the internal variability of the AMOC will decrease (Cheng et al., 2016). The CMIP5 projections were completed by comparing the period 2100–2300 in each future forcing scenario to the preindustrial control. Using MPI-GE, we can determine the time dependence of the projected variability change. The 1% CO_2 run shows a 30% drop in the internal variability in the first 80 years of the simulation, stabilizing around 0.9 Sv. All three RCP scenarios show a drop in variability after 2000 and a stabilization after 2050, with the RCP2.6 stabilizing at the highest variability and RCP8.5 stabilizing at the lowest variability, with a similar stabilization value to the 1% CO_2 run. By using MPI-GE, we can clearly demonstrate that the changes in projected AMOC variability have a high time dependence.

When considering the interplay of the forced response and variability, we can see that even though the forced response of the AMOC varies in each of the future scenarios, all three scenarios could have the same AMOC at any given year due to internal variability. When considering the 30-year running mean, all three scenarios are very similar until 2050. By 2100, the scenarios have diverged; however, there is still overlap of the 30-year means of RCP2.6 and RCP4.5, and even RCP4.5 and RCP8.5 have a few ensemble members that could have a similar 30-year mean (Figure 10a). This demonstrates how internal variability is important in determining possible observable futures in a single-forcing scenario.

5.4. Assessing the Required Ensemble Size

Deser et al. (2012) have previously used the 40-member Community Climate System Model 3 ensemble with SRES A1B forcing to ask the question of whether the forced response can be estimated with fewer than 40 members. When considering SLP trends from 2005 to 2060, they find a wide variety of SLP responses in individual realizations across the 40 members and argue that this demonstrates the need for an ensemble of size 20–30 members to accurately quantify the forced response. Given that this estimate of necessary ensemble size is close to the actual ensemble size of 40 members, a different answer may be found with a larger ensemble. We use the 100 members from MPI-GE to investigate this question and determine whether this result is robust when a larger ensemble is used. The SLP trend and variability patterns from MPI-GE (Figures 11a and 11b) over the period 2007–2099 in RCP4.5 are qualitatively similar to the results of Deser et al. (2012), indicating that both models have a similar forced response and variability of the future trend in SLP and that we can use MPI-GE to build on the results of Deser et al. (2012).

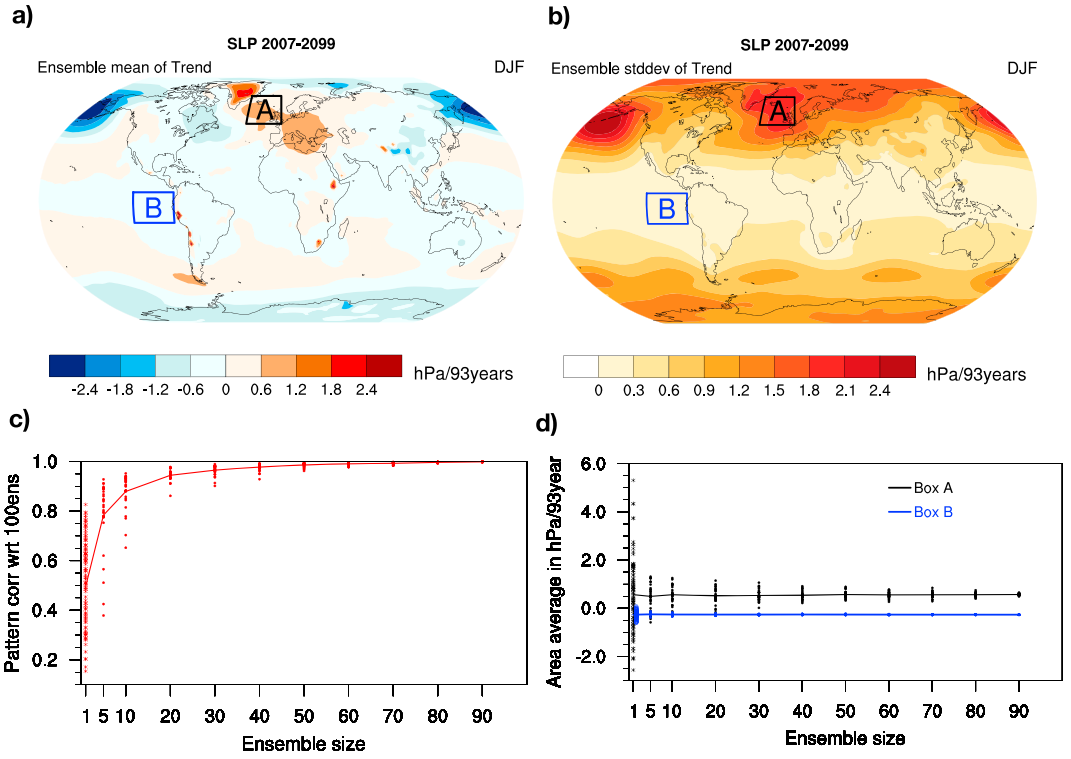

Figure 11. The Max Planck Institute Grand Ensemble (a) mean of the SLP trend and (b) standard deviation of the SLP trend for the period 2007–2099 from representative concentration pathway 4.5. (c) Pattern correlation of the 100-member ensemble mean of the SLP trend (red) with smaller ensembles sizes (randomly selected 30 times) and (d) the area average (in box A; black and box B; blue) ensemble mean of the SLP trend, with increasing ensemble size (randomly selected 30 times). The units of the trends and standard deviation of the trends are hPa/93 years. SLP = sea level pressure.

To assess the ensemble size needed to robustly isolate the pattern of the forced trend in SLP, we compare the ensemble mean trend of a subset of the ensemble to the 100-member ensemble mean trend by computing a pattern correlation. A low pattern correlation indicates that a subset does not capture the pattern of the forced trend. A large spread of trends in the subsets indicates that the ensemble mean trend in the smaller ensemble is still dominated by internal variability. We compute the pattern correlation for different ensemble sizes by randomly subsampling the 100 members. For each ensemble size, the subsampling is repeated 30 times to investigate if an ensemble of a given size can robustly isolate the spatial pattern of the forced trend (Figure 11c). The ensemble size is sufficient when the pattern correlation for all random subsets of a given ensemble size is high. For the trend in SLP, we see that the pattern correlation increases with increasing ensemble size and the spread of different samples reduces, with around 40–50 members sufficient to capture the ensemble mean pattern of the full ensemble. Similar to Deser et al. (2012), the error in the pattern is much lower above 20–30 members than below.

We then investigate the ensemble size needed to quantify the magnitude of the forced trend in SLP. To do this, we investigate two different regions (boxes in Figure 11; A: high variability and B: low variability). We estimate the "true" magnitude of the forced trend using the full ensemble of 100 members. To estimate the reliability of trend magnitudes for smaller ensemble sizes, we use resampling from the 100-member ensemble to create a set of 30 smaller ensembles for each ensemble size. We find that more ensemble members are needed in box A than in box B based on the faster reduction of the spread of subsets with increasing ensemble size (Figure 11d).

The true trend in box A is positive (0.56 hPa/93 years), while the true trend in box B is negative (−0.26 hPa/93 years). We find that with small ensembles, the trend in both boxes A and B could be estimated as either positive or negative (Figure 11d), meaning that with only a small ensemble, the sign of the forced trend could be misidentified. To correctly determine the sign of the forced trend, five members are needed in box B, and

40 members are needed in box A. To determine the actual ensemble size needed to quantify the trend, one must first decide on the size of the acceptable error. The error is estimated as the largest difference between the true trend and the subsampled trends for each ensemble size. In box B, the error is already small at 10 members, while in box A, there is still a noticeable error at 50 members. In box A, it is unclear whether the reduction of the error above 50 members is due to resampling or because the ensemble size is large enough.

As we approach the maximum ensemble size ($n = 100$), the overlap between different random samples increases; therefore, the spread is reduced. This reduction of the spread due to resampling impedes identification of a spread reduction due to a larger ensemble size. Further research on this topic and the effects of resampling is needed to determine whether this limitation in the interpretation can be overcome and how we can use larger ensembles to quantify the error in smaller ensembles. It is clear, however, that in box A, the error is reduced by having 40–50 members compared to 20–30, demonstrating the utility of using a 100-member ensemble for this analysis.

This example demonstrates how the 100-member ensemble can be used to estimate the ensemble size needed for a specific application. While MPI-GE can be used to address this question, the answer will depend on both the question asked, how large an error is acceptable to the user, and likely the model used. When the ensemble size needed approaches the size of the ensemble itself, it can be difficult to determine whether the ensemble is large enough or the apparent error is only reduced because of resampling from a limited sample. MPI-GE is by far the largest ensemble currently available and is hence currently the best tool to investigate the ensemble size problem when ensemble sizes needed approach the actual size of smaller ensemble projects. We suggest that MPI-GE can be used to inform how to design new ensembles or the process of choosing which ones of the currently available ensembles might be suitable for a given application.

6. Summary and Conclusions

MPI-GE has been presented, and its power has been demonstrated. First, due to its large size of 100 members, events with long return periods and quantities with high internal variability can be investigated. The initialization strategy means that most quantities can be investigated from the beginning of each simulation because the distribution of internal variability is adequately sampled from the beginning of the ensemble. Second, MPI-GE is the only large ensemble currently available with three future scenarios and a 1% CO_2 simulation allowing investigation into the targets set by the Paris Agreement and early twentieth century warming, something that could not be done with previous ensembles due to their later start dates.

We have demonstrated in this paper three ways to evaluate a large ensemble using observations. The first method can be used for observations with a long time series of observational coverage and uses where the observations sit within the ensemble spread as well as a rank histogram to evaluate the model. The second method is appropriate when good observational coverage is only available for a short time period (such as the satellite era). Here we demonstrate how to compare a single observational estimate with a histogram from a large ensemble. The third method provides a novel way to assess where on the globe internal variability in the model agrees well with observations by considering the agreement of the whole distribution. Additionally, this allows us to delve into specific regions and determine why there is a disagreement between the model and the observations and determine if the forced response and internal variability are realistic. We have additionally provided an example of how the model can be used to contextualize observations, by looking at the decadal variability of Arctic sea ice trends.

MPI-GE has then been used to complete four novel analyses that can only be undertaken using a large ensemble. The first addresses the question of whether there is a pathway dependence of temperature and precipitation responses under differing future scenarios, something that can only currently be addressed using the multiple future scenarios of MPI-GE. We find regional pathway dependence for precipitation, but not for temperature, that is larger than the model's internal variability. The second analysis asks whether there are forced changes in the highly variable atmospheric circulation. We demonstrate that these changes could already be occurring and that the tropospheric response dominates the Northern Pacific, whereas both the stratospheric and tropspheric changes are important in the North Atlantic. The third analysis asks whether changes in AMOC variability are time-dependent. We demonstrate that the projected decrease in AMOC variability largely occurs in the first half of the 21st century, indicating a strong time dependence. Finally, we give an example of how MPI-GE can be used to investigate the ensemble size needed for a given problem and demonstrate its utility for a problem such as forced SLP trends, where the ensemble size needed

appears to be close to or larger than the ensemble size that is available for other large ensembles. For quantities that need fewer ensemble members, we recommend that multiple large ensembles should be used to make such an estimate to account for possible model dependence of forced trends.

Overall, due to its large size and multiple scenarios, MPI-GE is a powerful tool that can be used to address uncertainties both due to internal variability and the unknown future pathway. Much can be learnt from using this ensemble alone, in combination with observations and with other large ensemble projects. The data are now publicly available, and we urge potential users to access it. Future studies that combine multiple large ensembles and in particular compare the magnitude of model uncertainty to internal variability will be vital to additionally address model uncertainty and to build on the work completed with single-model ensembles.

Acknowledgments
We thank the Max Planck Society for the core funding that made this project possible. We are indebted to T. Schulthess and the Swiss National Computing Centre (CSCS) for providing the computational resources for the historical simulations and the 1% CO_2 experiment. The RCP scenario simulations were performed with the facilities at the German Climate Computing Centre (DKRZ). We would like to thank Karsten Peters, Katharina Berger, Heinz-Dieter Hollweg and Fabian Wachsmann for their work in making the data publicly available. We also thank Veronika Gayler for her work with the JSBACH data and Irene Stemmler for her input on the HAMOCC data. Additionally, we thank Florian Ziemen for conducing an internal review, Helmuth Haak for his input on the difference between MPI-ESM and MPI-ESM1.1, in the ocean, and Thorsten Mauritsen for providing the equilibrium climate sensitivity. We thank Gábor Drótos and Tamás Bódai as well as the two anonymous reviewers for their comments on this manuscript. Nicola Maher was supported by the Alexander von Humboldt Foundation. Yohei Takano and Lena Boysen are supported by the European Union's Horizon 2020 research and innovation program under grant agreement 641816 (CRESCENDO). Rohit Ghosh and Elisa Manzini are partly supported by the European Union's Horizon 2020 research and innovation program under grant agreement 727852 (Blue-Action). Information on the publication of the Max Planck Institute Grand Ensemble (MPI-GE) output can be found on our website (https://www.mpimet.mpg.de/en/grand-ensemble/).

References

Andrews, T., Gregory, J. M., Webb, M., & Taylor, K. E. (2012). Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere-ocean climate models. *Geophysical Research Letters*, *39*, L09712. https://doi.org/10.1029/2012GL051,607

Barnes, E. A., & Polvani, L. (2013). Response of the midlatitude jets, and of their variability, to increased greenhouse gases in the CMIP5 models. *Journal of Climate*, *26*(18), 7117–7135.

Bengtsson, L., & Hodges, K. I. (2018). Can an ensemble climate simulation be used to separate climate change signals from internal unforced variability? *Climate Dynamics*, *52*, 3553–3573. https://doi.org/10.1007/s00382-018-4343-8

Bindoff, N., Stott, P., AchutaRao, K., Allen, M., Gillett, N., Gutzler, D., et al. (2013). Detection and attribution of climate change: From global to regional. In T. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. Allen, J. Boschung, et al. (Eds.), *The physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* pp. 867–952). Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

Bittner, M., Schmidt, H., Timmreck, C., & Sienz, F. (2016). Using a large ensemble of simulations to assess the northern hemisphere stratospheric dynamical response to tropical volcanic eruptions and its uncertainty. *Geophysical Research Letters*, *43*, 9324–9332. https://doi.org/10.1002/2016GL070587

Branstator, G., & Selten, F. (2009). "Modes of variability" and climate change. *Journal of Climate*, *22*(10), 2639–2658. https://doi.org/10.1175/2008JCLI2517.1

Cheng, J., amd Chiang, W., & Zhang, D. (2013). Atlantic Meridional Overturning Circulation (AMOC) in CMIP5 models: RCP and historical simulations. *Journal of Climate*, *26*, 7187–7197.

Cheng, J., Liu, Z., Zhang, S., Liu, W., Dong, L., Liu, P., & Li, H. (2016). Reduced interdecadal variability of Atlantic Meridional Overturning Circulation under global warming. *Proceedings of the National Academy of Sciences*, *113*(12), 3175–3178.

Collins, M., Knutti, R., Arblaster, J., Dufresne, J., Fichefet, T., Friedlingstein, P., et al. (2013). Long-term climate change: Projections, commitments and irreversibility. In T. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. Allen, J. Boschung, et al. (Eds.), *Climate change 2013: The physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* pp. 1029–1136). Cambridge, United Kingdom and New York, USA: Cambridge University Press.

Dai, A., & Bloecker, C. E. (2019). Impacts of internal variability on temperature and precipitation trends in large ensemble simulations by two climate models. *Climate Dynamics*, *52*(1), 289–306. https://doi.org/10.1007/s00382-018-4132-4

Daron, J. D., & Stainforth, D. A. (2013). On predicting climate under climate change. *Environmental Research Letters*, *8*(3), 034021. https://doi.org/10.1088/1748-9326/8/3/034021

Deser, C., Phillips, A., Bourdette, V., & Teng, H. (2012). Uncertainty in climate change projections: The role of internal variability. *Climate Dynamics*, *38*(3), 527–546. https://doi.org/10.1007/s00382-010-0977-x

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

Dessler, A. E., Mauritsen, T., & Stevens, B. (2018). The influence of internal variability on Earth's energy balance framework and implications for estimating climate sensitivity. *Atmospheric Chemistry and Physics*, *18*(7), 5147–5155. https://doi.org/10.5194/acp-18-5147-2018

Diffenbaugh, N. S., Swain, D. L., & Touma, D. (2015). Anthropogenic warming has increased drought risk in california. *Proceedings of the National Academy of Sciences*, *112*(13), 3931–3936. https://doi.org/10.1073/pnas.1422385112

Drótos, G., Bódai, T., & Tél, T. (2017). Probabilistic concepts in a changing climate: A snapshot attractor picture. *Journal Climate*, *28*, 3275–3288.

Fasullo, J. T., & Nerem, R. S. (2016). Interannual variability in global mean sea level estimated from the CESM large and last millennium ensembles. *Water*, *8*(11), 491. https://doi.org/10.3390/w8110491

Fetterer, F., Knowles, K., Meier, W., Savoie, A., & Windnagel, A. (2017). Sea ice index version 3. NSIDC National Snow and Ice Data Center, https://doi.org/10.7265/N5K072F8

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S., Collins, W., et al. (2013). Evaluation of climate models. In T. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. Allen, J. Boschung, et al. (Eds.), *Climate change 2013: The physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* pp. 741–866). Cambridge: Cambridge University Press.

Frankcombe, L. M., England, M. H., Kajtar, J. B., Mann, M. E., & Steinman, B. A. (2018). On the choice of ensemble mean for estimating the forced signal in the presence of internal variability. *Journal of Climate*, *31*(14), 5681–5693. https://doi.org/10.1175/JCLI-D-17-0662.1

Frankcombe, L. M., England, M. H., Mann, M. E., & Steinman, B. A. (2015). Separating internal variability from the externally forced climate response. *Journal of Climate*, *28*(20), 8184–8202. https://doi.org/10.1175/JCLI-D-15-0069.1

Frankignoul, C., Gastineau, G., & Kwon, Y.-O. (2017). Estimation of the SST response to anthropogenic and external forcing and its impact on the Atlantic Multidecadal Oscillation and the Pacific Decadal Oscillation. *Journal of Climate*, *30*(24), 9871–9895. https://doi.org/10.1175/JCLI-D-17-0009.1

Fu, Q., Lin, P., Solomon, S., & Hartmann, D. L. (2015). Observational evidence of strengthening of the Brewer-Dobson circulation since 1980. *Journal of Geophysical Research: Atmospheres*, *120*, 10,214–10,228. https://doi.org/10.1002/2015JD023657

Gibson, P., Perkins-Kirkpatrick, S., Alexander, L., & Fischer, E. (2017). Comparing Australian heat waves in the CMIP5 models through cluster analysis. *Journal of Geophysical Research: Atmospheres*, *122*, 3266–3281. https://doi.org/10.1002/2016JD025878

Giorgetta, M. A., Jungclaus, J., Reick, C. H., Legutke, S., Bader, J., Böttinger, M., et al. (2013). Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5. *Journal of Advances in Modeling Earth Systems*, *5*, 572–597. https://doi.org/10.1002/jame.20038

Goll, D., Brovkin, V., Liski, J., Raddatz, T., Thum, T., & Todd-Brown, K. (2015). Strong dependence of CO_2 emissions from anthropogenic land cover change on soil carbon parametrization and initial land cover. *Global Biogeochemical Cycles*, *29*, 1511–1523. https://doi.org/10.1002/2014GB004988

Hagemann, S., & Stacke, T. (2014). Impact of the soil hydrology scheme on simulated soil moisture memory. *Climate Dynamics*, *44*, 1731–1750.

Hasselmann, K. (1976). Stochastic climate models. Part 1. *Theory, Technical Bulletin of the Registry of Medical Technologistsllus*, *28A*, 473–485.

Hawkins, E., Smith, R. S., Gregory, J. M., & Stainforth, D. A. (2016). Irreducible uncertainty in near-term climate projections. *Climate Dynamics*, *46*(11), 3807–3819. https://doi.org/10.1007/s00382-015-2806-8

Hawkins, E., & Sutton, R. (2009). Decadal predictability of the Atlantic Ocean in a coupled GCM: Forecast skill and optimal perturbations using linear inverse modeling. *Journal of Climate*, *22*(14), 3960–3978. https://doi.org/10.1175/2009JCLI2720.1

Hedemann, C., Mauritsen, T., Jungclaus, J., & Marotzke, M. (2017). The subtle origins of surface-warming hiatuses. *Nature Climate Change*, *7*, 336–339. https://doi.org/10.1038/nclimate3274

Herein, M., Drótos, G., Haszpra, T., Márfy, J., & Tél, T. (2017). The theory of parallel climate realizations as a new framework for teleconnection analysis. *Scientific Reports*, *7*, 44529.

Ilyina, T., Six, K. D., Segschneider, J., Maier-Reimer, E., Li, H., & Núñez-Riboni, I. (2013). Global ocean biogeochemistry model HAMOCC: Model architecture and performance as component of the MPI-Earth system model in different CMIP5 experimental realizations. *Journal of Advances in Modeling Earth Systems*, *5*, 287–315. https://doi.org/10.1029/2012MS000178

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 American Meteorological Society*, *96*(8), 1333–1349. https://doi.org/10.1175/BAMS-D-13-00255.1

Kidston, J., Scaife, A. A., Hardiman, S. C., Mitchell, D. M., Butchart, N., Baldwin, M. P., & Gray, L. J. (2015). Stratospheric influence on tropospheric jet streams, storm tracks and surface weather. *Nature Geoscience*, *8*, 433–440.

Kirchmeier-Young, M. C., Zwiers, F. W., & Gillett, N. P. (2017). Attribution of extreme events in Arctic sea ice extent. *Journal of Climate*, *30*(2), 553–571. https://doi.org/10.1175/JCLI-D-16-0412.1

Kwon, Y.-O., Camacho, A., Martinez, C., & Seo, H. (2018). North Atlantic winter eddy-driven jet and atmospheric blocking variability in the Community Earth System Model version 1 Large Ensemble simulations. *Climate Dynamics*, *51*, 3275–3289.

Lehner, F., Deser, C., & Terray, L. (2017). Toward a new estimate of "time of emergence" of anthropogenic warming: Insights from dynamical adjustment and a large initial-condition model ensemble. *Journal of Climate*, *30*, 7739–7756.

Li, H., & Ilyina, T. (2018). Current and future decadal trends in the oceanic carbon uptake are dominated by internal variability. *Geophysical Research Letters*, *45*, 916–925. https://doi.org/10.1002/2017GL075,370

Lin, L., Wang, Z., Xu, Y., & Fu, Q. (2016). Sensitivity of precipitation extremes to radiative forcing of greenhouse gases and aerosols. *Geophysical Research Letters*, *43*, 9860–9868. https://doi.org/10.1002/2016GL070869

Lin, L., Wang, Z., Xu, Y., Fu, Q., & Dong, W. (2018). Larger sensitivity of precipitation extremes to aerosol than greenhouse gas forcing in CMIP5 models. *Journal of Geophysical Research: Atmospheres*, *123*, 8062–8073. https://doi.org/10.1029/2018JD028821

Loeb, N. G., Doelling, D. R., Wang, H., Su, W., Nguyen, C., Corbett, J. G., et al. (2018). Clouds and the Earth's Radiant Energy System (CERES) Energy Balanced and Filled (EBAF) Top-of-Atmosphere (TOA) Edition-4.0 Data Product. *Journal of Climate*, *31*(2), 895–918. https://doi.org/10.1175/JCLI-D-17-0208.1

Maher, N., Matei, D., Milinski, S., & Marotzke, J. (2018). ENSO change in climate projections: Forced response or internal variability? *Geophysical Research Letters*, *45*, 11,390–11,398. https://doi.org/10.1029/2018GL079764

Maher, N., McGregor, S., England, M. H., & Sen Gupta, A. (2015). Effects of volcanism on tropical variability. *Geophysical Research Letters*, *42*, 6024–6033. https://doi.org/10.1002/2015GL064751

Manzini, E., Karpechko, A. Y., Anstey, J., Baldwin, M. P., Black, R. X., Cagnazzo, C., et al. (2014). Northern winter climate change: Assessment of uncertainty in CMIP5 projections related to stratosphere-troposphere coupling. *Journal of Geophysical Research: Atmospheres*, *119*, 7979–7998. https://doi.org/10.1002/2013JD021403

Manzini, E., Karpechko, A. Y., & Kornblueh, L. (2018). Nonlinear response of the stratosphere and the North Atlantic-European climate to global warming. *Geophysical Research Letters*, *45*, 4255–4263. https://doi.org/10.1029/2018GL077826

Maroon, E., Kay, J., & Karnauskas, K. (2018). Influence of the Atlantic Meridional Overturning Circulation on the Northern Hemisphere surface temperature response to radiative forcing. *Journal of Climate*, *31*, 9207–9224.

Marotzke, J. (2019). Quantifying the irreducible uncertainty in near-term climate projections. *Wiley Interdisciplinary Reviews: Climate Change*, vol. *10*, pp. e563. https://doi.org/10.1002/wcc.563

Marotzke, J., & Forster, P. M. (2015). Forcing, feedback and internal variability in global temperature trends. *Nature*, *517*, 565–570.

Marsland, S. J., Haak, H., Jungclaus, J. H., Latif, M., & Röske, F. (2003). The Max Planck Institute global ocean/sea ice model with orthogonal curvilinear coordinates. *Ocean Modelling*, *5*, 91–127.

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-ESM 1.2) and its response to increasing CO_2. *Journal of Advances in Modeling Earth Systems*, *11*, 998–1038. https://doi.org/10.1029/2018MS001400

McKinnon, K. A., & Deser, C. (2018). Internal variability and regional climate trends in an observational large ensemble. *Journal of Climate*, *31*(17), 6783–6802. https://doi.org/10.1175/JCLI-D-17-0901.1

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

Morice, C., Kennedy, J., Rayner, N., & Jones, P. (2012). Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *Journal of Geophysical Research*, *117*, D08101. https://doi.org/10.1029/2011JD017187

Niederdrenk, A. L., & Notz, D. (2018). Arctic sea ice in a 1.5°C warmer world. *Geophysical Research Letters*, *45*, 1963–1971. https://doi.org/10.1002/2017GL076159

Notz, D. (2015). How well must climate models agree with observations? *Philosophical Transactions of the Royal Society A: Mathematical Physical and Engineering Sciences*, *373*, 20140164. https://doi.org/10.1098/rsta.2014.0164
Notz, D. (2017). Arctic sea ice seasonal-to-decadal variability and long-term change. *PAGES Magazine*, *25*(1), 14–19.
Notz, D., & Marotzke, J. (2012). Observations reveal external driver for Arctic sea-ice retreat. *Geophysical Research Letters*, *39*, L08502. https://doi.org/10.1029/2012GL051,094
Olonscheck, D., Mauritsen, T., & Notz, D. (2019). Arctic sea-ice variability is primarily driven by atmospheric temperature fluctuations. *Nature Geoscience*, *12*, 430–434. https://doi.org/10.1038/s41561-019-0363-1
Olonscheck, D., & Notz, D. (2017). Consistently estimating internal climate variability from climate model simulations. *Journal of Climate*, *30*(23), 9555–9573. https://doi.org/10.1175/JCLI-D-16-0428.1
Pendergrass, A., Lehner, F., & Sanderson, B. (2015). Does extreme precipitation intensity depend on the emissions scenario? *Geophysical Research Letters*, *42*, 8767–8774. https://doi.org/10.1002/2015GL065854
Plesca, E., Grützun, V., & Buehler, S. A. (2018). How robust is the weakening of the Pacific Walker Circulation in CMIP5 idealized transient climate simulations? *Journal of Climate*, *31*(1), 81–97. https://doi.org/10.1175/JCLI-D-17-0151.1
Rädel, G., Mauritsen, T., Stevens, B., Dommenget, D., Matei, D., Bellomo, K., & Clement, A. (2016). Amplification of El Niño by cloud longwave coupling to atmospheric circulation. *Nature Geoscience*, *9*, 106–110.
Reick, C. H., Raddatz, T., Brovkin, V., & Gayler, V. (2013). Representation of natural and anthropogenic land cover change in MPI-ESM. *Journal of Advances in Modeling Earth Systems*, *5*, 459–482. https://doi.org/10.1002/jame.20022
Risbey, J., Lewandowsky, S., Langlais, C., Monselesan, D., O'Kane, T., & Oreskes, N. (2014). Well-estimated global surface warming in climate projections selected for ENSO phase. *Nature Climate Change*, *4*, 835–840.
Rodgers, K. B., Lin, J., & Frölicher, T. L. (2015). Emergence of multiple ocean ecosystem drivers in a large ensemble suite with an Earth system model. *Biogeosciences*, *12*(11), 3301–3320.
Sanderson, B. M., Oleson, K. W., Strand, W. G., Lehner, F., & O'Neill, B. C. (2018). A new ensemble of GCM simulations to assess avoided impacts in a climate mitigation scenario. *Climatic Change*, *146*(3), 303–318. https://doi.org/10.1007/s10584-015-1567-z
Screen, J., Gillett, N. P., Stevens, D., Marshall, G., & Roscoe, H. (2009). The role of eddies in the Southern Ocean temperature response to the Southern Annular Mode. *Journal of Climate*, *22*, 806–818.
Segschneider, J., Beitsch, A., Timmreck, C., Brovkin, V., Ilyina, T., Jungclaus, J., et al. (2013). Impact of an extremely large magnitude volcanic eruption on the global climate and carbon cycle estimated from ensemble Earth System Model simulations. *Biogeosciences*, *10*, 669–687.
Sen Gupta, A., Jourdain, N. C., Brown, J. N., & Monselesan, D. (2013). Climate drift in the CMIP5 models. *Journal of Climate*, *26*(21), 8597–8615. https://doi.org/10.1175/JCLI-D-12-00521.1
Seviour, W. (2017). Weakening and shift of the Arctic stratospheric polar vortex: Internal variability or forced response? *Geophysical Research Letters*, *44*, 3365–3373. https://doi.org/10.1002/2017GL073071
Simpson, I., Hitchcock, P., Seager, R., Wu, Y., & Callaghan, P. (2018). The downward influence of uncertainty in the northern hemisphere stratospheric polar vortex response to climate change. *Journal of Climate*, *31*(16), 6371–6391.
Smith, A., & Jahn, A. (2019). Definition differences and internal variability affect the simulated Arctic sea ice melt season. *The Cryosphere*, *13*, 1–20.
Stephens, G., O'Brien, D., Webster, P., Pilewski, P., Kato, S., & Li, J. (2015). The albedo of Earth. *Reviews of Geophysics*, *53*, 141–163. https://doi.org/10.1002/2014RG000449
Stevens, B. (2015). Rethinking the lower bound on aerosol radiative forcing. *Journal of Climate*, *28*(12), 4794–4819. https://doi.org/10.1175/JCLI-D-14-00656.1
Stevens, B., Giorgetta, M., Esch, M., Mauritsen, T., Crueger, T., Rast, S., et al. (2013). Atmospheric component of the MPI-M Earth System Model: ECHAM6. *Journal of Advances in Modeling Earth Systems*, *5*, 146–172. https://doi.org/10.1002/jame.20015
Stolpe, M. B., Medhaug, I., Sedláček, J., & Knutti, R. (2018). Multidecadal variability in global surface temperatures related to the Atlantic Meridional Overturning Circulation. *Journal of Climate*, *31*(7), 2889–2906. https://doi.org/10.1175/JCLI-D-17-0444.1
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.
Suárez-Gutiérrez, L., Li, C., Thorne, P. W., & Marotzke, J. (2017). Internal variability in simulated and observed tropical tropospheric temperature trends. *Geophysical Research Letters*, *44*, 5709–5719. https://doi.org/10.1002/2017GL073798
Sutton, R. T., & Hodson, D. L. R. (2005). Atlantic ocean forcing of North American and European summer climate. *Science*, *309*(5731), 115–118. https://doi.org/10.1126/science.1109496
Swart, S., Fyfe, J., Hawkins, E., Kay, J. E., & Jahn, A. (2015). Influence of internal variability on Arctic sea-ice trends. *Nature Climate Change*, *5*, 86–89.
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., & Wehner, M. F. (2018). Benefits of mitigation for future heat extremes under RCP4.5 compared to RCP8.5. *Climatic Change*, *146*(3), 349–361. https://doi.org/10.1007/s10584-016-1605-5
Thompson, D. W. J., Barnes, E. A., Deser, C., Foust, W. E., & Phillips, A. S. (2015). Quantifying the role of internal climate variability in future climate trends. *Journal of Climate*, *28*(16), 6443–6456. https://doi.org/10.1175/JCLI-D-14-00830.1
von Känel, L., Frölicher, T. L., & Gruber, N. (2017). Hiatus-like decades in the absence of equatorial Pacific cooling and accelerated global ocean heat uptake. *Geophysical Research Letters*, *44*, 7909–7918. https://doi.org/10.1002/2017GL073578
Wang, S.-Y. S., Zhao, L., Yoon, J.-H., Klotzbach, P., & Gillies, R. R. (2018). Quantitative attribution of climate effects on Hurricane Harvey's extreme rainfall in Texas. *Environmental Research Letters*, *13*(5), 054014. https://doi.org/10.1088/1748-9326/aabb85
Wittenberg, A. T., Rosati, A., Delworth, T. L., Vecchi, G. A., & Zeng, F. (2014). ENSO modulation: Is it decadally predictable? *Journal of Climate*, *27*(7), 2667–2681. https://doi.org/10.1175/JCLI-D-13-00577.1
Zelle, H., Jan van Oldenborgh, G., Burgers, G., & Dijkstra, H. (2005). El Niño and greenhouse warming: Results from ensemble simulations with the NCAR CCSM. *Journal of Climate*, *18*(22), 4669–4683. https://doi.org/10.1175/JCLI3574.1
Zhang, L., Han, W., & Sienz, F. (2018). Unraveling causes for the changing behavior of the tropical Indian Ocean in the past few decades. *Journal of Climate*, *31*(6), 2377–2388. https://doi.org/10.1175/JCLI-D-17-0445.1