Climate Dynamics (2021) 57:2557–2580
https://doi.org/10.1007/s00382-021-05821-w

Exploiting large ensembles for a better yet simpler climate model evaluation

Laura Suarez-Gutierrez[1] · Sebastian Milinski[1] · Nicola Maher[1]

Received: 30 November 2020 / Accepted: 18 May 2021 / Published online: 29 May 2021

Abstract
We use a methodological framework exploiting the power of large ensembles to evaluate how well ten coupled climate models represent the internal variability and response to external forcings in observed historical surface temperatures. This evaluation framework allows us to directly attribute discrepancies between models and observations to biases in the simulated internal variability or forced response, without relying on assumptions to separate these signals in observations. The largest discrepancies result from the overestimated forced warming in some models during recent decades. In contrast, models do not systematically over- or underestimate internal variability in global mean temperature. On regional scales, all models misrepresent surface temperature variability over the Southern Ocean, while overestimating variability over land-surface areas, such as the Amazon and South Asia, and high-latitude oceans. Our evaluation shows that MPI-GE, followed by GFDL-ESM2M and CESM-LE offer the best global and regional representation of both the internal variability and forced response in observed historical temperatures.

Keywords Climate model evaluation · Large ensembles · SMILEs · Climate models · Forced response · Internal variability · Surface temperatures

1 Introduction

Observations reflect how the real-world climate system responds to changing natural and anthropogenic external forcings, as well as how the system fluctuates due to its own chaotic internal variability. Similarly, individual climate model simulations also are a combination of the simulated forced response in the model and its simulated internal variability. Therefore, discrepancies between observations and simulations may arise due to errors in the model's external forcing and its simulated response to this forcing, in its internal variability, or some combination of these factors (Notz 2015; Marotzke and Forster 2015; Suarez-Gutierrez et al. 2017). However, these discrepancies between observations and model simulations may also result from the observed and simulated fluctuations caused by internal variability being in different states (Notz 2015). In this paper, we provide an evaluation of how well models capture the internal variability and forced response in observed surface temperatures by applying a robust yet simple framework (Anderson 1996; Hamill 2001; Suarez-Gutierrez et al. 2018; Maher et al. 2019) that exploits the power of single model initial-condition large ensembles (SMILEs) from ten fully-coupled, comprehensive climate models.

SMILE experiments consist of many simulations of a single climate model under the same time-evolving external forcings, but starting from different initial conditions. This experimental design ensures that the simulations in the ensemble differ only due to the effect of internal variability, and allows a precise quantification of both the forced response, represented by the ensemble mean, and internal variability, represented by the ensemble's spread of deviations from this mean (Maher et al. 2019; Deser et al. 2012; Frankcombe et al. 2015). The sampling of internal variability is particularly important to capture low-probability events and events that show large deviations from the mean state, and is key to obtaining well-defined probability distributions (Suarez-Gutierrez et al. 2018). This is crucial not only to correctly capture the events at the tails of the distribution, but also to ensure that the forced response is not biased by an insufficient sampling of the different states of

✉ Laura Suarez-Gutierrez
laura.suarez@mpimet.mpg.de

[1] Max-Planck-Institut für Meteorologie, 20146 Hamburg, Germany

internal variability (Frankcombe et al. 2015, 2018; Milinski et al. 2020). Thus, using SMILEs we can determine whether observations fall within the now better-sampled range of the transient spread simulated by each climate model, potentially reconciling differences between model simulations and observations (Thorne et al. 2015; Hedemann et al. 2017). In the cases where discrepancies between model simulations and observations remain, we can now directly attribute whether these remaining discrepancies are caused by an incorrect simulation of internal variability, or rather by an incorrect simulation of how the climate system responds to external forcings, without relying on assumptions to separate both signals in the observations (Bittner et al. 2016; Suarez-Gutierrez et al. 2017; Smith and Jahn 2019).

Due to their experimental design, SMILEs are powerful tools for model evaluation that allow us to rethink and expand our methodologies beyond customary practices. One example of this is the methodological framework that we apply in our study. This framework (Suarez-Gutierrez et al. 2018; Maher et al. 2019; Suarez-Gutierrez et al. 2020a, b) relies on the precise characterization of simulated internal variability in SMILE experiments, which provide well-defined estimates of both the time-evolving forced response and the probability distribution of deviations from this mean state caused by internal variability. Also, it eliminates the need of filtering or detrending techniques to separate the effect of internal variability from the forced response either in the models or in observations. The evaluation of the adequacy of climate models in this framework is based on a simple approach: whether observations are distributed evenly across the whole spread of the ensemble, and whether they generally stay within the limits of this spread (Hamill 2001; Marotzke and Forster 2015; Suarez-Gutierrez et al. 2017, 2018; Maher et al. 2019).

In the ideal case in which real-world variability and forced response are perfectly simulated by a model, and the observational record is sufficiently long, observations would occur across the whole simulated ensemble spread with similar frequency, and the simulated ensemble spread would generally cover the range of observed variability. Thus, to determine that an ensemble adequately captures observations two conditions should be met: first, observed values should fall across the ensemble with no preferred frequency, and second only occasionally outside of its limits. Together, these two conditions provide a robust metric for model evaluation, based on whether the range of well-defined simulated climate states in an ensemble adequately captures the long-term trajectory and the possible deviations from this trajectory caused by internal variability in the real-world climate system. Furthermore, if this is not the case, we can identify whether discrepancies between the range of the simulated climate states and observations arise because of an incorrect simulated forced response, or because the simulated internal variability in the model over- or underestimates the observed internal variability.

State-of-the-art model evaluation frameworks (Flato et al. 2013; IPCC, SR15 2018) are often constrained to mean state comparisons, or to relying on detrended quantities and assumptions for filtering and isolating the observed variability and forced response (Gleckler et al. 2008; Frankcombe et al. 2018; Tokarska et al. 2020). Similarly, previous studies evaluating internal variability are constrained to using standard deviations as a proxy (Schär et al. 2004; Lehner et al. 2017; McKinnon et al. 2017; Bengtsson and Hodges 2019), thus relying on assumptions regarding the shape of the full probability distributions and overlooking the evaluation of its higher-order moments. Here, we go beyond previous model evaluation efforts by assessing whether the whole simulated distribution, including its tails, agrees well with observations. In this sense, our framework resembles probabilistic forecast verification techniques in the climate prediction literature (Anderson 1996; Hamill 2001; Annan and Hargreaves 2010). By considering the whole spread of the ensemble we are able to implicitly assess the higher-order moments of the distribution, offering a more appropriate evaluation of the simulated representation of the magnitude and frequency of observed estimates, including when they are extreme (Suarez-Gutierrez et al. 2020b). Ultimately, we can then condense this information to determine how many climate models succeed at capturing the observed internal variability and the climate system's response to external forcings over each region of the world. Thus, we provide a framework to assess models fitness-for-purpose, determining in which regions models can adequately simulate the observed climate, and why, in other regions, they cannot.

SMILEs of a large number of fully-coupled climate models have only recently become widely available (Maher et al. 2019; Deser et al. 2020). Therefore, most previous studies using SMILEs to evaluate the agreement between model simulations and observations are based on a limited number of three or fewer SMILEs (Suarez-Gutierrez et al. 2017; Maher et al. 2018; Schaller et al. 2018; von Trentini et al. 2020), and most are also limited to conventional evaluation methods (Maher et al. 2018; Schaller et al. 2018). Similarly, previous studies using evaluation frameworks that consider whole ensemble distributions to evaluate the agreement of several climate models with observations are based on multi-model ensembles such as the Coupled Model Intercomparison Project (CMIP; Taylor et al. 2012), which have a limited number of simulations for each model and do not allow for a clean separation between simulated forced response and internal variability (Annan and Hargreaves 2010; Marotzke and Forster 2015). In addition, previous multi-model evaluation frameworks may favour models with larger internal variability, since they are more likely to capture observations by chance (Beusch et al. 2020), thus indicating that a robust

multi-model comparison and evaluation of internal variability across climate models is crucial for assessing model performance. Here, we provide a multi-model comparison of the well-sampled transient internal variability and forced response in SMILEs from ten comprehensive, fully-coupled climate models from both the CMIP5 and CMIP6 generations, and the first multi-model evaluation of how well these models capture the internal variability and forced response in observations.

2 Data and methods

2.1 Climate model simulations and observational data

We include SMILEs from a broad range of climate models: CanESM2 (Kirchmeier-Young et al. 2017), CanESM5 (Swart et al. 2019), CESM(Hurrell et al. 2013; Kay et al. 2015), CSIRO-MK3.6 (Jeffrey et al. 2013), GFDL-CM3 (Sun et al. 2018), GFDL-ESM2M (Rodgers et al. 2015), IPSL-CM5A (Frankignoul et al. 2017), IPSL-CM6A (Boucher et al. 2020), MIROC6 (Tatebe et al. 2019), and MPI-ESM (Maher et al. 2019). Each SMILE comprises of several simulations for one fully coupled climate model that differ only in their initial state, and evolve under one specific set of forcing conditions. However, the ensembles differ in the number of simulations included (from 20 up to 100 members), in how sensitive the model is to increasing CO_2 (Equilibrium Climate Sensitivity, ECS, values of 2.4 K to more than 5 K), in the method used for the initialization of their members (from micro atmospheric perturbations to different initial states sampled from the control simulation), in the generation of forcing scenarios used (CMIP5 to CMIP6), and in the forcing scenario (from only historical to historical extended with a high emissions scenario such as RCP8.5). When possible, historical model simulations are extended with one available future forcing scenario to cover the observational record. More details can be found in Table 1, and in previous studies (Maher et al. 2019; Deser et al. 2020) and references therein. Observed surface temperature data from the HadCRUT4.6 (Morice et al. 2012) dataset for the period of 1850–2020 are used for comparison to the SMILE simulations.

We use a total of 481 simulations to evaluate how SMILEs of ten different climate models capture the internal variability and the response to external forcings in the HadCRUT4.6 observed surface temperatures. We define surface temperatures as the annually averaged near-surface 2 m air temperature anomaly over land grid cells, and sea surface temperature over the ocean. Global mean surface temperature (GMST) is defined as the global average of these surface temperatures. All simulated data are regridded to the coarser resolution of HadCRUT observations, and subsampled to grid boxes where observations are available. All data shown are anomalies calculated with respect to the climatological baseline defined by the period 1961–1990 in each ensemble member and observations. This approach removes potential mean biases in the simulations. However, it does not remove time or phase dependent biases in how the simulated climate responds to external forcing conditions arising in different periods. These remaining biases are referred to as biases in the forced response throughout the paper. Choosing a different reference period may lead to biases of different signs appearing in different periods. However, we maintain the period of 1961–1990 as climatological reference to ensure contributions from all SMILEs and the most complete observational coverage (Jones et al. 2012; Morice et al. 2012).

Table 1 Details of SMILE experiments included

SMILE	Members	Years	Gen.	Forcing	ECS	Reference
CanESM2	50	1950–2020	CMIP5	Hist+RCP8.5	3.7 K	Kirchmeier-Young et al. (2017)
CanESM5*	50	1850–2014	CMIP6	Hist	5.7 K	Swart et al. (2019)
CESM-LE	35	1920–2020	CMIP5	Hist+RCP8.5	4.1 K	Kay et al. (2015)
CSIROMK3.6	30	1850–2020	CMIP5	Hist+RCP8.5	4.1 K	Jeffrey et al. (2013)
GFDL-CM3	20	1920–2020	CMIP5	Hist+RCP8.5	4.0 K	Sun et al. (2018)
GFDL-ESM2M	30	1950–2020	CMIP5	Hist+RCP8.5	2.4 K	Rodgers et al. (2015)
IPSL-CM5A	30	1941–2020	CMIP5	Hist+RCP8.5	4.1 K	Jebri et al. (2020)
IPSL-CM6A*	31	1850–2014	CMIP6	Hist	4.5 K	Boucher et al. (2020)
MIROC6*	50	1850–2020	CMIP6	Hist+SSP2-4.5	2.6 K	Tatebe et al. (2019)
MPI-GE	100	1850–2020	CMIP5	Hist+RCP4.5	2.8 K	Maher et al. (2019)

Experiment name, number of members, simulated years used, forcing generation, forcing scenarios, and Equilibrium Climate Sensitivity (ECS) of SMILE experiments included in our study. All experiments include historical forcing (Hist) until 2005 for CMIP5 or until 2014 for CMIP6. CMIP6 generation SMILEs are marked by a star in the table and in all figures. When possible, simulations are extended using one future forcing scenario. ECS refers to the equilibrium temperature response to the doubling of CO_2 (Andrews et al. 2012; Hurrell et al. 2013; Swart et al. 2019; Maher et al. 2019)

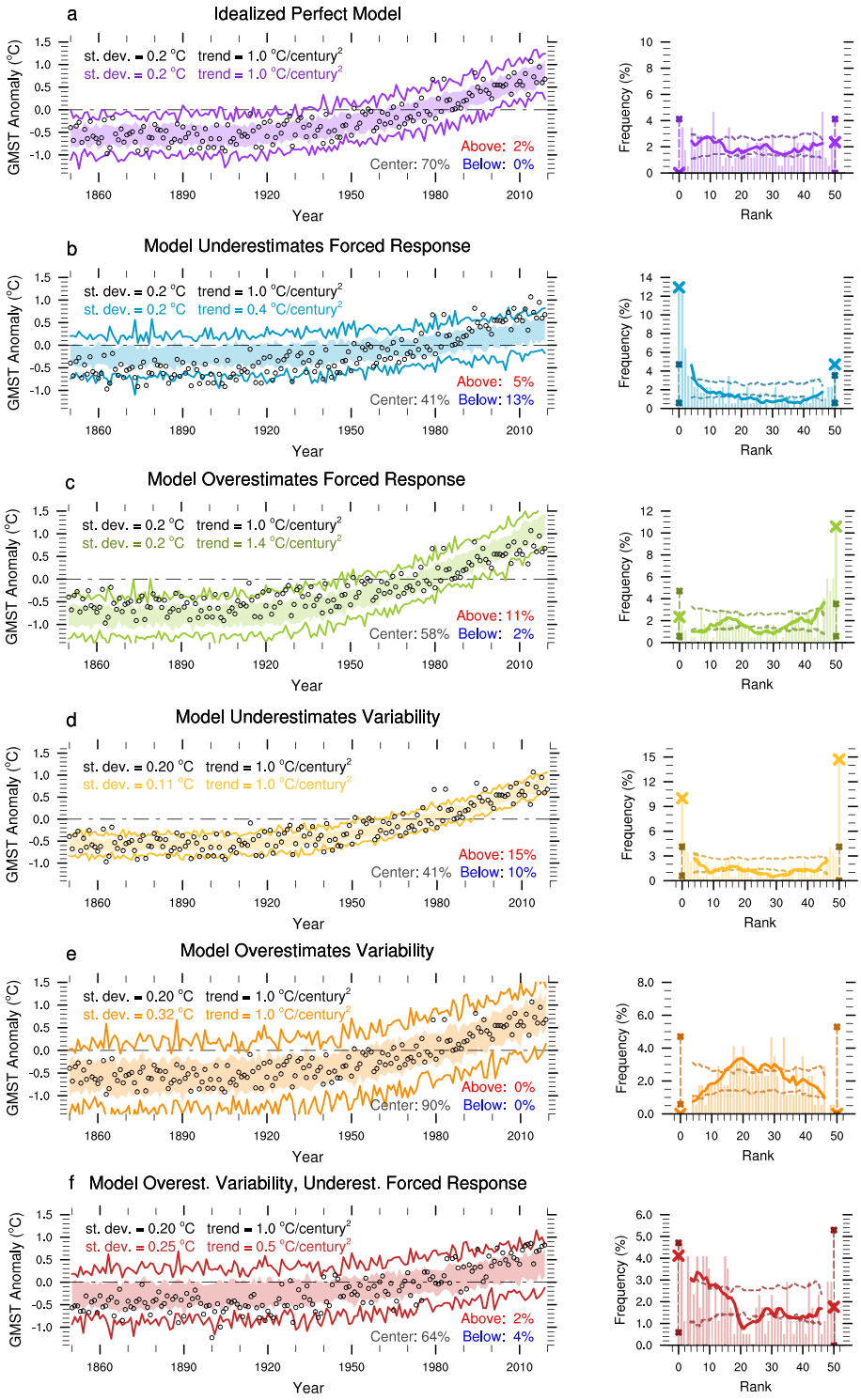


Fig. 1 Idealized examples of time series and rank histograms. Time series of idealized annual GMST anomalies relative to the period 1961–1990 by synthetic ensembles (color) and observations (black). The data follows normal distributions with different standard deviations and added quadratic trend terms after year 1900. Lines represent ensemble maxima and minima, and shading represents the central 75th percentile range (12.5th to 87.5th percentiles). Percentages represent observations within the 75th percentile central range (gray), above the ensemble maximum (red) and below the ensemble minimum (blue). Rank histograms represent the frequency of each place observations would take in a list of ensemble members ordered by ascending GMST values. Lines illustrate the rank histogram's slope, as the mean rank frequency over a centered 6-bin window for observations (solid lines, light colors), and the perfect model range, as the rank slopes for each ensemble member treated as observations (5th–95th percentile; dashed, dark colors). Crosses represent the frequency of minimum (0) and maximum (number of members) ranks for observations (light colors), and for the perfect model range (5th–95th percentile; dark colors). Bin sizes are 1 rank

In this study we evaluate climate models against one single observational product, HadCRUT4.6 (Morice et al. 2012; Cowtan and Way 2014). We recommend users who wish to apply this framework to determine which SMILEs better capture observations for their variable and region of interest to base this evaluation on a wider range of observational products, combining point-level surface datasets and reanalyses with gridded statistically processed datasets such as HadCRUT4.

2.2 Interpretation of our evaluation framework

Our evaluation framework determines how well models capture the internal variability and forced response in observations based on a simple metric: whether the range of well-defined simulated climate states in an ensemble adequately captures the long-term trajectory as well as the possible deviations from this trajectory caused by internal variability in the observations. To determine to which extent this evaluation metric is fulfilled, we first perform a time series and rank frequency analysis, here based on spatially averaged quantities such as GMST, and second a spatial analysis to evaluate this metric at the grid cell level. The basis of the methodological evaluation framework demonstrated in this paper was first applied to evaluate European summer temperature and precipitation in MPI-GE in Suarez-Gutierrez et al. (2018), Supporting Information Fig. S1, S2, and S3; and further expanded globally for MPI-GE in Maher et al. (2019) for annual mean temperatures, and in Suarez-Gutierrez et al. (2020b) for summer maximum temperatures, as well as in Suarez-Gutierrez et al. (2020a) for temperatures over North America in six SMILEs. Here, we extend this framework and its theoretical justification and implications, and apply it to all currently available SMILE experiments. In this section we explain the two core evaluation analyses in our framework, and elaborate on their interpretation based on idealized and specific examples.

2.2.1 Time series and rank frequency analysis

The times series and rank frequency analysis in our framework identifies different possible model biases in capturing the shape and range of the observed distribution, highlighting these biases with different rank histogram shapes. Here we illustrate these biases using synthetic data to represent idealized 50-member ensemble simulations and observations (Fig. 1). The time series (Fig. 1, left column) illustrate the temporal evolution of observations against the ensemble simulations, represented by the ensemble maxima and minima and the central 75th percentile ensemble range. In these time series we also highlight three main indicators summarizing how observations are distributed across the ensemble spread: the frequency with which observations occur above the ensemble maxima, below the ensemble minima, and within the central ensemble range. This indicators will be the base of the spatial evaluation analysis in the following section.

The rank frequency analysis (Anderson 1996; Hamill 2001; Annan and Hargreaves 2010) represents with which frequency the observed GMST anomalies take each place, or rank, in a list of ensemble members ordered by ascending GMST values for each year (Fig. 1, right column). The rank is 0 if the observed GMST value for a given year is lower than the ensemble minimum for that year, thus lower than each GMST simulated by all the ensemble members that year. If the observed value is higher than all simulated values for that year, the rank is n, with n the number of ensemble members. For a long enough observational record that is adequately simulated, observations should take all ranks with no preferred frequency, resulting in a flat rank histogram. By contrast, sloped rank histograms indicate that either the observed variability or forced response are not correctly simulated by the model.

To determine to which extent the flatness of the rank histogram and the distribution of rank frequencies can be affected by internal variability while still indicating an adequate representation of the observed distribution, we introduce the perfect model rank range. This perfect model range in the rank histograms (Fig. 1, dark dashed lines and crosses in right column panels) highlights the range of rank histogram slope fluctuations that are possible due to internal variability and to limited sample size for a model that perfectly captures observations. This is characterized by the range of rank histograms that each ensemble member would generate if it was treated as the observations (90% confidence range, 5th–95th percentiles). Thus, if the actual rank frequency slope from observations occurs substantially outside of this range for any given rank, we can robustly determine that the distribution in the model deviates from the observed distribution beyond what can be explained by internal variability, for the given sample size of the ensemble and observational record length, and therefore the ensemble does not capture observations adequately.

We exemplify this analysis here using synthetic data representing idealized model ensembles and observations with different ranges of internal variability and forced response, as characterized by the standard deviation in their distributions and the added quadratic trend terms. For comparison, we include an idealized case where the model perfectly captures observations, with both the ensemble and observations drawn from the same identical distribution (Fig. 1a). For this case of a model that perfectly captures observations, we find that observations are indeed well distributed across the ensemble spread, and occur outside of the ensemble limits only occasionally. This is then illustrated by a rank histogram shape that, although not completely flat, is well within the perfect model range. Note that this 90% confidence range may also be matched or marginally exceeded by chance even by perfectly-performing ensembles.

Beyond the perfectly-performing ensemble, we illustrate four types of individual model biases: forced response underestimation (Fig. 1b), forced response overestimation (Fig. 1c), variability underestimation (Fig. 1d) and variability overestimation (Fig. 1e) in the model compared to observations. Lastly, we also include one example of the

combined effect of two biases: variability overestimation and forced response underestimation (Fig. 1f).

For the under- and overestimated forced response examples we see, as expected, diverging warming trends in the ensembles and observations, resulting respectively in disproportionately high maximum or minimum rank frequencies, and in sloped rank histograms (Fig. 1b, c). We also see that, due to the choice of reference climatological period of 1961-1990, models that misrepresent the forced response exhibit sign-changing biases for different periods. This translates into observations clustering around the ensemble minimum in the early record, versus clustering around the ensemble maximum in the late record, or vice versa, in these cases (Fig. 1b, c). It is important to note that, in practice, this forced response does not only reflect the warming effect of increasing concentrations of green house gases, but also the cooling effect of atmospheric aerosols, which exhibit high inter-model differences and could play a role in masking or amplifying these time-dependent biases in different models (Kiehl 2007; Tokarska et al. 2020).

In contrast, for the next two examples, although models and observations evolve in time in a similar manner, they exhibit systematically different distributions due to the misrepresentation of internal variability. For an ensemble that underestimates the internal variability in observations, we find that the ensemble fails to capture the more extreme observations at the tails of the distribution. This leads to too high frequencies for both minimum and maximum ranks occurring simultaneously, resulting in a concave rank histogram shape and observations occurring beyond the ensemble limits with disproportionately high frequency (Fig. 1d). For an ensemble that overestimates internal variability we find the opposite shape, with too high mid-rank frequencies resulting in a convex rank histogram shape (Fig. 1e). In this case, observations cluster in the central 75th percentile range of the ensemble for 90% of the time, and the ensemble simulates events that are systematically more extreme than those observed. Note that this variability evaluation refers to the overall variability in annually averaged temperatures, and not to the degree of variability arising on specific timescales. Thus, this may overlook, for example, potential decadal to multidecadal variability biases in the models (e.g., such as those identified in England et al. 2014 and McGregor et al. 2014). Such biases in simulated internal variability on different timescales could be assessed with a time series and rank frequency evaluation based on moving decadal averaged temperatures or temperature trends in the corresponding timescales.

Lastly, we exemplify a model with a simultaneous overestimation of the internal variability and underestimation of the forced response. In this case, we find a diverging temporal evolution in the warming rate in observations compared to the ensemble, that is still covered by the large simulated variability range in the ensemble spread. This translates into a sloped rank histogram with too high frequency for low ranks (Fig. 1f). These two biases could be most robustly disentangled by repeating the rank frequency analysis for separate periods with different forced response contributions (i.e. earlier record, climatological reference period, and late record), or potentially by shifting the climatological reference period. For this bias combination we find that observations do not fall outside of the ensemble limits beyond what can be explained by internal variability, as opposed to the case of a model with overestimated forced response and similar internal variability (as in the individual example in Fig. 1c). This occurs due to the large internal variability in the ensemble, which is sufficient to capture the range of observations adequately despite the underestimation of the forced response. However, even if the indicators of the frequency of observed anomalies beyond the ensemble limits do not highlight a bias in this case, this bias can be robustly identified by the rank frequency analysis. The rank histogram for this example is outside of the perfect model range, and indicates that observations are not evenly distributed across the ensemble and indeed not adequately represented, highlighting the usefulness of the rank frequency analysis on more complex cases.

Note that all of these sample biases are based on normally distributed data. For non-normally distributed variables, sloped rank histograms could also indicate a discrepancy between the shape or skewness of the tails of the simulated and observed probability distributions.

2.2.2 Spatial evaluation analysis

The next line of analysis in our evaluation framework is to summarize these concepts to evaluate how different the ensembles capture the variability and forced response in observations at the grid cell level. We illustrate the interpretation of this spatial evaluation based on selected examples for the 50-member CanESM5 ensemble (Fig. 2).

First, we evaluate how often observed surface temperatures occur outside the ensemble limits in each grid cell. For a model that perfectly captures observations, how frequently observations could occur, by chance, outside of the ensemble limits depends inherently on the size of the ensemble. For a 50-member ensemble, 1-in-50 year events occur on average every simulated year, indicating that observations may exceed this limit on average roughly twice per century, or 2% of the time, due to the effect of internal variability alone (Fig. 1a). For the smallest ensemble considered here of 20 members, observations could exceed this limit by chance roughly 5% of the years on average, even for a model that perfectly captures observations.

These frequency estimations are based on averages over idealized, infinitely long records. The observational record

used in our study has a maximum length of roughly 170 years, and is restricted to much shorter lengths over certain locations (Morice et al. 2012). Therefore, how frequently observations fall outside of the ensemble can be affected substantially by internal variability, with up to twice as high frequencies as the averaged estimates even for a perfect model and a complete record (Fig. 1a), and more so for relatively shorter periods. To allow for a simpler comparison across ensembles of different sizes and simulation lengths, as well as across regions with different observational coverage, we use a fixed and slightly permissive frequency threshold of 10%. Thus, we define regions where models show a biased representation of observations when at least 10% of the observations lie below the ensemble minimum (blue shading in Fig. 2, top panel) or above the ensemble maximum (red shading in Fig. 2, top panel).

For the cases when observations occur both below and above the ensemble limits beyond this 10% threshold across the whole simulation length, this indicates that the model does not sufficiently capture the observed variability (e.g., Fig. 1d). When observations occur during specific periods mostly above the ensemble limits, and mostly below during another period, it indicates a sign-changing bias, with the model under- and overestimating the forced response respectively over these different periods (e.g., Fig. 1b and c). Note however that, depending on the length of such periods, this type of fluctuation may also occur as a result of different phases of low-frequency variability modes in observations, as well as from a changing response to aerosol forcing. Observations may also occur only above or only below the ensemble limits in some years across either the whole simulation length, pointing to a bias in the shape of the probability distribution, or concentrated during specific periods, pointing to a bias in the response to external forcings.

Second, we identify the regions where observed surface temperatures do not sufficiently cover the ensemble range and do not occur across the whole ensemble spread. For this, we determine where observations cluster within the central range of the simulated ensembles too frequently (corresponding to the central colored bounds between the 12.5th to 87.5th percentiles in Fig. 1). Ideally, for perfectly simulated internal variability and forced response, i.e., observations that occur across the ensemble with no preferred frequency and exhibit a flat rank histogram, observed values would lie within the central 75th percentile bounds of the ensemble (12.5th–87.5th percentiles) around 75% of the time. We highlight regions where the models overestimate internal variability compared to observations when observations occur in the central ensemble bounds more than 80% of the time (e.g., Fig. 1e). This bias indicates that the simulated distribution is systematically wider than the distribution of observed values, and that extremes at the tails of the simulated distribution are systematically more extreme than those observed. Note that this type of bias can only be robustly identified for periods when both the simulated and observed distributions are centered around a comparable mean state, and when evaluated over a period long enough to sufficiently cover the observed range of internal variability up to multi-decadal scales.

Lastly, the white regions represent the areas where none of these biases occur to a substantial degree. This means that in these white areas the ensemble captures observations adequately, with less than 10% of all observations occurring either above or below the ensemble limits, and less than 80% of all observations clustering within the central 75th percentile range of the ensemble. In these areas the models simulate a mean forced response and deviations from this mean that are comparable to those in observations for the whole length of their simulations (e.g., Fig. 1a). Thus, the total fraction of area where each model exhibits an unbiased representation of observations gives us an overarching ranking for our evaluation of which models most adequately capture the forced response and internal variability in observed surface temperatures.

Here we exemplify the different behaviors that arise in our evaluation with time series at the grid cell level (Fig. 2a-d) for annual mean surface temperatures at four locations (marked a-d in Fig. 2, top panel). These points are located, respectively, over the Southern Ocean (Fig. 2a), South America (Fig. 2b), the North Atlantic (Fig. 2c) and Europe (Fig. 2d).

Over the southern ocean, red and blue hatching indicates that observations occur both above and below the ensemble with too high frequencies (Fig. 2a). This behavior is caused by too little variability over that region in CanESM5, and potentially also by insufficient interannual sampling in the observations for some periods. Similarly, in the South American point observations also occur both above and below the ensemble limits with too high frequencies (Fig. 2b). However, in this case this is caused by observations that are warmer than the model simulations during the 19th and 20th centuries, but colder in recent decades (Fig. 2b). This indicates that, while the model captures the variability range adequately, it overestimates the forced warming over this region.

Over the North Atlantic we find a region of substantially overestimated variability in CanESM5 compared to observations (Fig. 2, top panel). Over this area, CanESM5 simulates surface temperature probability distributions that are systematically wider than observed, and simulated extreme anomalies are up to more than 2°C higher than the observed maxima, with observations clustering in the central ensemble range (Fig. 2c). In contrast, over Europe we find that CanESM5 does not exhibit any of the considered biases to a substantial degree (Fig. 2d). This is reflected by observations that occur with no preferred frequency across the whole

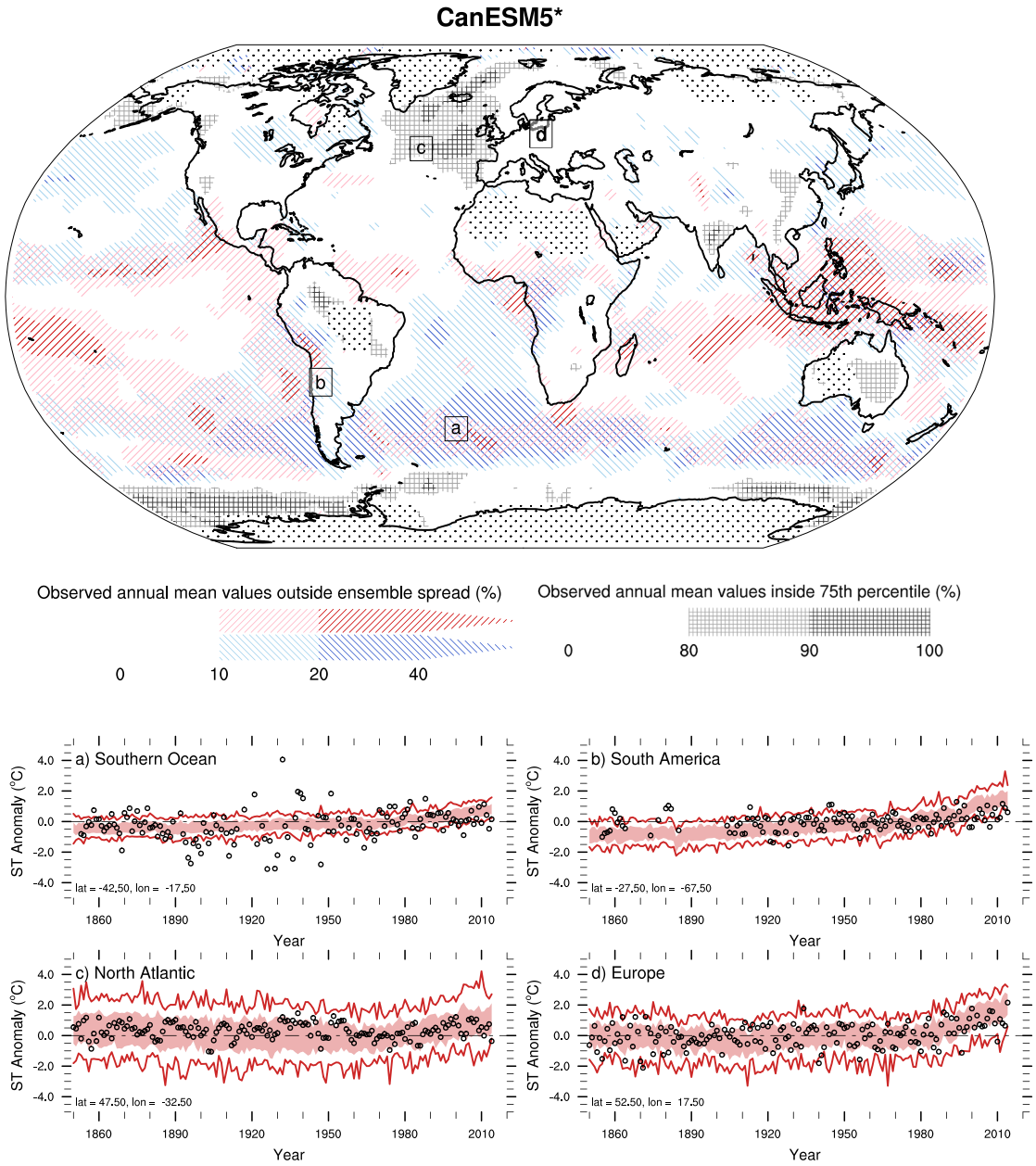


Fig. 2 Evaluation of internal variability and forced response in annual mean surface temperatures. Illustration of the evaluation of annual surface temperature (ST) anomalies at the grid-cell level in CanESM5 simulations compared to HadCRUT4 observed anomalies globally (top panel; same as top left panel in Fig. 5) and in four different grid cells (**a**–**d**). Red shading represents where observations are larger than the ensemble maximum; while blue shading represents where they are smaller than the ensemble minimum, both for more than 10% (light color) or 20% (dark color) of the time. Gray hatching represents where observations cluster within the 75th percentile bounds of the ensemble (12.5th to 87.5th percentiles) more than 80% (light color) or 90% (dark color) of the time. Dotted areas represent where observations are available for less than 10 years, and therefore excluded from our analysis. Frequencies are normalized to percentage. Panels **a**–**d** show ST time series simulated by CanESM5 (colored) against HadCRUT4 observed anomalies (black circles) for the period 1850–2014 for four grid-cells. Colored lines represent ensemble maxima and minima, shading represents the ensemble spread within the 75th percentile bounds (12.5th to 87.5th percentiles). Anomalies are relative to the period 1961–1990. Model output data are regridded to match the observational grid

simulated spread, including at its limits, but only occasionally beyond them (Fig. 2d). This indicates that CanESM5 is able to adequately capture both the forced response changes and internal variability range in annual surface temperatures in the historical record over Europe.

Note that this spatial evaluation may result in non-significant results at the grid-cell level due to the decreased signal-to-noise ratio and potential differences in observational coverage. We recommend a combination of this spatial analysis with the time series and rank frequency analysis in the previous section based on averages over multiple grid-cells for a more robust evaluation over the regions of interest.

3 Results

3.1 Global time series and rank histogram analysis

Global mean surface temperature (GMST) is arguably not only the most prominent and policy-relevant metric of climate change, but also one of the most robustly and extensively observed variables in the climate system. These reasons make GMST the ideal variable for our evaluation. GMST fluctuates around its long-term transient forced response due to internal variability, as all other variables in the climate system. Therefore, it is not appropriate to expect observations to match one single model simulation nor the ensemble mean at any given time. Ideally, for a model that adequately captures both the internal variability in the observed climate system and its response to external forcings, observations should occur across the whole ensemble spread of simulations with uniform frequency, and mostly within the ensemble limits.

To test whether models meet these two criteria of adequacy in capturing observations we perform time series and rank frequency analyses for GMST anomalies in the ensembles compared to HadCRUT4.6 (Morice et al. 2012) observations (Figs. 3 and 4). We consider GMST anomalies with respect to the climatological reference period of 1961–1990. The rank frequency analysis (Anderson 1996; Hamill 2001; Annan and Hargreaves 2010) represents with which frequency these observed GMST anomalies take each place in a list of ensemble members ordered by ascending GMST values for each year (Figs. 3 and 4, right columns). For rank 0, the observed GMST value for a given year is lower than each GMST simulated by all the ensemble members for that year; while for rank n, with n the number of ensemble members, the observed value is higher than all simulated GMST values. For a long enough observational record that is adequately simulated, GMST observations should take all ranks with no preferred frequency, resulting in a flat rank histogram. By contrast, sloped rank histograms indicate that either the variability or forced response in GMST observations are not correctly simulated by the model.

From the time series analysis we can readily determine that GMST observations generally occur within the range simulated by the ten models in our study (Figs. 3 and 4, left column). We find that the largest discrepancies occur in recent decades, with substantially higher simulated forced warming responses compared to observations for the models CanESM2, CanESM5, GFDL-CM3, and IPSL-CM5A. In contrast, CESM-LE, GFDL-ESM2M, IPSL-CM6A and MPI-GE show the best overall agreement with observations throughout the length of their simulations. CanESM2 and CanESM5 show substantially higher warming signals during recent decades than those observed, but are able to capture the observed evolution of GMST for most of the observational record adequately (Figs. 3 and 4). Due to the choice of climatological reference period (1961–1990), some models show simulated GMST warming responses similar to those observed in recent decades, but GMST anomalies that are generally higher than observations during the 19th and 20th centuries. This is the case for CSIRO-MK3.6 (Fig. 4) and MIROC6 (Fig. 3), indicating that these models exhibit less forced warming than observed over the last century.

Lastly, GFDL-CM3 and IPSL-CM5A show discrepancies during both early and recent periods; with GMSTs that are either higher or lower than those observed during the twentieth century, and substantially higher during the twenty-first century. This indicates a misrepresentation of historical forced warming throughout the length of the simulations. We repeat this evaluation for the period of 1950–2014 common to all SMILEs, shown in the Supplementary Information (SI) Fig. S.1 and S.2, to account for the different simulation lengths and the different percentage of this length covered by the climatological reference period across different models, and find comparable results for model performance. Ideally, longer simulation lengths that extend further into the past and a common climatological reference period in the earlier preindustrial period of the late 19th century would be preferred to identify biases in the forced response more robustly across different models.

The cases where observations occur outside of the ensemble limits with a higher frequency than can be attributed to internal variability can result from either biases in the simulated internal variability or in the forced response. If these occurrences outside the ensemble are distributed throughout the entirety of the record, the observed variability is either underestimated by the model, or the ensemble size is too small to cover the full range of possible climate states. However, if these occurrences are clustered around certain periods, this likely indicates that the ensemble fails to capture the observed response to external forcings. The former can be seen for the case of CESM-LE, while the clustering of low ranks for observed anomalies around recent decades can be seen for CanESM2 and CanESM5 (Figs. 3, 4). CanESM5 and CanESM2 have relatively flat rank histograms, but disproportionately large occurrences of rank 0 during recent decades. For models that have higher equilibrium climate sensitivities (ECS;

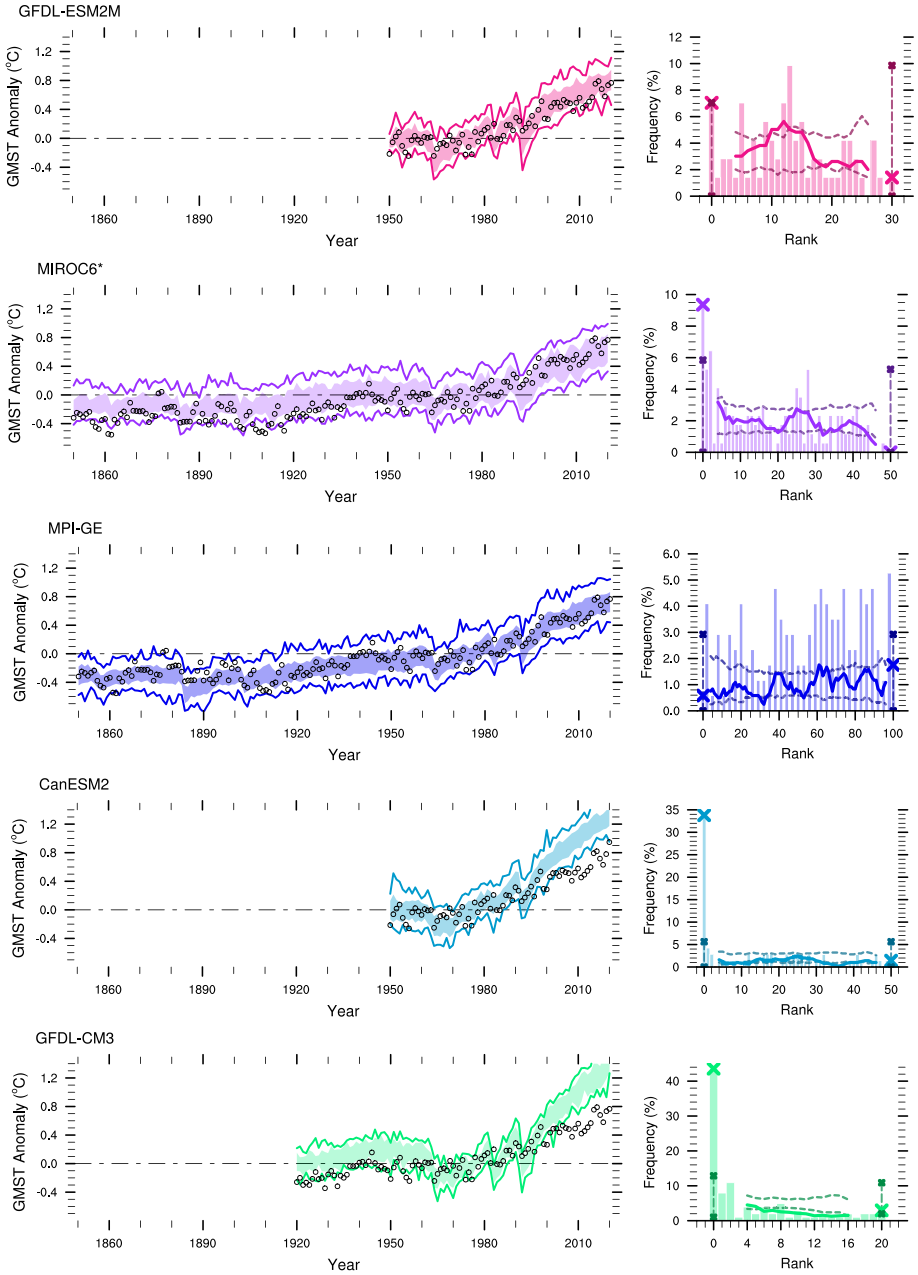


Fig. 3 Time series and rank histograms of annual GMST anomalies. Time series of annual GMST anomalies simulated by each SMILE (color) and GMST HadCRUT4 observed anomalies (black circles) ordered by increasing ECS (left column). Lines represent ensemble maxima and minima, and shading represents the ensemble spread within the 75th percentile bounds (12.5th to 87.5th percentiles). Rank histograms represent the frequency of each place that HadCRUT4 GMST observations would take in a list of ensemble members ordered by ascending GMST values (right column). Lines illustrate the rank histogram's slope, as the mean rank frequency over a centered 6-bin window for HadCRUT4 observations (solid lines, light colors), and for the 90% confidence perfect model range as the 5th to 95th percentile range of rank slopes for each ensemble member treated as observations (dashed, dark colors). Crosses represent the frequency of minimum (0) and maximum (number of members) ranks for observations (light colors), and the perfect model range as 5th to 95th percentile in frequency (dark colors). Bin sizes are 1 rank, except for MPI-GE where bin size for ranks 1 to n-1 is 3 ranks to aid visualization. Anomalies are relative to the period 1961–1990

Table 1), such as CanESM2, CanESM5, GFDL-CM3 and IPSL-CM5A (Figs. 3, 4), the high 0-rank frequencies in recent decades indicate that the ECS and forced warming response in these models is likely larger than in observations (Jiménez-de-la-Cuesta and Mauritsen 2019; Tokarska et al. 2020), resulting in discrepancies between observed and simulated GMST values.

We also find that observed GMST anomalies generally do not cluster in the central 75th percentile range of any of the simulated ensemble spreads (colored shading in Figs. 3, 4). This indicates that the observed GMST variability is not systematically overestimated by any of the models in our study. In the case that this bias were present, a concave rank histogram would indicate that the ensemble overestimates the observed variability and simulates wider probability distributions and events that are systematically more extreme than those observed (Fig. 1e). However, for shorter observational records, this may also occur due to the record not being long enough to sufficiently sample low-probability events with long return periods at the same rate as the ensemble (Keller and Hense 2011; Suarez-Gutierrez et al. 2020b).

We introduce perfect-model rank frequency tests (Data and Methods Sect. 2.2.1) to assess how the flatness of the rank histogram and the distribution of rank frequencies can be affected by internal variability. For this, we treat each

ensemble member as if it were the observations, and compute the rank histogram of each singular ensemble member against the new n-1 ensemble. We can then identify the range where the slope of the observational rank histogram for a model that perfectly captures observations may fluctuate due to the effect of internal variability for a given model and simulation length. If the rank frequencies from observations occur substantially outside of this range for any ranks, we can determine that the distribution in the model deviates from observations beyond what can be explained by internal variability and therefore does not capture observations adequately. We find that most ensembles, including CanESM2, CanESM5, CSIRO.MK3.6, GFDL-ESM3, IPSL-CM5A, and MIROC6, exceed this range due to too high rank 0 and rank n frequencies that are caused by biases in the forced response beyond what can be attributed to internal variability.

In contrast, only four ensembles offer rank histograms sufficiently within their perfect-model range to determine that they provide an adequate representation of observations: CESM-LE, GFDL-ESM2M, IPSL-CM6A, and MPI-GE. The ensembles CESM-LE (Fig. 4) and GFDL-ESM2M (Fig. 3) exhibit rank histograms marginally outside of this range, which could happen by chance (e.g., Fig. 1a) and could be attributed to the relatively short simulation length (Keller and Hense 2011). Longer simulation lengths and larger ensemble sizes would be beneficial to robustly determine whether CESM-LE and GFDL-ESM2M are indeed capturing the observed internal variability and forced response in GMST adequately.

For IPSL-CM6A (Fig. 4) our results show that the rank histogram of observations is well within the range of deviations caused by internal variability. This is also true when the climatological reference period is shifted to the nineteenth century (not shown) and when only the 1950–2014 period is considered (SI Fig. S.2). However, the high ECS in IPSL-CM6A combined with the fact that observations occur always in the lower half of the ensemble for the last two decades points toward a potential forced response bias in the model that may be identifiable only when more scenario data and future observations become available. Lastly, MPI-GE presents both a reasonably flat rank histogram and a low frequency of occurrences outside the ensemble limits both within the perfect-model ranges (Fig. 3). This indicates that, from the SMILEs considered in our study, MPI-GE captures observed GMSTs most adequately, performing well in simulating both the internal variability and forced response in observed GMST during the entirety of the period of 1850–2019.

Our findings reveal that while an overestimated forced warming response in several models causes the largest discrepancies between observed and simulated GMSTs, internal variability in GMSTs is not systematically over- or underestimated by any of the models considered. Two of the four ensembles with the most adequate GMST representation, GFDL-ESM2M and MPI-GE, are also two of the three ensembles from models with the lower ECS (2.4 K and 2.8 K, respectively; Table 1). In contrast, the other two ensembles with adequate GMST representation have higher ECS, of 4.1 K for CESM-LE and 4.5 K for IPSL-CM6A. These values are well within the range of sensitivities of models with substantially overestimated forced warming, and indicate that observed historical surface temperatures can be reproduced with a high-ECS model. Note however that this relationship between warming during the historical record and climate sensitivity can be substantially affected by the large uncertainty in aerosol cooling (Kiehl 2007; Tokarska et al. 2020).

3.2 Where climate models perform well

The next aspect of our evaluation framework is to apply these concepts to evaluate how different climate models capture the observed variability and forced response in annual surface temperatures at the grid cell level (Fig. 5).In this analysis we highlight four different metrics to assess how observations are distributed across the simulated ensemble spread at each grid cell (Data and Methods Section 2.2.2). First, we evaluate how often observed surface temperatures occur beyond the ensemble limits, and define regions where models show a biased representation of observations when at least 10% of the observations lie below the ensemble minimum (blue shading in Fig. 5) or above the ensemble maximum (red shading in Fig. 5). Second, we identify the regions where observed surface temperatures do not sufficiently cover the whole ensemble spread. For this, we determine that when observations cluster within the central 75th percentile range of the simulated ensembles (corresponding to the central colored bounds in Figs. 3 and 4) more than 80% of the time, it indicates that the models overestimate internal variability compared to observations in the area. This bias implies that the simulated distribution is systematically wider than the distribution of observed values, and that extremes at tails of the simulated distribution are systematically more extreme than those observed. The effect of choosing a more permissive threshold of 85% for this variability overestimation bias can be seen in SI Fig. S.3.

Lastly, the white regions represent where none of these biases occur to a substantial degree. This means that in these areas the ensemble captures observations adequately, with less than 10% of all observations occurring either above or below the ensemble limits, and less than 80% of all observations clustering within the central 75th percentile range of the ensemble. Thus, the total fraction of white area for each mode represents over how much area the models simulate a mean forced response and deviations from this mean that are comparable to observations for the whole length of their

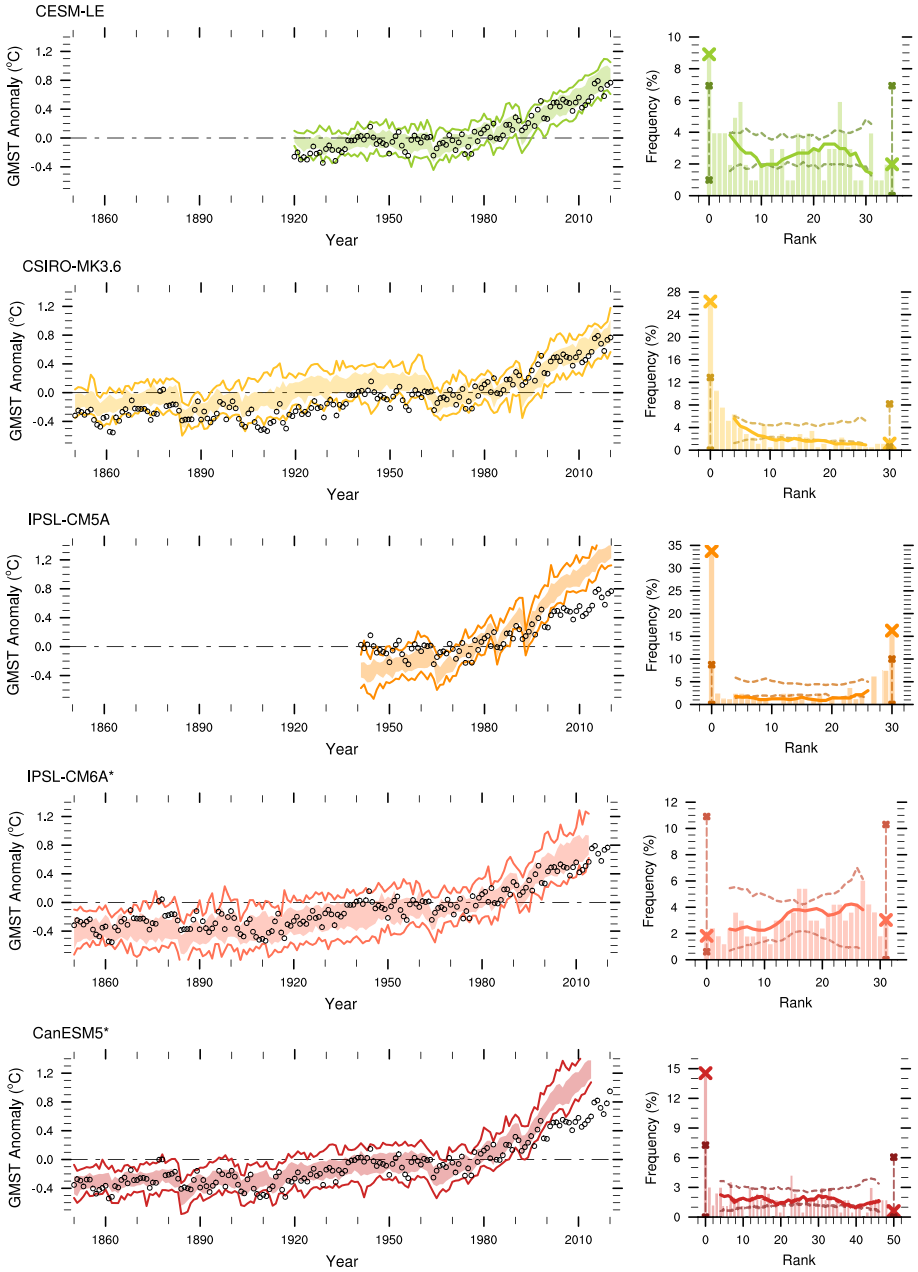


Fig. 4 Time series and rank histograms of annual GMST anomalies, continued. Time series of annual GMST anomalies simulated by each SMILE (color) and GMST HadCRUT4 observed anomalies (black circles) ordered by increasing ECS (left column). Lines represent ensemble maxima and minima, and shading represents the ensemble spread within the 75th percentile bounds (12.5th to 87.5th percentiles). Rank histograms represent the frequency of each place that HadCRUT4 GMST observations would take in a list of ensemble members ordered by ascending GMST values (right column). Lines illustrate the rank histogram's slope, as the mean rank frequency over a centered 6-bin window for HadCRUT4 observations (solid lines, light colors), and for the 90% confidence perfect model range as the 5th to 95th percentile range of rank slopes for each ensemble member treated as observations (dashed, dark colors). Crosses represent the frequency of minimum (0) and maximum (number of members) ranks for observations (light colors), and the perfect model range as 5th to 95th percentile in frequency (dark colors). Bin sizes are 1 rank. Anomalies are relative to the period 1961–1990

simulations, and gives us an overarching ranking for our evaluation.

In this spatial evaluation we find that observations tend to occur outside the ensemble limits more frequently, and over larger areas, than they tend to cluster in the central range of the ensembles (Fig. 5). This indicates that the ensemble spreads of most models do not sufficiently capture the range of observed surface temperatures over large areas. This is either because they fail to capture the forced response, or because they insufficiently sample or underestimate internal variability on regional scales. In contrast, over smaller areas, some models simulate both positive and negative surface temperature anomalies in any given year that are systematically more extreme than those observed, indicating that they overestimate the observed internal variability in these regions. Such a clustering of the observations in the center of the ensembles indicates that an overestimation of internal variability occurs for several models over land-surface areas in India, South East Asia, or Central North and South America, as well as near the sea-ice edges in high-latitude oceans, over the North Atlantic, and over parts of the tropical Pacific and Indian Oceans. By contrast, observations occur outside the ensemble spreads with high frequency for all models over the Southern Ocean, and for some models also over the Southern Hemisphere oceans or the Maritime Continent.

This evaluation shows that surface temperatures over the Southern Ocean are not adequately simulated by any of the models considered in our study, even with the better sampling of internal variability provided by SMILEs (Fig. 5). Over large areas, more than 40% of the observed anomalies occur both above and particularly below the ensemble limits for all models. The fact that observations tend to occur predominantly below the ensemble minimum, in combination with simulated temperatures being generally higher than observed in the region in recent decades (SI Fig. S.9), indicates that some of the discrepancies may be explained by models warming more and at a faster pace than is observed over the Southern Ocean. We find that this warming bias, previously identified in individual simulations in the CMIP5 multimodel ensemble (Hyder et al. 2018), remains outside of the range of the better-sampled internal variability in SMILE experiments, and it is still present in the CMIP6 models considered.

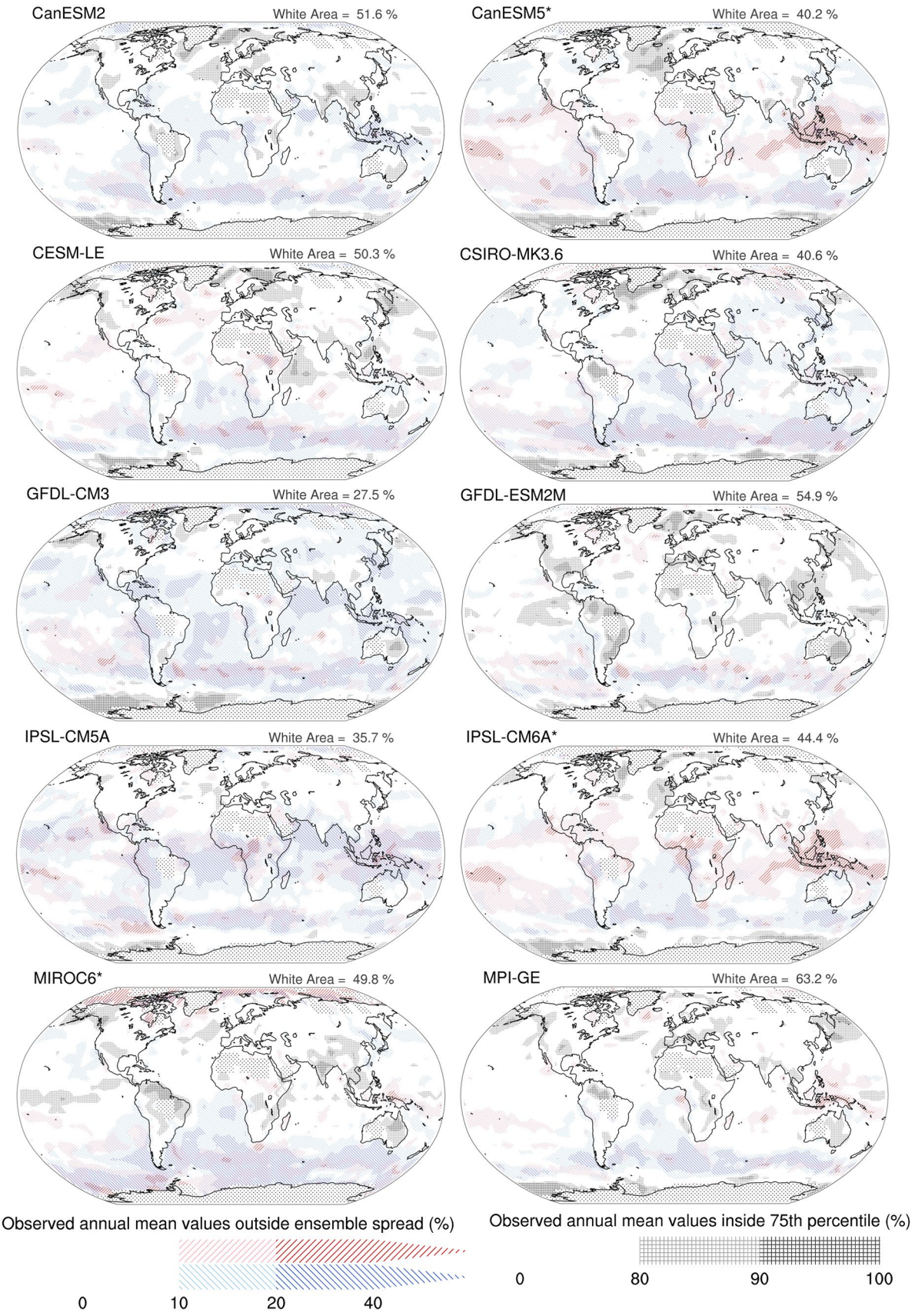
In addition to this potential warming bias, an underestimation of variability over the Southern Ocean may also be a contributing factor to these discrepancies. This could result from the relatively coarse model spatial resolution in all SMILEs considered here, which does not allow an explicit simulation of ocean eddies. A poor representation of ocean eddies can result in an underestimation of the observed variability in surface temperatures through an incorrect representation of eddy mixing and cold water upwelling (Screen et al. 2009; Frenger et al. 2015). Moreover, coarse resolution can further affect the surface temperature mean state and variability through an incorrect representation of the ocean-sea-ice interactions, wind location, deep water formation, and the strength and spatial variability of the overturning circulation of the Southern Ocean (Stössel et al. 2015; Gutjahr et al. 2019).

Lastly, the observational sparsity in the Southern Ocean region may also be a contributing factor. In-situ observations over the Southern Ocean are both spatially and temporally sparse, clustering in the austral summer months and in different regions over different periods (Fig. 2a). The inconsistent sampling of different phases of variability in observed surface temperatures over the Southern Ocean may result in a range of observed variability that is artificially reduced, thus hiding even larger discrepancies than shown here. However, this inconsistent sampling could also result in an overestimation of the observed annual variability caused by annual averages that are biased due to unbalanced sampling across different seasons (Fig. 2a). Longer and better-sampled observational records combined with SMILEs of higher resolution eddy-resolving climate models may be required to robustly determine the source of the remaining discrepancies between observed and simulated surface temperatures over the Southern Ocean.

Our evaluation shows that the ensembles with the most adequate representations of surface temperatures on regional scales, as measured by the largest area of no substantial biases, are MPI-GE, followed by GFDL-ESM2M, CanESM2, CESM-LE, and MIROC6 (Fig. 5). To account for the effect of the double to almost triple ensemble size of the 100-member MPI-GE compared to the other ensembles considered here, we replicate this analysis for ensemble sizes limited to the first 30 ensemble members (SI Fig. S.4). This evaluation for comparable ensemble sizes still highlights a similar set of SMILEs as exhibiting the largest fraction of non-biased area, but in a now different order. The MPI-GE limited to its first 30 members exhibits larger areas where the ensemble does not sufficiently cover the range of variability in observed surface temperatures, particularly in the Southern Hemisphere. By limiting the ensemble size to a comparable range, the 30-member GFDL-ESM2M ensemble surpasses MPI-GE in fraction of unbiased area by roughly 6%, and offers the largest area of adequate representation of surface temperatures, followed by CESM-LE and MPI-GE.

From the ensembles with the most adequate representation of observed temperatures, the one that most frequently overestimates the observed variability in surface temperatures is the one with the fewest members, GFDL-ESM2M (Fig. 5). This result stands in contrast to the expectation that larger ensemble sizes may result in larger ensemble spreads. We find that whereas most of the models considered show similar magnitude and patterns of internal variability, GFDL-ESM2M shows more variability over the low and mid latitude land areas than most other models (SI Fig. S.5). This result also holds when the ensemble size differences are accounted for (SI Fig. S.6). GFDL-ESM2M simulates temperature distributions that are systematically too wide, with observations clustering in the center of the simulated distribution, over large land areas, such as India, South East Asia or the majority of America, but also over the oceans. This behavior occurs for MIROC6 and CanESM2 over relatively large land-surface areas, for CESM-LE over smaller regions over the Indian and Pacific Oceans and the high latitudes, and is even less marked for MPI-GE (Fig. 5). A more in-depth comparison of the patterns and magnitude of the simulated internal variability in different SMILEs and the effect of ensemble size can be found in the Supplementary Information (SI Fig. S.5 and S.6).

For the rest of the models evaluated, we find that observed annual mean surface temperature anomalies fall with high frequency below the ensemble minima over large regions (Fig. 5). This is the case particularly for CanESM2, CESM-LE, CSIRO-MK3.6, GFDL-CM3, and IPSL-CM5A. This occurs over smaller areas for the CMIP6 models CanESM5, IPSL-CM6A and MIROC6, which show observations lower than the ensemble minima clustered around the Southern Ocean and South Atlantic. For CanESM5 and IPSL-CM6A,


CanESM2
White Area = 51.6 %
CanESM5*
White Area = 40.2 %
CESM-LE
White Area = 50.3 %
CSIRO-MK3.6
White Area = 40.6 %
GFDL-CM3
White Area = 27.5 %
GFDL-ESM2M
White Area = 54.9 %
IPSL-CM5A
White Area = 35.7 %
IPSL-CM6A*
White Area = 44.4 %
MIROC6*
White Area = 49.8 %
MPI-GE
White Area = 63.2 %
Observed annual mean values outside ensemble spread (%)
0
10
20
40
Observed annual mean values inside 75th percentile (%)
0
80
90
100

◂**Fig. 5** Evaluation of internal variability and forced response in annual surface temperatures. Evaluation of annual surface temperature anomalies simulated by different SMILEs compared to HadCRUT4 observed anomalies. Red shading represents where observed anomalies are larger than the ensemble maximum; while blue shading represents where observed anomalies are smaller than the ensemble minimum, both for more than 10% (light color) or 20% (dark color) of the time. Gray hatching represents where observations cluster within the 75th percentile bounds of the ensembles (12.5th to 87.5th percentiles) more than 80% (light color) or 90% (dark color) of the time. White Area represents the percentage of total area included in the analysis that exhibits no substantial biases for each SMILE. Dotted areas represent where observations are available for less than 10 years, and therefore excluded from our analysis. Anomalies are relative to the period 1961–1990. Model output is regridded to match the observational grid

as well as for IPSL-CM5A, observations occur above all ensemble members over large areas of the tropical oceans and the Maritime Continent. For MIROC6, observations are higher than the ensemble maxima over the northern polar regions. This behavior is not present in any other of the models considered, and may arise due to an overestimation of ice cover illustrated by a combination of low variability (SI Fig. S.5 and S.6) and a cold mean bias in recent decades (SI Fig. S.9) in MIROC6 over the northern polar region.

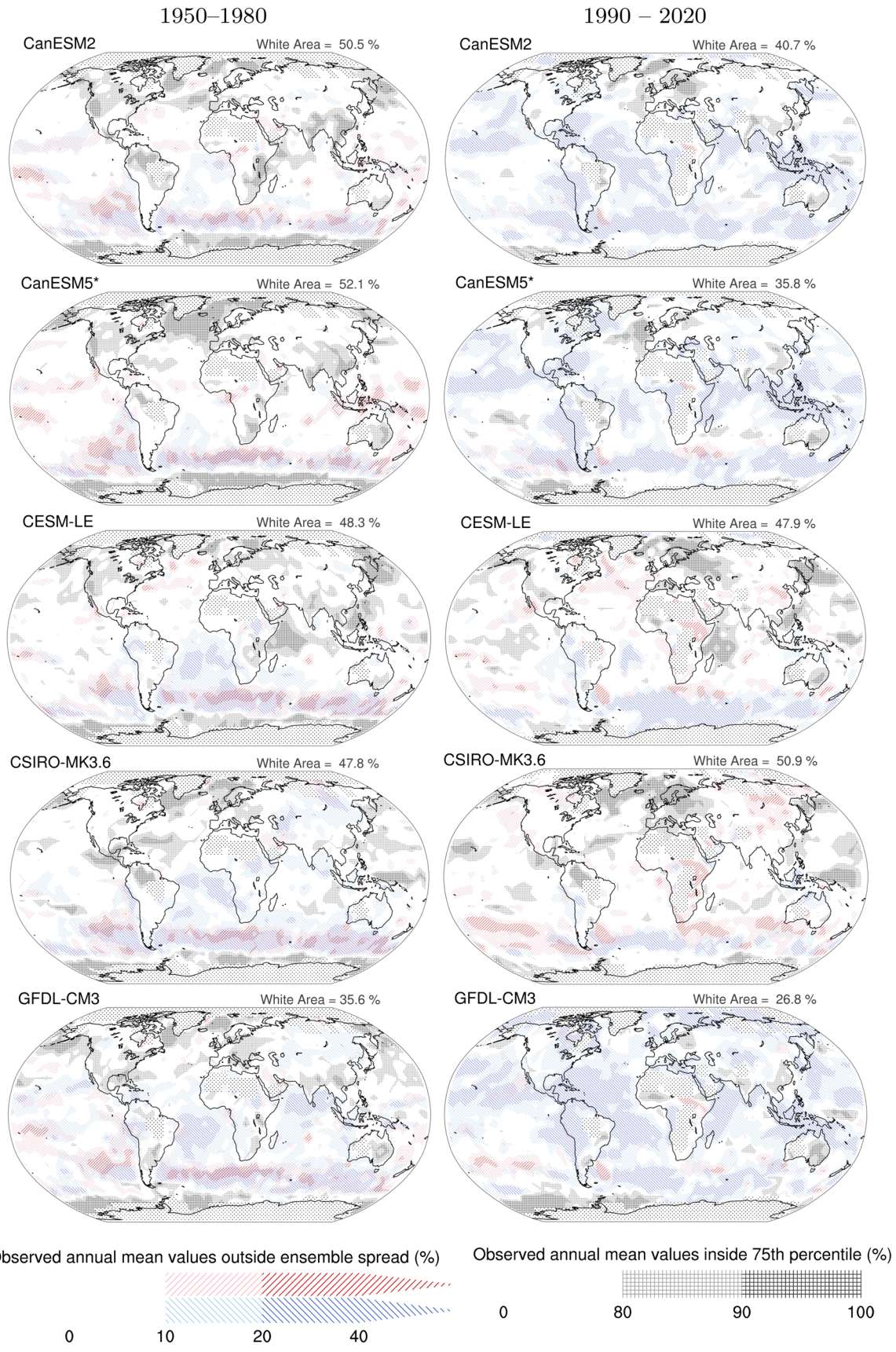
3.3 Disentangling forced response and internal variability biases

In this section we refine our analysis to disentangle forced response biases from biases in internal variability in our spatial evaluation. A high frequency of occurrence of observations outside the ensemble spread evaluated during the whole observational record may indicate either a bias in the forced response or in internal variability. To distinguish between these effects, we repeat this analysis both for recent decades (1990–2020) and for a period of the same length earlier in the 20th century (1950–1980; Figs. 6 and 7). Although other effects such as differences in aerosol forcing and respective cooling effects in the observations (Kiehl 2007; Tokarska et al. 2020) and decadal to multidecadal variability in the observations (England et al. 2014; McGregor et al. 2014) may generate differences between these two periods and cannot be excluded without further analysis, we expect that biases that originate from a incorrect representation of internal variability are comparable between the two periods, whereas biases that originate from an incorrect representation of the forced response are period-dependent. Because the forced response depends on time-changing external forcings, models that overestimate this response in recent decades may show no biases in earlier periods, or may show biases that change in sign, i.e., simulated temperature anomalies that are lower than observed before the climatological reference period, and higher afterwards. Note however that period-depending cooling aerosol forcings in the models may also affect these time-changing biases, and could compete with the warming greenhouse gas forcing, leading to masked biases.

On the other hand, the cases where observations still occur consistently both below and above the ensemble limits also during these shorter periods indicate an underestimation of internal variability. Ideally we would use an even earlier period as a proxy for pre-industrial conditions in this comparison. However, we restrict our analysis to the period starting in 1950 to ensure contributions from all SMILEs. Note that during these roughly 30-year periods, observations appear to cluster in the center of the ensembles with too high frequencies for more models and over larger areas, because these periods are not sufficiently long to identify the variability overestimation bias in annual surface temperatures robustly.

This period-based analysis highlights the models with the largest differences between these two periods as those most affected by biases in the forced response: CanESM2, CanESM5, GFDL-CM3, and IPSL-CM5A (Figs. 6, 7). These models showing period-dependent biases in the forced response also overestimate the observed warming response in GMST in recent decades (Figs. 3, 4). In the early period, CanESM2 and CanESM5 (Fig. 6) exhibit percentages of unbiased area that are comparable to those from other models with adequate representations of surface temperatures during the whole observational record, such as GFDL-ESM2M (Fig. 5). They exhibit some regions of overestimated variability, as well as observations occurring both above and below the ensemble limits in the Southern Ocean (Fig. 6). In the later period, this percentage drops substantially and observed temperatures are lower than the ensemble minima of CanESM2 and CanESM5 almost over all locations on the globe (Fig. 6). This drastic change between the two periods indicates that these two models are sufficiently able to capture the variability range in observed surface temperatures under relatively low levels of historical global warming, but simulate surface temperatures systematically higher than those observed under higher historical warming levels.

For CanESM5, with longer simulations starting in 1850, we also identify high percentages of observations that are warmer than all ensemble members during the late ninteenth and early twentieth Centuries, in particular over the tropics and Southern Hemisphere oceans (not shown). Therefore, additionally to the higher than observed temperature anomalies in recent decades, this model also shows lower than observed temperature anomalies in the early historical period, before the climatological reference period. This sign-changing bias behavior is also present for IPSL-CM5A (Fig. 7) and to some extent for CSIRO-MK3.6, GFDL-CM3 (Fig. 6) and IPSL-CM6A (Fig. 7). With the exception of


1950–1980
1990 – 2020
CanESM2
White Area = 50.5 %
CanESM2
White Area = 40.7 %
CanESM5*
White Area = 52.1 %
CanESM5*
White Area = 35.8 %
CESM-LE
White Area = 48.3 %
CESM-LE
White Area = 47.9 %
CSIRO-MK3.6
White Area = 47.8 %
CSIRO-MK3.6
White Area = 50.9 %
GFDL-CM3
White Area = 35.6 %
GFDL-CM3
White Area = 26.8 %
Observed annual mean values outside ensemble spread (%)
0
10
20
40
Observed annual mean values inside 75th percentile (%)
0
80
90
100

◂**Fig. 6** Evaluation of internal variability and forced response in surface temperatures for different periods. Evaluation of annual surface temperature anomalies simulated by different SMILEs compared to HadCRUT4 observed anomalies as in Fig. 5, for the period of 1950–1980 (left column) and 1990–2020 (right column), except for CMIP6 SMILE CanESM5, which ends in 2014. Anomalies are relative to the period 1961–1990

the Southern Ocean, this indicates that in the regions where observations occur both above and below the ensemble limits for these models when analyzing the whole observational record (Fig. 5), it is due to the misrepresentation of the observed forced warming by the models, and not due to an underestimation of internal variability. For CSIRO-MK3.6, we find that although the percentage of white area does not change substantially between the two periods, the bias patterns change substantially from observations occurring mostly below the ensemble in the early period, to above the ensemble in the recent decades, highlighting forced response biases. For IPSL-CM6A, we also find high percentages of white areas for both periods, higher than when the whole observational record is considered (Fig. 5), that decrease in the 19th and early 20th centuries (not shown), indicating a potential forced response bias in the model.

The ensemble MPI-GE, and to some extend also CESM-LE, GFDL-ESM2M, and MIROC6, exhibit similar biases for both periods over similar regions (Figs. 6, 7). This indicates that the biases for these models shown in Fig. 5 are largely not period-dependent, and therefore not dominated by errors in the forced response; but rather dominated by biases in the shape of the simulated probability distributions caused by over- or underestimations in the simulated internal variability. Our findings highlight that most models offer a reasonable representation of internal variability over large areas, although they exhibit similar biases over similar regions: an overestimation of the observed temperature variability over land-surface areas such as Central South America, and an underestimation of the observed variability over the Southern Ocean. In contrast, forced response biases occur over much larger areas, and are the dominating source of discrepancies between several models and observations also on regional scales.

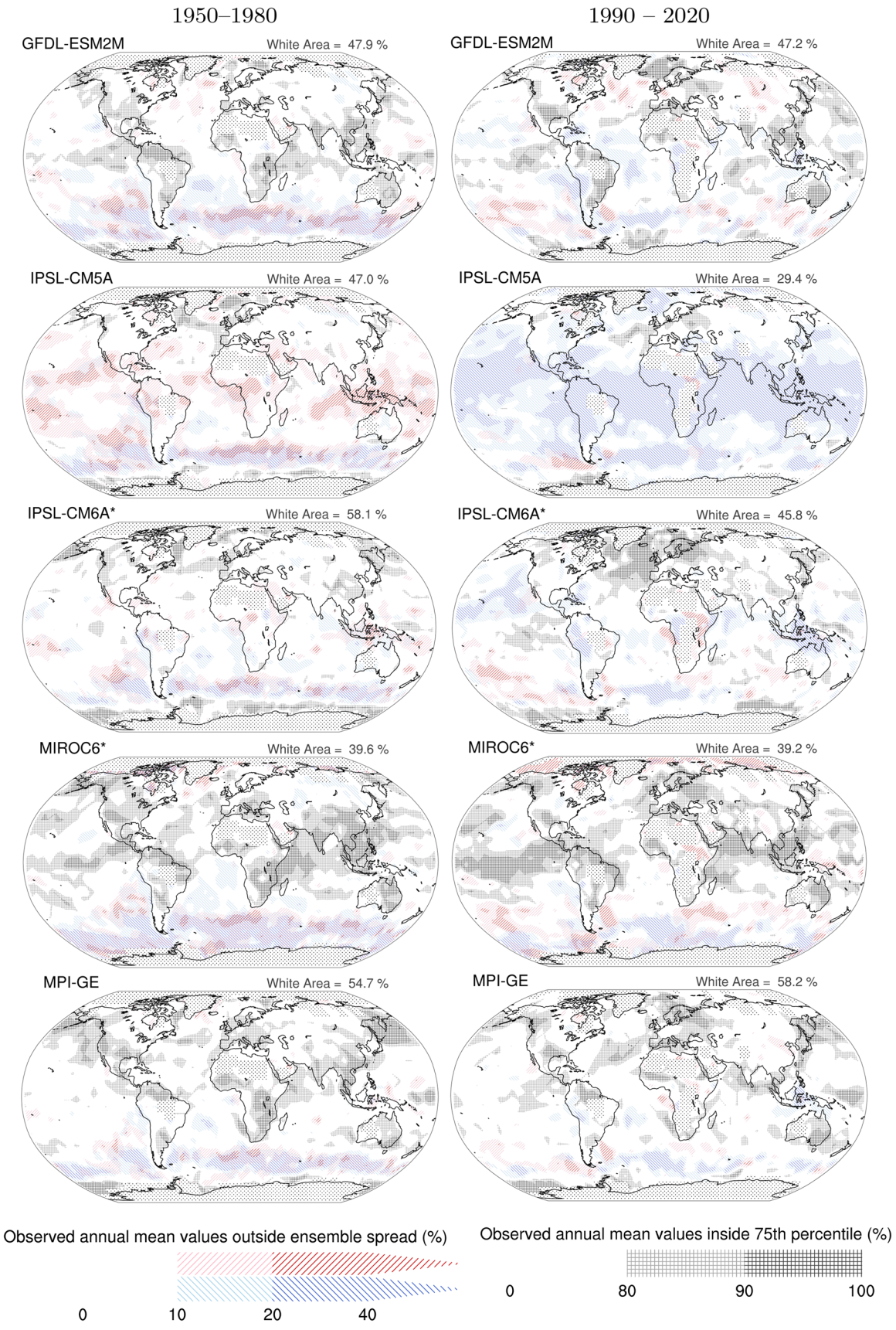
3.4 How many climate models adequately capture observations

We summarize these results to identify regions where most models adequately capture both the variability and forced response in surface temperatures for each region during the entirety of the observational record, versus regions where most models do not (Fig. 8a). We define regions where the ensembles capture observations adequately as areas where they do not show any of the considered biases to a substantial degree. This includes that less than 10% of all observations occur either above or below the ensemble limits, and less than 80% of all observations occur within the central 75th percentile bounds of the ensemble (white areas in Fig. 5). We find that the highest number of models that adequately simulate observed surface temperatures according to these criteria, a maximum of nine out of ten models, do so over the North Atlantic, Tropical Eastern Pacific, and northern mid to high latitude land surfaces. Over the Southern Ocean, none of the climate models considered in our study succeeds at adequately simulating surface temperatures; while fewer than three models offer adequate surface temperature simulations over the South Atlantic or the Maritime Continent.

We repeat this evaluation again for two different periods during the 20th century (1950–1980; Fig. 8b), and during recent decades (1990–2020; Fig. 8c). By doing so, we can distinguish between models that can adequately capture the variability and forced response in observed surface temperatures under conditions comparable to a pre-industrial state, versus models that do so under higher atmospheric concentrations of greenhouse gases. We find that substantially fewer models offer adequate simulations of surface temperatures in recent decades compared to both 1950–1980 (Fig. 8d), and to the whole observational record (Fig. 8a). This indicates that the overestimated forced warming response in several models substantially reduces the number of models that adequately simulate surface temperatures at regional scales, particularly over Northern Europe or the North Atlantic and Pacific Oceans (Fig. 8d). To a lesser extent, we also find that the number of models that capture the observed surface temperature increases over the South Atlantic, Southern Ocean and the sea-ice covered regions around Antarctica in recent decades (Fig. 8d), likely due to recent observational sampling improvements in these areas. In contrast, a similar number of models simulate surface temperature adequately during both periods over the South Atlantic, Indian and Southern Oceans, and South Asia. Our results show that there is no region of the globe where all models offer an adequate simulation of the internal variability and forced response in observed annually averaged surface temperatures for the entire observational record.

4 The importance of robustly evaluating internal variability

By combining SMILE experiments with our rank-based evaluation framework we can for the first time determine whether the range of well-sampled internal variability in a model is adequate, or rather an under- or overestimation of the variability in observations. Therefore, our framework allows us to determine, more robustly than ever before, to what extent the range of events that are possible under


1950–1980
1990 – 2020
GFDL-ESM2M
White Area = 47.9 %
GFDL-ESM2M
White Area = 47.2 %
IPSL-CM5A
White Area = 47.0 %
IPSL-CM5A
White Area = 29.4 %
IPSL-CM6A*
White Area = 58.1 %
IPSL-CM6A*
White Area = 45.8 %
MIROC6*
White Area = 39.6 %
MIROC6*
White Area = 39.2 %
MPI-GE
White Area = 54.7 %
MPI-GE
White Area = 58.2 %
Observed annual mean values outside ensemble spread (%)
0
10
20
40
Observed annual mean values inside 75th percentile (%)
0
80
90
100

◂**Fig. 7** Evaluation of internal variability and forced response in surface temperatures for different periods, continued. Evaluation of annual surface temperature anomalies simulated by different SMILEs compared to HadCRUT4 observed anomalies as in Fig. 5, for the period of 1950–1980 (left column) and 1990–2020 (right column), except for CMIP6 SMILE IPSL-CM6A, which ends in 2014. Anomalies are relative to the period 1961–1990

specific climatic conditions in a model represents the range of events that would be possible in the real world, including low-probability extreme events. This is useful not only for evaluating the reliability of the range of events in future projections; but also to robustly determine the likelihood of events that have occurred in the past, either under real-world or alternative forcing conditions. This makes evaluations such as ours crucial for further improving upon current detection and attribution efforts. Ultimately, we can now determine which model ensembles simulate an internal variability that is closest to the internal variability in the real world, thus providing us with a choice of simulated proxies for the unobservable real-world internal variability. In turn, this allows us to improve our understanding of the real-world internal variability by studying these proxies in a setting with the cleanest and most robust analysis of this signal currently possible.

An important highlight of our evaluation is that while the forced response in annual surface temperature varies widely across models, most SMILEs present fairly similar patterns and magnitudes of internal variability. They show larger variability over land than over the oceans, and in the Northern Hemisphere than in the Southern Hemisphere. IPSL-CM5A exhibits the highest variability, followed by GFDL-CM3 and GFDL-ESM2M. MIROC6 exhibits the lowest variability over the high latitudes, while exhibiting a mid- and low-latitude variability range comparable to other ensembles. Furthermore, both variability as well as performance in representing the range of observed temperatures increase when including more ensemble members for all models, indicating that 30 members is not sufficient to offer fully saturated ensemble spreads, even for relatively low variability magnitudes such as for annually averaged surface temperatures. However, we do not find a relationship between larger ensemble sizes and larger internal variability across models; and the largest ensemble, MPI-GE, does not exhibit the largest variability when its 100 available ensemble members are included. This means that, beyond certain quasi-saturation limits, increasing ensemble size does not necessarily increase the range of internal variability. Therefore, our results indicate that internal variability appears to be a model property independent of ensemble size, beyond a certain number of ensemble members.

5 Conclusions

Using a robust yet simple model evaluation framework, we assess how climate models represent the internal variability and response to external forcings in observed historical temperatures. With this methodological framework, we can exploit the power of SMILE experiments to determine whether real-world observations are well distributed within the now well-sampled range of climate states simulated by each model. This allows us to attribute discrepancies between model simulations and observations to either biases in the simulated forced response or in the simulated internal variability, without the need to separate both signals in the observations. Thus, we can now determine whether comprehensive climate models capture the long-term trajectory of the climate system, as well as the range of possible fluctuations from this trajectory caused by internal variability in any given region and time period. Such an evaluation, both in terms of a model's forced response and range of internal variability, allows us to assess model performance more robustly than ever before, and thus to appropriately select which models are the best fit for different analysis in different regions of the globe, for studying current and past climate states, as well as future climate projections (Krinner and Flanner 2018).

Our evaluation of global mean surface temperatures shows that while some models fail to capture the long-term response to external forcing, none of them systematically under- or overestimate the range of internal variability in GMST. Most models show good agreement between the ranges of simulated and observed GMST, but several models show warming signals substantially higher than observed during recent decades. Thus, our findings indicate that these models, namely CanESM2, CanESM5, GFDL-CM3, and IPSL-CM5A, overestimate recent forced warming (Jiménez-de-la Cuesta and Mauritsen 2019; Tokarska et al. 2020) beyond the range of plausible fluctuations caused by internal variability. From all SMILEs, the 100-member MPI-GE offers the most adequate representation of both the internal variability and forced response in observed GMST during the entire historical record, followed by IPSL-CM6A, CESM-LE, and GFDL-ESM2M. Two of the models that capture GMST most adequately, GFDL-ESM2M and MPI-GE, are also two of the three models with the lowest ECS values. In contrast, CESM-LE and IPSL-CM6A illustrate that models with a much higher ECS can still adequately capture the observed historical surface temperatures.

Using our evaluation framework we can directly identify regions where models under- or overestimate internal variability, as well as where they exhibit regional biases in the forced response compared to observations. Models capture

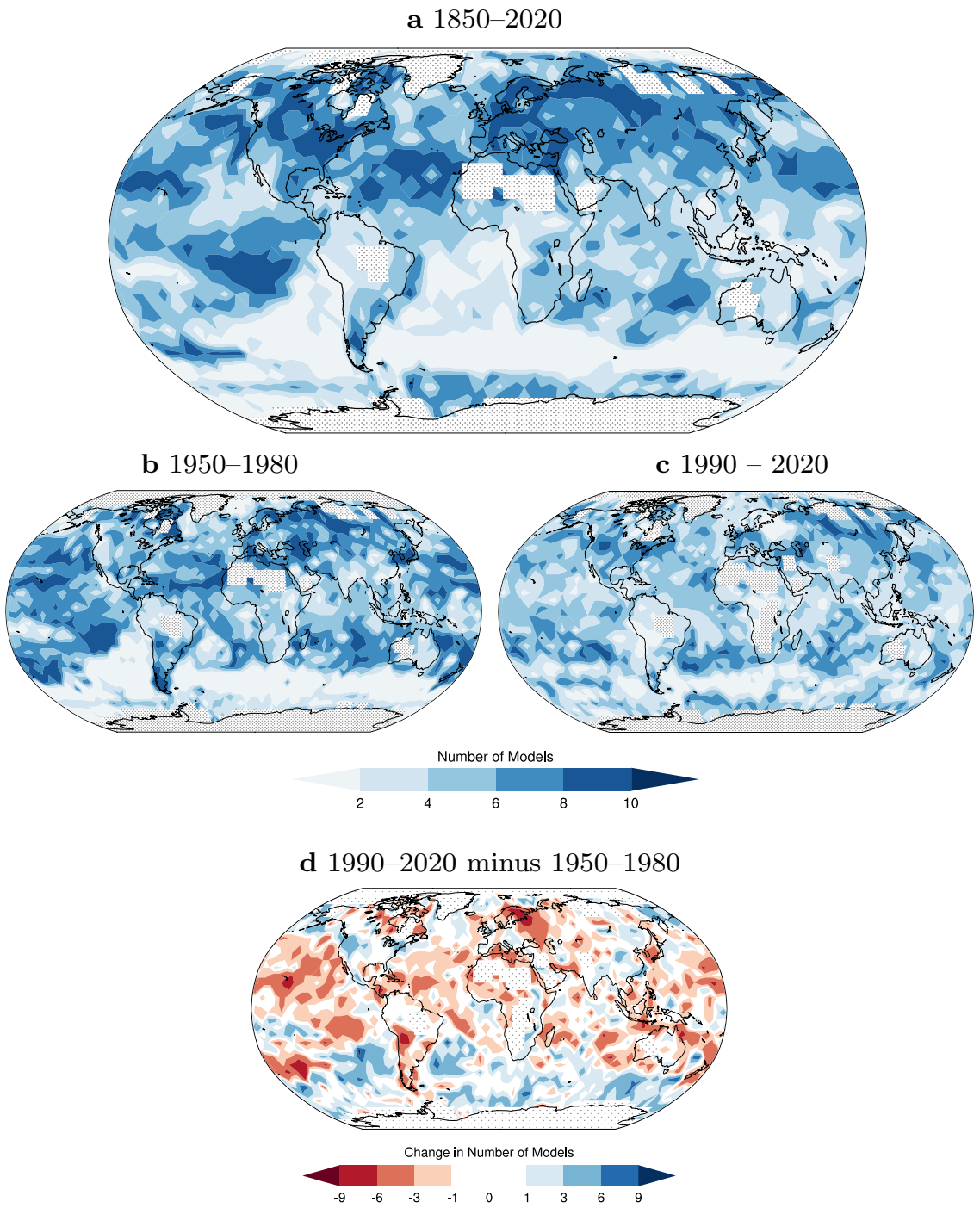


Fig. 8 Number of models that adequately represent observed surface temperatures. Number of models that adequately represent the combined effect of internal variability and forced response in annual surface temperature HadCRUT4 observed anomalies for the period of **a** 1850–2020, **b** 1950–1980, and **c** 1990–2020, and **d** the difference between the number of models in 1990–2020 minus in 1950–1980. We consider that models capture observations adequately when less than 10% of all observed anomalies fall either above or below the ensemble limits, while less than 80% of all observed anomalies fall within the central 75th percentile bounds of the ensemble, for each grid cell. Dotted areas represent regions where observations are available for less than 10 years

the range of observed temperatures adequately over land in the Northern Hemisphere; while capturing this range less adequately near the sea ice edges and over the Southern Ocean. Observations occur outside the ensemble limits over most of the globe, and in particular over the Southern Hemisphere oceans. In contrast, observations cluster in the central bounds of the ensembles indicating overestimated variability over similar regions for most models: the sea ice edges

near the poles, and the low and middle latitude land-surface areas. Our findings show that observations tend to occur outside the ensemble limits more than they tend to cluster in the central bounds of the ensembles. This means that models fail to capture observations due to forced response biases or underestimated variability more frequently, and over larger areas, than they overestimate this variability due to simulated extremes that are systematically more intense than those observed. Therefore, simulated annual surface temperature extremes are less likely biased due to models overestimating their intensity, and more likely biased due to models underestimating their intensity and misrepresenting forced changes.

On regional scales, the ensembles MPI-GE, GFDL-ESM2M, MIROC6, and CESM-LE capture the observed variability and forced response in historical surface temperatures most adequately, both in early as well as in recent periods. This indicates that, according to our evaluation metrics, MPI-GE, GFDL-ESM2M, and CESM-LE are the most adequate ensembles to investigate future projections of surface temperatures both globally averaged and globally at the grid-cell level. Our results show that the 100-member MPI-GE offers a representation of the range of observed temperatures that is adequate over larger areas than the 30-member GFDL-ESM2M and 35-member CESM-LE; however, the performance of all three ensembles is comparable when limiting ensemble size to the first 30 members. Over the North Atlantic, Tropical Eastern Pacific, and northern mid- to high-latitude land areas we find the highest number of models that adequately simulate observed surface temperatures, a maximum of nine out of ten models. Over the Southern Ocean, none of the models considered succeeds at adequately capturing observed surface temperatures; while fewer than three models do so over the South Atlantic or the Maritime Continent. In recent decades, fewer models offer adequate simulations of surface temperatures than compared to earlier periods. This occurs due to the overestimation of the recent forced warming in some models which also occurs at regional levels, and indicates that climate projections from these models would likely also overestimate future warming beyond what can be explained by internal variability.

Our novel perspective on model evaluation provides new ways of testing the performance of climate models. Consequently, it also offers new confidence in historical simulations and projections for the long-term climate response to changing forcing in the future, as well as on the simulated range of fluctuations around that response. We can now robustly yet simply determine which models best capture the real-world climate, and assess whether models under- or overestimate the forced response and internal variability in observed variables. The robustness of our framework comes from the unique experimental design of SMILE experiments. Its simplicity comes from fully exploiting the tools that the climate science community has available, taking our methodologies away from previously necessary assumptions and limitations and towards the next generation of climate model evaluation.

Supplementary Information The online version supplementary material available at https://doi.org/10.1007/s00382-021-05821-w.

Acknowledgements We acknowledge all the modelling groups that developed and produced the simulations used in this study, as well as the the Climatic Research Unit (University of East Anglia) in conjunction with the Hadley Centre (UK Met Office) for developing and facilitating the observational compilations used. We also acknowledge the Deutsches Klimarechenzentrum (DKRZ) for providing the necessary computational resources to carry out this work. We would like to thank Lydia Keppler and Veit Lüschow for sharing their knowledge and expertise about the Southern Ocean. Lastly, we thank Clara Deser, Wolfgang A. Müller, Jochem Marotzke, and two anonymous reviewers for providing insightful remarks and suggestions that helped improve this manuscript.

Author Contributions L.S.G., N.M and S.M. developed the analysis. L.S.G. performed the analysis and wrote the manuscript. N.M and S.M. gathered and processed the model data used. All authors contributed to the discussion of the results and the manuscript at all stages.

Funding Open Access funding enabled and organized by Projekt DEAL. This research was supported by the Max Planck Society for the Advancement of Science and by the German Ministry of Education and Research (BMBF) under the ClimXtreme project DecHeat (Grant number 01LP1901F; L.S.G.)

Availability of data and material The model simulations used in this article can be accessed in the Multi-Model Large Ensemble Archive of the US CLIVAR Working Group on Large Ensembles (Deser et al. 2020), in the Earth System Grid Federation database (Cinquini et al. 2014), or by contacting the authors or the respective modelling groups.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Code availability The scripts used to perform this analysis and other supporting information that may be useful in reproducing the author's work are archived by the Library and Information Service at the Max Planck Institute for Meteorology and are freely available by contacting publications@mpimet.mpg.de. The figures in this article were created using the NCAR Command Language (NCAR 2019; Version 6.6.2).

References

Anderson JL (1996) A method for producing and evaluating probabilistic forecasts from ensemble model integrations. J Clim 9(7):1518–1530. https://doi.org/10.1175/1520-0442(1996)009<1518:AMFPAE>2.0.CO;2

Andrews T, Gregory JM, Webb MJ, Taylor KE (2012) Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere-ocean climate models. Geophys Res Lett 39(9). https://doi.org/10.1029/2012GL051607

Annan JD, Hargreaves JC (2010) Reliability of the CMIP3 ensemble. Geos Res Lett 37:L02703. https://doi.org/10.1029/2009GL041994

Bengtsson L, Hodges KI (2019) Can an ensemble climate simulation be used to separate climate change signals from internal unforced variability? Clim Dyn 52(5):3553–3573. https://doi.org/10.1007/s00382-018-4343-8

Beusch L, Gudmundsson L, Seneviratne SI (2020) Crossbreeding CMIP6 Earth System Models with an emulator for regionally optimized land temperature projections. Geophys Res Lett 47(15):e2019GL086812. https://doi.org/10.1029/2019GL086812

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. Geophys Res Lett 43:9324–9332. https://doi.org/10.1002/2016GL070587

Boucher O, Servonnat J, Albright AL, Aumont O, Balkanski Y, Bastrikov V, Bekki S, Bonnet R, Bony S, Bopp L, Braconnot P, Brockmann P, Cadule P, Caubel A, Cheruy F, Codron F, Cozic A, Cugnet D, D'Andrea F, Davini P, de Lavergne C, Denvil S, Deshayes J, Devilliers M, Ducharne A, Dufresne JL, Dupont E, Éthé C, Fairhead L, Falletti L, Flavoni S, Foujols MA, Gardoll S, Gastineau G, Ghattas J, Grandpeix JY, Guenet B, Guez L, Guilyardi E, Guimberteau M, Hauglustaine D, Hourdin F, Idelkadi A, Joussaume S, Kageyama M, Khodri M, Krinner G, Lebas N, Levavasseur G, Lévy C, Li L, Lott F, Lurton T, Luyssaert S, Madec G, Madeleine JB, Maignan F, Marchand M, Marti O, Mellul L, Meurdesoif Y, Mignot J, Musat I, Ottlé C, Peylin P, Planton Y, Polcher J, Rio C, Rochetin N, Rousset C, Sepulchre P, Sima A, Swingedouw D, Thiéblemont R, Traore AK, Vancoppenolle M, Vial J, Vialard J, Viovy N, Vuichard N (2020) Presentation and evaluation of the IPSL-CM6A-LR climate model. JAMES 12(7):e2019MS002010. https://doi.org/10.1029/2019MS002010

Cinquini L, Crichton D, Mattmann C, Harney J, Shipman G, Wang F, Ananthakrishnan R, Miller N, Denvil S, Morgan M, Pobre Z, Bell GM, Doutriaux C, Drach R, Williams D, Kershaw P, Pascoe S, Gonzalez E, Fiore S, Schweitzer R (2014) The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data. Fut Gen Comput Syst 36:400–417. https://doi.org/10.1016/j.future.2013.07.002. https://www.sciencedirect.com/science/article/pii/S0167739X13001477

Cowtan K, Way RG (2014) Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. Q J R Meteorol Soc 140(683):1935–1944. https://doi.org/10.1002/qj.2297

Jiménez-de-la-Cuesta D, Mauritsen T (2019) Emergent constraints on Earth's transient and equilibrium response to doubled CO2 from post-1970s global warming. Nat Geosci 12(11):902–905. https://doi.org/10.1038/s41561-019-0463-y

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

Deser C, Lehner F, Rodgers KB, Ault T, Delworth TL, DiNezio PN, Fiore A, Frankignoul C, Fyfe JC, Horton DE, Kay JE, Knutti R, Lovenduski NS, Marotzke J, McKinnon KA, Minobe S, Randerson J, Screen JA, Simpson IR, Ting M (2020) Insights from Earth System Model Initial-condition Large Ensembles and future prospects. Nat Clim Chang. https://doi.org/10.1038/s41558-020-0731-2

England MH, McGregor S, Spence P, Meehl GA, Timmermann A, Cai W, Gupta AS, McPhaden MJ, Purich A, Santoso A (2014) Recent intensification of wind-driven circulation in the Pacific and the ongoing warming hiatus. Nat Clim Chang 4(3):222–227. https://doi.org/10.1038/nclimate2106

Flato GJ, Marotzke J, Abiodun B, Braconnot P, Chou SC, Collins W, Cox P, Driouech F, Emori S, Eyring V, Forest C, Gleckler P, Guilyardi E, Jakob C, Kattsov V, coauthors (2013) In: Stocker TF, Qin D, Plattner G-K, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, Midgley PM (eds) Evaluation of climate models. Cambridge University Press, Cambridge, pp 741–866. https://doi.org/10.1017/CBO9781107415324.020

Frankcombe LM, England MH, Mann ME, Steinman BA (2015) Separating internal variability from the externally forced climate response. J Clim 28(20):8184–8202. https://doi.org/10.1175/JCLI-D-15-0069.1

Frankcombe LM, England MH, Kajtar JB, Mann ME, Steinman BA (2018) On the choice of ensemble mean for estimating the forced signal in the presence of internal variability. J Clim 31(14):5681–5693. https://doi.org/10.1175/JCLI-D-17-0662.1

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

Frenger I, Münnich M, Gruber N, Knutti R (2015) Southern ocean eddy phenomenology. J Geophys Res Oceans 120(11):7413–7449. https://doi.org/10.1002/2015JC011047

Gleckler PJ, Taylor KE, Doutriaux C (2008) Performance metrics for climate models. J Geophys Res Atmos 113(D6). https://doi.org/10.1029/2007JD008972

Gutjahr O, Putrasahan D, Lohmann K, Jungclaus JH, von Storch JS, Brüggemann N, Haak H, Stössel A (2019) Max planck institute earth system model (MPI-ESM1.2) for the high-resolution model intercomparison project (HighResMIP). Geosci Model Dev 12(7):3241–3281. https://doi.org/10.5194/gmd-12-3241-2019

Hamill TH (2001) Interpretation of rank histograms for verifying ensemble forecasts. Mon Weather Rev 129:550–560. https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2

Hedemann C, Mauritsen T, Jungclaus J, Marotzke J (2017) The subtle origins of surface-warming hiatuses. Nat Clim Chang 7:336–339. https://doi.org/10.1038/nclimate3274

Hurrell JW, Holland MM, Gent PR, Ghan S, Kay JE, Kushner PJ, Lamarque JF, Large WG, Lawrence D, Lindsay K, Lipscomb WH, Long MC, Mahowald N, Marsh DR, Neale RB, Rasch P, Vavrus S, Vertenstein M, Bader D, Collins WD, Hack JJ, Kiehl J, Marshall S (2013) The Community Earth System Model: A framework for collaborative research. BAMS 94(9):1339–1360. https://doi.org/10.1175/BAMS-D-12-00121.1

Hyder P, Edwards JM, Allan RP, Hewitt HT, Bracegirdle TJ, Gregory JM, Wood RA, Meijers AJS, Mulcahy J, Field P, Furtado K, Bodas-Salcedo A, Williams KD, Copsey D, Josey SA, Liu C, Roberts CD, Sanchez C, Ridley J, Thorpe L, Hardiman SC, Mayer M, Berry DI, Belcher SE (2018) Critical Southern Ocean climate model biases traced to atmospheric model cloud errors. Nat Commun 9(1):3625. https://doi.org/10.1038/s41467-018-05634-2

IPCC (2018) Global warming of 1.5C. An IPCC Special Report on the impacts of global warming of 1.5C above pre-industrial levels and related global greenhouse gas emission pathways. In: Stocker TF et al (eds) The context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty. IPCC, Cambridge Univ Press, Cambridge

Jebri B, Khodri M, Echevin V, Gastineau G, Thiria S, Vialard J, Lebas N (2020) Contributions of internal variability and external forcing to the recent trends in the Southeastern Pacific and Peru-Chile upwelling system. J Clim 33(24):10555–10578. https://doi.org/10.1175/JCLI-D-19-0304.1

Jeffrey S, Rotstayn LD, Collier M, Dravitzki SM, Hamalainen C, Moeseneder C, Wong K, Syktus J (2013) Australia's CMIP5 submission using the CSIRO-Mk 3.6 model. Aust Meteorol Oceanogr J 63(1):1–13. https://doi.org/10.22499/2.6301.001

Jones PD, Lister DH, Osborn TJ, Harpham C, Salmon M, Morice CP (2012) Hemispheric and large-scale land-surface air temperature variations: an extensive revision and an update to 2010. J Geophys Res Atmos. https://doi.org/10.1029/2011JD017139

Kay JE, Deser C, Phillips A, Mai A, Hannay C, Strand G, Arblaster JM, Bates SC, Danabasoglu G, Edwards J, Holland M, Kushner P, Lamarque JF, Lawrence D, Lindsay K, Middleton A, Munoz E, Neale R, Oleson K, Polvani L, Vertenstein M (2015) The community earth system model (CESM) large ensemble project: a community resource for studying climate change in the presence of internal climate variability. BAMS 96(8):1333–1349. https://doi.org/10.1175/BAMS-D-13-00255.1

Keller JD, Hense A (2011) A new non-gaussian evaluation method for ensemble forecasts based on analysis rank histograms. Meteorol Z 20(2):107–117. https://doi.org/10.1127/0941-2948/2011/0217

Kiehl JT (2007) Twentieth century climate model response and climate sensitivity. Geophys Res Lett 34(22). https://doi.org/10.1029/2007GL031383

Kirchmeier-Young MC, Zwiers FW, Gillett NP (2017) Attribution of extreme events in arctic sea ice extent. J Clim 30(2):553–571. https://doi.org/10.1175/JCLI-D-16-0412.1

Krinner G, Flanner MG (2018) Striking stationarity of large-scale climate model bias patterns under strong climate change. Proc Natl Acad Sci 115(38):9462–9466. https://doi.org/10.1073/pnas.1807912115. https://www.pnas.org/content/115/38/9462

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. J Clim 30(19):7739–7756. https://doi.org/10.1175/JCLI-D-16-0792.1

Maher N, Matei D, Milinski S, Marotzke J (2018) ENSO change in climate projections: forced response or internal variability? Geophys Res Lett 45(20):11,390–11,398: https://doi.org/10.1029/2018GL079764

Maher N, Milinski S, Suarez-Gutierrez L, Botzet M, Dobrynin M, Kornblueh L, Kröger J, Takano Y, Ghosh R, Hedemann C, Li C, Li H, Manzini E, Notz D, Putrasahan D, Boysen L, Claussen M, Ilyina T, Olonscheck D, Raddatz T, Stevens B, Marotzke J (2019) The max planck institute grand ensemble: enabling the exploration of climate system variability. JAMES 11(7):2050–2069. https://doi.org/10.1029/2019MS001639

Marotzke J, Forster PM (2015) Forcing, feedback and internal variability in global temperature trends. Nature 517:565–U291. https://doi.org/10.1038/nature14117

McGregor S, Timmermann A, Stuecker MF, England MH, Merrifield M, Jin FF, Chikamoto Y (2014) Recent walker circulation strengthening and pacific cooling amplified by atlantic warming. Nat Clim Chang 4(10):888–892. https://doi.org/10.1038/nclimate2330

McKinnon KA, Poppick A, Dunn-Sigouin E, Deser C (2017) An observational large ensemble to compare observed and modeled temperature trend uncertainty due to internal variability. J Clim 30(19):7585–7598. https://doi.org/10.1175/JCLI-D-16-0905.1

Milinski S, Maher N, Olonscheck D (2020) How large does a large ensemble need to be? Earth Syst Dyn 11(4):885–901. https://doi.org/10.5194/esd-11-885-2020

Morice CP, Kennedy JJ, Rayner NA, Jones PD (2012) Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: the HadCRUT4 data set. J Geophys Res Atmos. https://doi.org/10.1029/2011JD017187

NCAR, Boulder, Colorado: UCAR/NCAR/CISL/TDD (2019) The ncar command language (version 6.5.0) [software]. https://doi.org/10.5065/D6WD3XH5

Notz D (2015) How well must climate models agree with observations? Philos Trans R Soc A 373(2052):20140164. https://doi.org/10.1098/rsta.2014.0164

Rodgers KB, Lin J, Frölicher TL (2015) Emergence of multiple ocean ecosystem drivers in a large ensemble suite with an earth system model. Biogeoscience 12(11):3301–3320. https://doi.org/10.5194/bg-12-3301-2015

Schaller N, Sillmann J, Anstey J, Fischer EM, Grams CM, Russo S (2018) Influence of blocking on northern European and western Russian heatwaves in large climate model ensembles. Environ Res Lett 13(5):054015. https://doi.org/10.1088/1748-9326/aaba55

Schär C, Virale PL, Lüthi D, Frei C, Häberli C, Liniger MA, Appenzeller C (2004) The role of increasing temperature variability in European summer heatwaves. Nature 427:332–336. https://doi.org/10.1038/nature02300

Screen JA, Nathan GP, Stevens DP, Marshall GJ, Howard RK (2009) The role of eddies in the southern ocean temperature response to the southern annular mode. J Clim 22(3):806–818. https://doi.org/10.1175/2008JCLI2416.1

Smith A, Jahn A (2019) Definition differences and internal variability affect the simulated arctic sea ice melt season. The Cryosphere 13(1):1–20. https://doi.org/10.5194/tc-13-1-2019

Stössel A, Notz D, Haumann FA, Haak H, Jungclaus J, Mikolajewicz U (2015) Controlling high-latitude southern ocean convection in climate models. Ocean Model 86:58–75. https://doi.org/10.1016/j.ocemod.2014.11.008

Suarez-Gutierrez L, Li C, Thorne PW, Marotzke J (2017) Internal variability in simulated and observed tropical tropospheric temperature trends. Geophys Res Lett 44:5709–5719. https://doi.org/10.1002/2017GL073798

Suarez-Gutierrez L, Li C, Müller WA, Marotzke J (2018) Internal variability in European summer temperatures at 1.5C and 2C of global warming. Environ Res Lett 44:5709–5719. https://doi.org/10.1002/2017GL073798

Suarez-Gutierrez L, Maher N, Milinski S (2020a) Evaluating the internal variability and forced response in large ensembles. CLIVAR Var 18(2):27–35. https://doi.org/10.5065/0DSY-WH17

Suarez-Gutierrez L, Müller WA, Li C, Marotzke J (2020b) Hotspots of extreme heat under global warming. Clim Dyn 55(3):429–447. https://doi.org/10.1007/s00382-020-05263-w

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. J Clim 31(19):7823–7843. https://doi.org/10.1175/JCLI-D-18-0134.1

Swart NC, Cole JNS, Kharin VV, Lazare M, Scinocca JF, Gillett NP, Anstey J, Arora V, Christian JR, Hanna S, Jiao Y, Lee WG, Majaess F, Saenko OA, Seiler C, Seinen C, Shao A, Solheim L, von Salzen K, Yang D, Winter B (2019) The Canadian earth system model version 5 (CanESM5.0.3). Geosci Model Dev 2019:1–68. https://doi.org/10.5194/gmd-2019-177

Tatebe H, Ogura T, Nitta T, Komuro Y, Ogochi K, Takemura T, Sudo K, Sekiguchi M, Abe M, Saito F, Chikira M, Watanabe S, Mori M, Hirota N, Kawatani Y, Mochizuki T, Yoshimura K, Takata K, O'ishi R, Yamazaki D, Suzuki T, Kurogi M, Kataoka T, Watanabe M, Kimoto M (2019) Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in MIROC6. Geosci Model Dev 12(7):2727–2765. https://doi.org/10.5194/gmd-12-2727-2019

Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experiment design. Bull Am Meteorol Soc 93:485–498. https://doi.org/10.1175/BAMS-D-11-00094.1

Thorne PW, Outten S, Bethke I, Seeland O (2015) Investigating the recent apparent hiatus in surface temperature increases: 2. Comparison of model ensembles to observational estimates. J Geos Res Atmos 120:8597–8620. https://doi.org/10.1002/2014JD022805

Tokarska KB, Stolpe MB, Sippel S, Fischer EM, Smith CJ, Lehner F, Knutti R (2020) Past warming trend constrains future warming in CMIP6 models. Sci Adv. https://doi.org/10.1126/sciadv.aaz9549

von Trentini F, Aalbers EE, Fischer EM, Ludwig R (2020) Comparing interannual variability in three regional Single-Model Initial-condition Large Ensembles (SMILEs) over Europe. Earth Syst Dyn 11(4):1013–1031. https://doi.org/10.5194/esd-11-1013-2020