Verdy, A., S. Dutkiewicz, M. J. Follows, J. Marshall, and A. Czaja, 2007: Carbon dioxide and oxygen fluxes in the Southern Ocean: Mechanisms of interannual variability. *Global Biogeochem. Cycles*, **21**, doi:10.1029/2006GB002916.

Wanninkhof, R., and Coauthors, 2013: Global ocean carbon uptake: Magnitude, variability and trends. *Biogeosciences*, **10**, 1983–2000, doi:10.5194/bg-10-1983-2013.

Woods, J. D., 1985: The World Ocean Circulation Experiment. *Nature*, **314**, 501–511, doi:10.1038/314501a0.

Yin, J. H., 2005: A consistent poleward shift of the storm tracks in simulations of 21st century climate. *Geophys. Res. Lett.*, **32**, doi:10.1029/2005GL023684.

Zanchettin, D., C. Timmreck, H.-F. Graf, A. Rubino, S. Lorenz, K. Lohmann, K. Krüger, and J. Jungclaus, 2012: Bi-decadal variability excited in the coupled ocean– atmosphere system by strong tropical volcanic eruptions. *Climate Dyn.*, **39**, 419–444, doi:10.1007/s00382-011-1167-1.

Evaluating the internal variability and forced response in Large Ensembles

Laura Suarez-Gutierrez, Nicola Maher, and Sebastian Milinski

Max-Planck-Institut für Meteorologie, Germany

Surface temperatures and all variables in the climate system fluctuate around their long-term evolving forced state due to the chaotic effect of internal variability. Real-world observations offer only one amongst many possible combinations of these fluctuations, making it difficult to distinguish the effect of internal variability from the forced response to external drivers. In contrast, initial-condition large ensembles (LEs) consist of up to hundreds of simulations of a single climate model under the same time-evolving external forcing conditions, which differ only due to the effect of chaotic internal variability. This means that when large enough LEs allow a precise quantification of both the time-evolving forced response, represented by the ensemble mean, and the internal variability, represented by the spread of possible fluctuations around this mean.

Due to their design, LEs allow for a more effective climate model evaluation. We can use LEs to determine whether observations fall within the ensemble spread simulated by each model. We exploit this potential of LEs to evaluate how well climate models capture the internal variability and forced response in observations, without the need to separate both quantities in the observations, by applying a methodological evaluation framework based on probabilistic forecast verification (Hamill 2001; Suarez-Gutierrez et al. 2018; Maher et al. 2019). This evaluation framework allows us to determine model performance more robustly than before, by assessing whether current climate models capture the long-term trajectory of the climate system as well as the possible range of fluctuations around this trajectory caused by internal variability in any given region and time period.

Here, we use this framework to evaluate historical near-surface air temperatures over North America in LEs from six comprehensive fully-coupled climate models in the Multi-Model Large Ensemble Archive (MMLEA; Deser et al. 2020) provided by the US CLIVAR Working Group on LEs: CanESM2, CESM-LE, CSIRO-MK3.6,GFDL-CM3, GFDL-ESM2M, MPI-GE; as well as in the Observational LE (OBS-LE). In contrast to the six model LEs, OBS-LE is a statistical product that combines the simulated time-evolving forced response from CESM-LE with a synthetic statistical estimate of internal variability derived from observations from the Berkeley Earth Surface Temperature (BEST) dataset for temperature (McKinnon and Deser 2018).

Results

Time series and rank histogram analysis

For the hypothetical case of an LE that perfectly represents the combined effect of the real-world forced response and internal variability, a sufficiently long sample of observations should fall across all of the ensemble spread with no preferred frequency, and mainly occur within the ensemble maximum and minimum limits. We evaluate this by computing time series and rank histograms of annually averaged North American near-surface air temperature anomalies (SAT) with respect to the reference period of 1961–1990 compared to CRUTEM4 observations (Figure 1).

The time series in Figure 1 show the ensemble maxima and minima as well as the 75th percentile central ensemble range (i.e., 12.5th to 87.5th percentile range), together with observations. The rank histograms shown in Figure 1 represent the frequency with which observations take each place in a list of ensemble members ordered by ascending SAT values for each year (Hamill 2001). The rank is zero if the observed SAT for a given year is lower than each SAT simulated by all the ensemble members for that year. If the observed SAT is higher than all simulated SATs, the rank is n, the number of ensemble members. For a long enough observational record that is adequately simulated, observations should occur in all ranks with uniform frequency, thus resulting in a flat rank histogram. In contrast, a non-flat rank histogram indicates a model bias in either the variability or forced response. This is the case for CanESM2, CSIRO-MK3.6, and GFDL-CM3, which show sloped rank histograms with disproportionately large low-rank frequencies. Thus, these ensembles overestimate the historical forced warming compared to observations. Observations occur frequently in the lower half of these ensembles, or below the ensemble minima, either during the entire observational record as for GFDL-CM3, or only in recent decades or early historical period, as for CanESM2 and CSIRO-MK3.6 respectively. The remaining LEs — CESM-LE, GFDL-ESM2M, MPI-GE, and OBS-LE — show relatively flat rank histograms. This indicates that these LEs cover the time-evolving observational spread in North American SATs adequately, with observations occurring uniformly across the ensemble spreads and mostly within the ensemble limits.

The LEs with longer simulation lengths, CESM-LE and in particular CSIRO-MK3.6 and MPI-GE, also appear to have larger SAT variability in the 19th and early 20th Centuries than in recent decades. This variability, represented by the ensemble spread of SAT, decreases in recent decades to maximum to minimum annual SAT ranges of around 1.5 to 2.0 °C, of similar magnitude across all LEs. We also find year-to-year variability in the ensemble maxima and minima SAT larger than 0.5 °C across all six climate models LEs.

By contrast, and due to its experimental design, OBS-LE shows substantially less year-to-year variability in the ensemble maxima and minima. This could arise from the large ensemble size of 1,000 members resulting in the saturation of the SAT ensemble spread on yearly timescales. However, this year-to-year variability remains comparatively low when only the first 100 members of OBS-LE are considered, and is also lower than the variability that we could expect from normally distributed data (not shown), indicating a potential under-sampling of the distribution tails. This suggests

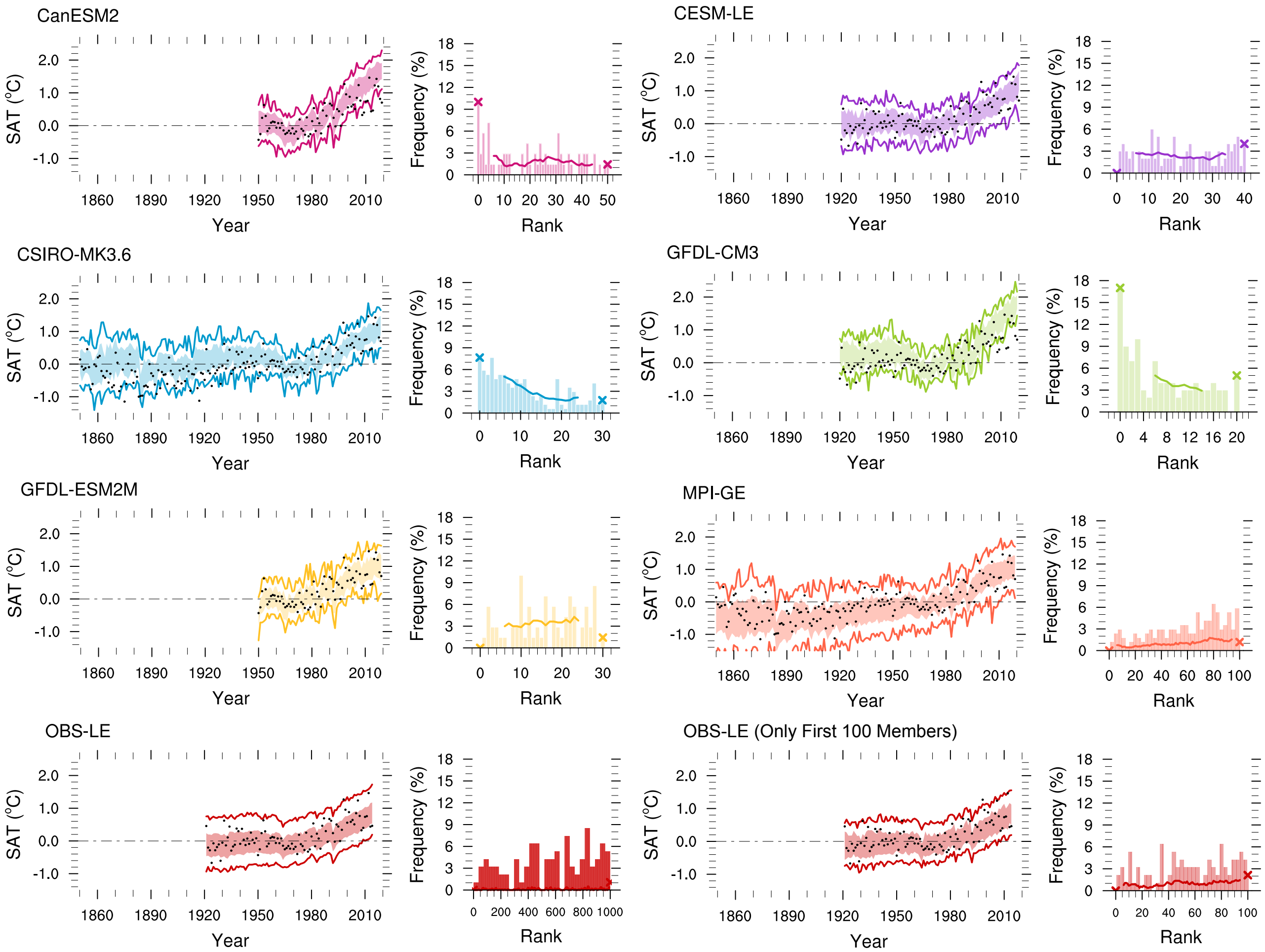


Figure 1: Time series and rank histograms of annual SAT over North America. Time series of annual land-surface SAT anomalies simulated by each LE (colored) and CRUTEM4 observed anomalies (black circles) for the period 1850–2019 (left column). Lines represent ensemble maxima and minima, shading represents the central ensemble range within the 75th percentile (12.5th to 87.5th percentiles). Rank histograms show the frequency of each place that CRUTEM4 observations would take in a list of ensemble members ordered by ascending SAT values (right column). Crosses represent the frequency of minimum (0) and maximum (number of members; n) ranks; lines illustrate the histogram's slope as the moving 10-rank mean. Frequencies are normalized to percentage. Bin sizes are 1 rank, except for MPI-GE and OBS-LE where bin sizes from ranks 1 to n-1 are 3 and 37 ranks, respectively, to aid visualization. Anomalies are relative to the period 1961–1990. Temperature anomalies are averaged over land-surface grid cells where observations are available in the [17.5–52.5°N, 62.5°W–127.5°W] domain.

that OBS-LE may underestimate the intensity of the most extreme SAT events. This could result from the lack of sufficiently large samples of observed low-probability events, due to relatively short observational record, which leads to not only the potential underestimation of the intensity of extreme events but also complicates the robust estimation of their likelihood.

Although this comparatively low year-to-year variability might indicate that OBS-LE underestimates the intensity of low-probability events at the tails of the ensemble distribution, OBS-LE offers the most adequate representation of the combination of the internal variability and forced response in observed SAT over North America throughout the historical observational record. The climate model LEs that capture both quantities in observations most adequately are GFDL-ESM2M, CESM-LE and MPI-GE.

Spatial representation of the combined forced response and internal variability in observations

Based on the concepts in the previous section, we now evaluate how different LEs capture the internal variability and forced response in observations at the grid-cell level by identifying three different possible biases (Figure 2). First, we evaluate how often observations lie either below or above the ensemble limits in each grid cell. We distinguish between regions where 5% or more of the time observations fall below the ensemble minimum (blue shading) and above the ensemble maximum (red shading). If only one of these biases occurs in a region, the model respectively over- or under-estimates the forced response in observations. Alternatively, such a bias could also be caused by a bias in the skewness of the probability distribution for non-normally distributed variables. If both of these biases occur at the same location, this means that observations fall below and also above the ensemble limits, either over the entire period of analysis (indicating the model does not sufficiently capture the observed variability) or during specific periods (indicating a likely change in the sign of the model bias over time).

The third metric of model performance highlights regions where observations cluster more than expected within the central 75th percentile range of the simulated ensembles. For the ideal case in which observations are uniformly distributed across the ensemble and exhibit a flat rank histogram, observed values would lie within the central 75th percentile ensemble range (12.5th–87.5th percentiles) around 75% of the time. Here we identify areas where observations occur in the central ensemble range more than 80% of the time (gray shading in Figure 2), indicating that the model overestimates internal variability. This bias results in simulated extreme events at the tails of the ensemble distribution that are systematically more intense than observed. Note that this type of bias can only be robustly identified when the simulated distribution adequately captures the forced response in observations, and when evaluated over a period long enough to sufficiently sample the timescales of internal variability under study.

White areas without any shading in Figure 2 indicate that none of the three biases occurs to a substantial degree, indicating that the ensembles simulate a time-evolving forced response and range of variability around this response that are comparable to those in observations for the whole length of their simulations. Thus, in these areas, our evaluation framework indicates that the models adequately capture the forced response and internal variability in observed surface temperatures. The percentage of white areas over North America represents areas with no substantial biases for each LEs (upper right corners in Figure 2), and indicates that OBS-LE, with 85.9%, offers the most adequate spatial representation of the combined internal variability and forced response in observed historical SAT. MPI-GE, with 46.6% of white areas, offers the best representation of historical SAT over North America amongst the model LEs, followed by CanESM2, CESM-LE, and GFDL-ESM2M.

The predominance of blue shading over red shading in Figure 2 for CanESM2, and especially CSIRO-MK3.6 and GFDL-CM3, indicates that observations fall below the ensemble minima more frequently and over larger regions than they fall above the ensemble maxima for these models. These are the same models that show overestimated forced warming compared to observations in Figure 1. Observations exceed the ensemble maxima over the East Coast and Gulf of Mexico area for CESM-LE and CSIRO-MK3.6 (red shading in Figure 2b and c), indicating that these ensembles underestimate the intensity of warm near-surface air temperature extremes in these areas.

Over the Caribbean Islands and the Baja California Peninsula, observations occur both below and above ensemble limits with high frequency (overlapping blue and red shading in Figure 2). This indicates that models underestimate observed SAT variability in these

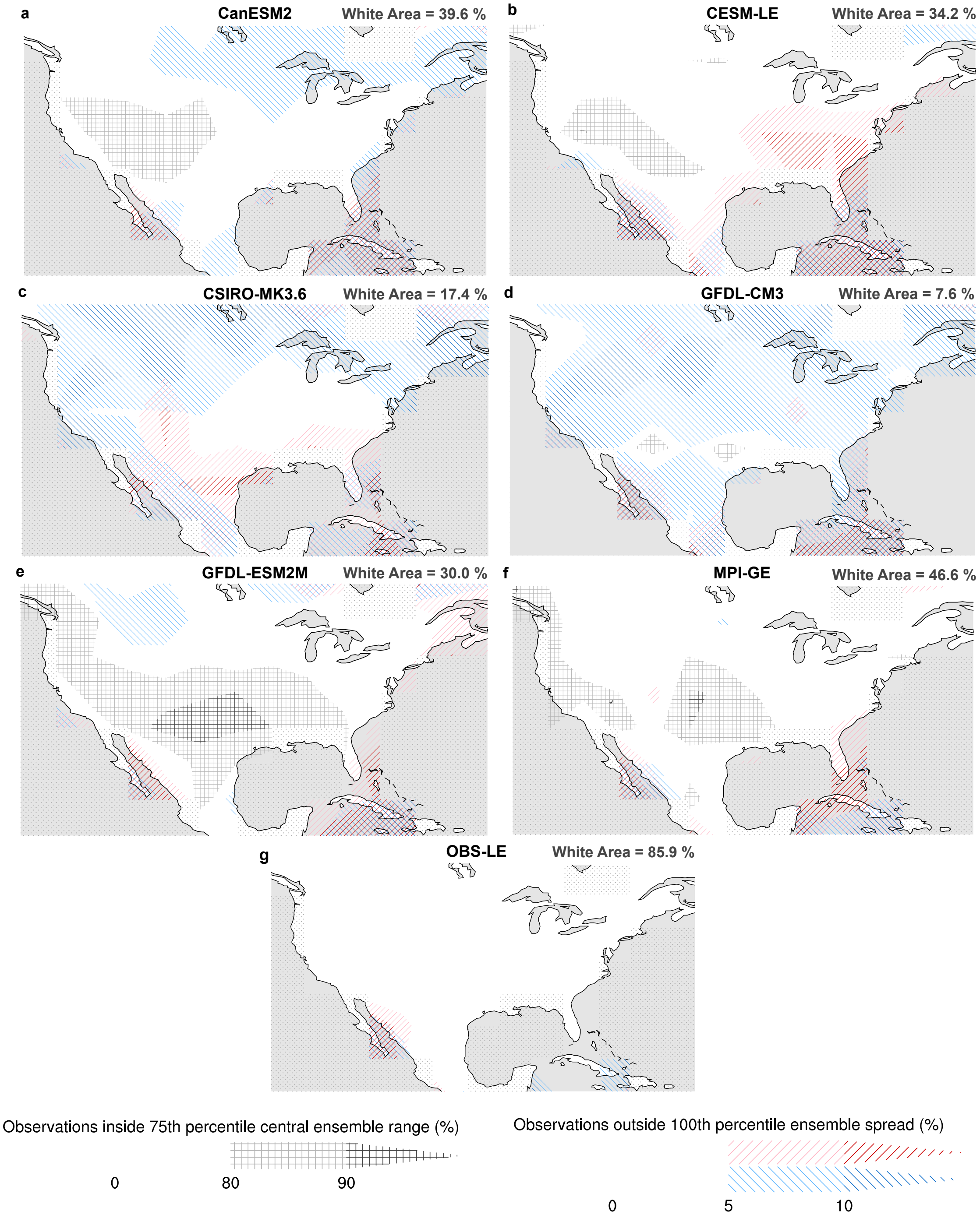


Figure 2: Evaluation of internal variability and forced response in annual SATs. Evaluation of annual SAT anomalies simulated by different LEs compared to CRUTEM4 observed anomalies from 1850, or each LE starting year, until 2019. Red shading represents the percentage of time that the observed yearly anomaly is larger than the ensemble maximum; blue shading represents the percentage of time that the observed yearly anomaly is lower than the ensemble minimum. Gray hatching represents how often observations cluster within the 75th percentile bounds of the ensembles (12.5th to 87.5th percentiles). Dotted areas are excluded from our analysis due to CRUTEM4 observations being available for less than 10 years. Percentages of white area in the upper right corners represent the percentual area of North America where none of these biases occur to a substantial degree for each LEs. Anomalies are relative to the period 1961–1990. Model output data are regridded to match the observational grid.

regions, likely due to the effect of model resolution and complex orography in confounding land versus ocean in these grid cells. Lastly, observations cluster in the central ensemble ranges of several models, including CanESM2, CESM-LE, MPI-GE, and especially GFDL-ESM2M, over the West Coast and Central US, indicating that these models overestimate the variability in these regions (gray shading in Figure 2).

OBS-LE shows no substantial biases over North America, with the exception of the underestimation of SAT variability over the Baja California Peninsula (Figure 2g). Our results indicate that OBS-LE offers the most adequate spatial representation of the internal variability and forced response in observed historical SAT over this region. In agreement with the results in Mckinnon and Deser 2018 for 50-year trends, we find that OBS-LE shows only minor biases in annual SATs over most of the Northern Hemisphere; while it exhibits underestimated annual SAT variability compared to observations over large areas in the low latitudes (not shown). Over these regions, OBS-LE fails to cover the observed variability range in SATs, with observed extreme anomalies beyond the OBS-LE maximum and minimum values over more than 10% of the years. This could result from a combination of the comparatively lower variability at the tails of the OBS-LE distribution identified in Figure 1, that could be more prominent in these areas, as well as an increased spatial and temporal observational sparsity in these regions that could affect the statistical processing used to generate OBS-LE.

Comparison of internal variability

Following our evaluation of the forced response and internal variability in LEs, we can now determine which LEs provide the most realistic simulations of internal variability in annual SAT over North America to better estimate the internal variability in the real world. Here, we measure the magnitude of internal variability in the model LEs and OBS-LE as the 2.5th to 97.5th percentile ensemble spread averaged over the period 1950–1990 (Figure 3a-g). We restrict this analysis to the period 1950–1990 to ensure contributions from all LEs and to minimize the

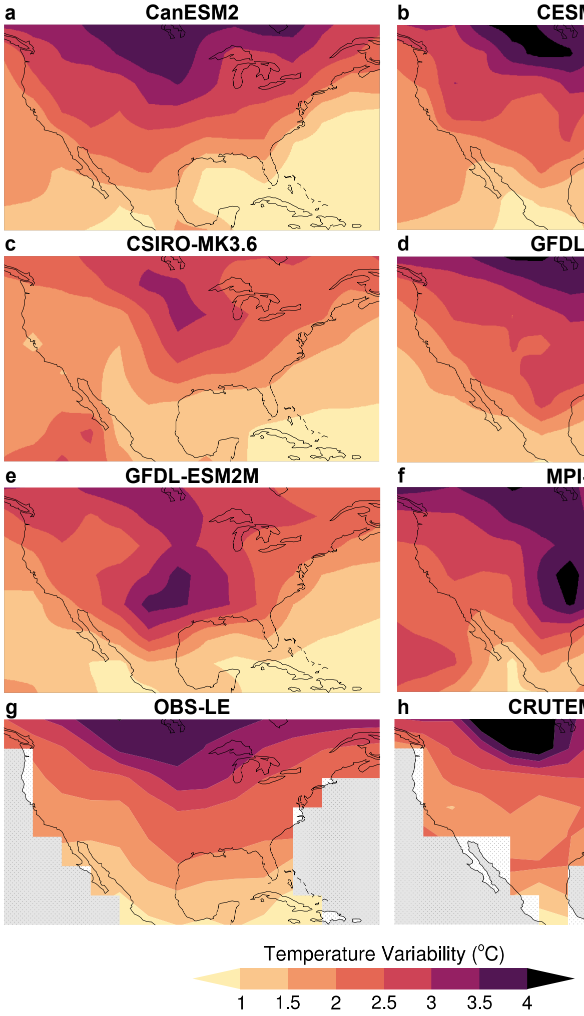


Figure 3: Variability in annual surface temperatures. (a-g) Ensemble spread for annual SAT anomalies by different model LEs and OBS-LE averaged for the period of 1950–1990 measured as the difference between the 2.5th to 97.5th annual SAT percentiles. (h) SAT spread in CRUTEM4 observations measured as the difference between the 2.5th to 97.5th percentiles of the whole distribution for the period of 1950–1990 of annual SAT anomalies. Simulated data are regridded to match the observational grid.

potential effect of the forced response. For comparison, we compute the range of internal variability in CRUTEM4 observations, estimated directly as the 2.5th to 97.5th percentile range in the distribution of non-detrended observed annual SAT anomalies in the same period (Figure 3h). The simulated amplitude of SAT internal variability ranges from 2 to 4°C over most of the continental land area across the different model LEs (Figure 3a-f) and OBS-LE (Figure 3g). The observational estimate (Figure 3h) is generally larger than the LE estimates, in particular over the northern part of the domain, and decreases more steeply with decreasing latitude. However, unlike the LEs estimates, the observational estimate of internal variability could be affected by the confounding effect of the forced response in observations.

OBS-LE, which by design most adequately captures the forced response and internal variability in SAT over North America, exhibits a stratified pattern with variability increasing polewards, which is not completely captured by any of the LEs. Two of the four LEs that most adequately capture North American SATs, GFDL-ESM2M and MPI-GE (Figure 3e and f), simulate hotspots of too high SAT variability over the central United States and Gulf of Mexico region, in agreement with the areas of overestimated variability in Figure 2. These hotspots exhibit SAT variability ranges of 3.5°C to more than 4.0°C, almost twice as large as the SAT variability in other LEs. This indicates that in these areas, these two ensembles simulate annual mean SAT extremes systematically more intense than those observed, possibly due to an overestimation of the cold tail of the distribution during the summer months (not shown).

Summary and conclusions

We use a novel framework exploiting the power of large ensembles to evaluate historical temperatures over North America in six comprehensive, fully-coupled climate models, as well as in the observational ensemble OBS-LE. This framework is based on a simple approach: evaluating whether observations occur evenly across the ensemble spread of simulations, and whether they occur mainly within the limits of this spread. Our evaluation shows that the experimental design in OBS-LE results in the most adequate representation of the combined effect of the forced response and internal variability in observed temperatures over North America. The climate model LEs that provide the best representation according to our metrics are MPI-GE, CanESM2, CESM-LE, and GFDL-ESM2M, suggesting that these LEs are the best choice for investigating future temperature projections over this region. Our evaluation framework highlights MPI-GE as the model LE that most adequately captures the combined forced response and internal variability in observed North American surface temperatures for the period 1850–2019, with the largest area with no substantial biases, 46.6% of the North American region.

Several models show similar biases over similar regions, such as an overestimation of temperature variability in Central North America and an underestimation of variability over the Caribbean Islands and the Baja California Peninsula, likely due the combination of model resolution and complex orography in these regions. Some models overestimate recent forced warming over North America beyond the range of plausible fluctuations caused by internal variability. Our results show that models do not consistently over- or underestimate internal variability in surface temperatures, and that models that perform adequately over one region will not necessarily do so in another.

Overall, this evaluation framework provides a new and more robust approach to determine model performance, allowing users to decide which models are most appropriate for their variable and region of interest, by highlighting which models offer the most adequate representation of the real-world internal variability and forced response.

Acknowledgments

We acknowledge the modeling groups that created the model data used here, as well as the US CLIVAR Working Group on Large Ensembles for facilitating

access to these data in the Multi-Model Large Ensemble Archive (Deser et al. 2020). We would also like to acknowledge the groups that developed and facilitated the observational compilations used here: the Climatic Research Unit (University of East Anglia) in conjunction with the Hadley Centre, UK Met Office (Jones et al. 2012).

Data and methods

We include LEs from six coupled climate models in the US CLIVAR MMLEA (Deser et al. 2020) as well as the synthetic product OBS-LE based on observations (Table 1). Each of the climate model LEs comprises 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 their number of simulations, in how sensitive the model is to increasing CO2, or in the method used for the initialization of their members. When available, historical simulations are extended with one available future forcing scenario to cover the entire observational record. We also use surface temperature observations from the CRUTEM4 (Jones et al. 2012) dataset for comparison to the LE simulations. All simulated data are regridded to match the coarser resolution of CRUTEM4 observations and transformed to anomalies with respect to the 1960–1991 climatological period.

The methodological framework demonstrated in this paper was first used in Suarez-Gutierrez et al. 2018 to evaluate European summer temperature and precipitation in MPI-GE; and further expanded to evaluate global annual mean temperatures in Maher et al. 2019, and global summer maximum temperatures in Suarez-Gutierrez et al. 2020.

Table 1: Details of LE experiments analysed from the Multi-Model Large Ensemble Archive (Deser et al. 2020). Experiment name, number of members, simulated years used, forcing scenarios and references of LE experiments included. All experiments include historical forcing (Hist.) until 2005, except for OBS-LE, which is based on historical observed temperatures (see McKinnon and Deser, 2018). Historical simulations are extended beyond 2005 using one future forcing scenario.

LE Experiment	Members	Years	Forcing	Reference
CanESM2	50	1950-2018	Hist + RCP8.5	Kirchmeier-Young et al. 2017
CESM-LE	40	1920-2018	Hist + RCP8.5	Kay et al. 2015
CSIRO-MK3.6	30	1850-2018	Hist + RCP8.5	Jeffrey et al. 2013
GFDL-CM3	20	1920-2018	Hist + RCP8.5	Sun et al. 2018
GFDL-ESM2M	30	1950-2018	Hist + RCP8.5	Rodgers et al. 2015
MPI-GE	100	1850-2018	Hist + RCP8.5	Maher et al. 2019
OBS-LE	1000	1920-2014	Hist + RCP8.5	McKinnon et al. 2017

References

Deser, C., and Coauthors, 2020: Insights from Earth system model initial-condition large ensembles and future prospects. *Nat. Climate Change*, **10**, 277-286, doi:10.1038/s41558-020-0731-2.

Hamill, T. H., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560, doi:10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

Jeffrey S., L. Rotstayn, M. Collier, S. Dravitzki, C. Hamalainen, C. Moeseneder, K. Wong, and J. Syktus, 2013: Australia's CMIP5 submission using the CSIRO-Mk 3.6 model. *Aust. Meteor. Oceanogr. J.*, **63**, 1–13, doi:10.22499/2.6301.001.

Jones, P. D., D. H. Lister, T. J. Osborn, C. Harpham, M. Salmon, and C. P. Morice, 2012: Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010. *J. Geophys. Res.: Atmos.* **117**, doi:10.1029/2011JD017139.

Kay, J. E., and Coauthors, 2015: The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bull. Amer. Meteor. Soc.*, **96**, 1333–1349, doi:10.1175/BAMS-D-13-00255.1.

Kirchmeier-Young, M. C., F. W. Zwiers, and N. P. Gillett, 2017: Attribution of extreme events in Arctic Sea ice extent. *J. Climate,* **30**, 553–571, doi:10.1175/JCLI-D-16-0412.1.

McKinnon, K. A. and C. Deser, 2018. Internal Variability and Regional Climate Trends in an Observational Large Ensemble. *Journal of Climate* **31**, 6783–6802 doi:10.1175/JCLI-D-17-0901.1.

Maher, N., and Coauthors, 2019: The Max Planck Institute Grand Ensemble: Enabling the exploration of climate system variability. *J. Adv. Model. Earth Syst.,* **11**, 2050–2069, doi:10.1029/2019MS001639.

Rodgers, K. B., J. Lin, and T. L. Frölicher, 2015: Emergence of multiple ocean ecosystem drivers in a large ensemble suite with an Earth system model. *Biogeosciences,* **12**, 3301–3320, doi:10.5194/bg-12-3301-2015.

Suarez-Gutierrez, L., W. A. Müller, C. Li, and J. Marotzke, 2018: Internal variability in European summer temperatures at 1.5 °C and 2 °C of global warming. *Environ. Res. Lett.,* **44**, 5709–5719, doi:10.1088/1748-9326/aaba58.

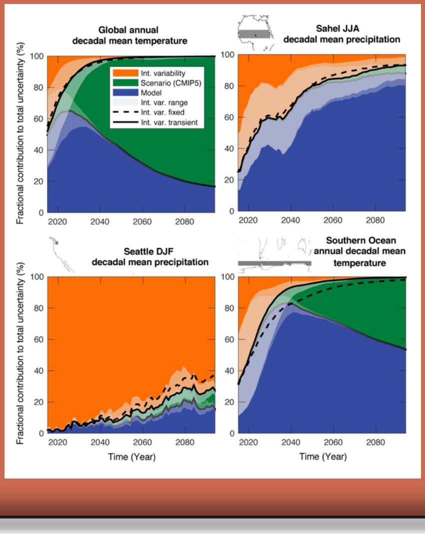
Suarez-Gutierrez, L., W. A. Müller, C. Li, and J. Marotzke, 2020: Hotspots of extreme heat under global warming. *Climate Dyn.,* **55**, 429-447, doi:10.1007/s00382-020-05263-w.

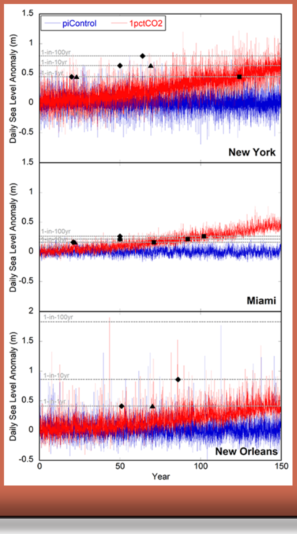
Sun, L., M. Alexander, and C. Deser, 2018: Evolution of the global coupled climate response to Arctic Sea ice loss during 1990–2090 and its contribution to climate change. *J. Climate,* **31**, 7823–7843, doi:10.1175/JCLI-D-18-0134.1.

Submit a Research Highlight

US CLIVAR aims to feature the latest research results from the community. Check out the collection of research highlights and consider contributing.

Learn more here


Global annual decadal mean temperature
Sahel JJA decadal mean precipitation
Seattle DJF decadal mean precipitation
Southern Ocean annual decadal mean temperature
Int. variability
Scenario (CMIP5)
Model
Int. var. range
Int. var. fixed
Int. var. transient
Fractional contribution to total uncertainty (%)
Time (Year)


piControl
1pctCO2
1-in-100yr
1-in-10yr
1-in-1yr
New York
Miami
New Orleans
Daily Sea Level Anomaly (m)
Year