Which climate models capture the observed internal variability and forced response?

Laura Suarez-Gutierrez*, Nicola Maher*, and Sebastian Milinski*

* Max Planck Institute for Meteorology, Hamburg, Germany

@DrLauraSuarez

1. Introduction

We apply a **novel methodological framework** utilizing the power of single model initial condition large ensembles (SMILEs) for evaluating how fully-coupled climate models capture the observed internal variability and the observed response in the climate system to natural and anthropogenic external forcings. As a test case for this framework we use surface temperatures.

Our framework[1,2,3] is based on a simple approach: assessing whether **observations are well distributed** across the whole ensemble spread of simulated states, and whether they generally stay **within this ensemble spread**.

This method relies on the robust characterization of internal variability in SMILEs, that provides **well-defined evolving mean climate and probability distributions** of deviations around this mean state caused by internal variability. This allows us to **distinguish** what causes discrepancies between models and observations: an incorrect simulated **forced response,** or rather an over or underestimated simulated **internal variability.**

With this framework we are not constrained to mean state comparisons, detrended quantities, assumptions for isolating the observed forced response, or to evaluating internal variability by using standard deviations as a proxy. In contrast, we can directly quantify **whether the whole distribution, including its tails, agrees well with what is observed.**

2. Data

SMILEs consist of many simulations of exactly the same climate model run under the same external forcings, but starting from different initial conditions. This design ensures that the simulations differ only due to internal variability, and together offer a sampling of the climate system in different flavors of internal variability and its response to external forcings that is robust and precise.

We use SMILE surface temperature simulations from ten CMIP5 and CMIP6 models (Table 1) and HadCRUT4 observations[4].

SMILE	Members	Years	Forcing	ECS
CanESM2[5]	50	1950-2018	Hist. + RCP8.5	3.7 K
CanESM5*[6]	50	1850-2014	Hist.	5.7 K
CESM-LE[7]	40	1920-2018	Hist. + RCP8.5	4.1 K
CSIRO[8]	30	1850-2018	Hist. + RCP8.5	4.1 K
GFDL-CM3[9]	20	1920-2018	Hist. + RCP8.5	4.0 K
GFDL-ESM2M[10]	30	1950-2018	Hist. + RCP8.5	2.4 K
IPSL-CM5A[11]	30	1941-2018	Hist. + RCP8.5	4.1 K
IPSL-CM6A*	31	1850-2014	Hist.	4.5 K
MIROC6*[12]	50	1850-2018	Hist. + SSP2-4.5	2.6 K
MPI-GE[2]	100	1850-2018	Hist. + RCP4.5	2.8 K

Table 1: Details of SMILE experiments used including name, number of members, simulated years used, forcing conditions, and available Equilibrium Climate Sensitivity (ECS). CMIP6 models are marked with a star. All experiments include historical forcing (Hist.) of CMIP5 until 2005, or CMIP6 until 2014 and are extended using one future forcing scenario when available.

Large ensembles provide better yet simpler model evaluation frameworks.

They allow us to reassess whether observations occur within the better-sampled range of simulated climate states; without the need to make assumptions about the observed variability and forced response.

3. Global mean temperature time series and rank histogram analysis

→ **Internal variability in GMST is captured well by most models, but some overestimate forced warming.**

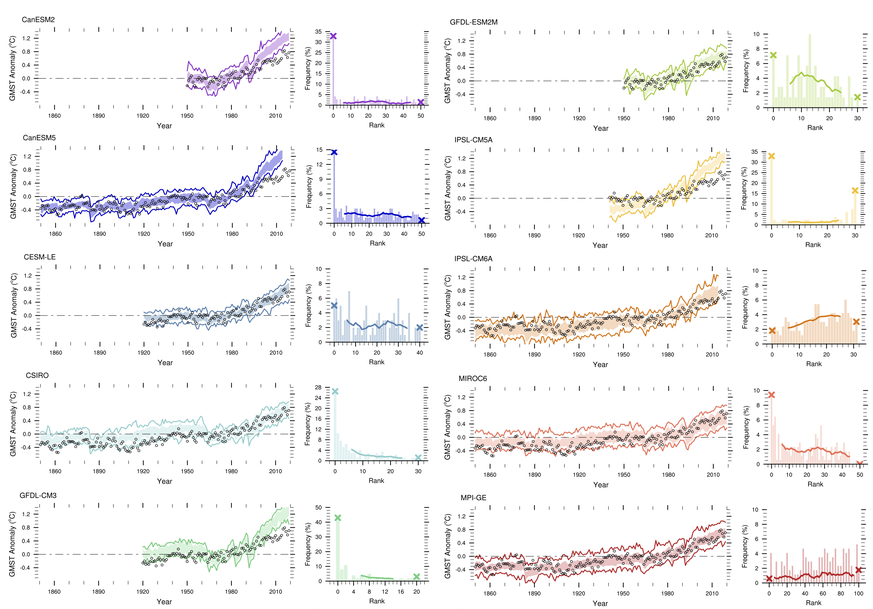

Figure 1: Time series and rank histograms of annual GMST anomalies by each SMILE (colored) and HadCRUT4(3) observed anomalies (black circles) for the period. Colored lines represent ensemble maxima and minima, shading represents the ensemble spread within the central 75th percentile bounds (12.5th to 87.5th percentiles). Rank histograms represent the frequency of each rank of HadCRUT4 GMST observations shown as a member of each SMILE. Crosses mark the frequency of minimum (0) and maximum (number of members) ranks, while lines illustrate the histogram's slope, as the mean rank frequency over a centered 10-rank window. All anomalies are relative to the period of 1961-1990.

We use time series and rank* histograms to assess whether global mean surface temperature (GMST) observations occur within the ensemble spread with uniform frequency (Fig. 1).

*Rank = the place that the observations would take in a list of members ordered by ascending GMST values for each year. Is 0 when observed GMST is lower than all GMSTs simulated for that year; and N when it is higher than all simulated GMST, with N the number of members.

The shape of the rank histogram illustrates whether the observed variability and forced response are adequately simulated:

- Flat histograms indicate that both the observed variability and forced response are well captured → CESM-LE, GFDL-ESM2M, MPI-GE
- High rank 0 frequencies indicate overestimated forced warming if they occur clustered in time → CanESM2, CanESM5, GFDL-CM3, IPSL-CM5A
- Sloped histograms may indicate both a bias in variability affecting the shape or skewness of the distribution, or a bias in the forced response → CSIRO, IPSL-CM6A, MIROC6
- Concave or convex histograms indicate that internal variability is respectively under or overestimated → not found

4. Where do climate models perform well?

Using this framework we can identify regions where:

- observed internal variability is overestimated (gray) or underestimated (blue and red)
- observed forced response is not adequately captured due to overestimated (only blue) or underestimated (only red) simulated warming, or to forced response bias changing signs over different periods (blue and red)

We find that, for most models, observations occur outside the ensemble limits (blue or/and red) more frequently than they tend to cluster in the central ensemble bounds (gray). This means that these models fail to capture the observed forced response and/or underestimate the observed variability in surface temperatures more frequently than they tend to overestimate the observed internal variability (Fig. 2).

The areas over the North Atlantic, Tropical Eastern Pacific, and northern hemisphere land regions are where most models, a maximum of nine, adequately simulate observed surface temperatures. While over the Southern Ocean, none of the models considered offers an adequate representation.

→ **CESM-LE, GFDL-ESM2M and MPI-GE offer the best representation of the observed internal variability and forced response in surface temperatures.**

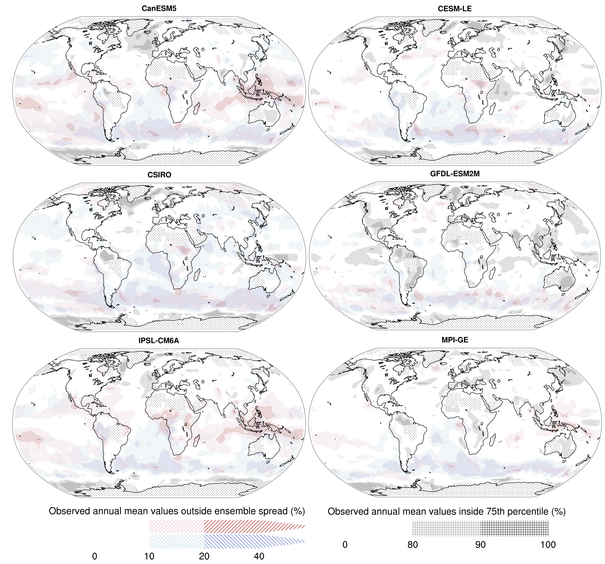

Figure 2: Evaluation of variability and forced response in surface temperature annual anomalies simulated by selected SMILEs against HadCRUT4 observations. Red shading marks regions where observations are larger than the ensemble maximum, while blue shading marks where observations are smaller than the ensemble minimum for more than 10% to 20% of the time. Gray hatching marks regions where observations cluster within the central 75th percentile bounds of the ensembles (12.5th to 87.5th percentiles). Dotted areas represent regions where observations are available for less than 10 years. Simulated data are regridded to match the observational grid (~5°).

References

1. Suarez-Gutierrez et al., 2018. Internal variability in European summer temperatures at 1.5C and 2C of global warming. ERL.
2. Maher et al., 2019. The Max Planck Institute Grand Ensemble: Enabling the Exploration of Climate System Variability. JAMES.
3. Hamill, 2001. Interpretation of rank histograms for verifying ensemble forecasts. Mon. Weath. Rev.
4. Morice et al., 2012. Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The hadCRUT4 data set. JGR:Atmos.
5. Kirchmeier-Young et al., 2017. Attribution of Extreme Events in Arctic Sea Ice Extent. J of Climate.
6. Swart et al., 2019. The Canadian Earth System Model version 5 (CanESM5.0.3). Geosci Model Dev.
7. Kay et al. 2015. The Community Earth System Model (CESM) Large Ensemble Project: A Community Resource for Studying Climate Change in the Presence of Internal Climate Variability. BAMS
8. Jeffrey et al., 2013. Australia's CMIP5 submission using the CSIRO-Mk 3.6 model. AMOJ.
9. Sun et al., 2018. Evolution of the Global Coupled Climate Response to Arctic Sea Ice Loss during 1990-2090 and ts Contribution to Climate Change. J of Climate.
10. Rodgers et al., 2015. Emergence of multiple ocean ecosystem drivers in a large ensemble suite with an Earth system model. Biogeosci.
11. Frankignoul et al., 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 of Climate.
12. Tatebe et al, 2019. Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in MIROC6. Geosci Model Dev.

Max-Planck-Institut für Meteorologie

Do you have any comments or questions?

Would you like more details?

Laura.suarez@mpimet.mpg.de

MAX-PLANCK-GESELLSCHAFT