Geophysical Research Letters

RESEARCH LETTER
10.1029/2019GL086773

Broad Consistency Between Observed and Simulated Trends in Sea Surface Temperature Patterns

Dirk Olonscheck[1], **Maria Rugenstein**[1], **and Jochem Marotzke**[1]

[1]Max Planck Institute for Meteorology, Hamburg, Germany

Key Points:
- Observed and simulated SST trends are consistent in at least 90% of the global ocean area since the midtwentieth century
- This consistency is obtained over larger areas in CMIP6 models than in CMIP5 models
- Observed regional trends that lie at the edge of the simulated distribution of internal variability are hard to interpret

Supporting Information:
- Supporting Information S1
- Text S1

Correspondence to:
D. Olonscheck,
dirk.olonscheck@mpimet.mpg.de

Citation:
Olonscheck, D., Rugenstein, M., & Marotzke, J. (2020). Broad consistency between observed and simulated trends in sea surface temperature patterns. *Geophysical Research Letters*, *47*, e2019GL086773. https://doi.org/10.1029/2019GL086773

Received 3 JAN 2020
Accepted 28 APR 2020
Accepted article online 7 MAY 2020

Abstract Using seven single-model ensembles and the two multimodel ensembles CMIP5 and CMIP6, we show that observed and simulated trends in sea surface temperature (SST) patterns are globally consistent when accounting for internal variability. Some individual ensemble members simulate trends in large-scale SST patterns that closely resemble the observed ones. Observed regional trends that lie at the outer edge of the models' internal variability range allow two nonexclusive interpretations: (a) Observed trends are unusual realizations of the Earth's possible behavior and/or (b) the models are systematically biased but large internal variability leads to some good matches with the observations. The existing range of multidecadal SST trends is influenced more strongly by large internal variability than by differences in the model formulation or the observational data sets.

Plain Language Summary Climate model simulations agree well with the observed evolution of the global mean sea surface temperature, but their ability to realistically represent changes in the patterns of sea surface temperatures has been questioned. We show with an unprecedented number of simulations from different models and different initial conditions that the observed and simulated changes in SST patterns are consistent in most regions of the ocean. For each model, a few individual simulations recreate the observed patterns. In some regions, the observed changes may be an extreme realization of the Earth's possible behavior. Alternatively, structural model errors may be hidden by the large range of possible realizations.

1. Should Simulated SST Patterns Resemble Observed SST Patterns?

Over the historical period, global mean sea surface temperatures (SSTs) mostly increased in response to rising atmospheric greenhouse gas concentrations (Hartmann et al., 2013). The spatial pattern of SST is, however, nonuniform and changing in time. Recent research has shown that spatial and temporal changes in SST patterns strongly influence the magnitude of atmospheric radiative feedbacks (Paynter & Frölicher, 2015; Gregory & Andrews, 2016; Zhou et al., 2016; Armour, 2017) and that simulations with prescribed observed SST patterns show lower estimates of climate sensitivity than historical simulations with a dynamic and full depth ocean (Andrews et al., 2018; Marvel et al., 2018). Given this relevance of SST patterns, it has been debated whether global climate model formulation and mean state biases cause changes in SST patterns that are inconsistent with observations (e.g., Sohn et al., 2016; Zhou et al., 2016; Coats & Karnauskas, 2017, 2018; Kostov et al., 2018; Marvel et al., 2018; Seager et al., 2019). Trends in SST patterns are an emergent property of climate models and not tuned for (Hourdin et al., 2017). Since the climate system is chaotic, we do not expect the models to be in sync with the observations on a regional scale at any specific point in time. However, the trust in climate models would decline if they were not able to reproduce the observed structures of SST pattern change. We show that observed 30- and 20-year trends in the historical period are globally consistent with multiple single-model large ensembles and the multimodel ensembles of the fifth and sixth phases of the Coupled Model Intercomparison Project (CMIP5 and CMIP6).

In the recent debate, the equatorial Pacific and the Southern Ocean received special attention. For the equatorial Pacific, Coats and Karnauskas (2017) suggested that SST trends in CMIP5 are inconsistent with the observed strengthening of the equatorial Pacific east-to-west SST gradient over recent decades, possibly due to the errors in simulating the strengthening of the Equatorial Undercurrent (Coats & Karnauskas, 2018). Seager et al. (2019) argued that CMIP5 and one single-model large ensemble fail to reproduce the observed 60-year SST trends in the Nino3.4 region and the observed strengthening of the equatorial Pacific SST gradient. The study concluded that this failure is due to the systematic cold bias in the models' equatorial Pacific

cold tongue regions. In contrast, Bordbar et al. (2017) argued that the observed strengthening of the equatorial Pacific SST gradient is still within the range of the CMIP5 models. For the Southern Ocean SSTs, Kostov et al. (2018) showed that CMIP5 models have not reproduced the observed 1979–2014 regional cooling and found that the different model responses to both the Southern Annular Mode and greenhouse gas forcing contribute to the inconsistency. Böning et al. (2008) suggested that climate models hardly generate the observed Southern Ocean changes because of the coarse oceanic resolution. On the other hand, Hobbs et al. (2016) and Zhang et al. (2019) argued that the observed changes are small and not robustly detectable in the range of CMIP5 model differences, implying consistency between models and observations. These contrasting findings on regional scales question whether or not simulated SST trends are consistent with observations globally.

We make use of the unprecedented opportunity to estimate internal variability from the large number of historical simulations in seven single-model initial condition large ensembles (Deser et al., 2020) and the ongoing CMIP6 (Eyring et al., 2016; section 2). We quantify the consistency of observed and simulated trends on the grid scale for the entire ocean area for different periods in the twentieth century (section 3) and show that single ensemble members reproduce the large-scale observed trends in SST patterns (section 4). We conclude with a discussion of the limitations of interpreting the currently available observational data sets and model ensembles (section 5).

2. Models as Known and Observations as Unknown Distributions

We use three fully interpolated observational data sets: COBE-SST2 (1850–2014, 1° × 1°; Hirahara et al., 2013), ERSSTv5 (1854–2014, 2° × 12°; Huang et al., 2017), and HadISST2 (version HadISST2.1.0.0, 1850–2010, 1° × 1°; Titchner & Rayner, 2014) to account for the differences in observational estimates (Diamond & Bennartz, 2015; Carella et al., 2018). HadISST2 is a 10-member ensemble that accounts for structural uncertainties in SST and sea ice concentration (available at https://www.metoffice.gov.uk/hadobs/hadisst/data/hadisst2/). The data sets are mostly based on the same in situ data but differ in bias correction and interpolation algorithms (Yasunaka & Hanawa, 2011; Kennedy, 2014). Thus, we have 12—closely related—estimates of the otherwise unknown probability density function (PDF) of the real world. The PDFs of regional climate change in the real world might be even broader than the ones from the climate models (Laepple & Huybers, 2014a, 2014b; Cassou et al., 2018; Kravtsov et al., 2018).

We use seven single-model initial condition large ensembles, namely, CanESM2 (Kirchmeier-Young et al., 2016), CESM-LE (Kay et al., 2014), CSIRO-Mk3-6-0 (Jeffrey et al., 2013), GFDL-ESM2M (Rodgers et al., 2015), IPSL-CM6A-LR (Boucher et al., 2020), MIROC6 (Tatebe et al., 2019), and MPI-GE (Maher et al., 2019). The number of realizations from each single-model ensemble are given in Figure 1. The realizations are started from different initial conditions, and therefore, the single-model ensemble spread represents differences that arise solely because of internal variability in the simulated climate. We further use 116 simulations from 33 CMIP5 models (Taylor et al., 2012) and 148 simulations from 27 CMIP6 models (available as of December 2019, Eyring et al., 2016). The different generations of models cover slightly different time periods: CMIP6 and the single-model large ensembles CSIRO-Mk3-6-0, IPSL-CM6A-LR, and MIROC6 cover the period 1850–2014, which we will use in all figures. CESM-LE starts in 1920; CanESM2 and GFDL-ESM2M start in 1950. The historical simulations from CMIP5, CESM-LE, and MPI-GE end in 2005. The climate models substantially differ in their representation of internal variability (Hawkins et al., 2016; Olonscheck & Notz, 2017; Deser et al., 2020). We conform all observational data sets and model simulations to a spatial resolution of 1° × 1° by bilinear interpolation and analyze annual means.

We choose four characteristic time periods that represent the historical evolution of SSTs, differentiated by their global warming behavior (Figure 1; Cahill et al., 2015; Rahmstorf et al., 2017; Yu & Ruggieri, 2019): the early warming in the twentieth century 1915–1944 (Brönnimann, 2009; Hegerl et al., 2018), the early hiatus 1945–1974 (Hegerl et al., 2019), the recent warming 1975–2004, and the recent hiatus 1995–2014 (Fyfe & Gillett, 2014; Medhaug et al., 2017). The period 1995–2014 overlaps with the period 1975–2004 but is of special interest as it reflects the most recent SST changes (Forster et al., 2019; Swart et al., 2019; Tokarska et al., 2020). All trends over the respective time periods are calculated by least squares linear regression. Our analysis is qualitatively not sensitive to the exact start and end dates of the periods. We use the different periods to test whether climate models can represent different observed trend patterns and whether a similar global signal (warming versus hiatus) is expressed in a similar spatial SST pattern.

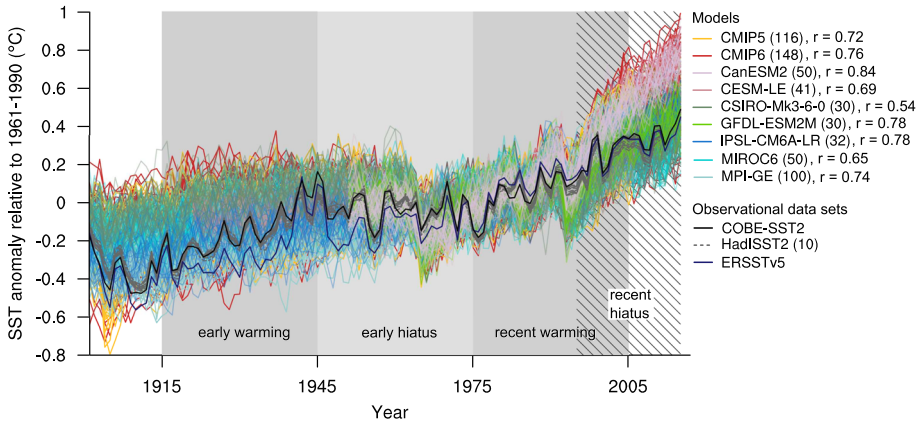

Figure 1. Evolution of the annual global mean sea surface temperature (SST) anomaly for 1900–2014 relative to 1961–1990 in CMIP5, CMIP6, seven single-model large ensembles, and three observational estimates. The number of members for each multimodel or single-model ensemble is shown in brackets. For each ensemble, the averaged correlation, *r*, between each ensemble member and each observational estimate is shown. The range of the correlation between the realizations of each model ensemble and the observational estimates is given in Table S1.

3. Global and Local Consistency of SST Changes

The global mean SST evolution in models matches the observations with a multimodel-averaged Pearson linear correlation coefficient r of 0.72 as expected from the well-discussed global temperature evolution (e.g., Hartmann et al., 2013; Marotzke & Forster, 2015; Rahmstorf et al., 2017) (Figure 1; see Table S1 in the supporting information for the spread in r across ensemble members). The observational data sets fall within the spread of CMIP5, CMIP6, and most single-model ensembles. During the last decade, the model ensembles show much stronger warming than observed (Oudar et al., 2018; Tokarska et al., 2020; see Table S2).

The observed local SST trends, shown in Figure 2, mirror the global SST evolution: In the warming periods, large regions warm coherently, while during the hiatus periods regional but spatially structured warming and cooling cancel out globally. The two warming periods differ mostly in the Kuroshio region, the extratropical South Pacific and the Southern Ocean; the differences are reflected by a pattern correlation of only $r = 0.36$ between the two periods. The differences in the patterns of the two warming periods can be due to the large observational uncertainty in the early warming period (see Figures S2, S4, and S5 for a comparison between the different observational data sets), due to differences in the forcing that drives the warming or due to multidecadal to centennial variability (e.g., Latif et al., 2013; Stevens, 2015; Hegerl et al., 2019).

The two hiatus periods differ more than the two warming periods: The early hiatus shows a warming trend of more than 0.1 K/decade over 23% of the ocean area, compensated by widespread cooling trends especially in the North Atlantic. The recent hiatus, in contrast, shows more pronounced patterns of change with a warming trend of more than 0.1 K/decade over 52% of the ocean area and largely compensating cooling trends in the eastern Pacific, parts of the Southern Ocean, and the North Atlantic. These substantial differences between the early and recent hiatus are reflected by a pattern correlation of only $r = 0.06$. Although SST patterns are not the only determining factor for hiatuses (Cohen et al., 2012; Li et al., 2015), the low-pattern correlation confirms that hiatuses can be realized with very different SST patterns (Hedemann et al., 2017; Känel et al., 2017; Medhaug et al., 2017).

We evaluate the consistency of simulations and observations in Figure 2. We define consistency between each model ensemble and the observational data sets as

$$\varphi = \frac{t_{mean} - t_{obs}}{\sigma} \tag{1}$$

where t_{mean} is the ensemble mean trend, interpreted as the simulated forced response, t_{obs} is the trend from the observational data, and σ is the standard deviation calculated across each ensemble (Olonscheck & Notz, 2017). Each term is calculated for the 30- and 20-year trends discussed above. We evaluate equation (1) for each single-model large ensemble and for CMIP5 and CMIP6. For CMIP5 and CMIP6 we use only the first ensemble member of each model ("r1i1p1" method; e.g., Colman & Hanson, 2013; Cox et al., 2018).

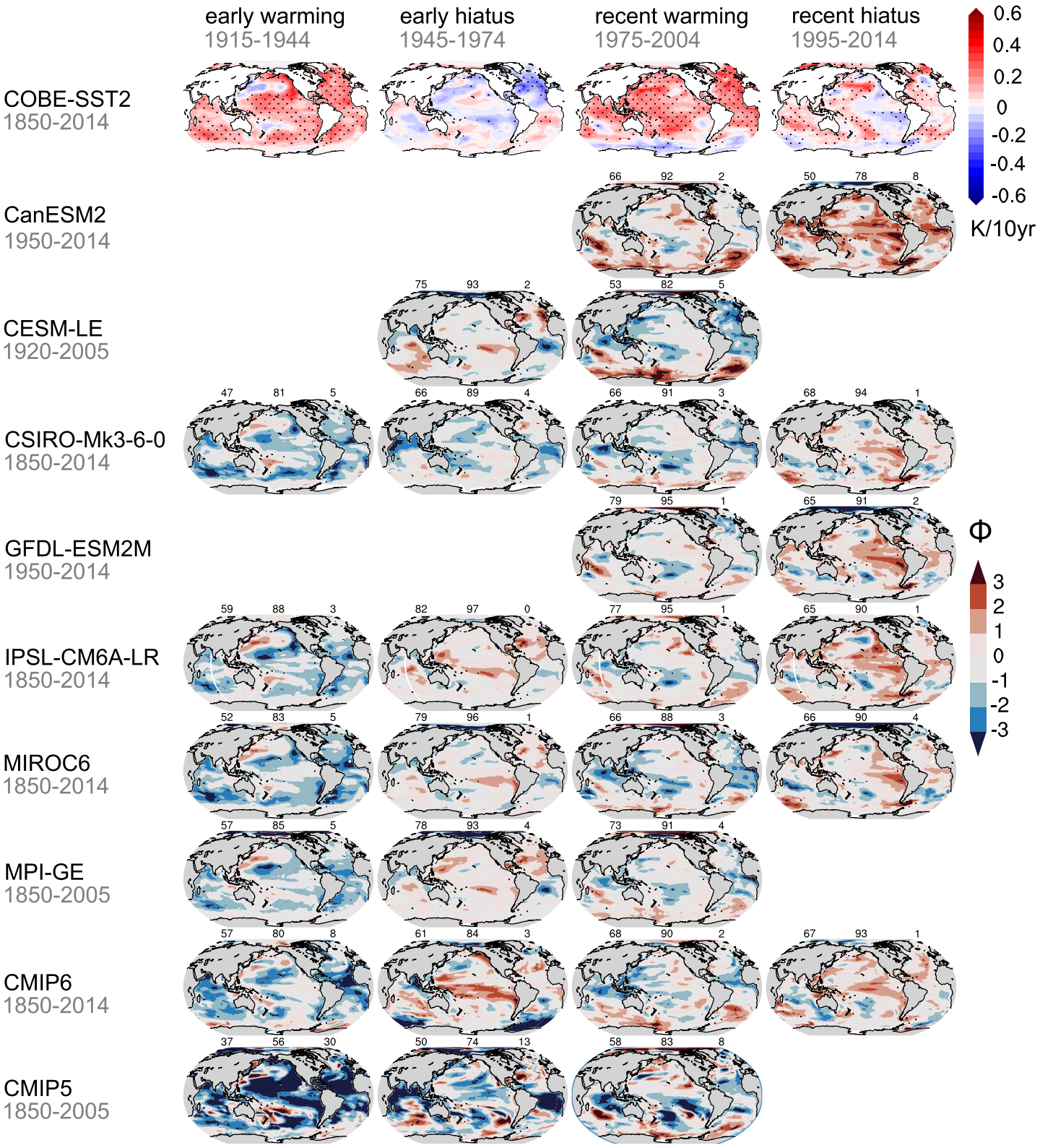

Figure 2. Observed SST trends for the four periods from COBE-SST2 (top row) and consistency as measured by equation (1) (other rows). Colors indicate that the observed trend lies below (red) or above (blue) the ensemble mean trend. The darker the color, the farther the observations are from the ensemble mean. Observations that lie outside ±3 ensemble standard deviations indicate that the model cannot reproduce the observed trend. The numbers on top of each panel show from left to right the percentage fraction of the ocean area in which the observed trend is within ±1, within ±2, and outside ±3 standard deviations of the simulated ensemble mean trend. Figures S4 and S5 show the consistency of simulated trends when compared to ERSSTv5 and HadISST2, respectively.

This approach samples the differences in model physics and to some degree the internal variability between models but does not emphasize the models that contributed a large number of ensemble members. Figure S6 shows the same analysis for taking into account all CMIP5 and CMIP6 ensemble members.

φ measures how far the observed trend deviates from the ensemble mean trend normalized by the model-specific standard deviation σ. The approach assumes a Gaussian fit to the distribution of trends in each ensemble. Figures S7 and S8 show each term of equation (1) for all model ensembles. The internal variability σ differs between the model ensembles, but the overall magnitude is similar in each model ensemble (Figure S8). Large-scale differences in internal variability occur only in the multimodel ensembles, with CMIP6 having a substantially larger ensemble standard deviation than CMIP5. Large-scale coherent patterns in φ mean that the model ensemble mean trend is weaker or stronger than in observations. We do not expect the ensemble mean to reproduce the observations, because the observations are drawn from the unknown PDF of SST pattern evolution of the real world, including internal variability. The simulations are consistent with the observations where the observed trend deviates by less than $\pm 2\sigma$ from the simulated trend. In contrast, the simulations are considered inconsistent with the observations where the observed trend deviates by more than $\pm 3\sigma$ from the ensemble mean. When averaged across the four time periods and all single-model ensembles, we find that in 90% of the ocean area the observed trends from COBE-SST2 are consistent with $\pm 2\sigma$ of the simulated distribution of trends (85% for ERSSTv5 and 88% for HadISST2). In contrast, in 3% of the ocean area the observed trends from COBE-SST2 are inconsistent with the simulated trends (5% for ERSSTv5 and 4% for HadISST2). This number is larger than the 0.3% of $\pm 3\sigma$ outliers we expect statistically, but compared to the uncertainty in the observations (Kennedy, 2014) this number does not allow the conclusion that the models are systematically inconsistent with the observations.

The consistency between observed and simulated trends differs for the four characteristic periods. For the early warming period, all model ensembles show too negative trends in large coherent regions in the subtropical and midlatitudinal oceans, reflecting that the models' global warming trend during this period is too weak (see also Figure 1 and Table S2). In contrast, for both the early hiatus and the recent warming period, most ensembles show small biases with a large-scale consistency. Exceptions for the recent warming period are CanESM2 and CESM-LE, which do not simulate the observed Southern Ocean cooling trend, while other model ensembles do. For the recent hiatus period, the most coherent inconsistency across model ensembles is the missing cooling trend in the eastern Pacific (Coats & Karnauskas, 2017; Seager et al., 2019). However, only CanESM2 shows regions where no ensemble member simulates a cooling trend as strong as observed, and 8% of the ocean area shows a deviation larger than $\pm 3\sigma$. For all other model ensembles the possibility remains that the observed cooling trend in the eastern equatorial Pacific over the period 1995–2014 is an extreme realization of the very large internal variability, which is captured by a few simulations (indicated by red in Figure 2, indicating deviations between 2 and 3σ).

4. Global and Regional Pattern Correlations and Model Improvement

To give a sense of the difference in observational data sets, we correlate the SST patterns from COBE-SST2 with those from ERSSTv5 and HadISST2 for the four characteristic periods (Figure 3). The pattern correlation ranges from 0.63 to 0.87 in the four trend periods, substantially smaller than 1 (see also section 2 and Figure S2).

We do not expect the pattern correlations between observed and simulated SST trends to be close to 1, because (1) the internal variability in the model simulations is not necessarily synchronized in time with the observed variability, certainly not everywhere at the same time; (2) uncertain historical forcings such as aerosols, volcanoes, and greenhouse gas concentrations are used in the simulations; and (3) the number of ensemble members could be too small to simulate a trend pattern that perfectly matches the observed one (Frankcombe et al., 2018; Milinski et al., 2019). From correlating the simulated trends in SST patterns of each ensemble member with the trend pattern from COBE-SST2, we find that every model ensemble simulates a broad range of pattern correlations, including negative correlations. This range of pattern correlations from single-model ensembles highlights the large variety of SST patterns that is realizable from internal variability alone. All ensembles show the highest pattern correlations in the recent warming period 1975–2004 when several ensemble members reach pattern correlations larger than 0.5. The high correlations are due to similar warming trends in observations and models (see Figure 1 and Table S2) and a rather uniform warming

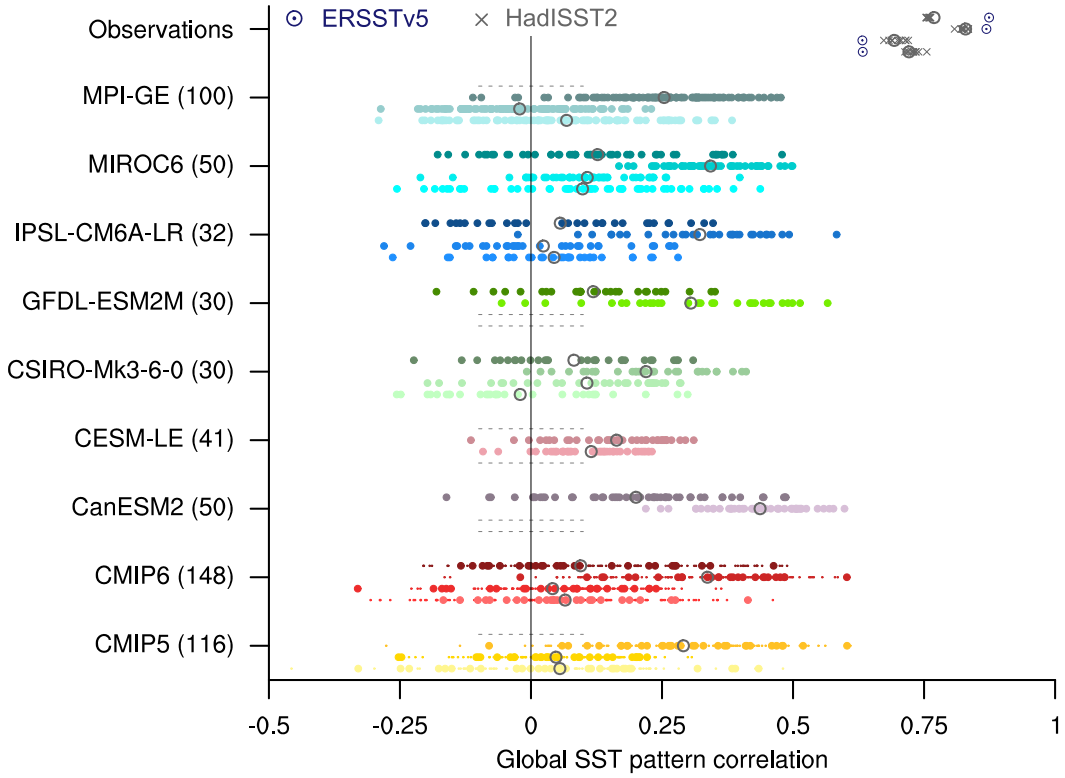

Figure 3. Global pattern correlation of each realization from CMIP5, CMIP6, the single-model large ensembles, and two observational estimates compared to the observational estimate COBE-SST2. The four rows (bottom to top) for each ensemble correspond to the early warming, the early hiatus, the recent warming, and the recent hiatus. Gray circles show the ensemble mean, and the dashed horizontal lines indicate missing data in the simulations. For CMIP6 and CMIP5, the larger filled dots show the first ensemble member of each model and the small dots all other ensemble members.

pattern. The ensemble members with the highest pattern correlations resemble or even closely match the large-scale SST patterns for the recent warming and the recent hiatus period (Figure S3).

We now investigate three ocean regions that are prone to large model trend biases and large observational uncertainty, namely, the equatorial Pacific (Figure 4a; Bordbar et al., 2017; Coats & Karnauskas, 2017; Seager et al., 2019), the Southern Ocean (Figure 4b; Latif et al., 2013; Hobbs et al., 2016; Kostov et al., 2018; Zhang et al., 2019), and the North Atlantic (Figure 4c; Ting et al., 2009; Ruiz-Barradas et al., 2018; Hand et al., 2019). The changes in the east-to-west SST gradient in the equatorial Pacific (3°S–3°N, 120–180°E minus 3°S–3°N, 85–155°W) range from strongly negative to strongly positive for all periods and each single-model ensemble, while the SST gradient slightly strengthens with time in the observations. Although for the recent hiatus, the observed positive trend is close to the upper bound of some model ensembles in agreement with Seager et al. (2019), we find individual members in all model ensembles that show SST gradients as strong as observed.

The SST trends of the Southern Ocean (40–90°S) relative to the 90°N to 40°S mean show strong changes over time both in the observations and the models. The Southern Ocean warms less than the global mean ocean during the two warming periods and the recent hiatus period but more than the global mean ocean during the early hiatus period. Notably, these significant changes of observed Southern Ocean SST trends throughout the four trend periods are reflected by the ensemble mean changes of many model ensembles (black circles in Figure 4b), which points to a forced response. Consequently, for all four trend periods, the observed SST trends are within the range of internal variability from the simulated trends of all model ensembles, except for CESM-LE for the recent warming period (Figure 2). Hence, we find a general consistency between observed and simulated Southern Ocean SST trends.

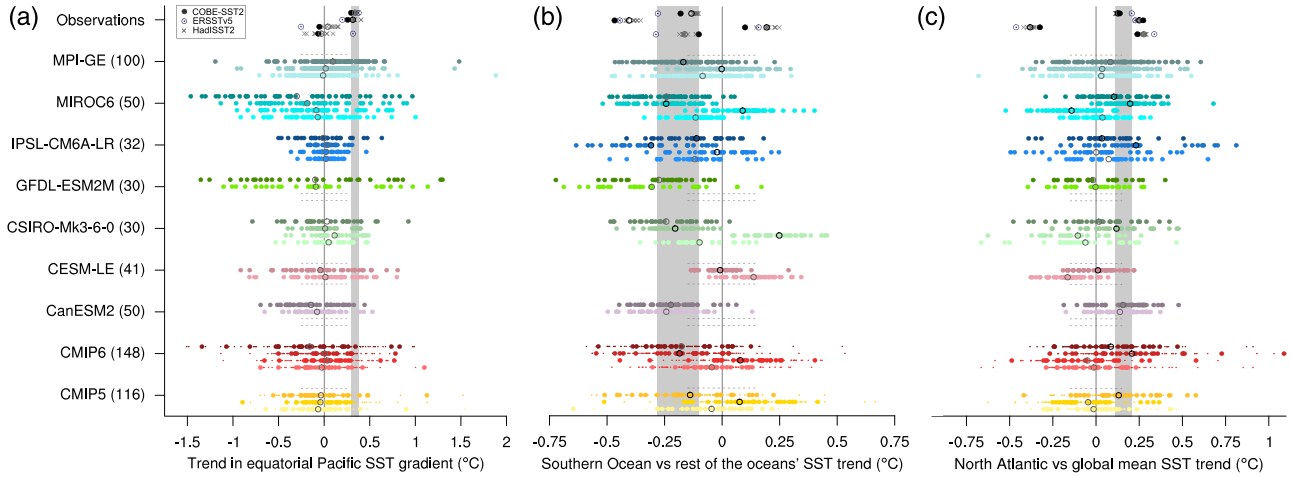

Figure 4. Trends in regional SST indices of each realization from CMIP5, CMIP6, the single-model large ensembles, and three observational estimates for (a) the tropical Pacific east-to-west SST gradient, (b) the Southern Ocean, and (c) the North Atlantic compared to the global mean ocean SST. The four rows for each ensemble correspond to (bottom to top) the early warming, early hiatus, recent warming, and recent hiatus. Dashed horizontal lines indicate unavailability of data for that model. The ensemble mean trend is shown as black circle when significantly different to the previous trend period, and as gray circle otherwise. Significance is tested with a Student's *t* test. For CMIP5 and CMIP6, the large filled dots show the first ensemble member of each model ("r1i1p1" method) and the small dots all other ensemble members. The light gray area highlights the observed range during the recent hiatus period.

Finally, and in contrast to the Southern Ocean, the observed SST trends in the North Atlantic (80–20°N,70–0°W) relative to the global mean ocean show strong warming during the early and recent warming period as well as in the recent hiatus period but a strong cooling during the early hiatus period (Ting et al., 2009; Taboada & Anadn, 2012; Hand et al., 2019). While this evolution is well represented by MIROC6, the strong cooling trend in the early hiatus period is at the lower bound of the internal variability range of all other model ensembles.

Overall, the comparison between observed and simulated SST trends shows that the trends in SST patterns are consistent in regions that are important for ocean heat uptake (e.g., Frölicher et al., 2014; Garuba et al., 2018; Kostov et al., 2014; Watanabe et al., 2013) and that influence the global atmospheric feedback parameter (e.g., Zhou et al., 2016; Frey et al., 2017; Andrews & Webb, 2018; Dong et al., 2019; Bloch-Johnson et al., 2020; Rugenstein et al., 2020). The consistency holds in these regions for the different periods representative of global temperature change. The coherently simulated large range of regional SST trends for all model ensembles highlights that internal variability can be more important for explaining regional SST trends than differences between models.

CMIP6 better represents the observed trends in SST patterns than CMIP5. For the local SST trends during all periods, the ocean area that is inconsistent with the observed trends decreases by about three quarters from CMIP5 to CMIP6 (see percentage on the top right of each panel in Figure 2). The area of consistency mostly increases even more when all realizations of each model are considered (Figure S6). The more widespread consistency of SST trends in CMIP6 compared to CMIP5 is in part caused by the larger multimodel spread in multidecadal SST trends in CMIP6 compared to CMIP5 (Figure S8) but might also point to either improvements in the model formulation or a stronger tuning of global mean temperatures to match the historical record in the recent generation of model intercomparison.

5. Limitations, Interpretation, and Conclusions

Our study is limited to quantifying the distance between the observations and model simulations (φ in equation (1) and Figure 2). An alternative comparison could include assigning observational uncertainties and estimates of internal variability of the real world at the grid point level. The influence of the mean state biases of models can be assessed with dedicated simulations and an increased understanding of local process (e.g., Richter, 2015; Sohn et al., 2016; Seager et al., 2019). Our study strengthens the argument for evaluating the models' realism of internal variability to better compare observed and simulated surface temperature and SST trends (McKinnon & Deser, 2018; Deser et al., 2020).

Benefiting from the large amount of model output, we demonstrate that current generation climate models are able to simulate the observed patterns of SST trends for characteristic historical time periods, such as the early and recent warming periods and the early and recent hiatuses. We find:

1. Observed and simulated trends in SST patterns are broadly consistent (Figures 2–4). Individual ensemble members simulate trends in SST patterns that resemble the observed patterns (Figure S3).
2. Internal variability strongly influences trends in SST patterns and can explain model differences of simulated 30- and 20-year SST trends better than differences in the model formulation or differences in the observational data sets.
3. No region is robustly inconsistent with the observations across multiple model ensembles. However, the observations lie toward the outer edge of the simulated distributions in similar regions in all model ensembles (Figure 2). Two nonexclusive interpretations are (a) observed trends are an unusual realization of the Earth's possible behavior and/or (b) the models are systematically biased but the large internal variability covers that bias and leads to occasional good matches with the observations in a large enough ensemble.
4. If ensemble mean and observations show the same direction of change, such as in the Southern Ocean and the North Atlantic throughout the historical period (Figures 4b and 4c, black circles), this suggests a forced pattern response. If ensemble mean and observations show the opposite direction of change as in the equatorial Pacific (Figure 4a), this could be by observational chance or indicative of a systematic structural model bias.
5. For CMIP6, the observational estimates fall within the simulated ranges over a larger area than for CMIP5 (Figure 2). This could be caused by the larger internal variability, improvements in the model formulation, or stronger tuning to the observations in CMIP6 compared to CMIP5.

Acknowledgments
This study was supported by the European Union's Horizon 2020 research and innovation program under Grant Agreement 820829 (CONSTRAIN project). M. R. is funded by the Alexander von Humboldt Foundation. We thank Clara Deser and Mohammad Hadi Bordbar for their helpful reviews, Christopher Hedemann for an internal review, Martin Stolpe for constructive comments, and the participants of the CONSTRAIN kickoff meeting for lively discussions. We acknowledge the U.S. CLIVAR Working Group on Large Ensembles for providing the archive of multimodel large ensembles. The model output used in this study can be accessed from the Multi-Model Large Ensemble Archive (http://www.cesm.ucar.edu/projects/community-projects/MMLEA/) and from the Earth System Grid Federation (https://esgf-node.llnl.gov/projects/cmip5/ and https://esgf-node.llnl.gov/projects/cmip6/). The observational data sets can be downloaded here (COBE-SST2: https://www.esrl.noaa.gov/psd/data/gridded/data.cobe2.html website, ERSSTv5: https://www.esrl.noaa.gov/psd/data/gridded/data.noaa.ersst.v5.html website, and HadISST2: https://www.metoffice.gov.uk/hadobs/hadisst2/). Data and scripts used in this study are available online (publications@mpimet.mpg.de).

References

Andrews, T., Gregory, J. M., Paynter, D., Silvers, L. G., Zhou, C., Mauritsen, T., & Titchner, H. (2018). Accounting for changing temperature patterns increases historical estimates of climate sensitivity. *Geophysical Research Letters*, *45*, 8490–8499. https://doi.org/10.1029/2018GL078887

Andrews, T., & Webb, M. J. (2018). The dependence of global cloud and lapse rate feedbacks on the spatial structure of tropical Pacific warming. *Journal of Climate*, *31*(2), 641–654. https://doi.org/10.1175/JCLI-D-17-0087.1

Armour, K. C. (2017). Energy budget constraints on climate sensitivity in light of inconstant climate feedbacks. *Nature Climate Change*, *7*, 331–335. https://doi.org/10.1038/nclimate3278

Bloch-Johnson, J., Rugenstein, M., & Abbot, D. S. (2020). Spatial radiative feedbacks from internal variability uusing multiple regression. *Journal of Climate*, *33*, 4121–4140. https://doi.org/10.1175/JCLI-D-19-0396.s1

Böning, C. W., Dispert, A., Visbeck, M., Rintoul, S. R., & Schwarzkopf, F. U. (2008). The response of the Antarctic Circumpolar Current to recent climate change. *Nature Geoscience*, *1*(12), 864–869. https://doi.org/10.1038/ngeo362

Bordbar, M. H., Martin, T., Latif, M., & Park, W. (2017). Role of internal variability in recent decadal to multidecadal tropical Pacific climate changes. *Geophysical Research Letters*, *44*, 4246–4255. https://doi.org/10.1002/2016GL072355

Brönnimann, S. (2009). Early twentieth-century warming. *Nature Geoscience*, *2*(11), 735–736. https://doi.org/10.1038/ngeo670

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., & Vuichard, N. (2020). Presentation and evaluation of the IPSL-CM6A-LR climate model. *Journal of Advances in Modeling Earth System*.

Cahill, N., Rahmstorf, S., & Parnell, A. C. (2015). Change points of global temperature. *Environmental Research Letters*, *10*(8), 084,002. https://doi.org/10.1088/1748-9326/10/8/084002

Carella, G., Kennedy, J. J., Berry, D. I., Hirahara, S., Merchant, C. J., Morak-Bozzo, S., & Kent, E. C. (2018). Estimating sea surface temperature measurement methods using characteristic differences in the diurnal cycle. *Geophysical Research Letters*, *45*, 363–371. https://doi.org/10.1002/2017GL076475

Cassou, C., Kushnir, Y., Hawkins, E., Pirani, A., Kucharski, F., Kang, I. S., & Caltabiano, N. (2018). Decadal climate variability and predictability: Challenges and opportunities. *Bulletin of the American Meteorological Society*, *99*(3), 479–490. https://doi.org/10.1175/BAMS-D-16-0286.1

Coats, S., & Karnauskas, K. B. (2017). Are simulated and observed twentieth century tropical Pacific Sea surface temperature trends significant relative to internal variability? *Geophysical Research Letters*, *44*, 9928–9937. https://doi.org/10.1002/2017GL074622

Coats, S., & Karnauskas, K. B. (2018). A role for the equatorial undercurrent in the ocean dynamical thermostat. *Journal of Climate*, *31*(16), 6245–6261. https://doi.org/10.1175/JCLI-D-17-0513.1

Cohen, J. L., Furtado, J. C., Barlow, M., Alexeev, V. A., & Cherry, J. E. (2012). Asymmetric seasonal temperature trends. *Geophysical Research Letters*, *39*, 1–7. https://doi.org/10.1029/2011GL050582

Colman, R. A., & Hanson, L. I. (2013). On atmospheric radiative feedbacks associated with climate variability and change. *Climate Dynamics*, *40*(1), 475–492. https://doi.org/10.1007/s00382-012-1391-3

Cox, P. M., Huntingford, C., & Williamson, M. S. (2018). Emergent constraint on equilibrium climate sensitivity from global temperature variability. *Nature*, *553*(7688), 319–322. https://doi.org/10.1038/nature25450

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

Diamond, M. S., & Bennartz, R. (2015). Occurrence and trends of eastern and central Pacific El Nino in different reconstructed SST data sets. *Geophysical Research Letters*, *42*, 10,375–10,381. https://doi.org/10.1002/2015GL066469

Dong, Y., Proistosescu, C., Armour, K. C., & Battisti, D. S. (2019). Attributing historical and future evolution of radiative feedbacks to regional warming patterns using a Green's function approach: The preeminence of the western Pacific. *Journal of Climate*, *32*, 5471–5491. https://doi.org/10.1175/JCLI-D-18-0843.1

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Forster, P. M., Maycock, A. C., McKenna, C. M., & Smith, C. J. (2019). Latest climate models confirm need for urgent mitigation. *Nature Climate Change*, *10*, 7–10. https://doi.org/10.1038/s41558-019-0660-0

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

Frey, W. R., Maroon, E. A., Pendergrass, A. G., & Kay, J. E. (2017). Do Southern Ocean cloud feedbacks matter for 21st century warming? *Geophysical Research Letters*, *44*, 12,447–12,456. https://doi.org/10.1002/2017GL076339

Frölicher, T. L., Sarmiento, J. L., Paynter, D. J., Dunne, J. P., Krasting, J. P., & Winton, M. (2014). Dominance of the Southern Ocean in anthropogenic carbon and heat uptake in CMIP5 models. *Journal of Climate*, *28*(2), 862–886. https://doi.org/10.1175/JCLI-D-14-00117.1

Fyfe, J. C., & Gillett, N. P. (2014). Recent observed and simulated warming. *Nature Climate Change*, *4*(3), 150–151. https://doi.org/10.1038/nclimate2111

Garuba, O. A., Lu, J., Liu, F., & Singh, H. A. (2018). The active role of the ocean in the temporal evolution of climate sensitivity. *Geophysical Research Letters*, *45*, 306–315. https://doi.org/10.1002/2017GL075633

Gregory, J. M., & Andrews, T. (2016). Variation in climate sensitivity and feedback parameters during the historical period. *Geophysical Research Letters*, *43*, 3911–3920. https://doi.org/10.1002/2016GL068406

Hand, R., Keenlyside, N. S., Omrani, N. E., Bader, J., & Greatbatch, R. J. (2019). The role of local sea surface temperature pattern changes in shaping climate change in the North Atlantic sector. *Climate Dynamics*, *52*(1), 417–438. https://doi.org/10.1007/s00382-018-4151-1

Hartmann, D., Tank, A. K., Rusticucci, M., Alexander, L., Brnnimann, S., Charabi, Y., & Zhai, P. (2013). Observations: Atmosphere and surface. In T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, et al. (Eds.), *In: Climate change 2013: The physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

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

Hedemann, C., Mauritsen, T., Jungclaus, J., & Marotzke, J. (2017). The subtle origins of surface-warming hiatuses. *Nature Climate Change*, *7*(5), 336–339. https://doi.org/10.1038/nclimate3274

Hegerl, G. C., Brnnimann, S., Cowan, T., Friedman, A. R., Hawkins, E., Iles, C., & Undorf, S. (2019). Causes of climate change over the historical record. *Environment Research Letters*, *14*(12), 123006. https://doi.org/10.1088/1748-9326/ab4557

Hegerl, G. C., Brnnimann, S., Schurer, A., & Cowan, T. (2018). The early 20th century warming: Anomalies, causes, and consequences. *WIREs Climate Change*, *9*(4), e522. https://doi.org/10.1002/wcc.522

Hirahara, S., Ishii, M., & Fukuda, Y. (2013). Centennial-scale sea surface temperature analysis and its uncertainty. *Journal of Climate*, *27*(1), 57–75. https://doi.org/10.1175/JCLI-D-12-00837.1

Hobbs, W., Curran, M., Abram, N., & Thomas, E. R. (2016). Century-scale perspectives on observed and simulated Southern Ocean sea ice trends from proxy reconstructions. *Journal of Geophysical Research: Oceans*, *121*, 7804–7818. https://doi.org/10.1002/2016JC012111

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J. C., Balaji, V., Duan, Q., & Williamson, D. (2017). The art and science of climate model tuning. *Bulletin of the American Meteorological Society*, *98*(3), 589–602. https://doi.org/10.1175/BAMS-D-15-00135.1

Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., & Zhang, H. M. (2017). Extended Reconstructed Sea Surface Temperature, version 5 (ERSSTv5): Upgrades, validations, and intercomparisons. *Journal of Climate*, *30*(20), 8179–8205. https://doi.org/10.1175/JCLI-D-16-0836.1

Jeffrey, S., Rotstayn, L. D., Collier, M., Dravitzki, S. M., Hamalainen, C., Moeseneder, C., & Syktus, J. (2013). Australia's CMIP5 submission using the CSIRO-Mk3.6 model.

Känel, L., Frölicher, T. L., & Gruber, N. (2017). Hiatus-like decades in the absence of equatorial Pacific cooling and accelerated global ocean heat uptake. *Geophysical Research Letters*, *44*, 7909–7918. https://doi.org/10.1002/2017GL073578

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., & Vertenstein, M. (2014). The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bulletin America Meteorology Society*, *96*(8), 1333–1349. https://doi.org/10.1175/BAMS-D-13-00255.1

Kennedy, J. J. (2014). A review of uncertainty in in situ measurements and data sets of sea surface temperature. *Reviews of Geophysics*, *52*, 1–32. https://doi.org/10.1002/2013RG000434

Kirchmeier-Young, M. C., Zwiers, F. W., & Gillett, N. P. (2016). Attribution of extreme events in Arctic sea ice extent. *Journal of Climate*, *30*(2), 553–571. https://doi.org/10.1175/JCLI-D-16-0412.1

Kostov, Y., Armour, K. C., & Marshall, J. (2014). Impact of the Atlantic meridional overturning circulation on ocean heat storage and transient climate change. *Geophysical Research Letters*, *41*, 2108–2116. https://doi.org/10.1002/2013GL058998

Kostov, Y., Ferreira, D., Armour, K. C., & Marshall, J. (2018). Contributions of greenhouse gas forcing and the Southern Annular Mode to historical Southern Ocean surface temperature trends. *Geophysical Research Letters*, *45*, 1086–1097. https://doi.org/10.1002/2017GL074964

Kravtsov, S., Grimm, C., & Gu, S. (2018). Global-scale multidecadal variability missing in state-of-the-art climate models. *npj Climate and Atmospheric Science*, *1*(1), 34. https://doi.org/10.1038/s41612-018-0044-6

Laepple, T., & Huybers, P. (2014a). Global and regional variability in marine surface temperatures. *Geophysical Research Letters*, *41*, 2528–2534. https://doi.org/10.1002/2014GL059345

Laepple, T., & Huybers, P. (2014b). Ocean surface temperature variability: Large model-data differences at decadal and longer periods. *Proceedings of the National Academy of Sciences*, *111*(47), 16,682–16,687. https://doi.org/10.1073/pnas.1412077111

Latif, M., Martin, T., & Park, W. (2013). Southern Ocean sector centennial climate variability and recent decadal trends. *Journal of Climate*, *26*(19), 7767–7782. https://doi.org/10.1175/JCLI-D-12-00281.1

Li, C., Stevens, B., & Marotzke, J. (2015). Eurasian winter cooling in the warming hiatus of 1998–2012. *Geophysical Research Letters*, *42*, 8131–8139. https://doi.org/10.1002/2015GL065327

Maher, N., Milinski, S., Suarez-Gutierrez, L., Botzet, M., Dobrynin, M., Kornblueh, L., & Marotzke, J. (2019). The Max Planck Institute grand ensemble: Enabling the exploration of climate system variability. *Journal of Advances in Modeling Earth Systems*, *11*, 2050–2069. https://doi.org/10.1029/2019MS001639

Marotzke, J., & Forster, P. M. (2015). Forcing, feedback and internal variability in global temperature trends. *Nature*, *517*(7536), 565–570. https://doi.org/10.1038/nature14117

Marvel, K., Pincus, R., Schmidt, G. A., & Miller, R. L. (2018). Internal variability and disequilibrium confound estimates of climate sensitivity from observations. *Geophysical Research Letters*, *45*, 1595–1601. https://doi.org/10.1002/2017GL076468

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

Medhaug, I., Stolpe, M. B., Fischer, E. M., & Knutti, R. (2017). Reconciling controversies about the global warming hiatus. *Nature*, *545*(7652), 41–47. https://doi.org/10.1038/nature22315

Milinski, S., Maher, N., & Olonscheck, D. (2019). How large does a large ensemble need to be? *Earth System Dynamics Discussions*, 1–19, in review. https://doi.org/10.5194/esd-2019-70

Olonscheck, D., & Notz, D. (2017). Consistently estimating internal climate variability from climate model simulations. *Journal of Climate*, *30*(23), 9555–9573. https://doi.org/10.1175/JCLI-D-16-0428.1

Oudar, T., Kushner, P. J., Fyfe, J. C., & Sigmond, M. (2018). No impact of anthropogenic aerosols on early 21st century global temperature trends in a large initial-condition ensemble. *Geophysical Research Letters*, *45*, 9245–9252. https://doi.org/10.1029/2018GL078841

Paynter, D., & Frölicher, T. L. (2015). Sensitivity of radiative forcing, ocean heat uptake, and climate feedback to changes in anthropogenic greenhouse gases and aerosols. *Journal of Geophysical Research: Atmospheres*, *120*, 9837–9854. https://doi.org/10.1002/2015JD023364

Rahmstorf, S., Foster, G., & Cahill, N. (2017). Global temperature evolution: Recent trends and some pitfalls. *Environment Research Letters*, *12*(5), 054001. https://doi.org/10.1088/1748-9326/aa6825

Richter, I. (2015). Climate model biases in the eastern tropical oceans: Causes, impacts and ways forward. *WIREs Climate Change*, *6*(3), 345–358. https://doi.org/10.1002/wcc.338

Rodgers, K. B., Lin, J., & Frlicher, T. L. (2015). Emergence of multiple ocean ecosystem drivers in a large ensemble suite with an Earth system model. *Biogeosciences*, *12*(11), 3301–3320. https://doi.org/10.5194/bg-12-3301-2015

Rugenstein, M., Bloch-Johnson, J., Gregory, J., Andrews, T., Mauritsen, T., Li, C., & Knutti, R. (2020). Equilibrium climate sensitivity estimated by equilibrating climate models. *Geophysical Research Letters*, *47*, 1–12. https://doi.org/10.1029/2019GL083898

Ruiz-Barradas, A., Chafik, L., Nigam, S., & Hkkinen, S. (2018). Recent subsurface North Atlantic cooling trend in context of Atlantic decadal-to-multidecadal variability. *Tellus A: Dynamic Meteorology and Oceanography*, *70*(1), 1–19. https://doi.org/10.1080/16000870.2018.1481688

Seager, R., Cane, M., Henderson, N., Lee, D. E., Abernathey, R., & Zhang, H. (2019). Strengthening tropical Pacific zonal sea surface temperature gradient consistent with rising greenhouse gases. *Nature Climate Change*, *9*(7), 517–522. https://doi.org/10.1038/s41558-019-0505-x

Sohn, B. J., Lee, S., Chung, E. S., & Song, H. J. (2016). The role of the dry static stability for the recent change in the Pacific Walker circulation. *Journal of Climate*, *29*(8), 2765–2779. https://doi.org/10.1175/JCLI-D-15-0374.1

Stevens, B. (2015). Rethinking the lower bound on aerosol radiative forcing. *Journal of Climate*, *28*(12), 4794–4819. https://doi.org/10.1175/JCLI-D-14-00656.1

Swart, N. C., Cole, J. N. S., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., & Winter, B. (2019). The Canadian Earth System Model version 5 (CanESM5.0.3). *Geoscientific Model Development*, *12*(11), 4823–4873. https://doi.org/10.5194/gmd-12-4823-2019

Taboada, F. G., & Anadn, R. (2012). Patterns of change in sea surface temperature in the North Atlantic during the last three decades: Beyond mean trends. *Climatic Change*, *115*(2), 419–431. https://doi.org/10.1007/s10584-012-0485-6

Tatebe, H., Ogura, T., Nitta, T., Komuro, Y., Ogochi, K., Takemura, T., & Kimoto, M. (2019). Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in MIROC6. *Geoscientific Model Development*, *12*(7), 2727–2765. https://doi.org/10.5194/gmd-12-2727-2019

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

Ting, M., Kushnir, Y., Seager, R., & Li, C. (2009). Forced and internal twentieth-century SST trends in the North Atlantic. *Journal of Climate*, *22*(6), 1469–1481. https://doi.org/10.1175/2008JCLI2561.1

Titchner, H. A., & Rayner, N. A. (2014). The Met Office Hadley Centre sea ice and sea surface temperature data set, version 2: 1. Sea ice concentrations. *Journal of Geophysical Research: Atmospheres*, *119*, 2864–2889. https://doi.org/10.1002/2013JD020316

Tokarska, K. B., Stolpe, M. B., Sippel, S., Fischer, E. M., Smith, C. J., Lehner, F., & Knutti, R. (2020). Past warming trend constrains future warming in CMIP6 models. *Science Advances*, *6*(12), 1–13. https://doi.org/10.1126/sciadv.aaz9549

Watanabe, M., Kamae, Y., Yoshimori, M., Oka, A., Sato, M., Ishii, M., & Kimoto, M. (2013). Strengthening of ocean heat uptake efficiency associated with the recent climate hiatus. *Geophysical Research Letters*, *40*, 3175–3179. https://doi.org/10.1002/grl.50541

Yasunaka, S., & Hanawa, K. (2011). Intercomparison of historical sea surface temperature datasets. *International Journal of Climatology*, *31*(7), 1056–1073. https://doi.org/10.1002/joc.2104

Yu, M., & Ruggieri, E. (2019). Change point analysis of global temperature records. *International Journal of Climatology*, *39*(8), 3679–3688. https://doi.org/10.1002/joc.6042

Zhang, L., Delworth, T. L., Cooke, W., & Yang, X. (2019). Natural variability of Southern Ocean convection as a driver of observed climate trends. *Nature Climate Change*, *9*(1), 59–65. https://doi.org/10.1038/s41558-018-0350-3

Zhou, C., Zelinka, M. D., & Klein, S. A. (2016). Impact of decadal cloud variations on the Earth's energy budget. *Nature Geoscience*, *9*(12), 871–874. https://doi.org/10.1038/ngeo2828