Geophysical Research Letters

RESEARCH LETTER
10.1029/2019GL085378

Evaluating the Performance of Past Climate Model Projections

Zeke Hausfather[1], Henri F. Drake[2,3], Tristan Abbott[3], and Gavin A. Schmidt[4]

[1]Energy and Resources Group, University of California, Berkeley, CA, USA, [2]Massachusetts Institute of Technology/Woods Hole Oceanographic Institution Joint Program in Oceanography, Woods Hole, MA, USA, [3]Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA, [4]NASA Goddard Institute for Space Studies, Broadway, NY, USA

Key Points:
- Evaluation of uninitialized multidecadal climate model future projection performance provides a concrete test of model skill
- The quasi-linear relationship between model/observed forcings and temperature change is used to control for errors in projected forcing
- Model simulations published between 1970 and 2007 were skillful in projecting future global mean surface warming

Supporting Information:
- Supporting Information S1

Correspondence to:
Z. Hausfather,
hausfath@gmail.com

Citation:
Hausfather, Z., Drake, H. F., Abbott, T., & Schmidt, G. A. (2020). Evaluating the performance of past climate model projections. *Geophysical Research Letters*, *47*, e2019GL085378. https://doi.org/10.1029/2019GL085378

Received 16 SEP 2019
Accepted 26 NOV 2019
Accepted article online 4 DEC 2019

Abstract Retrospectively comparing future model projections to observations provides a robust and independent test of model skill. Here we analyze the performance of climate models published between 1970 and 2007 in projecting future global mean surface temperature (GMST) changes. Models are compared to observations based on both the change in GMST over time and the change in GMST over the change in external forcing. The latter approach accounts for mismatches in model forcings, a potential source of error in model projections independent of the accuracy of model physics. We find that climate models published over the past five decades were skillful in predicting subsequent GMST changes, with most models examined showing warming consistent with observations, particularly when mismatches between model-projected and observationally estimated forcings were taken into account.

Plain Language Summary Climate models provide an important way to understand future changes in the Earth's climate. In this paper we undertake a thorough evaluation of the performance of various climate models published between the early 1970s and the late 2000s. Specifically, we look at how well models project global warming in the years after they were published by comparing them to observed temperature changes. Model projections rely on two things to accurately match observations: accurate modeling of climate physics and accurate assumptions around future emissions of CO_2 and other factors affecting the climate. The best physics-based model will still be inaccurate if it is driven by future changes in emissions that differ from reality. To account for this, we look at how the relationship between temperature and atmospheric CO_2 (and other climate drivers) differs between models and observations. We find that climate models published over the past five decades were generally quite accurate in predicting global warming in the years after publication, particularly when accounting for differences between modeled and actual changes in atmospheric CO_2 and other climate drivers. This research should help resolve public confusion around the performance of past climate modeling efforts and increases our confidence that models are accurately projecting global warming.

1. Introduction

Physics-based models provide an important tool to assess changes in the Earth's climate due to external forcing and internal variability (e.g., Arrhenius, 1896; IPCC, 2013). However, evaluating the performance of these models can be challenging. While models are commonly evaluated by comparing "hindcasts" of prior climate variables to historical observations, the development of hindcast simulations is not always independent from the tuning of parameters that govern unresolved physics (Gettelman et al., 2019; Mauritsen et al., 2019; Schmidt et al., 2017). There has been relatively little work evaluating the performance of climate model projections over their future projection period (referred to hereafter as model projections), as much of the research tends to focus on the latest generation of modeling results (Eyring et al., 2019).

Many different sets of climate projections have been produced over the past several decades. The first time series projections of future temperatures were computed using simple energy balance models in the early 1970s, most of which were solely constrained by a projected external forcing time series (originally, CO_2 concentrations) and an estimate of equilibrium climate sensitivity from single-column radiative-convective equilibrium models (e.g., Manabe & Wetherald, 1967) or general circulation models (e.g., Manabe & Wetherald, 1975). Simple energy balance models have since been gradually sidelined in favor of

increasingly high resolution and comprehensive general circulation models, which were first published in the late 1980s (e.g., Hansen et al., 1988; IPCC, 2013; Stouffer et al., 1989).

Climate model projections are usefully thought about as predictions conditional upon a specific forcing scenario. We consider these to be projections of possible future outcomes when the intent was to use a realistic forcing scenario and where the realized forcings were qualitatively similar to the projection forcings. Evaluating model projections against observations subsequent to model development provides a test of model skill, and successful projections can concretely add confidence in the process of making projections for the future. However, evaluating future projection performance requires a sufficient period of time post-publication for the forced signal present in the model projections to be differentiable from the noise of natural variability (Hansen et al., 1988; Hawkins & Sutton, 2012).

Researchers have previously evaluated prior model projections from the Hansen et al. (1988) National Aeronautics and Space Administration Goddard Institute for Space Studies model (Hargreaves, 2010; Rahmstorf et al., 2007), the Stouffer et al. (1989) Geophysical Fluid Dynamics Laboratory model (Stouffer & Manabe, 2017), the IPCC First Assessment Report (FAR-IPCC, 1990; Frame & Stone, 2012), and the IPCC Third and Fourth Assessment reports (IPCC, 2001; IPCC, 2007; Rahmstorf et al., 2012). However, to-date there has been no systematic review of the performance of past climate models, despite the availability of warming projections starting in 1970.

This paper analyzes projections of global mean surface temperature (GMST) change, one of the most visible climate model outputs, from several generations of past models. GMST plays a large role in determining climate impacts, is tied directly to international-agreed-upon mitigation targets, and is one of the climate variables that has the most accurate and longest observational records. GMST is also the output most commonly available for many early climate models run in the 1970s and 1980s.

Two primary factors influence the long-term performance of model GMST projections: (1) the accuracy of the model physics, including the sensitivity of the climate to external forcings and the resolution or parameterization of various physical processes such as heat uptake by the deep ocean and (2) the accuracy of projected changes in external forcing due to greenhouse gases and aerosols, as well as natural forcing such as solar or volcanic forcing.

While climate models should be evaluated based on the accuracy of model physics formulations, climate modelers cannot be expected to accurately project future emissions and associated changes in external forcings, which depend on human behavior, technological change, and economic and population growth. Climate modelers often bypass the task of deterministically predicting future emissions by instead projecting a range of forcing trajectories representative of several plausible futures bracketed by marginally plausible extremes. For example, Hansen et al., 1988 consider a low-emissions extreme Scenario C with "more drastic curtailment of emissions than has generally been imagined," a high-emissions extreme Scenario A wherein emissions "must eventually be on the high side of reality," as well as a middle-ground Scenario B, which "is perhaps the most plausible of the three." More recently, the Representative Concentration Pathways (RCPs) used in CMIP5 and the IPCC AR5 report similarly includes a number of plausible scenarios bracketed by a low-emissions extreme Scenario RCP2.6 and a high-emissions extreme Scenario RCP8.5 (van Vuuren et al., 2011). Thus, an evaluation of model projection performance should focus on the relationship between the model forcings and temperature change, rather than simply assessing how well projected temperatures compare to observations, particularly in cases where projected forcings differ substantially from our best estimate of the subsequently observed forcings.

This approach—comparing the relationship between forcing and temperatures in both model projections and observations—can effectively assess the performance of the model physics while accounting for potential mismatches in projected forcing that climate modelers did not address at the time. In this paper we apply both a conventional assessment of the change in temperature over time and a novel assessment of the response of temperature to the change in forcing to assess the performance of future projections by past climate models compared to observations.

Climate modeling efforts have advanced substantially since the first modern single-column (Manabe & Strickler, 1964) and general circulation models (Manabe et al., 1965) of Earth's climate were published in the mid-1960s, resulting in continually improving model hindcast skill (Knutti et al., 2013; Reichler &

Kim, 2008). While these improvements have rendered virtually all of the models described here operationally obsolete, they remain valuable tools as they are in a unique position to have their projections evaluated by virtue of their decades-long postpublication projection periods.

2. Methods

We conducted a literature search to identify papers published prior to the early-1990s that include climate model outputs containing both a time series of projected future GMST (with a minimum of two points in time) and future forcings (including both a publication date and future projected atmospheric CO_2 concentrations, at a minimum). Eleven papers with 14 distinct projections were identified that fit these criteria. Starting in the mid-1990s, climate modeling efforts were primarily undertaken in conjunction with the IPCC process (and later, the Coupled Model Intercomparison Projects, CMIPs), and model projections were taken from models featured in the IPCC FAR (1990), Second Assessment Report (SAR-IPCC, 1996), Third Assessment Report (TAR-IPCC, 2001), and Fourth Assessment Report (AR4-IPCC, 2007).

The specific models projections evaluated were Manabe, 1970 (hereafter Ma70), Mitchell, 1970 (Mi70), Benson, 1970 (B70), Rasool & Schneider, 1971 (RS71), Sawyer, 1972 (S72), Broecker, 1975 (B75), Nordhaus, 1977 (N77), Schneider & Thompson, 1981 (ST81), Hansen et al., 1981 (H81), Hansen et al., 1988 (H88), and Manabe & Stouffer, 1993 (MS93). The energy balance model projections featured in the main text of the FAR, SAR, and TAR were examined, while the CMIP3 multimodel mean (and spread) was examined for the AR4 (multimodel means were not used as the primary IPCC projections featured in the main text prior to the AR4). Details about how each individual model projection was digitized and analyzed as well as assessments of individual models included in the first three IPCC reports can be found in the supporting information.

The AR4 projection was excluded from the main analysis in the paper as both the observational uncertainties and model projection uncertainties are too large over the short 2007–2017 period to draw many useful conclusions, and its inclusion makes the figures difficult to read. However, analyses including the AR4 projection can be found in the supporting information.

We assessed model projections over the period between the date the model projection was published and the end of 2017 or when the model projection ended in cases where model runs did not extend through 2017. An end date of 2017 was chosen for the analysis because the ensemble of observational estimates of radiative forcings we used only extends through that date.

Five different observational temperature time series were used in this analysis—National Aeronautics and Space Administration GISTEMP (Lenssen et al., 2019), National Oceanic and Atmospheric Administration GlobalTemp (Vose et al., 2012), Hadley/UEA HadCRUT4 (Morice et al., 2012), Berkeley Earth (Rohde et al., 2013), and Cowtan and Way (Cowtan & Way, 2014). The observational temperature records used do not present a completely like-to-like comparison with models, as models provide surface air temperature (SAT) fields while observations are based on SAT fields over land and sea surface temperature (SST) fields over the ocean. This means that the trends in the models used here are likely biased high compared to observations, as model blended field trends are about 7% (±5%) lower than model global SAT fields over the 1970–2017 period (Cowtan et al., 2015; Richardson et al., 2016). However, the absence of SST fields from the models analyzed here prevents a comparison of blended SAT/SST against observations.

We compared observations to climate model projections over the model projection period using two approaches: change in temperature versus time and change in temperature versus change in radiative forcing ("implied TCR"). We use an implied TCR metric to provide a meaningful model-observation comparison even in the presence of forcing differences. Implied TCR is calculated by regressing temperature change against radiative forcing for both models and observations, and multiplying the resulting values by the forcing associated with doubled atmospheric CO_2 concentrations, F_{2x}, (following Otto et al., 2013):

$$TCR_{implied} = F_{2x}\Delta T/\Delta F_{anthro}$$

We express implied TCR with units of temperature using a fixed value of $F_{2x} = 3.7$ W/m^2 (Vial et al., 2013). ΔF_{anthro} includes only anthropogenic forcings and excludes volcanic and solar changes to avoid

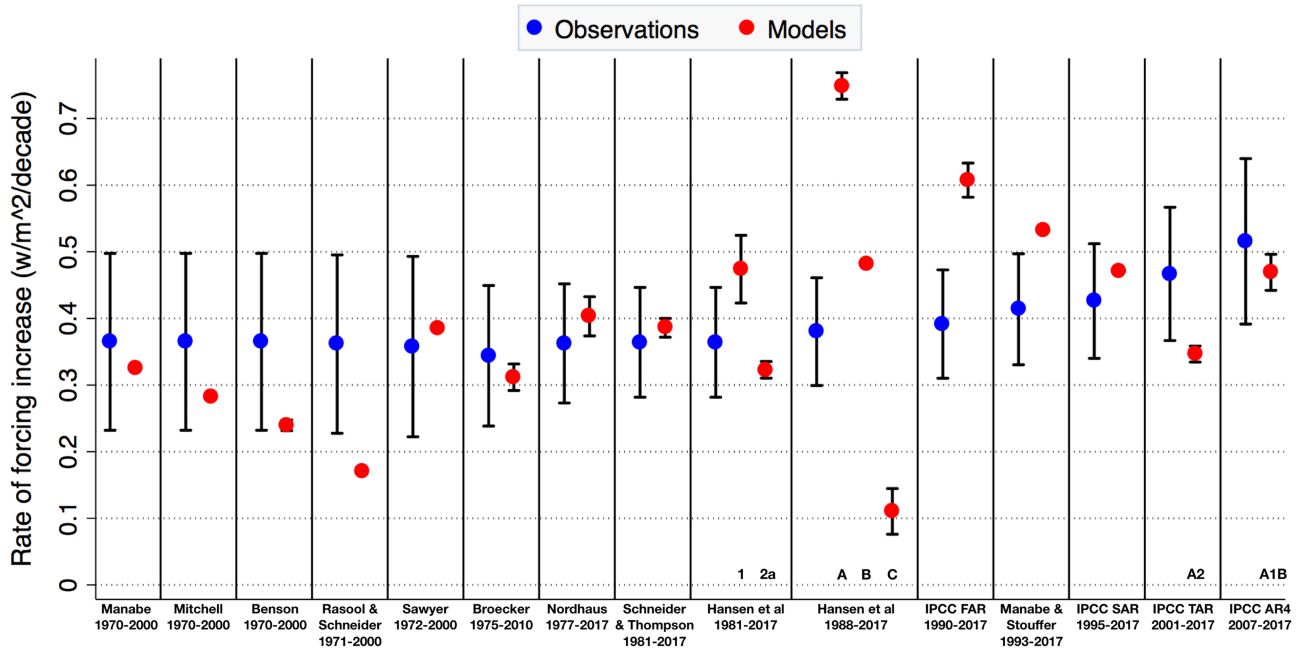

Figure 1. Rate of external forcing increase (in watts per meter squared per decade) in models and observations over model projection periods

introducing sharp interannual changes in forcing that would complicate the interpretation of TCR over shorter time periods. For the observational record, ΔF_{anthro} is based on a 1,000-member ensemble of observationally informed forcing estimates (Dessler & Forster, 2018). Model forcings are recomputed from published formulas and tables when possible and otherwise digitized from published figures (see supporting information section S2 for details). Instantaneous forcings rather than effective or efficacy-adjusted forcing are used, as those are all that is available for some early models (Hansen et al., 2005; Marvel et al., 2016; see supporting information section S1.0). Details on the approach used to calculate implied TCR can be found in supporting information section S1.2.

Comparing models and observations via implied TCR assumes a linear relationship between forcing and warming, an approach that has been widely used in prior analyses (Gregory et al., 2004; Otto et al., 2013). If forcing varies sufficiently slowly in time and deep ocean temperatures remain approximately constant, then a linear relationship is expected to hold with a constant of proportionality that depends on the strength of radiative feedbacks and ocean heat uptake (Held et al., 2010). In this regime, our implied TCR metric provides information about model physics and is unaffected by the time rate of change of forcing; moreover, previous studies have suggested that the temperature response to twentieth century anthropogenic forcing falls within this regime (Gregory & Forster, 2008; Gregory & Mitchell, 1997; Held et al., 2010).

However, sudden increases or decreases such as those associated with volcanic eruptions will not engender an equivalent immediate temperature response. For this reason, only anthropogenic forcings were used in estimating $TCR_{implied}$, as all models evaluated lacked additional volcanic events during their projection periods with the exception of Scenarios B and C of H88. Similarly, thermal inertia in the climate system can affect the relationship between temperature and external forcing if forcing increases sufficiently rapidly (Geoffroy et al., 2012). Scenarios where forcing is rapidly increasing will, all things being equal, tend to be further away from an equilibrium state than scenarios with more gradual increase after a given period of time (Rohrschneider et al., 2019) and thus have a lower implied TCR. With a few exceptions (e.g., RS71, H88 Scenarios A and C), however, most models evaluated had a rate of external forcing increase in the projection period within 1.3 times of the mean estimate of observational forcings and thus likely fall into the regime where implied TCR depends largely on radiative feedbacks and ocean heat uptake.

In this analysis we refer to model projections as consistent or inconsistent with observations based on a comparison of the differences between the two. Specifically, if the 95% confidence interval in the differences

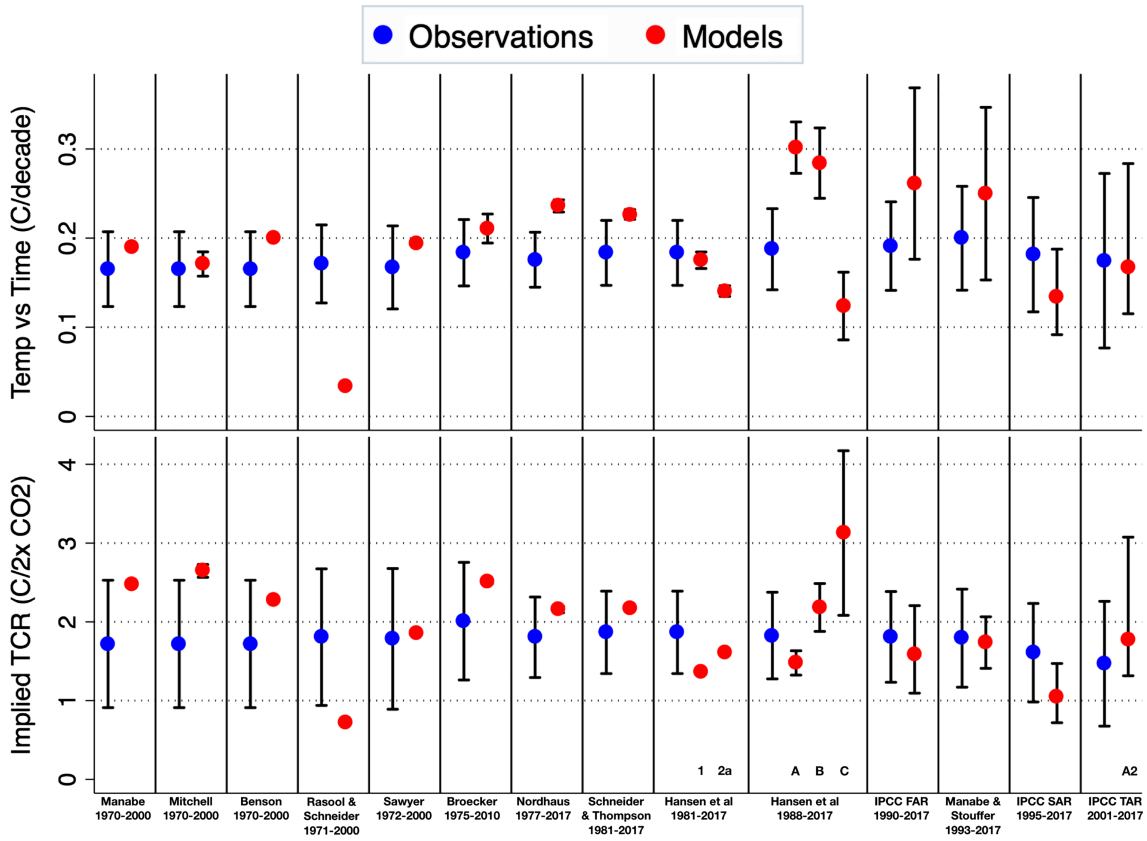

Figure 2. Comparison of trends in temperature versus time (top panel) and implied TCR (bottom panel) between observations and models over the model projection periods displayed at the bottom of the figure. Figure S1 shows a variant of this figure with the AR4 projections included

between the modeled and observed metrics includes 0, the two are deemed consistent; otherwise, they are inconsistent (Hausfather et al., 2017). Additionally, we follow the approach of Hargreaves (2010) in calculating a skill score for each model for both temperature versus time and implied TCR metrics. This skill score is based on the root-mean-square errors of the model projection trend versus observations compared to a zero-change null-hypothesis projection. See supporting information section S1.3 for details on calculating consistency and skill scores.

3. Results

A direct comparison of projected and observed temperature change during each historical model's projection period can provide an effective test of model skill, provided that model projection forcings are reasonably in-line with the ensemble of observationally informed estimates of radiative forcings. In about 9 of the 17 model projections examined, the projected forcings were within the uncertainty envelope of observational forcing ensemble. However, the remaining eight models—RS71, H81 Scenario 1, H88 Scenarios A, B, and C, FAR, MS93, and TAR—had projected forcings significantly stronger or weaker than observed (Figure 1). For the latter, an analysis comparing the implied TCR between models and observations may provide a more accurate assessment of model performance.

Comparisons between climate models and observations over model projection periods are shown in Figure 2 for both temperature versus time and implied TCR metrics (differences between models and observations are shown in Figure S2). Overall the majority of model projections considered were consistent with observations under both metrics. Using the temperature versus time metric, 10 of the 17 model projections show results consistent with observations. Of the remaining seven model projections, four project more warming than observed—N77, ST81, and H88 Scenarios A and B—while three project less warming than observed—RS71, H81 Scenario 2a, and H88 Scenario C.

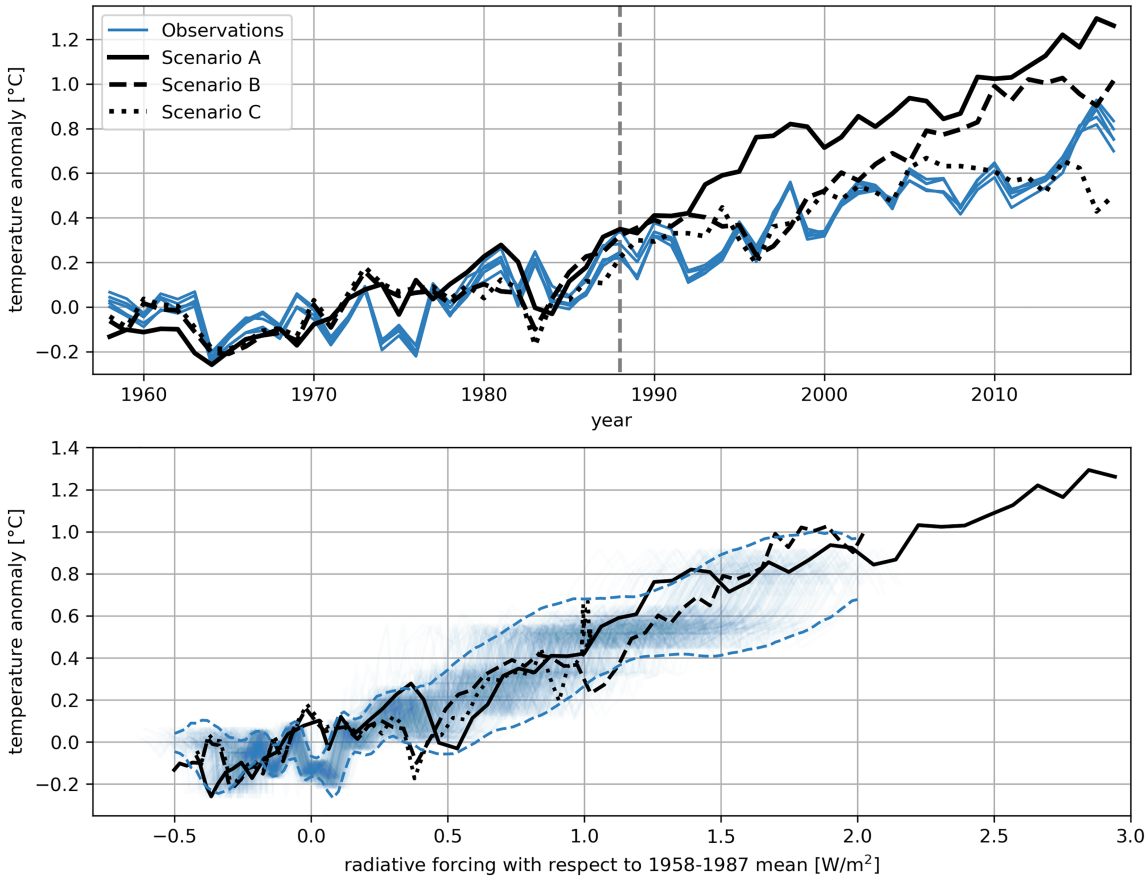

Figure 3. Hansen et al., 1988 projections compared with observations on a temperature versus time basis (top) and temperature versus external forcing (bottom). The dashed gray line in the top panel represent the start of the projection period. The transparent blue lines in the lower panel represent 500 random samples of the 5,000 combinations of the five temperature observation products and the 1,000 ensemble members of estimated forcings (the full ensemble is subsampled for visual clarity). The dashed blue lines show the 95% confidence intervals for the 5,000-member ensemble (see supporting information Text S1.4 for details). Anomalies for both temperature and forcing are shown relative to a 1958–1987 preprojection baseline.

When mismatches between projected and observed forcings are taken into account, a better performance is seen. Using the implied TCR metric, 14 of the 17 model projections were consistent with observations; of the three that were not, Mi70 and H88 Scenario C showed higher implied TCR than observations, while RS71 showed lower implied TCR (Schneider, 1975; see supporting information Text S2 for a discussion of the anomalously low-equilibrium climate sensitivity (ECS) model used in RS71).

A number of model projections were inconsistent with observations on a temperature versus time basis but are consistent once mismatches between modeled and observed forcings are taken into account. For example, whileN77 and ST81 projected more warming than observed, their implied TCRs are consistent with observations despite forcings within—though on the high end of—the ensemble range of observational estimates. Similarly, while H81 Scenario 2a projects less warming than observed, its implied TCR is consistent with observations.

A number of 1970s-era models (Ma70, Mi70, B70, B75, and N77) show implied TCR on the high end of the observational ensemble-based range. This is likely due to their assumption that the atmosphere equilibrates instantly with external forcing, which omits the role of transient ocean heat uptake (Hansen et al., 1985). However, despite this high implied TCR, a number of the models (e.g., Ma70, Mi70, B70, and B75) still end up providing temperature projections in-line with observations as their forcings were on the lower end of observations due to the absence of any non-CO_2 forcing agents in their projections.

In principle, the same underlying model should show consistent results for modestly different forcing scenarios under the implied TCR metric. However, the inconsistency of the H88 Scenario C is illustrative of

Table 1
Model Skill Scores Over the Projection Period, Where 1 Represents Perfect Agreement With Observations and Less Than 0 Represents Worse Performance Than a No-Change Null Hypothesis

Model	Timeframe	$\Delta T/\Delta t$ skill	$\Delta T/\Delta F$ skill
Ma70	1970–2000	0.84 [0.57 to 0.99]	0.51 [−0.11 to 0.94]
Mi70	1970–2000	0.91 [0.69 to 0.99]	0.41 [−0.26 to 0.90]
B70	1970–2000	0.78 [0.45 to 0.97]	0.63 [0.06 to 0.96]
RS71	1971–2000	0.19 [0.16 to 0.25]	0.42 [0.28 to 0.59]
S72	1972–2000	0.83 [0.49 to 0.99]	0.83 [0.43 to 0.98]
B75	1975–2010	0.85 [0.64 to 0.98]	0.72 [0.31 to 0.97]
N77	1977–2017	0.67 [0.44 to 0.84]	0.79 [0.48 to 0.98]
ST81	1981–2017	0.76 [0.53 to 0.94]	0.82 [0.52 to 0.98]
H81(1)	1981–2017	0.93 [0.81 to 0.99]	0.74 [0.59 to 0.93]
H81(2a)	1981–2017	0.77 [0.66 to 0.91]	0.87 [0.69 to 0.99]
H88(A)	1988–2017	0.38 [0.01 to 0.68]	0.81 [0.63 to 0.98]
H88(B)	1988–2017	0.48 [0.08 to 0.77]	0.79 [0.41 to 0.98]
H88(C)	1988–2017	0.66 [0.48 to 0.89]	0.28 [−0.46 to 0.84]
FAR	1990–2017	0.63 [0.29 to 0.87]	0.86 [0.68 to 0.99]
MS93	1993–2017	0.71 [0.20 to 0.97]	0.87 [0.61 to 0.99]
SAR	1995–2017	0.73 [0.58 to 0.95]	0.66 [0.49 to 0.91]
TAR	2001–2017	0.81 [0.15 to 0.98]	0.76 [−0.13 to 0.98]
AR4	2007–2017	0.56 [0.35 to 0.92]	0.60 [0.37 to 0.93]

Note. Both temperature versus time (ΔT/year) and implied TCR ($\Delta T/\Delta F$) median scores and uncertainties are shown.

the limitations of the implied TCR metric when the model forcings differ dramatically from observations, as Scenario C has roughly constant forcings after the year 2000.

The H88 model provides a helpful illustration of the utility of an approach that can account for mismatches between modeled and observed forcings. H88 was featured prominently in congressional testimony, and the recent thirtieth anniversary of the event in 2018 focused considerable attention on the accuracy of the projection (Borenstein & Foster, 2018; United States. Cong. Senate, 1988). H88's "most plausible" Scenario B overestimated warming experienced subsequent to publication by around 54% (Figure 3). However, much of this mismatch was due to overestimating future external forcing—particularly from CH_4 and halocarbons (Figure S3). When H88 Scenario B is evaluated based on the relationship between projected temperatures and projected forcings, the results are consistent with observations (Figures 2 and 3).

Skill score median estimates and uncertainties for both temperature versus time and implied TCR metrics are shown in Table 1 (see supporting information Text S1.3). A skill score of one represents perfect agreement between a model projection and observations, while a skill score of less than 0 represents worse performance than a no-change null-hypothesis projection.

The average of the median skill scores across all the model projections evaluated is 0.69 for the temperature versus time metric. Only three projections (RS71, H88 Scenario A, and H88 Scenario B) had skill scores below 0.5, while H81 Scenario 1 had the highest skill score of any model—0.93. Using the implied TCR metric, the average projection skill of the models was also 0.69. Models with implied TCR skill scores below 0.5 include Mi70, RS71, and H88 Scenario C, while MS93 had the highest skill score at 0.87. H88 Scenarios A and B and the IPCC FAR all performed substantially better under an implied TCR metric, reflecting the role of misspecified future forcings in their high-temperature projections. It is important to note that the skill score uncertainties for very short future projection periods—as in the case of the TAR and AR4—are quite large and should be treated with caution due to the combination of short-term temperature variability and uncertainties in the forcings.

A number of model projections had external forcings that poorly matched observational estimates due to the exclusion of non-CO_2 forcing agents. However, all models included projected future CO_2 concentrations, providing a common metric for comparison, and these are shown in Figure S4. Most of the historical climate model projections overestimated future CO_2 concentrations, some by as much as 40 ppm over current levels, with projected CO_2 concentrations increasing up to twice as fast as actually observed (Meinshausen, 2017). Of the 1970s climate model projections, only Mi70 projected atmospheric CO_2 growth in-line with observations. Many 1980s projections similarly overestimated CO_2, with only the Hansen 88 Scenarios A and B projections close to observed concentrations.

The first three IPCC assessments included projections based on simple energy balance models tuned to general circulation model results, as relatively few individual model runs were available at the time. From the AR4 onward IPCC projections were based on the multimodel mean and model spread. We examine individual models from the first three IPCC reports on both a temperature versus time and implied TCR basis in Figure S5.

4. Conclusions and Discussion

In general, past climate model projections evaluated in this analysis were skillful in predicting subsequent GMST warming in the years after publication. While some models showed too much warming and a few showed too little, most models examined showed warming consistent with observations, particularly when mismatches between projected and observationally informed estimates of forcing were taken into account. We find no evidence that the climate models evaluated in this paper have systematically overestimated or

underestimated warming over their projection period. The projection skill of the 1970s models is particularly impressive given the limited observational evidence of warming at the time, as the world was thought to have been cooling for the past few decades (e.g., Broecker, 1975; Broecker, 2017).

A number of high-profile model projections—H88 Scenarios A and B and the IPCC FAR in particular—have been criticized for projecting higher warming rates than observed (e.g., Michaels & Maue, 2018). However, these differences are largely driven by mismatches between projected and observed forcings. H88 A and B forcings increased 97% and 27% faster, respectively, than the mean observational estimate, and FAR forcings increased 55% faster. On an implied TCR basis, all three projections have high model skill scores and are consistent with observations.

While climate models have grown substantially more complex than the early models examined here, the skill that early models have shown in successfully projecting future warming suggests that climate models are effectively capturing the processes driving the multidecadal evolution of GMST. While the relative simplicity of the models analyzed here renders their climate projections operationally obsolete, they may be useful tools for verifying or falsifying methods used to evaluate state-of-the-art climate models. As climate model projections continue to mature, more signals are likely to emerge from the noise of natural variability and allow for the retrospective evaluation of other aspects of climate model projections.

Acknowledgments
Z. H. conceived the project, Z. H. and H. F. D. created the figures, and Z. H., H. F. D., T. A., and G. S. helped gather data and wrote the article text. A public GitHub repository with code used to analyze the data and generate figures and csv files containing the data shown in the figures is available online (https://github.com/hausfath/OldModels). Additional information on the code and data used in the analysis can be found in the supporting information. We would like to thank Piers Forster for providing the ensemble of observationally-informed radiative forcing estimates. No dedicated funding from any of the authors supported this project.

References

Arrhenius, S. (1896). On the influence of carbonic acid in the air upon the temperature of the ground. *Philosophical Magazine and Journal of Science*, *5*(41), 237–276.

Benson, G. S. (1970). Carbon dioxide and its role in climate change. *Proceedings of the National Academy of Sciences*, *67*(2), 898–899. https://doi.org/10.1073/pnas.67.2.898

Borenstein, S., & Foster, N. (2018). *Warned 30 years ago, global warming 'is in our living room'*. New York, NY: Associated Press. https://www.apnews.com/dbd81ca2a7244ea088a8208bab1c87e2 June 18, 2018. (last accessed Aug 22, 2019).

Broecker, W. (2017). When climate change predictions are right for the wrong reasons. *Climatic Change*, *142*(1-2), 1–6. https://doi.org/10.1007/s10584-017-1927-y

Broecker, W. S. (1975). Climatic change: Are we on the brink of a pronounced global warming? *Science*, *189*(4201), 460–463. https://doi.org/10.1126/science.189.4201.460

Cowtan, K., Hausfather, Z., Hawkins, E., Jacobs, P., Mann, M. E., Miller, S. K., et al. (2015). Robust comparison of climate models with observations using blended land air and ocean sea surface temperatures. *Geophysical Research Letters*, *42*, 6526–6534. https://doi.org/10.1002/2015GL064888

Cowtan, K., & Way, R. G. (2014). Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. *Quarterly Journal of the Royal Meteorological Society*, *140*, 1935–1944. https://doi.org/10.1002/qj.2297

Dessler, A. E., & Forster, P. M. (2018). An estimate of equilibrium climate sensitivity from interannual variability. *Journal of Geophysical Research: Atmospheres*, *123*, 8634–8645. https://doi.org/10.1029/2018JD028481

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., et al. (2019). Taking climate model evaluation to the next level. *Nature Climate Change*, *9*(2), 102–110. https://doi.org/10.1038/s41558-018-0355-y

Frame, D. J., & Stone, D. A. (2012). Assessment of the first consensus prediction on climate change. *Nature Climate Change*, *3*(4), 357–359. https://doi.org/10.1038/nclimate1763

Geoffroy, O., Saint-Martin, D., Olivié, D. J. L., Voldoire, A., Bellon, G., & Tytéca, S. (2012). Transient climate response in a two-layer energy-balance model. Part I: Analytical solution and parameter calibration using CMIP5 AOGCM experiments. *Journal of Climate*, *26*(6), 1841–1857. https://doi.org/10.1175/JCLI-D-12-00195.1

Gettelman, A., Hannay, C., Bacmeister, J. T., Neale, R. B., Pendergrass, A. G., Danabasoglu, G., et al. (2019). High climate sensitivity in the Community Earth System Model Version 2 (CESM2). *Geophysical Research Letters*, *46*, 8329–8337. https://doi.org/10.1029/2019GL083978

Gregory, J. M., & Forster, P. M. (2008). Transient climate response estimated from radiative forcing and observed temperature change. *Journal of Geophysical Research*, *113*, D23105. https://doi.org/10.1029/2008JD010405

Gregory, J. M., Ingram, W. J., Palmer, M. A., Jones, G. S., Stott, P. A., Thorpe, R. B., et al. (2004). A new method for diagnosing radiative forcing and climate sensitivity. *Geophysical Research Letters*, *31*, L03205. https://doi.org/10.1029/2003GL018747

Gregory, J. M., & Mitchell, J. F. B. (1997). The climate response to CO_2 of the Hadley Centre coupled AOGCM with and without flux adjustment. *Geophysical Research Letters*, *24*(15), 1943–1946. https://doi.org/10.1029/97GL01930

Hansen, J., Fung, I., Lacis, A., Rind, D., Lebedeff, S., Ruedy, R., et al. (1988). Global climate changes as forecast by Goddard Institute for Space Studies three-dimensional model. *Journal of Geophysical Research*, *93*, 9341–9364. https://doi.org/10.1029/JD093iD08p09341

Hansen, J., Johnson, D., Lacis, A., Lebedeff, S., Lee, P., Rind, D., & Russell, G. (1981). Climate impact of increasing atmospheric carbon dioxide. *Science*, *213*(4511), 957–966. https://doi.org/10.1126/science.213.4511.957

Hansen, J., Russell, G., Lacis, A., Fung, I., Rind, D., & Stone, P. (1985). Climate response times: Dependence on climate sensitivity and ocean mixing. *Science*, *229*(4716), 857–859. https://doi.org/10.1126/science.229.4716.857

Hansen, J., Sato, M., Ruedy, R., Nazarenko, L., Lacis, A., Schmidt, G. A., et al. (2005). Efficacy of climate forcings. *Journal of Geophysical Research*, *110*, D18104. https://doi.org/10.1029/2005JD005776

Hargreaves, J. C. (2010). (2010). Skill and uncertainty in climate models. *Wiley Interdisciplinary Reviews: Climate Change*, *1*, 556–564. https://doi.org/10.1002/wcc.58

Hausfather, Z., Cowtan, K., Clarke, D. C., Jacobs, P., Richardson, M., & Rohde, R. (2017). Assessing recent warming using instrumentally homogeneous sea surface temperature records. *Science Advances*, *3*(1). https://doi.org/10.1126/sciadv.1601207

Hawkins, E., & Sutton, R. (2012). Time of emergence of climate signals. *Geophysical Research Letters*, *39*, L01702. https://doi.org/10.1029/2011GL050087

Held, I. M., Winton, M., Takahashi, K., Delworth, T., Zeng, F., & Vallis, G. K. (2010). Probing the fast and slow components of global warming by returning abruptly to preindustrial forcing. *Journal of Climate*, *23*, 2418–2427. https://doi.org/10.1175/2009JCLI3466.1

IPCC (AR4) (2007). In S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, & H. L. Miller (Eds.), *Climate change 2007: The physical science basis, Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, UK and New York, NY: Cambridge University Press. ISBN 978-0-521-88009-1 (pb: 978-0-521-70596-7)

IPCC (AR5) (2013). In T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, & P. M. Midgley (Eds.), *Climate change 2013: The physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (p. 1535). Cambridge, UK and New York, NY: Cambridge University Press.

IPCC (FAR) (1990). Climate change: The IPCC scientific assessment. Report prepared by Working Group I. In J. T. Houghton, G. J. Jenkins, & J. J. Ephraums (Eds.), *Intergovernmental Panel on Climate Change* (p. 365). Cambridge, UK and New York, NY: Cambridge University Press.

IPCC (SAR) (1996). In J. T. Houghton, L. G. Meira Filho, B. A. Callander, N. Harris, A. Kattenberg, & K. Maskell (Eds.), *Climate change 1995: The science of climate change, Contribution of Working Group I to the Second Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, UK and New York, NY: Cambridge University Press. ISBN 0-521-56433-6 (pb: 0-521-56436-0)

IPCC (TAR) (2001). In J. T. Houghton, Y. Ding, D. J. Griggs, M. Noguer, P. J. van der Linden, X. Dai, K. Maskell, & C. A. Johnson (Eds.), *Climate change 2001: The scientific basis, Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, UK and New York, NY: Cambridge University Press. ISBN 0-521-80767-0 (pb: 0-521-01495-6)

Knutti, R., Masson, D., & Gettelman, A. (2013). Climate model genealogy: Generation CMIP5 and how we got there. *Geophysical Research Letters*, *40*, 1194–1199. https://doi.org/10.1002/grl.50256

Lenssen, N. J. L., Schmidt, G. A., Hansen, J. E., Menne, M. J., Persin, A., Ruedy, R., & Zyss, D. (2019). Improvements in the GISTEMP uncertainty model. *Journal of Geophysical Research: Atmospheres*, *124*, 6307–6326. https://doi.org/10.1029/2018JD029522

Manabe, S. (1970). The dependence of atmospheric temperature on the concentration of carbon dioxide. In S. F. Singer (Ed.), *Global effects of environmental pollution* (Chap. 3, pp. 25–29). Dordrecht: Springer.

Manabe, S., Smagorinsky, J., & Strickler, R. F. (1965). Simulated climatology of a general circulation model with a hydrologic cycle. *Monthly Weather Review*, *93*, 769–798. https://doi.org/10.1175/1520-0493(1965)093<0769:SCOAGC>2.3.CO;2

Manabe, S., & Stouffer, R. J. (1993). Century-scale effects of increased atmospheric CO_2 on the ocean–atmosphere system. *Nature*, *364*(6434), 215–218. https://doi.org/10.1038/364215a0

Manabe, S., & Strickler, R. F. (1964). Thermal equilibrium of the atmosphere with a convective adjustment. *Journal of the Atmospheric Sciences*, *21*, 361–385. https://doi.org/10.1175/1520-0469(1964)021<0361:TEOTAW>2.0.CO;2

Manabe, S., & Wetherald, R. T. (1967). Thermal equilibrium of the atmosphere with a given distribution of relative humidity. *Journal of the Atmospheric Sciences*, *24*(3), 241–259. https://doi.org/10.1175/1520-0469(1967)024<0241:TEOTAW>2.0.CO;2

Manabe, S., & Wetherald, R. T. (1975). The effects of doubling the CO_2 concentration on the climate of a general circulation model. *Journal of the Atmospheric Sciences*, *32*, 3–15. https://doi.org/10.1175/1520-0469(1975)032<0003:TEODTC>2.0.CO;2

Marvel, K., Schmidt, G. A., Miller, R. L., & Nazarenko, L. S. (2016). Implications for climate sensitivity from the response to individual forcings. *Nature Climate Change*, *6*, 386. Retrieved from. https://doi.org/10.1038/nclimate2888

Mauritsen, T., Bader, J., Becker, T., Behrens, J., Bittner, M., Brokopf, R., et al. (2019). Developments in the MPI-M Earth System Model version 1.2 (MPI-ESM 1.2) and its response to increasing CO_2. *Journal of Advances in Modeling Earth Systems*, *11*(4), 998–1038. https://doi.org/10.1029/2018MS001400

Meinshausen, M., Vogel, E., Nauels, A., Lorbacher, K., Meinshausen, N., Etheridge, D. M., et al. (2017). Historical greenhouse gas concentrations for climate modelling (CMIP6). *Geoscientific Model Development*, *10*(5), 2057–2116. https://doi.org/10.5194/gmd.10.2057.2017

Michaels, P., & Maue, R. (2018). Thirty years on, How well do global warming predictions stand up? The Wall Street Journal, June 21st.

Mitchell, J. M. (1970). A preliminary evaluation of atmospheric pollution as a cause of the global temperature fluctuation of the past century. In S. F. Singer (Ed.), *Global Effects of Environmental Pollution* (Chap. 12, pp. 139–155). Dordrecht: Springer.

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

Nordhaus, W. (1977). Strategies for the control of carbon dioxide (Cowles Foundation Discussion Papers). Cowles Foundation for Research in Economics, Yale University. Retrieved from https://econpapers.repec.org/RePEc:cwl:cwldpp:443

Otto, A., Otto, F. E. L., Boucher, O., Church, J., Hegerl, G., Forster, P. M., et al. (2013). Energy budget constraints on climate response. *Nature Geoscience*, *6*, 415. https://doi.org/10.1038/ngeo1836

Rahmstorf, S., Cazenave, A., Church, J. A., Hansen, J. E., Keeling, R. F., Parker, D. E., & Somerville, R. C. J. (2007). Recent climate observations compared to projections. *Science*, *316*(5825), 709–709. https://doi.org/10.1126/science.1136843

Rahmstorf, S., Foster, G., & Cazenave, A. (2012). Comparing climate projections to observations up to 2011. *Environmental Research Letters*, *7*(4), 44035. https://doi.org/10.1088/1748-9326/7/4/044035

Rasool, S. L., & Schneider, S. H. (1971). Atmospheric carbon dioxide and aerosols: Effects of large increases on global climate. *Science*, *173*(3992), 138–141. https://doi.org/10.1126/science.173.3992.138

Reichler, T., & Kim, J. (2008). How well do coupled models simulate today's climate? *Bulletin of the American Meteorological Society*, *89*, 303–312. https://doi.org/10.1175/BAMS-89-3-303

Richardson, M., Cowtan, K., Hawkins, E., & Stolpe, M. B. (2016). Reconciled climate response estimates from climate models and the energy budget of Earth. *Nature Climate Change*, *6*, 931. https://doi.org/10.1038/nclimate3066

Rohde, R., Muller, R. A., et al. (2013). A new estimate of the average Earth surface land temperature spanning 1753 to 2011. Geoinfor Geostat: An Overview 1:1. https://doi.org/10.4172/gigs.1000101

Rohrschneider, T., Stevens, B., & Mauritsen, T. (2019). On simple representations of the climate response to external radiative forcing. *Climate Dynamics.*, *53*(5-6), 3131–3145. https://doi.org/10.1007/s00382-019-04686-4

Sawyer, J. S. (1972). Man-made carbon dioxide and the "greenhouse" effect. *Nature*, *239*(5366), 23–26. https://doi.org/10.1038/239023a0

Schmidt, G. A., Bader, D., Donner, L. J., Elsaesser, G. S., Golaz, J. C., Hannay, C., et al. (2017). Practice and philosophy of climate model tuning across six U.S. modeling centers. *Geoscientific Model Development*, *10*, 3207–3223. https://doi.org/10.5194/gmd.10.3207.2017

Schneider, S. H. (1975). On the carbon dioxide–climate confusion. *Journal of the Atmospheric Sciences*, *32*, 2060–2066. https://doi.org/10.1175/1520-0469(1975)032<2060:OTCDC>2.0.CO;2

Schneider, S. H., & Thompson, S. L. (1981). Atmospheric CO_2 and climate: Importance of the transient response. *Journal of Geophysical Research*, *86*(C4), 3135–3147. https://doi.org/10.1029/JC086iC04p03135

Stouffer, R. J., & Manabe, S. (2017). Assessing temperature pattern projections made in 1989. *Nature Climate Change*, *7*(3), 163–165. https://doi.org/10.1038/nclimate3224

Stouffer, R. J., Manabe, S., & Bryan, K. (1989). Interhemispheric asymmetry in climate response to a gradual increase of atmospheric CO_2. *Nature*, *342*(6250), 660–662. https://doi.org/10.1038/342660a0

United States. Cong. Senate (1988). Committee on Energy and Natural Resources. Greenhouse Effect and Global Climate Change. Hearings, June 23, 1988. 100th Cong. 1st sess. Washington: GPO.

van Vuuren, D. P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., et al. (2011). The representative concentration pathways: An overview. *Climatic Change*, *109*(1), 5. https://doi.org/10.1007/s10584-011-0148-z

Vial, J., Dufresne, J.-L., & Bony, S. (2013). On the interpretation of inter-model spread in CMIP5 climate sensitivity estimates. *Climate Dynamics*, *41*(11), 3339–3362. https://doi.org/10.1007/s00382-013-1725-9

Vose, R. S., Arndt, D., Banzon, V. F., Easterling, D. R., Gleason, B., Huang, B., et al. (2012). NOAA's merged land–ocean surface temperature analysis. *Bulletin of the American Meteorological Society*, *93*(11), 1677–1685. https://doi.org/10.1175/BAMS-D-11-00241.1