# **AGU** PUBLICATIONS

# **Geophysical Research Letters**

# **RESEARCH LETTER**

10.1002/2015GL064888

#### **Key Points:**

- There is a systematic bias in model-observation comparisons from blending air and sea temperatures
- A further bias arises from using anomalies in regions where the sea ice boundary has changed
- Correcting these accounts for a quarter to half of the discrepancy between models and observations

#### Supporting Information:

Supporting Information S1

#### Correspondence to:

K. Cowtan, kevin.cowtan@york.ac.uk

#### Citation:

Cowtan, K., Z. Hausfather, E. Hawkins, P. Jacobs, M. E. Mann, S. K. Miller, B. A. Steinman, M. B. Stolpe, and R. G. Way (2015), Robust comparison of climate models with observations using blended land air and ocean sea surface temperatures, *Geophys. Res. Lett.*, *42*, 6526–6534, doi:10.1002/2015GL064888.

Received 10 JUN 2015 Accepted 25 JUL 2015 Accepted article online 29 JUL 2015 Published online 13 AUG 2015

©2015. American Geophysical Union. All Rights Reserved.

# Robust comparison of climate models with observations using blended land air and ocean sea surface temperatures

Kevin Cowtan<sup>1</sup>, Zeke Hausfather<sup>2</sup>, Ed Hawkins<sup>3</sup>, Peter Jacobs<sup>4</sup>, Michael E. Mann<sup>5</sup>, Sonya K. Miller<sup>5</sup>, Byron A. Steinman<sup>6</sup>, Martin B. Stolpe<sup>7</sup>, and Robert G. Way<sup>8</sup>

<sup>1</sup> Department of Chemistry, University of York, York, UK, <sup>2</sup>Energy and Resources Group, University of California, Berkeley, California, USA, <sup>3</sup>National Centre for Atmospheric Science, Department of Meteorology, University of Reading, Reading, UK, <sup>4</sup>Department of Environmental Science and Policy, George Mason University, Fairfax, Virginia, USA, <sup>5</sup>Department of Meteorology and Earth and Environmental Systems Institute, Pennsylvania State University, University Park, Pennsylvania, USA, <sup>6</sup>Department of Earth and Environmental Sciences, Large Lakes Observatory, University of Minnesota, Duluth, Duluth, Minnesota, USA, <sup>7</sup>Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland, <sup>8</sup>Department of Geography, University of Ottawa, Ottawa, Ontario, Canada

**Abstract** The level of agreement between climate model simulations and observed surface temperature change is a topic of scientific and policy concern. While the Earth system continues to accumulate energy due to anthropogenic and other radiative forcings, estimates of recent surface temperature evolution fall at the lower end of climate model projections. Global mean temperatures from climate model simulations are typically calculated using surface air temperatures, while the corresponding observations are based on a blend of air and sea surface temperatures. This work quantifies a systematic bias in model-observation comparisons arising from differential warming rates between sea surface temperatures and surface air temperatures over oceans. A further bias arises from the treatment of temperatures in regions where the sea ice boundary has changed. Applying the methodology of the HadCRUT4 record to climate model temperature fields accounts for 38% of the discrepancy in trend between models and observations over the period 1975–2014.

# 1. Introduction

Climate model projections of the global mean temperature response to future greenhouse gas emissions provide an important basis for decision making concerning mitigation and adaptation to climate change. However, model projections are subject to uncertainty in the size of the temperature response, which arises primarily from the scale of the amplifying effect of the cloud feedback and the temporal evolution of climate forcings [*Flato et al.*, 2013; *Andrews et al.*, 2012; *Sherwood et al.*, 2014]. Comparison of model projections against the observed rate of warming over recent decades can provide a test of the ability of models to simulate the transient evolution of climate. The comparison is complicated by the need to accurately simulate changes in atmospheric composition and solar radiation, as well as accounting for the unforced variability of the climate system [*Schmidt et al.*, 2014]. The fact that the observations fall at the lower end of the envelope of model simulations over the last decade has led to suggestions that climate model forecasts may overestimate the potential future warming resulting from increasing greenhouse gas concentrations [*Fyfe et al.*, 2013].

Observational records of global mean surface temperature are typically determined from air temperature measurements on land, blended with sea surface temperature (SST) observations measured in the top few meters of the ocean [*Morice et al.*, 2012; *Kennedy et al.*, 2011a]. Temperature records may be based on spatially incomplete data [*Morice et al.*, 2012; *Vose et al.*, 2012] or on data that have been infilled to provide an estimate of the global mean temperature [*Hansen et al.*, 2010; *Rohde et al.*, 2013; *Cowtan and Way*, 2014]. Observations of temperature are typically converted into anomalies (i.e., changes with respect to some baseline period) to allow observations from different environments to be meaningfully combined.

A homogenous global temperature record would ideally be based on a property which is independent of the surface type (land, ocean, or ice), such as air temperatures at a uniform height above the surface. However, sea surface temperature observations have historically been used in preference to marine air temperatures due to inhomogeneities in older marine air temperature data sets [*Kent et al.*, 2013]. Infilled temperature records

typically extrapolate air temperatures over sea ice, because the insulating effect of ice and snow isolates the air from the water [*Kurtz et al.*, 2011], an approach which is supported by observations [*Rigor et al.*, 2000], atmospheric reanalyses [*Simmons and Poli*, 2014], and satellite data [*Comiso and Hall*, 2014].

Global averages of the observational temperature records are typically compared to near-surface air temperature from an ensemble of climate model simulations (e.g., Intergovernmental Panel on Climate Change Fifth Assessment Report Working Group 1, Figure 9.8 [*Flato et al.*, 2013]). When comparing against spatially incomplete records, the model temperature fields may be masked to reduce coverage to match the observations or make the assumption that the observed regions are representative of the unobserved regions. This assumption may not hold for the last two decades of accelerated Arctic warming [*Simmons and Poli*, 2014; *Saffioti et al.*, 2015]. Although in some cases the model simulations were masked for coverage, most studies have used the surface air temperature field from models rather than blended land-ocean temperatures, with the notable exception of *Marotzke and Forster* [2015] and some attribution studies, e.g., *Knutson et al.* [2013].

A true like-with-like comparison would involve blending the air and sea surface temperature fields from the models in a manner consistent with the observational records. The purposes of this work are to evaluate the impact of comparing air temperatures from models with the blended observational data and to establish guidelines for the determination of blended temperature comparisons. These require changes both in the way global mean temperature from models is evaluated and ideally also in the preparation of blended observational data sets.

# 2. Data and Methods

The impact of using blended temperatures was evaluated for climate model simulations from the Coupled Model Intercomparison Project Phase 5 (CMIP5) archive [*Taylor et al.*, 2012] using a combination of the historical and Representative Concentration Pathway 8.5 (RCP8.5) emissions scenarios. The calculation of a gridded blended temperature record requires the surface air temperature ("tas" in CMIP5 nomenclature), sea surface temperature ("tos"), sea ice concentration ("sic"), and the proportion of ocean in each grid cell ("sftotf"). After eliminating incompatible data sets (Figure S1 in the supporting information), there were 84 useable model runs from 36 models. The Climate Data Operators software package (available at http://www.mpimet.mpg.de/cdo,version1.6.8) was used to convert all fields onto a standard  $1 \times 1^{\circ}$  grid, using distance-weighted interpolation to avoid the loss of coverage when interpolating fields containing missing values (however, similar results were obtained using nearest-neighbor interpolation or the native ocean grids).

For each model simulation, a global mean temperature series is calculated from the unblended surface air temperature field for comparison. A blended temperature field is then calculated using the air and sea surface temperature fields, using the land mask and sea ice concentration. In the blended temperature field, the air temperature for the whole grid cell is used as an estimate of the air temperature over land and sea ice, while the sea surface temperature is used for the proportion of the cell occupied by open water. Ideally, there would be separate simulated estimates for air temperature over land and ocean in fractional grid boxes, but these are not standard diagnostics in the CMIP5 models. The blended temperature field,  $T_{blend}$ , therefore takes the following form:

$$w_{\text{air}} = (1 - f_{\text{ocean}}) + f_{\text{ocean}} f_{\text{ice}}$$

$$T_{\text{blend}} = w_{\text{air}} T_{\text{air}} + (1 - w_{\text{air}}) T_{\text{ocean}}$$
(1)

where  $T_{air}$ ,  $T_{ocean}$ ,  $f_{ice}$ , and  $f_{ocean}$  correspond to the CMIP5 "tas," "tos," "sic," and "sftof" fields, respectively, and  $w_{air}$  is the land and sea ice fraction in a grid cell.

If a sea surface temperature or sea ice concentration cell is missing (e.g., for the Commonwealth Scientific and Industrial Research Organisation model sea surface temperatures are missing for ice cells),  $w_{air}$  is set to 1.0, ensuring that the blended temperature matches the air temperature. The difference between the latitude-weighted global means of the blended temperature and of the unblended air temperature provides a measure of the bias in the model-observation comparison.

Implicit assumptions in the implementation of the blending calculation may influence the results; therefore, three possible variants of the calculation were investigated:

1. The calculation may be performed over the whole globe, or alternatively, the fields may be masked to reduce coverage to that of the observational data. The full coverage calculation provides a measure of the

# **AGU** Geophysical Research Letters



**Figure 1.** Flowcharts describing the calculation of global mean temperature (*T*) from the original CMIP5 fields. Three different methods are illustrated: (a) air temperatures only (i.e., unblended), (b) blended absolute temperatures (no mask, variable ice), and (c) blended temperature anomalies (no mask, variable ice). The use of anomalies in Figure 1c involves reversal of the shaded steps; it will be shown that this significantly affects the results.

bias in a comparison with an infilled record, while the masked calculation provides a measure of the bias in a comparison with an incomplete coverage data set such as HadCRUT4 [*Morice et al.*, 2012].

- 2. The calculation may be performed using absolute temperatures, which are output by the climate model runs, or using temperature anomalies which are conventionally used for blending in the case of the observational record. In the latter case, anomalies are calculated with respect to the period 1961–1990 for consistency with HadCRUT4.
- 3. The blending calculation can be performed using the monthly varying sea ice cover or a fixed sea ice coverage in order to isolate any confounding effects due to the change of a grid cell from ice to open water. For the fixed sea ice case, sea surface temperatures are only used for grid cells for which the sea ice concentration is zero for the corresponding month of every year from 1961 onward. In this case the remaining grid cells are considered 100% sea ice and thus take the same value as in the unblended case.

These three options can be employed in any combination. The differences between the air-temperature-only calculation and two variants of the blended calculation (absolute versus anomaly based) are illustrated in Figure 1.

One further method was implemented with the aim of providing a better comparison to the HadCRUT4 temperature data. This requires reproducing the HadCRUT4 algorithm, the coarse HadCRUT4 grid, and the coverage of observations within each large grid cell. The steps are as follows:

- 1. The air and sea surface temperatures are converted to anomalies using the HadCRUT4 baseline period (1961–1990).
- 2. The air temperatures are masked to include only grid cells containing a nonzero land fraction.
- 3. Sea surface temperatures are masked to include only cells with no more than 5% sea ice. While the HadCRUT4 calculation does not explicitly take sea ice into account, observations from ships and buoys are confined to open water.
- 4. The remaining air and sea temperatures in each cell of the coarse 5×5° grid used by HadCRUT4 are averaged, omitting any values excluded by the previous steps.
- 5. The air and sea temperatures are masked to match the coverage of the air and sea temperatures in the HadCRUT4 data, respectively.
- 6. The temperatures are then blended: cells containing only an air or sea temperature take that value; otherwise, the air and sea temperatures are blended according to the land fraction for the grid cell. (As with



**Figure 2.** Difference between the global mean air temperature and the blended land-ocean temperatures for 84 CMIP5 model simulations combining the historical and RCP8.5 experiments. The differences are calculated using global coverage and blending absolute temperatures with variable sea ice. Temperature anomalies are relative to 1961–1990.

HadCRUT4, the land fraction is bounded by a minimum value of 0.25 for coastal cells so that air temperature observations on small islands are not eliminated.)

7. Following the HadCRUT4 convention, the global mean temperature is calculated from the mean of the cosine-weighted hemispheric means.

Improved compatibility between the model-derived temperatures and the observational data is achieved at a cost of complexity and of producing a set of model results which are only comparable to a specific observational data set.

# 3. Results

The difference between the global mean blended temperature and the global mean air temperature was determined for 36 CMIP5 models with 84 historical/RCP8.5 simulations, using global data (i.e., no coverage mask) and blending absolute temperatures with a variable sea ice boundary (Figure 2). The blended temperatures show consistently less change than air temperature, with blended temperatures lower than air temperatures over recent decades. Over the period 2009–2013 the difference between multimodel global mean blended and air temperatures is  $0.033 \pm 0.010^{\circ}$ C ( $1\sigma$ ) relative to 1961-1990, and this difference is estimated to increase in magnitude with time to  $0.18 \pm 0.04^{\circ}$ C by the year 2100.

The effect is broadly similar in magnitude across all the models both during the historical period and over the 21st century with the exception of the Beijing Climate Centre model, "bcc-csm." The different behavior of the "bcc-csm" model appears to arise from surface air temperature being almost equal to the skin temperature ("ts" in the CMIP5 nomenclature) in that model alone (Figure S2). Preindustrial control simulations were examined (where available) to determine whether model drift due to nonequilibrium initial conditions contributes to the difference between air and sea surface temperatures. In every case the difference between the blended and air temperature trends at the end of the control run was at least an order of magnitude smaller than the effect identified here (Figure S3).

The mean difference across all models between the global mean blended and global mean air temperatures was compared for the previously described variants of the blending calculation and for the HadCRUT4 method (Figure 3). The difference between the blended and air temperatures is greater when using anomalies (as in the observational record) than when using absolute temperatures. The reason arises from changes in the ice edge. As ice melts, grid cells switch from taking air temperatures to taking sea surface temperatures. When blending anomalies, the temperature anomaly is determined with respect to a period in the past when air temperatures



**Figure 3.** Difference between global mean blended temperature and air temperature, for different variants of the blending calculation, averaged over 84 historical + RCP8.5 simulations. Blended temperatures show less warming than air temperatures; hence, the sign of the difference is negative for recent decades. Results are shown for the four permutations of masked versus global and absolute temperatures versus anomalies (with variable sea ice in each case). Two additional series for the absolute and anomaly methods with fixed ice show that fixing the sea ice boundary eliminates the effect of using anomalies. The final series shows the HadCRUT4 method, which shows similar behavior to the other anomaly methods.

over the ice were lower, while the sea surface temperatures under the ice (constrained by the freezing point of seawater) are unchanged. Thus, the transition from air temperature anomaly (which is warmer than the baseline period) to sea surface temperature anomaly (which is roughly the same as during the baseline period) introduces a cool bias at the point when the ice melts (Figure S4).

When blending is performed using absolute temperatures, the blended temperature change is consistently around 95% of the air temperature change, for both the RCP8.5 scenario and the RCP4.5 scenario where temperatures have largely stabilized by 2100 (Figure S5). When blending is performed using temperature anomalies, the blended temperature change is reduced to about 91% of the air temperature change for the RCP8.5 scenario. The role of ice melt in the difference between blending absolute temperatures and temperature anomalies is confirmed by fixing the sea ice coverage; in this case both absolute and anomaly calculations give identical results (although the impact of blending is now underestimated due to the omission of large regions of formerly ice-covered ocean).

Masking the model data to match the HadCRUT4 observations reduces the difference between the global mean blended and air temperatures slightly when using anomalies and increases it slightly when using absolute temperatures. This behavior arises from the change in sign of the difference between the blended and air temperatures in ice melt cells between the anomaly and absolute cases (Figure S6).

When emulating the HadCRUT4 method, the difference between the air and blended temperatures is marginally greater than the result from the masked blended anomaly calculation. The difference arises primarily from the handling of ice edge cells. The coarse  $5 \times 5^{\circ}$  grid of the HadCRUT4 also contributes to spreading the effective area over which the ice edge plays a role.

The differences between the air and sea surface temperature changes are small compared to the uncertainties and bias corrections in the sea surface temperatures [*Kennedy et al.*, 2011b, 2011a], and so observational data are of limited use in detecting this bias. The comparison of daily sea surface temperatures to nighttime-only marine air temperatures is confounded by diurnal range effects as well as inhomogeneities in the observations, with the MOHMAT and HadNMAT2 marine air temperature data [*Rayner et al.*, 2003; *Kent et al.*, 2013]



**Figure 4.** Comparison of 84 RCP8.5 simulations against HadCRUT4 observations (black), using either air temperatures (red line and shading) or blended temperatures using the HadCRUT4 method (blue line and shading). The shaded regions represent the 90% range (i.e., from 5% to 95%) of the model simulations, with the corresponding lines representing the multimodel mean. (a) Anomalies derived from the unmodified RCP8.5 results and (b) the results adjusted to include the effect of updated forcings from *Schmidt et al.* [2014]. Temperature anomalies are relative to 1961–1990.

showing substantial differences to the SSTs not seen in the models (Figure S7). Similarly, uncertainties in the assimilated observations limit the utility of atmospheric reanalyses for this purpose (Figure S8).

What are the implications of using blended temperatures on a model-observation comparison for the CMIP5 models? Figure 4 shows a comparison of the 84 RCP8.5 model runs against the HadCRUT4 data, using either air or blended temperatures and the HadCRUT4 blending algorithm (i.e., with the HadCRUT4 coverage and averaging conventions). When using air temperatures, the HadCRUT4 data fall below the 90% range of climate model simulations for the years 2011–2013. When using the blended temperatures, the observations are at the lower end of the 90% range for 2011 and 2012 and within it for 2013.

The recent divergence between the models and the observations occurs after 1998, the period commonly associated with the so-called global warming "hiatus" [*Fyfe et al.*, 2013; *Fyfe and Gillett*, 2014; *Tollefson*, 2014]. Several contributory factors to the divergence have been identified, including an increase in moderate volcanic eruptions [*Solomon et al.*, 2011; *Ridley et al.*, 2014; *Santer et al.*, 2014a, 2014b], a reduction in solar activity, a decrease in stratospheric water vapor concentration [*Solomon et al.*, 2010], internal variability [*Meehl et al.*, 2011, 2013; *Trenberth and Fasullo*, 2013; *Kosaka and Xie*, 2013; *Mann et al.*, 2014; *Steinman et al.*, 2015; *Dai et al.*, 2015], and a bias due to the omission of the Arctic, which is warming more rapidly than projected by the models [*Cowtan and Way*, 2014; *Saffioti et al.*, 2015]. The contribution of internal variability to the remaining discrepancy between the models and the observations is beyond the scope of this analysis.

Using an impulse response model, *Schmidt et al.* [2014] estimate the temperature impact of the slower than predicted growth in forcing due to volcanoes, solar cycle, and also the possible cooling effect of an increase in aerosol emissions over the hiatus period. Other studies have found negligible or even a warming contribution of aerosols on hiatus temperature trends [*Regayre et al.*, 2014; *Gettelman et al.*, 2015; *Thorne et al.*, 2015], although *Schmidt et al.* [2014] include nitrate aerosols which are omitted from the other studies. The model outputs were also adjusted using the estimated impacts from *Schmidt et al.* [2014] due to volcanoes, solar cycle, and greenhouse emissions but not aerosols (Figure 4b). When using blended temperatures, the observations lie well within the 90% range of RCP8.5 runs for the whole of the last decade. Similar results are obtained from adjustments to the model temperatures derived using the Bern2.5D climate model of

intermediate complexity [*Huber and Knutti*, 2014]. Notably, *Thorne et al.* [2015] did not find a detectable reduction in the recent temperature increase when using updated forcings in a large ensemble of NorESM simulations.

The impact of using blended rather than air temperatures accounts for 27% of the difference between the models and the observations over the period 2009–2013. The adjustments by *Schmidt et al.* [2014] due to the overestimated forcings account for another 27% of the difference when omitting the tropospheric aerosol term or 41% of the difference when including aerosols. Over the period 1975–2014 the use of blended rather than air temperatures accounts for 38% of the difference in *trend* between the models and the observations (Table S1) or almost all of the difference if the last 5 years are omitted, consistent with the results of *Marotzke and Forster* [2015]. The model simulations suggest that the 40 year trend in HadCRUT4 is suppressed by  $0.017 \pm 0.004^{\circ}$ C/decade compared to an air temperature record.

Comparisons to the infilled reconstructions of *Cowtan and Way* [2014] and *Rohde et al.* [2013] require different variants of the blending calculation (Text S1) but lead to similar conclusions. Comparisons to the other temperature data sets will in turn require an appropriate choice of blending method or development of a custom method appropriate to that data set. The comparison will depend on explicit and/or implicit assumptions in the blending and anomaly calculations and is therefore best addressed by the record providers.

## 4. Discussion

These results have implications in three areas: first in the comparison of climate model ensembles to the observational record, second in estimating climate sensitivity, and third in the preparation of observational temperature records.

When comparing models to observations, the comparison should be strictly performed using blended land/ocean temperatures rather than air temperatures from the models. The size of the difference between the blended and air temperatures is sensitive to assumptions in the blending calculation and in particular whether blending is performed using absolute temperatures or anomalies. The most conservative approach is to blend absolute temperatures from the models (i.e., air temperature over land and ice, and sea surface temperature for the oceans), in which case the global mean blended temperatures will typically show 5% less warming than the air temperatures. However, the actual impact of the use of blended temperatures on the observational record is nearly twice as great owing to the blending of anomalies in the observational data.

Replication of the HadCRUT4 blending algorithm on the model outputs leads to a reduction in the model-observation divergence of  $0.056 \pm 0.015^{\circ}$ C over the years 2009–2013, or about a quarter of the divergence over that period. However the replication is not exact: for example, the results will depend on the climatology by which anomalies are calculated for ocean cells which were sea ice during the baseline period [*Rayner et al.*, 2006]. The comparison would also be further improved by the inclusion of a land-only surface air temperature field in future CMIP phases.

Comparison to other versions of the temperature record should ideally also involve reproducing the blending method for that particular observational data set. However, comparison to multiple observational data sets at the same time is then inconvenient, because the model ensemble will be different for each observational record. Alternatively, instead of modifying the model temperatures to match the methodology of a particular observational record, each observational record can be modified to produce an estimate of the global mean air temperature. The required correction is determined from the difference between the blended and air temperatures from the models using the methodology of the corresponding observational record. All the observational records may then be compared simultaneously.

Estimates of climate sensitivity, at least over decadal to centennial time scales, will be lower for blended temperatures than for air temperatures. Estimates of transient climate response (TCR) should therefore be quoted with an indication of whether the value was determined using observed air or blended temperatures and, in the case of blended temperatures, whether blending was performed using absolute temperatures or anomalies. In the case of blended absolute temperatures, TCR values are likely to be about 95% of those for air temperatures or 91% for blended anomalies. Estimates of TCR from the observational record are based on blended temperatures and thus are expected to underestimate TCR by about 10% in comparison to quoted figures for the models. There are two implications for observational records. First, a blended record from air temperatures over land and sea ice and sea surface temperatures over open ocean slightly underestimates the change in temperature diagnosed using global air temperatures alone. Second, the blending calculation should ideally be conducted with absolute temperatures to avoid introducing a cool bias due to the transformation of cells from sea ice to open water, particularly for infilled records. Otherwise, the approach of fixing the sea ice extent (Text S1) mitigates the problem at the cost of introducing a different but smaller bias. The new data set of *Karl et al.* [2015] incorporates adjustments to SSTs to match nighttime marine air temperatures [*Huang et al.*, 2015] and so *may* be more comparable to model air temperatures. The difference between air and sea surface temperature trends diagnosed here provides support for an increase in temperature trends when using marine air temperatures, as reported in *Karl et al.* [2015].

Finally, we emphasize that robust comparisons of observations and models require a like-with-like approach and encourage further development of appropriate diagnostics from model simulations to facilitate such comparisons.

## References

Andrews, T., J. M. Gregory, M. J. Webb, and K. E. Taylor (2012), Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere-ocean climate models, *Geophys. Res. Lett.*, 39, L09712, doi:10.1029/2012GL051607.

Comiso, J. C., and D. K. Hall (2014), Climate trends in the Arctic as observed from space, Wiley Interdiscip. Rev. Clim. Change, 5(3), 389–409, doi:10.1002/wcc.277.

- Cowtan, K., and R. G. Way (2014), Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends, Q. J. R. Meteorol. Soc., 140(683), 1935–1944, doi:10.1002/qj.2297.
- Dai, A., J. C. Fyfe, S. -P. Xie, and X. Dai (2015), Decadal modulation of global surface temperature by internal climate variability, *Nat. Clim. Change*, *5*(6), 555–559, doi:10.1038/nclimate2605.

Flato, G., et al. (2013), Evaluation of climate models, 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*, edited by T. F. Stocker et al., pp. 741–866, Cambridge Univ. Press, Cambridge, U. K., and New York.

Fyfe, J. C., and N. P. Gillett (2014), Recent observed and simulated warming, Nat. Clim. Change, 4(3), 150-151, doi:10.1038/nclimate2111.

Fyfe, J. C., N. P. Gillett, and F. W. Zwiers (2013), Overestimated global warming over the past 20 years, Nat. Clim. Change, 3(9), 767–769, doi:10.1038/nclimate1972.

Gettelman, A., D. Shindell, and J. Lamarque (2015), Impact of aerosol radiative effects on 2000–2010 surface temperatures, *Clim. Dyn.*, 1–15, doi:10.1007/s00382-014-2464-2.

Hansen, J., R. Ruedy, M. Sato, and K. Lo (2010), Global surface temperature change, *Rev. Geophys.*, 48, RG4004, doi:10.1029/2010RG000345. Huang, B., V. F. Banzon, E. Freeman, J. Lawrimore, W. Liu, T. C. Peterson, T. M. Smith, P. W. Thorne, S. D. Woodruff, and H.-M. Zhang (2015),

Extended Reconstructed Sea Surface Temperature version 4 (ERSST. v4). Part I: Upgrades and intercomparisons, J. Clim., 28, 911–930, doi:10.1175/JCLI-D-14-00006.1.

Huber, M., and R. Knutti (2014), Natural variability, radiative forcing and climate response in the recent hiatus reconciled, Nat. Geosci., 7, 651–656, doi:10.1038/ngeo2228.

Karl, T. R., A. Arguez, B. Huang, J. H. Lawrimore, J. R. McMahon, M. J. Menne, T. C. Peterson, R. S. Vose, and H. -M. Zhang (2015), Possible artifacts of data biases in the recent global surface warming hiatus, *Science*, *348*, 1469–1472, doi:10.1126/science.aaa5632.

Kennedy, J., N. Rayner, R. Smith, D. Parker, and M. Saunby (2011a), Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization, *J. Geophys. Res.*, *116*, D14104, doi:10.1029/2010JD015220. Kennedy, J., N. Rayner, R. Smith, D. Parker, and M. Saunby (2011b). Reassessing biases and other uncertainties in sea surface

temperature observations measured in situ since 1850: 1. Measurement and sampling uncertainties, J. Geophys. Res., 116, D14103, doi:10.1029/2010JD015218.

Kent, E. C., N. A. Rayner, D. I. Berry, M. Saunby, B. I. Moat, J. J. Kennedy, and D. E. Parker (2013), Global analysis of night marine air temperature and its uncertainty since 1880: The HadNMAT2 data set, J. Geophys. Res. Atmos., 118, 1281–1298, doi:10.1002/jgrd.50152.

Knutson, T. R., F. Zeng, and A. T. Wittenberg (2013), Multimodel assessment of regional surface temperature trends: CMIP3 and CMIP5 twentieth-century simulations, J. Clim., 26(22), 8709–8743, doi:10.1175/JCLI-D-12-00567.1.

Kosaka, Y., and S.-P. Xie (2013), Recent global-warming hiatus tied to equatorial Pacific surface cooling, *Nature*, *501*(7467), 403–407, doi:10.1038/nature12534.

Kurtz, N., T. Markus, S. Farrell, D. Worthen, and L. Boisvert (2011), Observations of recent Arctic sea ice volume loss and its impact on ocean-atmosphere energy exchange and ice production, J. Geophys. Res., 116, C04015, doi:10.1029/2010JC006235.

Mann, M. E., B. A. Steinman, and S. K. Miller (2014), On forced temperature changes, internal variability, and the AMO, *Geophys. Res. Lett.*, 41, 3211–3219, doi:10.1002/2014GL059233.

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

Meehl, G. A., J. M. Arblaster, J. T. Fasullo, A. Hu, and K. E. Trenberth (2011), Model-based evidence of deep-ocean heat uptake during surface-temperature hiatus periods, *Nat. Clim. Change*, 1(7), 360–364, doi:10.1038/nclimate1229.

Meehl, G. A., A. Hu, J. M. Arblaster, J. Fasullo, and K. E. Trenberth (2013), Externally forced and internally generated decadal climate variability associated with the Interdecadal Pacific Oscillation, J. Clim., 26(18), 7298–7310, doi:10.1175/JCLI-D-12-00548.1.

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones (2012), Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set, *J. Geophys. Res.*, *117*, D08101, doi:10.1029/2011JD017187.

Rayner, N., D. E. Parker, E. Horton, C. Folland, L. Alexander, D. Rowell, E. Kent, and A. Kaplan (2003), Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century, J. Geophys. Res., 108(D14), 4407, doi:10.1029/2002JD002670.

### Acknowledgments

The HadCRUT4 data are available from http://www.metoffice.gov.uk/hadobs/ hadcrut4/. The CMIP5 model outputs are available from http://pcmdi9. IInl.gov/esgf-web-fe/. The CDO package is available from https://code. zmaw.de/projects/cdo. Computer code is available from http://wwwusers.york.ac.uk/~kdc3/papers/ robust2015/. K.C. is grateful to the University of York for the provision of computing resources and to ETH-Zurich for data access. E.H. is funded by the UK Natural Environment Research Council. R.G.W. is funded by the Natural Sciences and Engineering Research Council of Canada. The authors would like to thank Gavin Schmidt, Reto Knutti, Markus Huber, John Kennedy, and Mark Richardson for data and advice.

The editor thanks two anonymous reviewers for their assistance in evaluating this paper.

Rayner, N., P. Brohan, D. Parker, C. Folland, J. Kennedy, M. Vanicek, T. Ansell, and S. Tett (2006), Improved analyses of changes and uncertainties in sea surface temperature measured in situ since the mid-nineteenth century: The HadSST2 dataset, J. Clim., 19(3), 446–469, doi:10.1175/JCLI3637.1.

Regayre, L., K. Pringle, B. Booth, L. Lee, G. Mann, J. Browse, M. Woodhouse, A. Rap, C. Reddington, and K. Carslaw (2014), Uncertainty in the magnitude of aerosol-cloud radiative forcing over recent decades, *Geophys. Res. Lett.*, 41, 9040–9049, doi:10.1002/2014GL062029.

Ridley, D., et al. (2014), Total volcanic stratospheric aerosol optical depths and implications for global climate change, *Geophys. Res. Lett.*, 41, 7763–7769, doi:10.1002/2014GL061541.

Rigor, I. G., R. L. Colony, and S. Martin (2000), Variations in surface air temperature observations in the Arctic, 1979–97, J. Clim., 13(5), 896–914, doi:10.1175/1520-0442(2000)013<0896:VISATO>2.0.CO;2.

Rohde, R., R. Muller, R. Jacobsen, S. Perlmutter, A. Rosenfeld, J. Wurtele, J. Curry, C. Wickham, and S. Mosher (2013), Berkeley Earth temperature averaging process, *Geoinfor. Geostat. An Overview*, 1(2), 1–13, doi:10.4172/2327-4581.1000103.

Saffioti, C., E. M. Fischer, and R. Knutti (2015), Contributions of atmospheric circulation variability and data coverage bias to the warming hiatus, *Geophys. Res. Lett.*, 42, 2385–2391, doi:10.1002/2015GL063091.

Santer, B. D., et al. (2014a), Volcanic contribution to decadal changes in tropospheric temperature, *Nat. Geosci.*, 7(3), 185–189, doi:10.1038/ngeo2098.

Santer, B. D., et al. (2014b), Observed multivariable signals of late 20th and early 21st century volcanic activity, *Geophys. Res. Lett.*, 42, 500–509, doi:10.1002/2014GL062366.

Schmidt, G. A., D. T. Shindell, and K. Tsigaridis (2014), Reconciling warming trends, *Nat. Geosci.*, 7(3), 158–160, doi:10.1038/ngeo2105. Sherwood, S. C., S. Bony, and J. -L. Dufresne (2014), Spread in model climate sensitivity traced to atmospheric convective mixing, *Nature*, 505(7481), 37–42, doi:10.1038/nature12829.

Simmons, A. J., and P. Poli (2014), Arctic warming in ERA-Interim and other analyses, Q. J. R. Meteorol. Soc., 141, 1147–1162, doi:10.1002/qj.2422.

Solomon, S., K. H. Rosenlof, R. W. Portmann, J. S. Daniel, S. M. Davis, T. J. Sanford, and G. -K. Plattner (2010), Contributions of stratospheric water vapor to decadal changes in the rate of global warming, *Science*, 327(5970), 1219–1223, doi:10.1126/science.1182488.

Solomon, S., J. S. Daniel, R. Neely, J.-P. Vernier, E. G. Dutton, and L. W. Thomason (2011), The persistently variable background stratospheric aerosol layer and global climate change, *Science*, 333(6044), 866–870, doi:10.1126/science.1206027.

Steinman, B. A., M. E. Mann, and S. K. Miller (2015), Atlantic and pacific multidecadal oscillations and Northern Hemisphere temperatures, Science, 347(6225), 988–991, doi:10.1126/science.1257856.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl (2012), An overview of CMIP5 and the experiment design, Bull. Am. Meteorol. Soc., 93(4), 485–498, doi:10.1175/BAMS-D-11-00094.1.

Thorne, P., S. Outten, I. Bethke, and Ø. Seland (2015), Investigating the recent apparent hiatus in surface temperature increases: Part 2. Comparison of model ensembles to observational estimates, J. Geophys. Res. Atmos., 120, doi:10.1002/2014JD022805.

Tollefson, J. (2014), Climate change: The case of the missing heat, *Nature, 505*(7483), 276–278, doi:10.1038/505276a.

Trenberth, K. E., and J. T. Fasullo (2013), An apparent hiatus in global warming?, *Earth's Future*, 1(1), 19–32, doi:10.1002/2013EF000165.
 Vose, R. S., et al. (2012), NOAA's merged land-ocean surface temperature analysis, *Bull. Am. Meteorol. Soc.*, 93(11), 1677–1685, doi:10.1175/BAMS-D-11-00241.1.