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

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

Corresponding author: Kevin Cowtan, Department of Chemistry, University of York, Heslington, York, YO10 5DD, United Kingdom. (kevin.cowtan@york.ac.uk)

[1]Department of Chemistry, University of

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.

York, Heslington, York, YO10 5DD, United

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 metres 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*,

Kingdom.

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 datasets [*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. IPCC AR5 WG1 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

---

[2]Energy and Resources Group, University

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 purpose of this work is 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 datasets.

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 datasets (Figure S1) there were 84 useable model runs from 36 models. The Climate Data Operators software package [*CDO*, 2015] was used to convert all fields onto a standard 1x1° grid, using distance weighted interpolation to avoid the loss of coverage when interpolating fields containing missing

of California Berkeley, Berkeley CA 94720.

values (however similar results were obtained using nearest neighbour 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:

$$\begin{aligned} w_{air} &= (1 - f_{ocean}) + f_{ocean} f_{ice} \\ T_{blend} &= w_{air} T_{air} + (1 - w_{air}) T_{ocean} \end{aligned} \quad (1)$$

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 CSIRO 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 mean of the blended temperature and the unblended air temperature provides a measure of the bias in the model-observation comparison.

---

[3]National Centre for Atmospheric Science,

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

Department of Meteorology, University of Reading, Reading. UK RG6 6BB.

[4]Department of Environmental Science and Policy, George Mason University, Fairfax VA 22030.

[5]Department of Meteorology and Earth and Environmental Systems Institute, Pennsylvania State University, University Park, Pennsylvania, USA.

[6]Department of Earth and Environmental Sciences, Large Lakes Observatory, University of Minnesota Duluth, Duluth, Minnesota, USA.

[7]Institute for Atmospheric and Climate Science, ETH Zurich, Universitaetstrasse 16, Zurich, Switzerland.

[8]Department of Geography, University of Ottawa, Ottawa, Canada, K1N 6N5.

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 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 dataset 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 onwards. 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 non-zero 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 5x5° 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 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 dataset.

3. Results

The difference between the global mean blended temperature and 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 multi-model global mean blended and air temperatures is 0.033 ± 0.010°C (1σ) relative to 1961-1990, and this difference is estimated to increase in magnitude with time to 0.18 ± 0.04°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 behaviour 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). Pre-industrial control simulations were examined (where available) to determine whether model drift due to non-equilibrium initial conditions contributes to the difference between air and sea surface temperature. 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 temperature 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 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, both for the RCP8.5 scenario and the RCP4.5 scenario where temperatures have largely stabilised 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 temperature slightly when using anomalies, and increases it slightly when using absolute temperatures. This behaviour arises from the

change in sign of the difference between the blended and air temperature 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 5x5° 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 change are small compared to the uncertainties and bias corrections in the sea surface temperatures [*Kennedy et al.*, 2011b, a], and so observational data are of limited use in detecting this bias. The comparison of daily sea surface temperatures to night-time 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] 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 falls below the 90% range of climate model simulations for the years 2011-2013. When using the blended temper-

atures, 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 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, b], 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 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 4(b). 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 ± 0.004°C/decade compared to an air temperature record with the same coverage, and 0.030 ± 0.011°C/decade compared to a global 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 (Supporting text S1), but lead to similar conclusions. Comparisons to the other temperature datasets will in turn require an appropriate choice of blending method or development of a custom method appropriate to that dataset. 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: firstly in the comparison of climate model ensembles to the observational record, secondly in estimating climate sensitivity, and thirdly 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 ± 0.015°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 dataset. However comparison to multiple observational datasets 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 temperature 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 timescales, 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. Firstly, 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. Secondly, 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 (Supporting Text S1) mitigates the problem at the cost of introducing a different but smaller bias. The new dataset 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 emphasise 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.

Acknowledgments.

The HadCRUT4 data are available from `http://www.metoffice.gov.uk/hadobs/hadcrut4/`

The CMIP5 model outputs are available from `http://pcmdi9.llnl.gov/esgf-web-fe/`

The CDO package is available from `https://code.zmaw.de/projects/cdo`

Computer code is available from `http://www-users.york.ac.uk/~kdc3/papers/robust2015/`

KC is grateful to the University of York for the provision of computing resources, and to ETH-Zurich for data access. EH is funded by the UK Natural Environment Research

309 Council. RGW is funded by the Natural Sciences and Engineering Research Council of
310 Canada.

311 The authors would like to thank Gavin Schmidt, Reto Knutti, Markus Huber, John
312 Kennedy and Mark Richardson for data and advice.

References

313 Andrews, T., J. M. Gregory, M. J. Webb, and K. E. Taylor (2012), Forcing, feedbacks and climate
314 sensitivity in cmip5 coupled atmosphere-ocean climate models, *Geophysical Research Letters*,
315 *39*, doi:10.1029/2012GL051607, l09712.

316 CDO (2015), Climate data operators, http://www.mpimet.mpg.de/cdo, version 1.6.8.

317 Comiso, J. C., and D. K. Hall (2014), Climate trends in the arctic as observed from space, *Wiley*
318 *Interdisciplinary Reviews: Climate Change*, *5*(3), 389–409, doi:10.1002/wcc.277.

319 Cowtan, K., and R. G. Way (2014), Coverage bias in the hadcrut4 temperature series and its
320 impact on recent temperature trends, *Quarterly Journal of the Royal Meteorological Society*,
321 *140*(683), 1935–1944, doi:10.1002/qj.2297.

322 Dai, A., J. C. Fyfe, S.-P. Xie, and X. Dai (2015), Decadal modulation of global surface
323 temperature by internal climate variability, *Nature Climate Change*, *5*(6), 555–559, doi:
324 10.1038/nclimate2605.

325 Flato, G., J. Marotzke, B. Abiodun, P. Braconnot, S. Chou, W. Collins, P. Cox, F. Driouech,
326 S. Emori, V. Eyring, et al. (2013), Evaluation of climate models, in *Climate Change 2013: The*
327 *Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of*
328 *the Intergovernmental Panel on Climate Change*, pp. 741–866, Cambridge University Press.

Fyfe, J. C., and N. P. Gillett (2014), Recent observed and simulated warming, *Nature Climate 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, *Nature Climate 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, *Climate Dynamics*, pp. 1–15, doi:10.1007/s00382-014-2464-2.

Hansen, J., R. Ruedy, M. Sato, and K. Lo (2010), Global surface temperature change, *Reviews of Geophysics*, *48*(4), 1–15, 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, *Journal of Climate*, *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, *Nature Geoscience*, pp. 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*, doi:10.1126/science.aaa5632, in press.

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, *Journal of Geophysical Research: Atmospheres (1984–2012)*, *116*, doi:10.1029/2010JD015220, d14104.

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, *Journal of Geophysical Research: Atmospheres (1984–2012)*, *116*, doi:10.1029/2010JD015218, d14103.

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, *Journal of Geophysical Research: Atmospheres*, *118*(3), 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, *Journal of Climate*, *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, *Journal of Geophysical Research: Oceans (1978–2012)*, *116*, C04,015, doi:10.1029/2010JC006235.

Mann, M. E., B. A. Steinman, and S. K. Miller (2014), On forced temperature changes, internal variability, and the amo, *Geophysical Research Letters*, *41*(9), doi:10.1002/2014GL059233, 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, *Nature Climate 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, *Journal of Climate*, *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, *Journal of Geophysical Research: Atmospheres (1984–2012)*, *117*(D8), doi:10.1029/2011JD017187, d08101.

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, *Journal of Geophysical Research: Atmospheres (1984–2012)*, *108*(D14), doi:10.1029/2002JD002670, 4407.

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, *Journal of Climate*, *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, *Geophysical Research Letters*, *41*(24), doi:10.1002/2014GL062029, 2014GL062029.

394 Ridley, D., S. Solomon, J. Barnes, V. Burlakov, T. Deshler, S. Dolgii, A. B. Herber, T. Nagai,
395 R. Neely, A. Nevzorov, et al. (2014), Total volcanic stratospheric aerosol optical depths and
396 implications for global climate change, *Geophysical Research Letters*, *41*(22), 7763–7769, doi:
397 10.1002/2014GL061541.

398 Rigor, I. G., R. L. Colony, and S. Martin (2000), Variations in surface air temperature ob-
399 servations in the arctic, 1979-97, *Journal of Climate*, *13*(5), 896–914, doi:10.1175/1520-
400 0442(2000)013¡0896:VISATO¿2.0.CO;2.

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

404 Saffioti, C., E. M. Fischer, and R. Knutti (2015), Contributions of atmospheric circulation vari-
405 ability and data coverage bias to the warming hiatus, *Geophysical Research Letters*, *42*(7),
406 2385–2391, doi:10.1002/2015GL063091, 2015GL063091.

407 Santer, B. D., C. Bonfils, J. F. Painter, M. D. Zelinka, C. Mears, S. Solomon, G. A. Schmidt,
408 J. C. Fyfe, J. N. Cole, L. Nazarenko, et al. (2014a), Volcanic contribution to decadal changes
409 in tropospheric temperature, *Nature Geoscience*, *7*(3), 185–189, doi:10.1038/ngeo2098.

410 Santer, B. D., S. Solomon, C. Bonfils, M. D. Zelinka, J. F. Painter, F. Beltran, J. C. Fyfe,
411 G. Johannesson, C. Mears, D. A. Ridley, et al. (2014b), Observed multivariable signals of late
412 20th and early 21st century volcanic activity, *Geophysical Research Letters*, *42*(2), 500–509,
413 doi:10.1002/2014GL062366, 2014GL062366.

414 Schmidt, G. A., D. T. Shindell, and K. Tsigaridis (2014), Reconciling warming trends, *Nature*
415 *Geoscience*, *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, *Quarterly Journal of the Royal Meteorological Society*, 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, *Bulletin of the American Meteorological Society*, *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, *Journal of Geophysical Research: Atmospheres*, 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., D. Arndt, V. F. Banzon, D. R. Easterling, B. Gleason, B. Huang, E. Kearns, J. H. Lawrimore, M. J. Menne, T. C. Peterson, et al. (2012), Noaa's merged land-ocean surface temperature analysis, *Bulletin of the American Meteorological Society*, *93*(11), 1677–1685, doi: 10.1175/BAMS-D-11-00241.1.

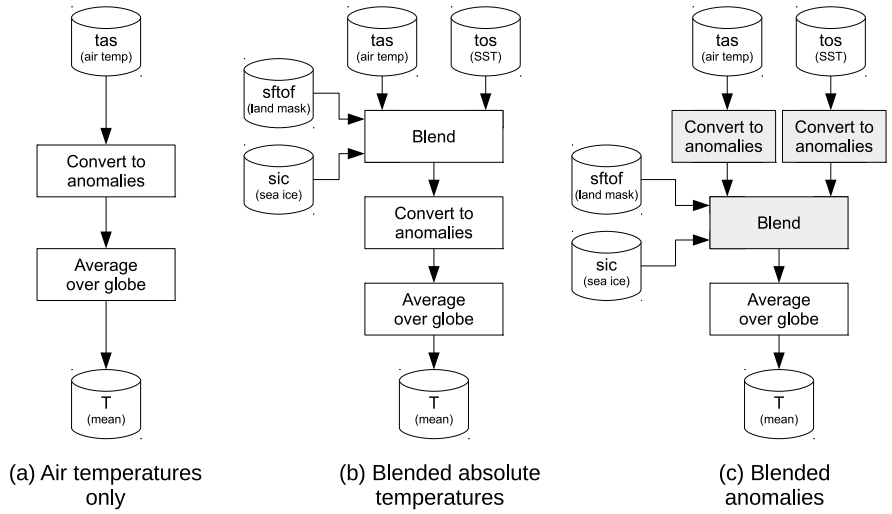

Figure 1. Flowcharts describing the calculation of global mean temperature (T) from the original CMIP5 fields. Three different methods are illustrated: (a) air temperature only (i.e. unblended). (b) blended absolute temperatures (no mask, variable ice). (c) blended temperature anomalies (no mask, variable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

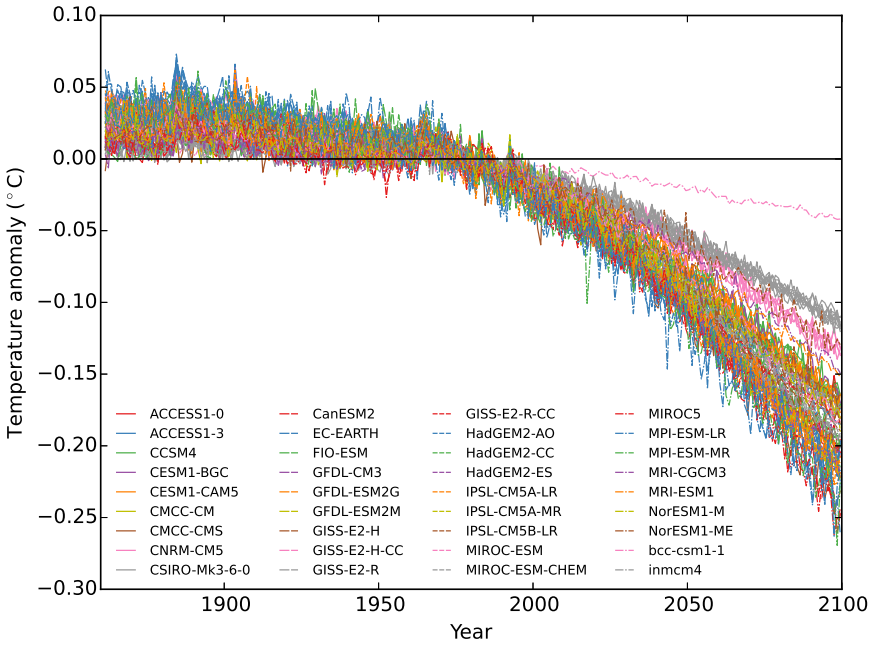

Figure 2. Difference between the global mean air temperature and blended land-ocean temperatures for 84 CMIP5 model simulations combining the historical and RCP8.5 experiments. The differences are calculated using global coverage and blending absolute temperatures with variable sea ice. Temperature anomalies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

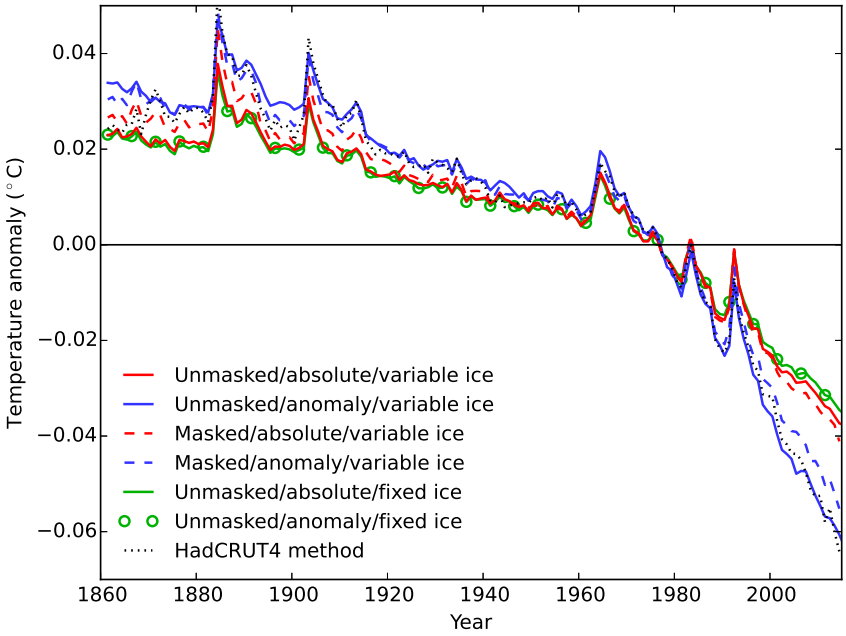

Figure 3. Difference between global mean blended temperature and air temperature, for different variants of the blending calculation, averaged over 84 historical + RCP8.5 simulations. Blended temperatures show less warming than air temperatures; hence the sign of the difference is negative for recent 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 behaviour to the other anomaly methods.

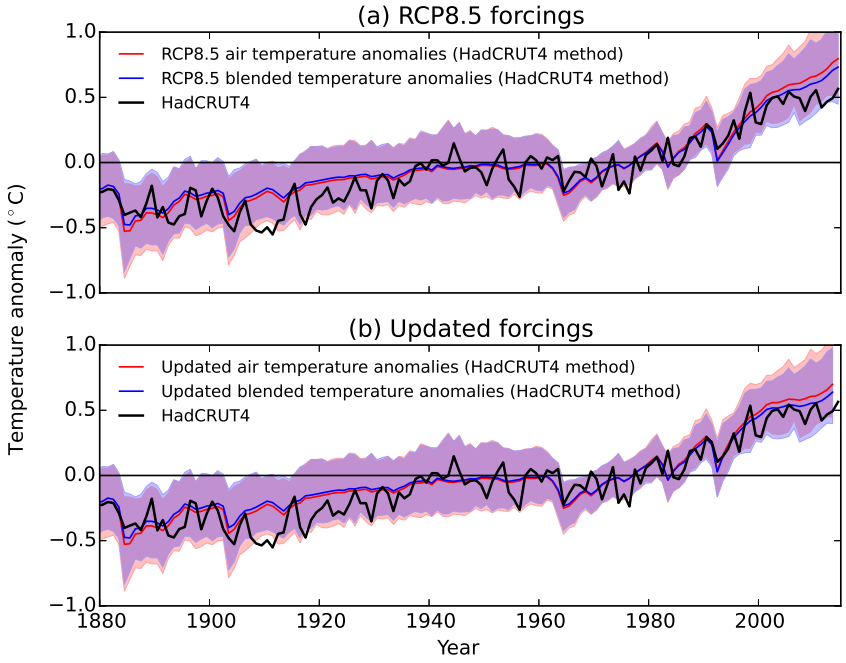

Figure 4. Comparison of 84 RCP8.5 simulations against HadCRUT4 observations (black), using either air temperatures (red line and shading) or blended temperatures using the HadCRUT4 method (blue line and shading). The shaded regions represent the 90% range (i.e. from 5-95%) of the model simulations, with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