Title
Improving the Observational Temperature Record

Author
Hausfather, Ezekiel Jon

Publication Date
2019

Peer reviewed|Thesis/dissertation

Improving the Observational Temperature Record

By

Ezekiel Jon Hausfather

A dissertation submitted in partial satisfaction

of the requirements for the degree of

Doctor of Philosophy

in

Energy and Resources

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Margaret Torn, Chair

Professor William Collins

Professor Lara Kueppers

Fall 2019

Abstract

Improving the Observational Temperature Record

by

Ezekiel Jon Hausfather

Doctor of Philosophy in Energy and Resources

University of California, Berkeley

Professor Margaret Torn, Chair

The observational temperature record is a critical part of our understanding of changes in Earth's climate. However, large uncertainties remain in our historical measurements of surface, ocean, and atmospheric temperatures. Many of these are introduced by changes in measurement techniques over time, such as changing instrumentation, time of observation, or changes to the surrounding environment not representative of the broader region. Reducing these uncertainties is important to improve our understanding of long-term climate change, and has implications for assessing the magnitude of inter-decadal climate variability, evaluating the performance of climate models, determining the remaining carbon budget to achieve mitigation targets, among other issues.

This dissertation is structured around four lead-authored papers that advance our understanding of the observational temperature record. The first paper, titled Quantifying the Effect of Urbanization on U.S. Historical Climatology Network Temperature Records, quantifies the extent to which changes in urban form surrounding measurement stations have biased long-term temperature records. By comparing temperature trends at urban and rural stations using four different proxy measures of urbanity, we find systematic differences between the raw (unadjusted) urban and rural temperature trends throughout the USHCN period of record. Based on these classifications, urbanization accounts for 14% to 21% of the rise in unadjusted minimum temperatures since 1895 and 6% to 9% since 1960. The homogenization process employed by NOAA effectively removes this urban signal such that it becomes insignificant during the last 50-80 years. In contrast, prior to 1930, only about half of the urban signal is removed. This suggests that biases in the land temperature record from urbanization are potentially significant, but can be effectively detected and removed when the network of observation stations is sufficiently dense to allow for neighbor-based pairwise homogenization.

The second paper is titled Evaluating the Impact of U.S. Historical Climatology Network Homogenization Using the U.S. Climate Reference Network. In this paper the homogenization of surface temperature records in the U.S. is assessed by comparing the old weather station network (USHCN) to a new state-of-the-art U.S. Climate Reference Network (USCRN). The new U.S. Climate Reference Network provides a homogenous set of surface temperature

observations that can serve as an effective empirical test of adjustments to raw USHCN stations. By comparing nearby pairs of USHCN and USCRN stations, we find that adjustments make both trends and monthly anomalies from USHCN stations much more similar to those of neighboring USCRN stations for the period from 2004-2015 when the networks overlap. These results improve our confidence in the reliability of homogenized surface temperature records.

The third paper, titled Assessing Recent Warming Using Instrumentally Homogeneous sea Surface Temperature Records, seeks to solve a substantial disagreement between warming rates in different Sea surface temperature (SST) records over the past two decades. SST records are subject to potential biases due to changing instrumentation and measurement practices. Significant differences exist between commonly-used composite sea surface temperature reconstructions from NOAA's Extended Reconstruction Sea Surface Temperature (ERSST), the Hadley Centre SST data set (HadSST3), and the Japanese Meteorological Agency's Centennial Observation-Based Estimates of SSTs (COBE-SST) in recent years. The update from ERSST version 3b to version 4 resulted in an increase in the SST trend estimate during the last 18 years from 0.07°C/decade to 0.12°C/decade, indicating a higher rate of warming in recent years and eliminating some of the apparent "pause" in global surface temperatures over that period. We show that ERSST version 4 trends generally agree with largely-independent, near-global and instrumentally-homogeneous SST measurements from floating buoys, Argo floats, and radiometer-based satellite measurements that have been developed and deployed during the past two decades. We find a large cooling bias in ERSSTv3b and smaller but significant cooling biases in HadSST3 and COBE-SST from 2003 to present with respect to most series examined. These results suggest that reported rates of SST warming in recent years have been underestimated in these three datasets due to biases in ship-based measurements.

The fourth paper, titled Evaluating the Performance of Past Climate Model Projections, looks at how well historical climate models published since 1970 have performed compared to observed temperature changes in the years after they were published. Climate models provide an important way to understand future changes in the Earth's climate. Model projections rely on two things to accurately match observations: accurate modeling of climate physics, and accurate assumptions around future emissions of CO2 and other factors affecting the climate. The best physics-based model will still be inaccurate if it projects future changes in emissions that differ from reality. To account for this, we look at how the relationship between temperature and atmospheric CO2 (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 CO2 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.

Work done in this dissertation has had a notable impact on our understanding and estimates of temperatures. This includes ensuring that urbanization is not biasing our record of land temperatures, testing the performance of land temperature homogenization, resolving

differences between ocean temperature records in recent decades, developing a novel sea surface temperature record to help better understand WW2-era uncertainties, and evaluating recent changes in ocean heat content. In an encouraging sign of the impact of our work, the new HadSST4 temperature product from the UK Met Office prominently features comparisons with the instrumentally homogenous sea surface temperature records we developed.

Similarly, the work that I and coauthors have undertaken has changed the approach used in evaluating the performance of GMST climate model projections, demonstrating the need to use common coverage and blended SAT/SST fields to ensure like-to-like comparisons with observations. Evaluating the future projections of old climate models improves our confidence that the current generation of models is accurately capturing the physical processes driving GMST change. This work on evaluating old climate models will be featured prominently in Chapter 1 of the IPCC 6th Assessment Report, where I serve as a contributing author.

CONTENTS

I. INTRODUCTION

1. ACCURATELY MEASURING THE CHANGING CLIMATE

Global temperature records have long been a critical part of our understanding of changes in Earth's climate. First estimated by Guy Callendar in 1938,[1] global temperatures have undergone many changes and improvements over the years, including assembling large datasets of historical observations and correcting for inhomogeneities introduced by changing instrumentation and measurement spatial and temporal characteristics.[2,3] The number and type of measurements has also expanded dramatically in recent decades, particularly after the advent of satellite-based temperature observations in the 1970s and the deployment of buoy and ARGO ocean measurement networks in recent decades.

Estimates of global temperatures tend to paradoxically suffer from both an underestimate and overestimate of uncertainty by users. Climate modelers and others who employ temperature products for comparisons often underappreciate both uncertainties in specific records and differences between observational records, as well as nuances regarding the spatial coverage of measurements and what things are actually being measured.[4] On the other hand, adjustments to temperature records or differences between temperature records have been portrayed in front of the United States Congress and in other public venues as calling into question the extent to which climate models are accurate or global warming is even occurring.[5]

Despite numerous groups of researchers around the world producing global temperature records, notable periods of disagreement and uncertainty remain. Figure I.1, below, shows a number of global surface temperature records (with a 5-year lagging mean smoothing applied to each). These smooth records reduce short-term variability associated with El Niño/Southern Oscillation (ENSO) events and reveal a number of different periods where records agree or diverge.

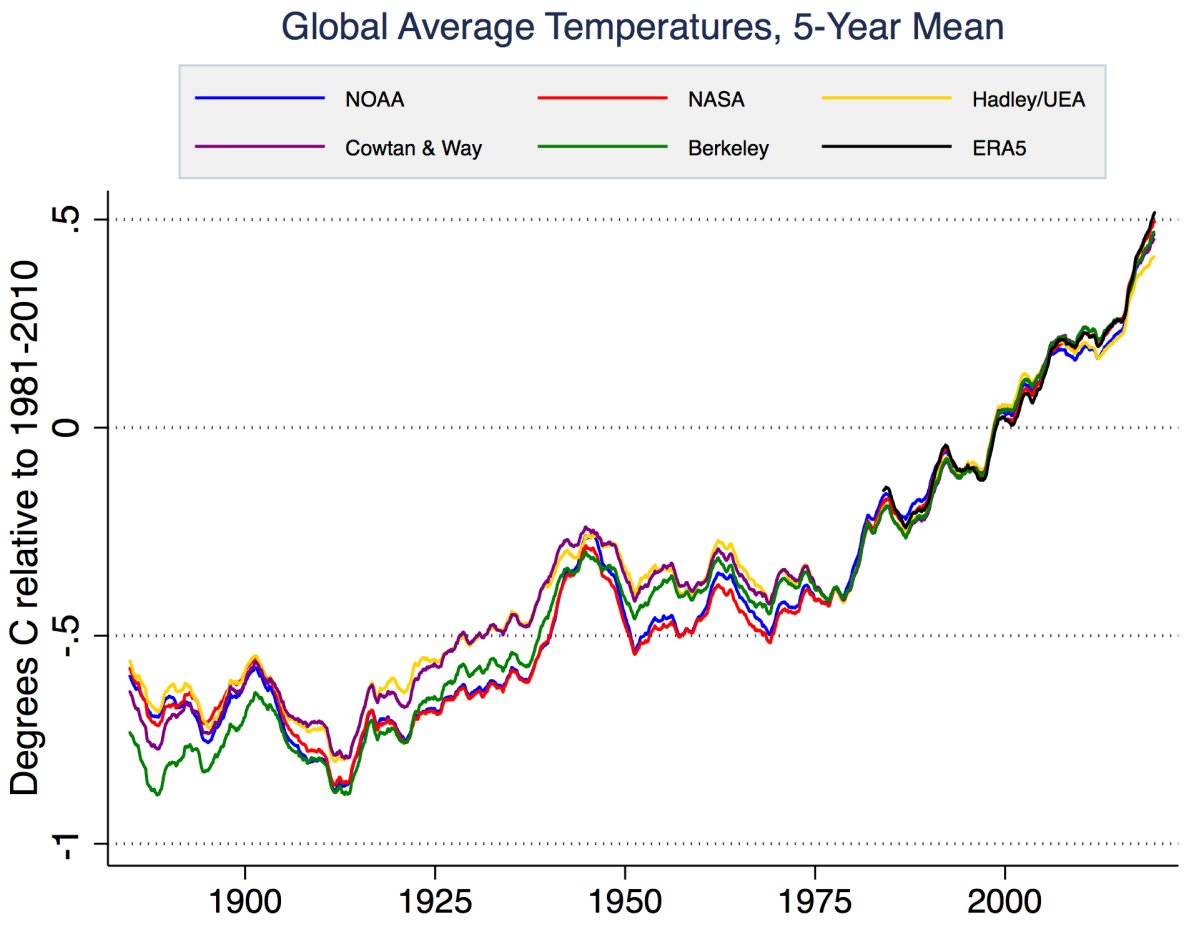


Figure I.1. *5-year lagging mean of global surface temperature from NASA's GISTEMP,[6] Hadley/UEA's HadCRUT4,[7] NOAA's GlobeTemp,[8] Cowtan and Way,[9] Berkeley Earth,[10] and ERA5[11] from January 1880 through October 2019.*

When examining the global temperature record, and a number of periods of substantive disagreement stand out. One particular divergence of interest both to policymakers and the climate modeling community occurs subsequent to the year 1998, when some estimates of surface temperatures show less warming between 1998 and 2014 (the so-called "hiatus" period), though these differences have become smaller in later revisions of some surface temperature records.[12]

There are large differences in temperature records both before and after the World War 2 (WW2) period, between 1920 and 1980. These emerge primarily from differences in sea surface temperatures (SSTs) between the UK's Hadley Centre HadSST record[13] and NOAA's ERSST record[14]. Smaller differences are present in the post-2000 period, driven by the differences

among ocean temperature records and by decisions about interpolation in regions of sparse coverage like the Arctic.[15,16]

Much of the differences between surface temperature records come down to adjustments made to raw observations to account for various measurement inhomogeneities. For land surface temperature measurements these include changing temporal and spatial observation availability, station moves,[17] instrument changes,[18] time of observation changes,[19] and changes in microscale or mesoscale characteristics like urbanization or other land use patterns.[20,21] Over the oceans much of the difference in records deals with changes in instrumentation, from wooden buckets to canvas buckets to engine room intake valves and finally to buoys.[3]

An example of the adjustments made to land and ocean components of NOAA's global surface temperature records are shown in Figure I.2, below. These are generally characteristic of the adjustments made to all surface temperature records, though notable differences arise over some specific periods as previously discussed.

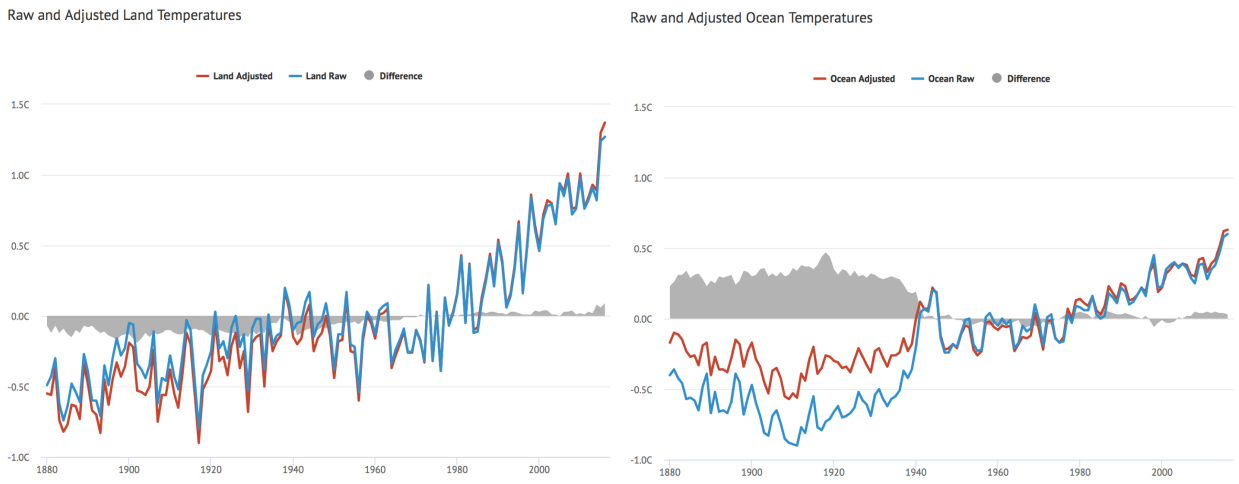


Figure I.2. *Raw and adjusted to global land temperatures (left) from NOAA GHCNv4 data[8] and ocean temperatures (right) from ICOADS and ERSSTv4.[14]*

Adjustments to land temperature records have a relatively modest impact from 1970 to present. Adjustments to land temperatures before 1970 reduce global temperatures, increasing the overall 1880-2017 land trend by around 16%. Adjustments to ocean temperatures have the opposite effect, increasing past temperatures prior to 1940 and reducing the 1880-2017 ocean

trend by around 36%. Overall 1880-2017 global temperature trends are reduced by around 20% due to adjustments, but trends from 1970 to present are increase by a modest 4%.

2. THE IMPORTANCE OF GLOBAL TEMPERATURE RECORDS

There are sizable differences between observational temperature records during a number of periods over the past 150 years. There are a number of different periods where the divergences in observational records are critically important. One that has gotten significant attention in recent years is the question about the extent to which there has been a detectable slowdown or "hiatus" in global temperatures post-1998. While the question is somewhat problematic given the choice of a large El Niño event at the start of the hiatus period, the extent to which a hiatus is detectable during the 1998-2014 period depends quite a bit on the choice of global temperature records.[22]

In particular, a 2015 paper in *Science* by NOAA's Tom Karl and colleagues received an immense amount of attention after they released a revised global temperature record that showed little detectable slowdown between 1998 and 2014.[23] This was primarily due to the inclusion of an updated SST record (moving from ERSSTv3b to ERSSTv4), which included corrections for a bias introduced by the transition from ship-based measurements to buoy-based measurements and weighted buoy-based measurements more in the resulting reconstruction.[14] This new temperature record differed noticeably from other SST records like HadSST3,[13] and resulted in considerably political controversy with the US Congress issuing subpoenas for the emails of the scientists involved. Thus understanding why SST records diverge during this period and determining which record is the most accurate has significant relevancy for both policymakers and the scientific community.

These post-2000 differences also have had a large impact on model-observation comparisons in recent years. Depending on the observational record used there may or may not be detectable divergences between model projections and observed temperatures.[22] The choice of observational record also impacts calculations of observationally-based estimates of transient climate response and equilibrium climate sensitivity.[24] While many observationally-based climate sensitivity estimates show lower sensitivity than seen in models,[25] some of this disagreement disappears when more spatially complete observational records are used and

when they are compared to model fields which similarly sample SSTs over the ocean and surface air temperatures over land.[24]

Another issue where uncertainty in observational temperature records is important is the estimation of carbon budgets. A central point in the debate about remaining allowable emissions is the amount of warming that has occurred since the late 1800s. A global temperature target of 1.5°C or 2°C above preindustrial depends a great deal on how preindustrial is actually defined. Unfortunately, the definition of preindustrial is quite vague, and the choice of both baseline and observational record has a big impact on remaining carbon budget.[26,27] The period prior to 1900 in particular is subject to large uncertainties, and different groups estimate a range of 1°C to 1.2°C warming in 2017 with respect to a 1880-1900 baseline, potentially reducing the allowable carbon budget to limit warming to 1.5°C by up to 40% depending on the observational record chosen. Better accounting for global temperature prior to 1800 as well as in more recent years can help narrow the uncertainty surrounding allowable future emissions.

3. EFFECTIVELY EVALUATING CLIMATE MODEL PERFORMANCE

Climate models are one of our most important tools to understand and project past and future changes to Earth's climate. Assessing the performance of climate models compared to observations can help identify where current models may be performing poorly, and can inform future model improvements. Understanding the historical performance of the current generation of climate models can also increase (or decrease) confidence in the accuracy of future projections.

While large differences exist among climate models, as represented in model ensembles-of-opportunity like CMIP5,[28] climate modelers can at times discount the large uncertainties present in some observational records. While uncertainties in modern global mean surface temperatures are relatively small, differences among observational records are still important when assessing short periods (such as the 1998-2014 period). Uncertainties become larger further back in time, with particularly large uncertainties in sea surface temperatures around the WW2-era and global temperatures prior to 1880. Large uncertainties also exist for tropospheric temperatures, with large differences between groups interpreting Microwave Sounding Unit (MSU) and Advanced Microwave Sounding Unit (AMSU) data.[29,30] Reanalysis products, which are increasingly used to

provide spatially and temporally-complete estimates across hundreds of different climate variables, may also be subject to temporal inhomogenites in the data they ingest.

Understanding and properly incorporating observational uncertainties is essential to properly evaluating climate model performance. In a number of cases – ocean heat content, for example – apparent mismatches between observations and model projections have been due to observational biases later corrected.[31]

When comparing models and observations it is also essential to make like-to-like comparisons of observational and model fields. For example, many observational temperature products have gaps associated with limited spatial coverage, particularly in regions like the Arctic and in the pre-satellite era.[9] Model fields should be masked to ensure the same temporal/spatial coverage as observations, instead of comparing globally complete estimates to more fragmentary ones.

Observations often comprise a combination of different measurement techniques. For example, the iconic observational global mean surface temperature (GMST) record is actually a combination of surface air temperatures (SAT) over the land and SSTs over the ocean.[4] While this was previously assumed to be comparable with the global surface air temperature field produced by climate models, work by myself and colleagues in 2015 found notable differences in the rate of warming between sea surface temperatures and surface air temperatures over the ocean in climate models. Creating blended SAT-over-land and SST-over-ocean model fields – along with masking models to observational coverage – explained around 40% of the apparent model-observational mismatch during the 1998-2014 "hiatus" period.[4]

Current-generation climate models are often compared to observations through "hindcasts" where observationally-based radiative forcing estimates are used to project historical temperatures from the mid-1800s onward. However, these hindcasts are not always an independent test of model skill. Some modeling groups have explicitly selected tunable parameters to improve GMST hindcast performance,[32] while others have implicitly done so, using poor hindcast performance as a reason to reassess parameter choices.[33]

Evaluating the performance of future projections from past climate models provides a more robust test of model skill, though it limits the assessable models to those produced at least 15 years ago when assessing GMST due to the internal variability.[34] In a recent paper we found that historical climate models show substantial skill in their future projections, with 10 of 17 model projections evaluated being statistically indistinguishable from observations on a

temperature vs time basis and 14 of 17 on an implied transient climate response (TCR) – e.g. the relationship between temperature change and forcing change.[34]

4. STRUCTURE OF THE DISSERTATION

This dissertation focuses on three distinct but related topic areas: land temperatures, ocean temperatures, and climate model/observation comparisons. Each topic area is the subject of one of the three dissertation chapters. Four published peer-reviewed papers are included in which I was the lead author: two on land temperatures, one on ocean temperatures, and one on model/observation comparisons.

Each chapter includes an introduction and discussion, in addition to the published papers. Each published paper includes its own separately numbered references and figures as appeared in print. References outside of the published papers are sequentially numbered throughout the document, and can be found in the additional references section at the end of the dissertation. A conclusions section at the end of the dissertation summarizes the results of each chapter and discusses general conclusions, outstanding questions, and additional research projects.

Finally, the dissertation appendix includes 15 additional papers either published or currently submitted on topics included in the dissertation on related subjects. These are divided into sections corresponding to the dissertation chapters, as well as an “other subjects” section for papers that do not fit under the three topic areas.

II. LAND TEMPERATURE HOMOGENIZATION

Land temperature observations come from weather stations located around the world. Different observational temperature products contain data from as little as ~5,000 stations (CRUTEM4)[34] to as many as ~32,000 (Berkeley Earth)[10]. In recent years, there has been a concerted effort by different groups including Berkeley Earth and ISTI[35] (both of which include the author as a member) to improve our collection of historical land temperature observations and increase the amount of data available for researchers to use.

Uncertainties associated with land temperature records arise from factors including instrument changes, station moves, time of observation changes, urbanization, and other changes in the environment surrounding station. These are generally detected and corrected via pairwise homogenization approaches that work by iteratively calculating difference series between each individual stations and its geographically proximate neighbors and identifying and removing localized breakpoints.[36] As long-term climate changes are not expected to result in significant localized heterogeneity in temperature changes, breakpoints found at individual stations but not in neighboring stations can generally be assumed to be localized biases (though this has limitations in regions with more sparse station coverage).

The research on land temperatures in this dissertation focuses on three areas:

1) Determining the impact of urbanization and other changes in station characteristics on land temperature records through comparison of trends in urban and rural stations.
2) Examining the efficacy of homogenization through comparisons with both homogenous reference networks and tests using synthetic data.
3) Collaborating with other researchers on the design of a global land temperature reference network that would provide more accurate measurements going forward.

1. URBAN HEAT ISLANDS AND OTHER MESOSCALE INFLUENCES

The structure of urban areas around the world has changed dramatically over the past century as global population has expanded and people have migrated from rural to urban areas. Similarly, land use has changed as agriculture has expanded (or contracted) in different regions

and forests have been cleared or allowed to regrow. When a weather station has continuous readings for 50 or 100 years, its local environment may experience large changes over that period.

Urbanization has long been considered a potential bias for temperature records.[37] The U.S. provides a dense network of long-lived measurement stations along with good metadata on urban characteristics, providing an opportunity to compare long-term warming differences between stations that are currently rural and those currently urban. While in principal stations in rural and urban areas should warm at roughly the same rate if their surrounding environments remain relatively unchanged, in practice many sites that were originally rural have been transformed into suburban or urban areas over the lifetime of the station (in the US, there are relatively few situations where the opposite would take place and areas would deurbanize over time, apart from cases where the station itself was moved).

In order to determine the extent to which changes in urban form bias temperature records, USHCN stations were categorized into urban or rural based on four different urbanity proxies based on remote sensing and other datasets for the square kilometer in which the station is located: the brightness of the location at night as measured by satellites, the percent of impermeable surface area, the growth in population density at the location between 1930 and 2000, and an urban boundaries database from the Global Rural-Urban Mapping Project.

By comparing both spatially gridded fields from urban and rural stations, as well as proximate pairs of urban and rural stations, we found that urbanization accounts for 14–21% of the rise in unadjusted minimum temperatures in the US since 1895 and 6–9% since 1960 (with smaller effects on maximum temperatures). This urbanization bias was effectively removed in the adjusted data, such that it was insignificant during the last 50-80 years (prior to that point the station network was sparser and only half the urbanization bias was detected and removed). We also produced a variant of the homogenized temperature record using only rural stations to homogenize, and found the results were nearly identical, suggesting that the composition of the network is sufficiently rural to limit the aliasing of urban heat island signals onto temperature trends during homogenization. The results were published in the Journal of Geophysical Research: Atmospheres and are included in full later in this chapter.[38]

One important finding of this work was that additional explicit urbanization bias adjustments – such as that done by NASA GISTEMP[6] – were not necessarily required if the data had undergone effective pairwise homogenization. However, pairwise homogenization only works

well when networks are sufficiently dense to detect localized breakpoints through difference series with neighboring stations. While the recent addition of tens of thousands of stations in global databanks used by NASA, NOAA, and Berkeley Earth will improve breakpoint detection in many parts of the world, there may still be some areas where the density is insufficient to detect and correct for urban-related biases and additional bias adjustments may be warranted.

2. ASSESSING THE EFFICACY OF HOMOGENIZATION

While the effect of homogenization on global land temperatures is relatively modest, resulting in an increase of warming trends since 1880 by around 16%, its impact in some regions is much larger. In the United States homogenization has a large effect on trends, roughly doubling the warming over the past century compared to raw temperature records, as shown in **Figure II.1** below.

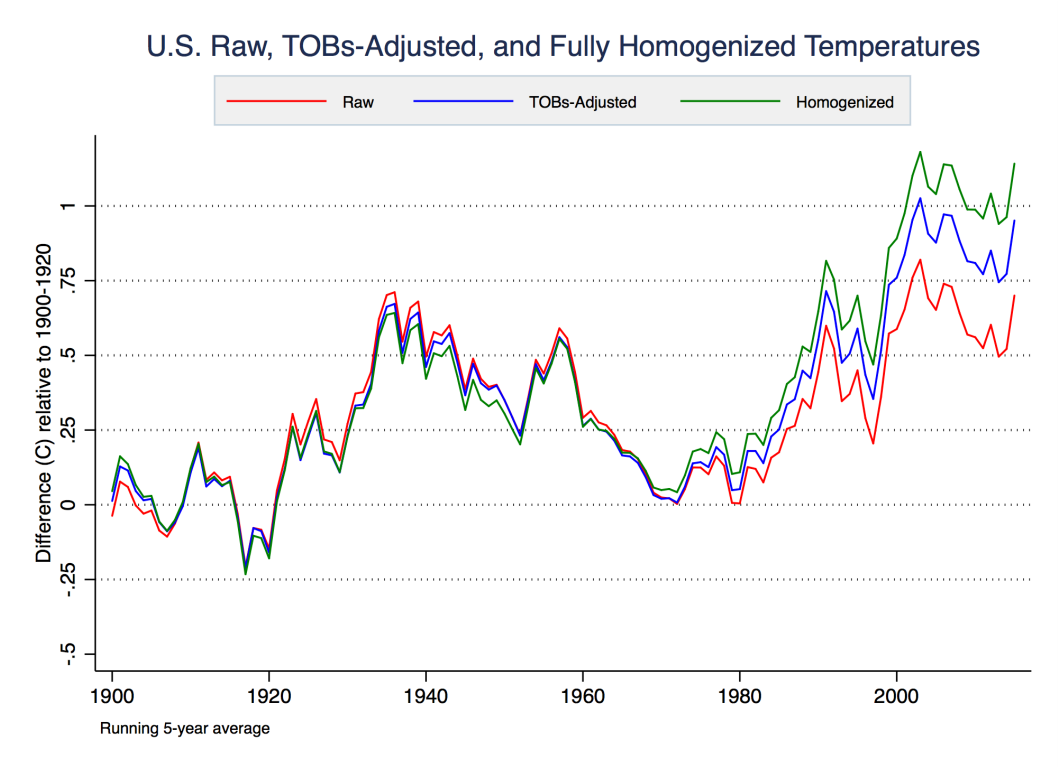


Figure II.1. *Raw, time of observation adjusted, and fully homogenized contiguous United States temperatures from the U.S. Historical Climatological Network.*[17]

Because of the large effect on the trend, these adjustments have proven controversial, resulting in multiple investigations of NOAA scientists' work by the U.S. Government Accountability Office (GAO). Independent evaluations of these approaches provide an important assessment of the U.S. temperature record, and can improve our confidence in their accuracy. To evaluate the effectiveness of these adjustments, we have taken a number of different approaches, including comparing raw and adjusted stations to the new U.S. Climate Reference Network (USCRN) and devising tests homogeneity algorithms using synthetic data.

2.1. REFERENCE NETWORK COMPARISONS

Existing temperature observation networks like the U.S. Historical Climatological Network (USHCN)[17] and its successor nClimDiv[39] (which is based on geographical regions called climate divisions) are subject to multiple inhomogeneities over time, with most stations having moved 2 to 3 times and changed instrumentation and time of observation at least once.

Many of these inhomogeneities were introduced to maximize the network's utility for short-term weather monitoring rather than long-term consistent climate observations. For example, the time of observation (e.g. the reset time of min/max thermometers) was changed at most US stations between 1960 and today due to a desire to conduct morning observations of rain gauges to avoid evaporation.

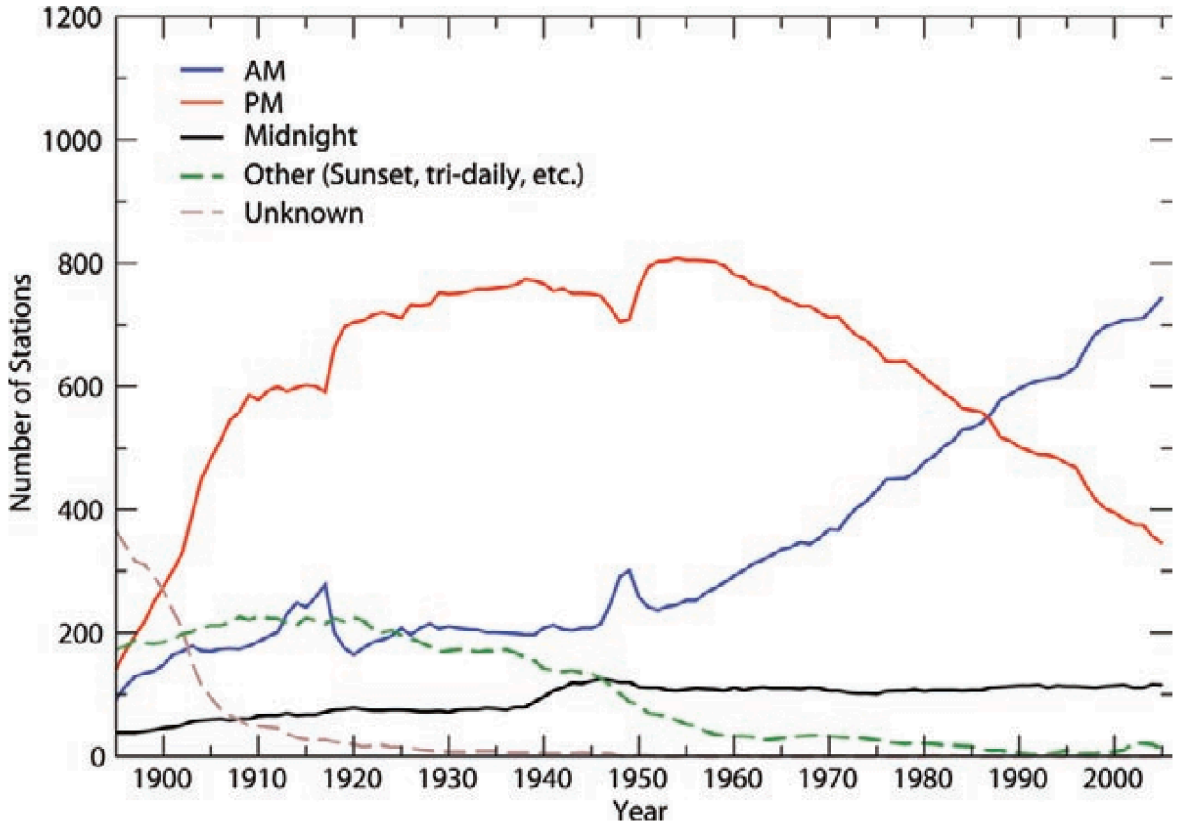


Figure II.2. *Time of observation at USHCN stations. From Menne et al 2009.*[17]

To avoid this problem going forward, the U.S. Climate Reference Network (USCRN) was established starting in 2001. USCRN stations are sited in pristine environments in rural areas away from any potential direct urban influence. Stations include three redundant temperature sensors that make measurements every 2 seconds and automatically report the data to a centralized server via satellite uplink. The USCRN is currently composed of 114 conterminous U.S. stations and has had sufficient station density and distribution to provide relatively good spatial coverage of the U.S. since the start of 2004.[40]

Figure II.3. *A picture of the three redundant temperature sensors at a pristinely sited U.S. Climate Reference Network station. Image from Diamond et al 2013.*[40]

After over a decade of operation, the period of overlap between the USCRN and other temperature observation networks is now sufficiently long to effectively assess the impact of temperature adjustments using the USCRN as an unbiased reference. We undertook a project to compare raw and adjusted USHCN stations to nearby USCRN stations in a paper published in Geophysical Research Letters in 2016 and included in full later in this chapter.[41] We found that adjustments bring raw station anomalies and trends much closer to nearby unbiased USCRN stations through an analysis of all proximate USHCN/USCRN pairs.

In nearly all cases adjustments serve to bring raw temperatures trends closer to those of the proximate USCRN station. Some residual trend differences remained in maximum temperatures, with USCRN stations warming faster than even adjusted USHCN stations. This is possibly due to differences in instrumentation, with the fan-aspirated solar shields employed by USCRN stations better capturing changes over the period than the USHCN instruments that include no fan.

The good level of agreement between adjusted USHCN stations and proximate USCRN stations increases our confidence that homogenization is effectively detecting and removing bias. While the current overlap period of 15 years is still relatively short, the CRN will provide a valuable resource going forward, giving access to known-unbiased temperatures for use in ensuring the accuracy of statistical homogenization approaches.

2.2. TESTS USING SYNTHETIC DATA

While comparisons to reference networks are a useful tool for validating homogenization, they are largely limited to the period post-2004 when the USCRN obtained U.S.-wide coverage. This was a period of limited systemic network changes (such as instrument changes or time-of-observation changes), and thus does not necessarily provide a thorough test of the performance of homogenization algorithms in more extreme circumstances.[41]

To provide a more detailed evaluation of where homogenization algorithms do and do not work effectively, we have used synthetic data where the "truth" is known and different types of biases are added. In most cases those running the homogenization algorithms are blinded to the actual true trends of the data they are evaluating until after the evaluation is completed. An example of the results of NOAA's pairwise homogenization algorithm applied to synthetic data with added trend biases is shown below in Figure II.4.

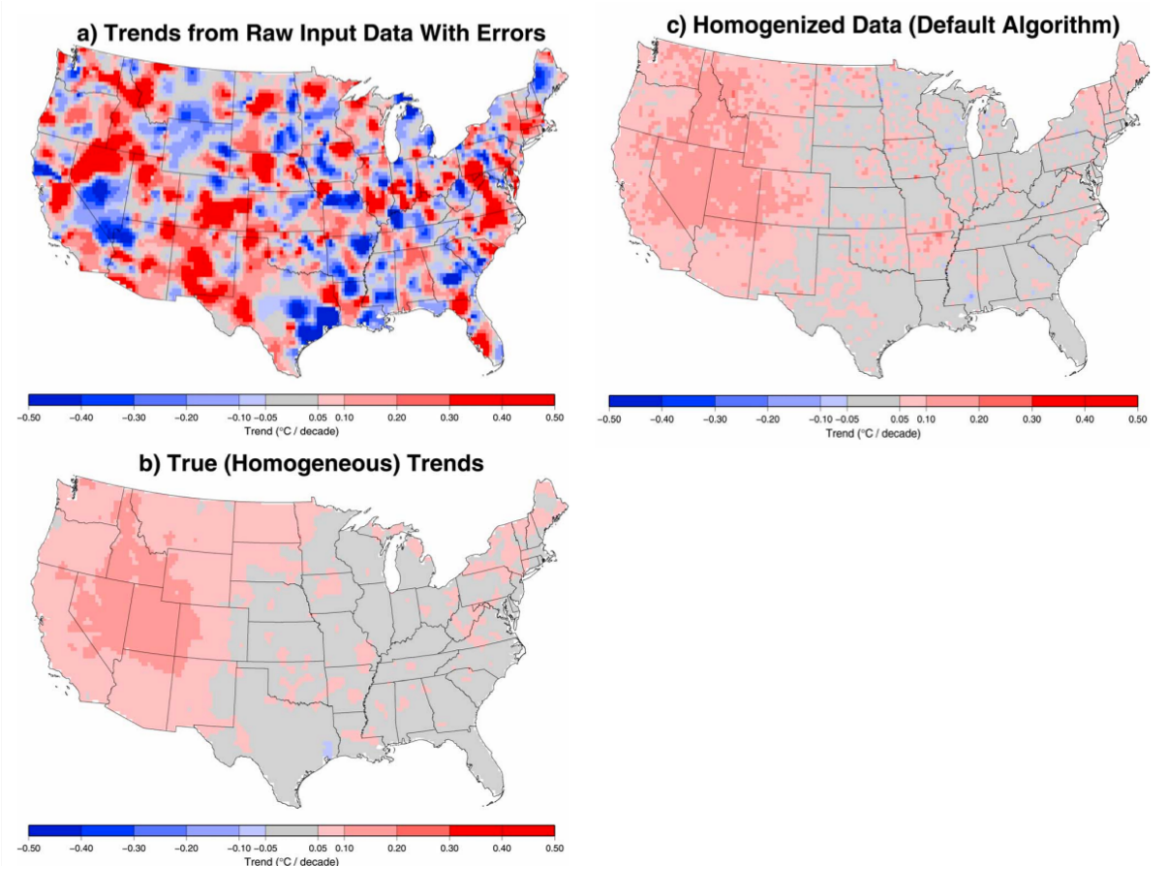


Figure II.4. *a) Synthetic data trends with added inhomogeneities, b) true trends in the synthetic data without inhomogeneities, c) data after being run through NOAA's pairwise homogenization algorithm. Figure via Williams et al 2012* [42] *and Hausfather et al 2014.*[43]

These tests using synthetic data have been applied to both NOAA and Berkeley Earth temperature records.[42,43] They show that homogenization is quite effective in detecting and removing biases. Performance is improved when metadata indicating potential breakpoints (e.g. station moves, instrument changes, time of observation changes, etc.) is available, but even in the absence of metadata localized biases can still be detected and removed. While station metadata is generally quite good in the US and Europe in recent decades, the same is not necessarily true for other parts of the world or in the more distant past. Thus the ability of statistical homogenization to operate in the absence of metadata indicating breakpoints is important.

Tests with synthetic data show that pairwise homogenization is effective as long as station-level inhomogenities are somewhat temporally stochastic. In other words, as long as a change in, say, observation does not happen simultaneously across the network, its effects can be detected and corrected through neighbor comparisons. This is one of the reasons why homogenizing satellite records is so much more difficult than surface records, for example, as the limited number of instruments and lack of redundancy makes cross-calibration and bias detection more difficult.

The International Surface Temperature Initiative has an ongoing program to develop benchmarks for homogeneity algorithm testing using synthetic data.[44,45] Early experiments have already found a diversity of performance across different groups around the world that provide statistical homogenization.[46]

3. TOWARDS A GLOBAL CLIMATE REFERENCE NETWORK

While the USCRN has been an invaluable resource in evaluating the adjustments made to the rest of the US weather stations, similar climate reference networks do not exist in most other parts of the world. However, as temperature anomalies are correlated over long distances, a much smaller network of only 160 stations well-distributed around the world's land area would provide sufficient coverage to calculate mean global land temperatures with reasonable

accuracy (more stations would be desirable for other climate variables such as precipitation). This network could be used to help assess the accuracy of regional weather station networks similar to the role the USCRN is playing in the United States.

An international team of surface temperature researcher – including the author of this dissertation – recently published a paper laying out what a global climate reference network would look like, in terms of siting, instrumentation, data management, and other factors.[46] The entire project would cost on the order of a few 10s of millions of dollars per decade, considerably less than a single satellite launch that only has a 5-10 year expected lifespan.

4. PAPER 1: QUANTIFYING THE EFFECT OF URBANIZATION ON U.S. HISTORICAL CLIMATOLOGY NETWORK TEMPERATURE RECORDS

Zeke Hausfather[1], Matthew J. Menne[2], Claude N. Williams[2], Troy Masters[3], Ronald Broberg[4], and David Jones[5]

[1]Energy and Resources Group, U°C Berkeley [2]NOAA/National Climatic Data Center [3]University of California at Los Angeles [4]University of Colorado, Colorado Springs [5]Climate Code Foundation

Journal of Geophysical Research: Atmospheres 118 (2), 481-494, 2013.

ABSTRACT

An assessment quantifying the impact of urbanization on temperature trends from the U.S. Historical Climatology Network (USHCN) is described. Stations were first classified as urban and non-urban (rural) using four different proxy measures of urbanity. Trends from the two station types were then compared using a pairing method that controls for differences in instrument type and via spatial gridding to account for the uneven distribution of stations. The comparisons reveal systematic differences between the raw (unadjusted) urban and rural temperature trends throughout the USHCN period of record according to all four urban classifications. Based on these classifications, urbanization accounts for 14% to 21% of the rise in unadjusted minimum temperatures since 1895 and 6% to 9% since 1960. The USHCN-Version 2 homogenization process effectively removes this urban signal such that it becomes insignificant during the last 50-80 years. In contrast, prior to 1930, only about half of the urban signal is removed. Accordingly, the NASA Goddard Institute for Space Studies urban-correction procedure has essentially no impact on USHCN version 2 trends since 1930, but effectively removes the residual urban-rural temperature trend differences for years before 1930 according to all four urban proxy classifications. Finally, an evaluation of the homogenization of USHCN temperature series using subsets of rural-only and urban-only reference series from the larger U.S. Cooperative Observer (Coop) Network suggests that the composition of Coop stations surrounding USHCN stations is sufficiently "rural" to limit the aliasing of urban heat island signals onto USHCN-Version 2 temperature trends during homogenization.

INTRODUCTION

Urbanization has long been recognized as having the potential to impact near-surface temperature readings by altering the sensible and latent heat fluxes in affected areas [e.g., *Mitchell*, 1953; *Oke*, 1982; *Arnfield*, 2003]. The concentration of high thermal mass impermeable surfaces in urbanized regions commonly leads to higher surface temperatures compared to those in less developed or rural areas, especially at night [*Oke*, 1982; *Parker*, 2010]. To mitigate the potential for an urban bias in temperature records used for climate monitoring, stations that comprise the U.S. Historical Climatology Network (USHCN) were selected to be largely from rural or small town locations [*Quinlan et al.*, 1987; *Menne et al.*, 2009]. Still, station locations tend to be correlated with inhabited areas. Relative to the percentage of total land area that is built up, "urban" observation stations are likely overrepresented in general, even in networks like the USHCN.

Given the potential for urban biases, a number of studies have been undertaken to quantify the impact of the "urban heat island" (UHI) signal on land surface air temperature trends globally (e.g., *Peterson et al.* 1999; *Parker*, 2006; *Jones et al.*, 2008; *Hansen et al.*, 2010) and regionally within the USA (e.g*., Kukla et al.*, 1986; *Karl et al.*, 1988; *Gallo et al.*, 1999, *Gallo and Owen*, 2002; *Peterson*, 2003; *Peterson and Owen*, 2005). Unfortunately, quantifying the impact of urbanization on temperature trends faces multiple confounding factors. For example, an instrument originally installed in an urban environment may well have warmer absolute temperatures than one in a nearby rural area, *ceteris paribus*, but will not necessarily show a higher trend over time unless the composition of the city or the microclimate around the sensor changes in such a way to cause the city observations to further diverge from temperatures at nearby rural locations [*Jones and Lister*, 2010], or the nature of urban land use leads to an amplifying of warm events whose frequency may change with time [*McCarthy et al.,* 2010]. It follows that urban heat island effects will lead to larger temperature trends compared to rural areas only if UHI-related factors cause incremental increases over rural temperatures during the period over which the trend is calculated [*Boehm*, 1998]. Moreover, cooling biases can be introduced into the temperature record when stations move from city centers to more rural areas on the urban periphery. This may have occurred, for example, during the period between about 1940 and 1960 when stations were moved from urban centers to newly constructed airports [*Hansen et al.*, 2001] and, in the case of the USHCN, airports, waste water treatment plants and other locations that lie outside the urban core [*National Climatic Data Center*, 2012].

Conversely, an instrument that is constructed in a relatively rural area that becomes more urban over time may exhibit a warming bias, and stations in small towns are not necessarily free of urban influences.

To further complicate matters, changes associated with urbanization may have impacts that affect both the meso-scale (10^2-10^4 meters) and the micro-scale (10^0-10^2 meters) signals. Small station moves (e.g., closer to nearby parking lots/buildings or to an area that favors cold air drainage) as well as local changes such as the growth of or removal of trees near the sensor may overwhelm any background UHI signal at the meso-scale [*Boehm*, 1998]. Notably, when stations are located in park-like settings within a city, the microclimate of the park can be isolated from the urban heat island "bubble" of surrounding built up areas [*Spronken-Smith and Oke,* 1998; *Peterson*, 2003]. Further, changes in observation practice such as time of observation and instrument changes may lead to artifacts (inhomogeneities) in the data record that complicate the quantification of urban heat island signals [*Peterson*, 2003], especially if these changes are correlated with urban form.

Here, an analysis is described whose aim is to quantify the potential UHI contribution to U.S. temperature trends by more fully controlling for external factors that impact the trends but are otherwise unrelated to urbanization. A range of estimates for the UHI contribution to average U.S. temperature trends is provided by making use of four separate ways to differentiate urban and rural station environments to help assess uncertainty associated with identifying urban environments. The impact of data homogenization on the UHI signal is also evaluated. Homogenization is necessary to account for shifts in the station-based data caused by historical changes in the circumstances behind surface temperature measurement (e.g., changes in instrument type, station relocations) rather than by true changes in the climate. The artifacts caused by these kinds of changes have large, systematic impacts on U.S. temperature trends [*Menne et al.* 2009; *Williams et al.* 2012]. Consequently, homogenized datasets are essential for evaluating temperature changes from the observational record [*Venema et al.* 2012; *Lawrimore et al*., 2011; *Hansen et al.* 2010; *Vose et al.*, in press]. Benchmarking the approach to homogenizing the U.S. monthly temperature data has essentially reaffirmed previous assessments regarding the nature and impact of these artifacts on USHCN temperature trends [*Williams et al.* 2012] .

Homogenization of the USHCN monthly version 2 temperature data does not specifically target changes associated with urbanization. Rather, the procedure used involves identifying and

accounting for shifts in the monthly temperature series that appear to be unique to a specific station--the assumption being that a spatially isolated and sustain shift in a station series is caused by factors unrelated to background climate variations [*Menne et al.* 2010]. Given that UHI-related changes may manifest as highly localized shifts or creeping changes, the focus in this analysis is to determine to what extent homogenization is removing apparent, local urban influences on the USHCN temperature record. Because homogenization may be removing local shifts caused by land use changes at non-urban stations, the same methodology used here could be applied to evaluating the impact of other types of land use changes.

The paper is organized as follows. Some additional background and motivation for the study are provided in section 2. The datasets and methods are discussed in section 3. Results are presented in section 4. Conclusions are provided in section 5.

BACKGROUND AND MOTIVATION

Motivation for assessing urban influences on temperature trends comes largely from interest in quantifying the contribution of urbanization in overall temperature trends relative to other factors. To that end, measures of ambient population [*Kukla et al.*, 1986] and satellite-derived nightlights [*Gallo et al.*, 1999] have been used to differentiate urban and rural environments. Using these measures, monthly temperatures from U.S. weather stations designated as urban have been found to have decadal trends as much as 0.12°C/decade higher than those classified as rural [*Kukla et al.* 1986]. Because differences of this magnitude represent a non-negligible fraction of the likely background climate change signal, *Karl et al.* [1988] developed a specific adjustment to control for the apparent contribution of the urban heat island signal in USHCN temperature data. After adjusting for shifts in the data associated with time of observation and other changes documented in station histories, the *Karl et al.* [1988] evaluation suggested that an additional urban bias was present in the USHCN average temperature of about 0.06°C during the period from 1900 to 1984. Essentially all of the bias was associated minimum temperatures in urban areas, which were about 0.13°C higher on average than rural areas; maximum temperatures appeared to have little urban bias.

The *Karl et al.* [1988] UHI correction was used to produce the USHCN (version 1) fully adjusted USHCN monthly temperature data until the release of version 2 [*Menne et al.* 2009]. As in version 1, the version 2 release includes bias adjustments for time of observation and other

station history changes, but version 2 also includes adjustments for changes (inhomogeneities) that are not documented in digital station histories (roughly 50% of all changes). Providing adjustments for both documented and undocumented station changes reduced the overall magnitude of minimum temperature trends from USHCN stations more than the fully adjusted version 1 temperatures even though version 1 contained the additional *Karl et al.* [1988] UHI adjustment. The reason for this may be that the more comprehensive homogenization in version 2 removes the impact of incremental, but previously unidentified step changes associated with meso and micro-scale urbanization factors, or, that signal arising from local UHI trend changes are sometimes aliased (i.e., inadvertently accounted for) onto estimates of the more comprehensive version 2 step-type adjustments [*Menne et al.* 2009]. In any case, the development of a method for identifying and adjusting undocumented shifts appeared to account for more than of the signal attributed to urban effects on minimum temperatures by *Karl et al.* [1988]. Thus, no separate UHI-specific correction was provided in USHCN version 2.

Another reason that the *Karl et al.* [1988] corrections were not used in version 2 is that they are monotonic functions of city population; that is, these adjustments always reduced minimum temperature trends based on the total population of the city. In contrast, *Hansen et al.* [1999, 2001, 2010] have used a nightlights-based method that forces urban (and "peri-urban") station trends to conform to surrounding rural trends in the NASA Goddard Institute for Space Studies (GISS) surface temperature analysis. In the process, the *Hansen et al.* approach actually increases the trend for about 40% of urban stations. The fact that so many urban trends are larger after the urban adjustment likely reflects the degree to which the confounding factors discussed above can mitigate or otherwise obscure potential urban heat island signals.

For the U.S. data contribution to the NASA GISS analysis*, Hansen et al.* [2001, 2010] use the USHCN data that has been adjusted by NOAA/NCDC for time of observation and station history changes, but apply their own UHI adjustment. The GISS urban adjustment reduced the otherwise adjusted USHCN version 1 temperature trends by an additional 0.15°C/century, more than twice that of *Karl et al.* [1988] method [*Hansen et al.* 2001] even though the NASA GISS UHI corrections are not monotonic. Using the USHCN version 2 adjusted data, the impact of the GISS UHI correction is on the order of 0.07°C/century [*Hansen et al.* 2010].

The differential impacts of these approaches to assessing and correcting for the UHI are indicative of the need to better frame the uncertainty of urban influences on temperature trends in the U.S. As noted more recently by *Peterson* [2003] and *Peterson and Owen* [2005], this

requires controlling for the many confounding issues like differences in instrumentation and other observation practices that may blur the urban signal. Whereas *Peterson* [2003] and *Peterson and Owen* [2005] focused primarily on a snapshot of mean urban-rural differences, here we build on their work by looking specifically at the time evolution of urban-rural differences. We use four rather than two proxy measures of urbanity and quantify the impact of data homogenization on the apparent UHI signal, focusing in particular on the potential magnitude of residual UHI contamination and whether there is evidence that homogenization transfers UHI bias from urban to non-urban station series.

METHODS

The Conterminous United States (CONUS) has some of the most dense, publicly available digital surface temperature data in the world with over 7000 Cooperative Observer (Coop) Network Program stations reporting daily maximum and minimum temperature for at least 10 of the network's 120-plus year history. A subset of 1218 stations, generally those with long records, comprises the USHCN [*Menne et al.*, 2009]. This highly sampled surface temperature field allows for the comparison of subsets of station data in a manner that avoids inherent biases due to changes in spatial coverage. The Coop Program also now maintains accurate geolocational information on the present location of observing stations, with coordinates expressed in degrees, minutes and seconds (roughly 30 meter accuracy) available for most stations. This also allows for the accurate indexing of current Coop station locations against high-resolution georeferenced datasets that are useful to delineating urban and non-urban areas.

Because there is not an obvious meso-scale metric that determines the impact of urban form on temperature in all situations, we examined four different measures of urbanity that are available as georeferenced datasets: satellite-derived nightlights, urban boundary delineations, percent of impermeable surfaces, and historical population growth during the period where high-resolution data is available (1930 to 2000). These four measures, which represent different snapshots of urban boundaries, were used to classify a station as urban or non-urban by retrieving the pixel values coincident with the each station's coordinates. In cases where the proxy for urban form involved continuous measurements (all but urban boundaries), a cutoff point to divide stations between urban and rural was chosen based on urban designations present in the literature (e.g.

Hansen et al., 2010 for nightlights; *Elvidge et al*. 2007 for impermeable surface area). Each of these proxies is described in section 3.a. below.

DATASETS USED TO CLASSIFY STATION TYPES

Satellite Nightlights

Satellite-derived brightness values associated with the COOP Network stations (including the USHCN) were taken from the Global Radiance Calibrated Nighttime Lights dataset produced by the Earth Observation Group using instruments flown on Defense Meteorological Satellite Program (DMSP) satellites. We used the data from the F16 satellite recorded between 2005-11-28 and 2006-12-24. The values we associate with each station are linearly interpolated from the 4 neighbor pixels in the image file and are converted to radiance by multiplying by 1.51586×10^{-10} giving a result in Watts sr^{-1} cm^{-2} [*Baugh et al.*, 2010]. To determine a radiance value threshold for designating urban stations that is consistent with the 32 microWatts/m^2/sr/micrometer used in *Hansen et al.* [2010] (who used data from *Imhoff et al.*, 1997), we divided radiance values by the optical bandwidth of the F16 satellite (0.7 micrometers), resulting in a cutoff of 14.78 (i.e, $32 \div 0.7 \times 1.51586$) as the equivalent value for the 2005-2006 satellite nightlight series. This is rounded to the nearest integer (15) for the purpose of assigning a cutoff to separate urban from non-urban pixels.

Urban Boundaries (GRUMP)

For the urban boundaries urbanity proxy, we use binary designations from the Global Rural-Urban Mapping Project (GRUMP), produced by the Center for International Earth Science Information Network (CIESIN) of the Earth Institute at Columbia University. GRUMP designations are based on the identification of urban areas using national census data (including the National Imagery and Mapping Agency database of populated places). GRUMP purports to identify cities and towns with populations exceeding 1,000 residents. Urban boundaries surrounding identified cities and towns are estimated based on DMSP Operational Linescan System (OLS) data from 1994-1995 as well as data from the Digital Chart of the World's Populated Places (DCW) [*Balk et al*., 2004].

Impermeable Surfaces

The Impervious Surface Area (ISA) for pixels coincident with Coop stations is taken from the Global Distribution and Density of Constructed Impervious Surfaces dataset produced by the Earth Observation Group. The 1km resolution data used for this study was derived from 30-meter ISA data generated by the US Geological survey as described in *Elvidge et al.* [2007]. The data product has a nominal date of 2000-2001 and represents the percentage of the surface area that is comprised of manmade structures such as roads, buildings and parking lots. Station latitude and longitude were used to reference the dataset and extract the percentage of impervious surface in the surrounding 1 km. To determine the urban/non-urban classification a cut off of ten percent was employed. As noted by *Schuler* [1994] and *Arnold and Gibbons* [1996], the impacts to hydrology typically begin above this figure. ISA values below ten percent were classified as rural. This approach is consistent with though somewhat more conservative than the recent work of *Potere et al.* [2009], who used a figure of twenty percent for detecting urban extent.

Population Growth

For the population growth proxy, we utilized Gridded 1 km Population Estimates for the Conterminous U.S., 1930-2000. This dataset was also used by *Peterson and Owen* [2005] and *Peterson* [2003] to classify USHCN stations into urban and non-urban categories. The gridded population was created using two U.S. Census Bureau data sets: The 2000 U.S. Census Bureau 1 km^2 population density grid for CONUS (*National Geophysical Data Center/NESDIS/NOAA*, 2002) and tabular U.S. Census county data [*U.S. Census Bureau*, 2002]. Urban sites were defined as those characterized by a 1930 to 2000 population growth of greater than or equal to 10 people per square kilometer, which yields similar sized numbers of urban and non-urban stations as shown in Table 1. While there is no available justification in the literature for this or any specific 1930-2000 population growth cutoff as a proxy for urbanization, this value was chosen to be reasonably conservative and to produce an urban/rural division generally in line with the other urbanity proxies. As Table 1 indicates, the GRUMP, Nightlights, and Population Growth urbanity proxies result in a relatively even distribution of stations in the rural and urban categories while the ISA proxies identifies the majority of stations as rural. Information on retrieving these datasets are provided as supplementary information.

Proxy Name	Urban Boundaries	Nightlights	ISA	Pop Growth

Rural Stations	608	594	857	685
Urban Stations	610	624	357	533

Table 1*: Number of USHCN stations classified by urbanity for each urbanity proxy. Note that four stations could not be classified using the ISA urbanity proxy due to dataset limitations.*

CALCULATION OF RURAL AND URBAN TEMPERATURE TREND DIFFERENCES

Urban-rural temperature differences were calculated by sub-setting the USHCN station data according to the urban/non-urban station classifications described above (for simplicity non-urban stations are referred to as rural). To examine the possible UHI signal present in the USHCN temperature record, we use two different but complimentary methods to compare urban and rural station temperatures: station pairing and spatial gridding.

Station Pairing Method

The station pairing method creates pairs of nearby urban and non-urban (rural) stations as classified by the four urban proxy measures. Pairs were created by forming all possible permutations of urban and rural stations, excluding those that were more than 161 kilometers (100 miles) apart; that had differing or unknown sensor equipment types (e.g. Maximum Minimum Temperature Sensors [MMTS] versus Liquid in Glass Thermometers in Cotton Region Shelters [CRS]); or cases in which both stations currently have MMTS sensors but installation dates differ by more than 5 years. This pairing method yields a set of proximate urban/rural station pairs for each classification method that should be relatively unaffected by bias introduced through sensor-type transitions [*Quayle et al.*, 1991; *Menne et al.*, 2009]. Time series of monthly maximum and minimum temperature anomalies relative to a 1961-1990 baseline were calculated for all urban and rural series. Difference series for each urban and rural station pairings were then created for the full period of the USHCN version 2 records (1895 to present).

More specifically, the approach in the station pairing method was to take all permutations of urban and rural stations and produce a set containing unique pairs but non-unique occurrences of individual urban and rural station series (see Table 2). For example, a specific urban station

would create distinct pairs with all surrounding rural stations within 100 kilometers with the same instrumentation type. To avoid overweighting regions with large numbers of adjacent urban and rural stations (and thus disproportionately more possible station pair combinations) we weight the urban-rural differences by the inverse of the number of stations pairs associated with each unique urban station. The mean urban-rural differences for unique urban stations are averaged for each month to obtain a best estimate of the underlying urban-rural temperature differences.

Proxy Name	Urban Boundaries	Nightlights	ISA	Pop Growth
Total Station Pairs	1684	1809	1446	1392
Unique Urban Stations	437	470	271	390

Table 2: *Number of total urban/rural station pairs and unique urban stations by urbanity proxy.*

The trend and confidence intervals for two periods, 1895-2010 and 1960-2010, are calculated from the station pair data using a weighted regression with clustered standard errors, with unique urban stations used for both the weighting and clustering. Standard errors are clustered by unique urban station because station pairs contain non-unique occurrences of individual urban and rural stations (e.g. a single urban station might be paired with four different rural stations), and treating each station pair as independent would result in erroneously narrow confidence intervals. As mentioned previously, each urban-rural pair is given a weight in the regression proportionate to the inverse of the number of station pairs that share the same urban station.

The station pairing method allows us to control for both spatial coverage and sensor type, avoiding potential complications introduced by differing locations of urban and rural stations as well as urban-correlated bias in the transition to MMTS sensors in the 1980s. The results will not necessarily be as representative of the entire CONUS temperature field as those produced by spatial gridding, however, as station pairing does not explicitly weight based on spatial coverage.

Spatial Gridding Method

The spatial gridding method is used to create separate gridded fields for the conterminous U.S. using the subsets of urban and rural station series (and separately for maximum and minimum temperatures) as classified by each urban proxy measure. Station temperatures are converted to anomalies relative to a 1961-1990 baseline period, and station series that fall within 2.5° latitude x 3.5° longitude grid cells are averaged together and each grid cell average is cosine weighted to produce a CONUS average time series. The CONUS average urban and rural station series are then differenced. Trends and confidence intervals for the urban-rural differences during the 1895-2010 and 1960-2010 periods are calculated by regressing against the date using an AR(1) model to account for autocorrelation.

The gridding method described above is commonly used by NOAA/NCDC to produce spatially averaged time series for climate monitoring. In addition to this method, results using the gridding method described in *Menne et al.* [2009; 2010] are provided as supplementary information.

USHCN VERSION 2 MONTHLY TEMPERATURE DATA

Urban-rural differences for mean monthly maximum and minimum temperatures were calculated using four different versions of the USHCN version 2 monthly temperature data. The four versions were used to help quantify the magnitude of the UHI in the underlying raw (unhomogenized) data, to isolate the impact of data homogenization on the UHI signal, and to evaluate impact of the GISS UHI correction when applied as an addition correction over and above homogenization. The dataset versions include

1) time of observation-only adjusted data (called TOB);
2) adjusted version 2 (TOB + pairwise homogenization adjustments; v2);
3) adjusted version 2 data produced by running the pairwise homogenization algorithm using (a) neighboring series classified only as rural (v2-rural neigh); and, (b) neighboring series classified only as urban (v2-urban neigh).
4) adjusted version 2 data with the GISS UHI correction (TOB + pairwise homogenization + GISS UHI adjustments; v2+step2)

Each of these variants is described below.

Time of Observation Bias-Adjusted data (TOB)

The TOB station series are the raw monthly temperature data adjusted only for the time-of-observation bias [*Schaal and Dale,* 1977; *Karl et al.*, 1986]. The time of observation bias is an artifact of the starting/ending hour for the 24-hour interval over which the maximum and minimum temperature occurred. This bias is unrelated to any physical artifacts associated with urbanization and only leads to biased trends when the time of observation changes through time. However, such changes are likely more prevalent at rural stations, which are commonly run by volunteer observers who have been systematically transitioning from afternoon to morning observation times [*Menne et al.*, 2009]. In order to remove the time of observation bias as a confounding factor in assessing UHI impacts, we use data adjusted according the method described by *Karl et al.* [1986] and *Vose et al.* [2003]. Results using completely unadjusted (raw) data are provided as supplementary information using the *Menne et al.* [2009; 2010] gridding method.

Data adjusted by the Pairwise Homogenization Algorithm (USHCN version 2)

Running the TOB-adjusted data through the Pairwise Homogenization Algorithm (PHA; *Menne and Williams*, 2009) produces the USHCN version 2 fully adjusted data [*Menne et al.*, 2009]. The PHA works by identifying and removing abrupt shifts in monthly temperature series that appear to be unique to a particular station. The shifts can be caused by small station moves, a change in instrumentation, or, possibly, from the local impacts of any kind of land use change. The shifts are identified via automated pairwise comparisons of monthly temperature series in which the relative homogeneity of a given station's series is evaluated by looking for breaks in differences series formed between the target station and a number of highly correlated neighboring series. The adjustments are based on the median shift magnitude calculated from pairwise temperature differences between the target and neighbors before and after the shift. For any particular target adjustment, the neighbor pool is drawn from those that appear to be homogeneous according to the PHA for a minimum period (24 months) before and after the target shift. The PHA does not specifically target urban station changes. Rather, the algorithm targets all shifts that appear to be unique to a particular station. Removing these local signals at all stations (rural and urban alike) produces temperature trend fields that more accurately reflect the general background climate signal than the raw data.

For version 2, USHCN monthly temperatures were compared to sets of highly correlated neighboring series within the larger Coop Network. Details regarding the mechanics of the PHA and evaluations of the algorithm's efficiency can be found in *Menne and Williams* [2009] and

Williams et al. [2012a]. Version 2.0 of the adjusted monthly data was released in 2008 based on PHA version "52d". Urban-rural differences in version 2.0 adjusted data are discussed below. An evaluation of the UHI signal in a new version of the dataset—termed version 2.5—is provided as supplementary information using the Menne et al. [2009, 2010] gridding method. Version 2.5 fully homogenized data are produced by algorithm version "52i", which contains some bug fixes relative to version 52d [Williams et al., 2012b].

To evaluate the potential for UHI bias to be transferred from urban Coop stations that may be used as neighbors in the homogenization of USHCN station records, we also ran the USHCN station series through the PHA using only Coop stations that were classified as rural in one case and using only stations classified as urban in the other according to the same set of four urban proxies.

Version 2 homogenized data with the additional NASA/GISS "GISTEMP" UHI correction

Finally, we apply the GISTEMP urban heat island adjustment (described *Hansen et al.*, 2010) to the version 2.0 series to see how it addresses any remaining urban-related signal from the homogenized monthly temperature records. The GISTEMP UHI correction adjusts the trend of stations classified as urban or peri-urban to match the trend of a distance-weighted composite record made from nearby rural stations. An urban station is adjusted only if there are at least three nearby rural stations with values that overlap at least two-thirds of the urban station's period of record. Periods and urban stations that fail the rural station requirement are excluded from the GISS analysis. Rural stations are ideally selected to be within 500 km of the urban station, but in some cases could be as far as 1000 km away to meet the selection requirement. Note that in performing this adjustment only rural stations from USHCN have been used. This contrasts with the usual GISTEMP analysis which will use any suitable rural stations in GHCN, possibly including stations not in USHCN (such as in Canada and Mexico). Given the spatial density of stations in USHCN we expect any differences in adjustment to be minimal.

The scheme for identifying stations as urban has changed in the history of the GISTEMP analysis (see *Hansen et al*., 1999; *Hansen et al*., 2001; *Hansen et al*., 2010); here we use nighttime radiances from the DMSP calibrated radiance product described earlier. The analysis was carried out using the ccc-gistemp software supplied by the Climate Code Foundation [*Barnes and Jones*, 2011]. The resulting version 2.0 series with the GISTEMP UHI correction should be essentially the same as the USHCN data used in NASA's GISTemp product, albeit with a slightly more up-to-date dataset used for determining nighttime brightness and separate

application of the Step 2 (UHI correction) process to average monthly minimum and maximum data rather than applying it to the mean monthly data only.

This analysis described above produces estimates of urban-rural differences for each month from 1895-2010 for mean monthly minimum and maximum temperatures for the TOB, v2, v2+Step 2, and v2-rural neigh/v2-urban neigh variants for each of the four urbanity proxy via both station pairing and spatial gridding methods, resulting in 64 different distinct urban-rural differences for each month.

RESULTS

Unhomogenized (TOB-Adjusted) Data

Figure 1, which summarizes the urban minus rural (urban-rural) trend differences for all data set versions, indicates that the USHCN unhomogenized (TOB-only adjusted) data contains significant urban warming signals ($p < 0.05$ for linear trend fit) over the period from 1895 to present in both the minimum and maximum temperatures according to nearly all urban classification and comparison methods (the exception being GRUMP and Nightlights maximum temperatures evaluated via spatial gridding).

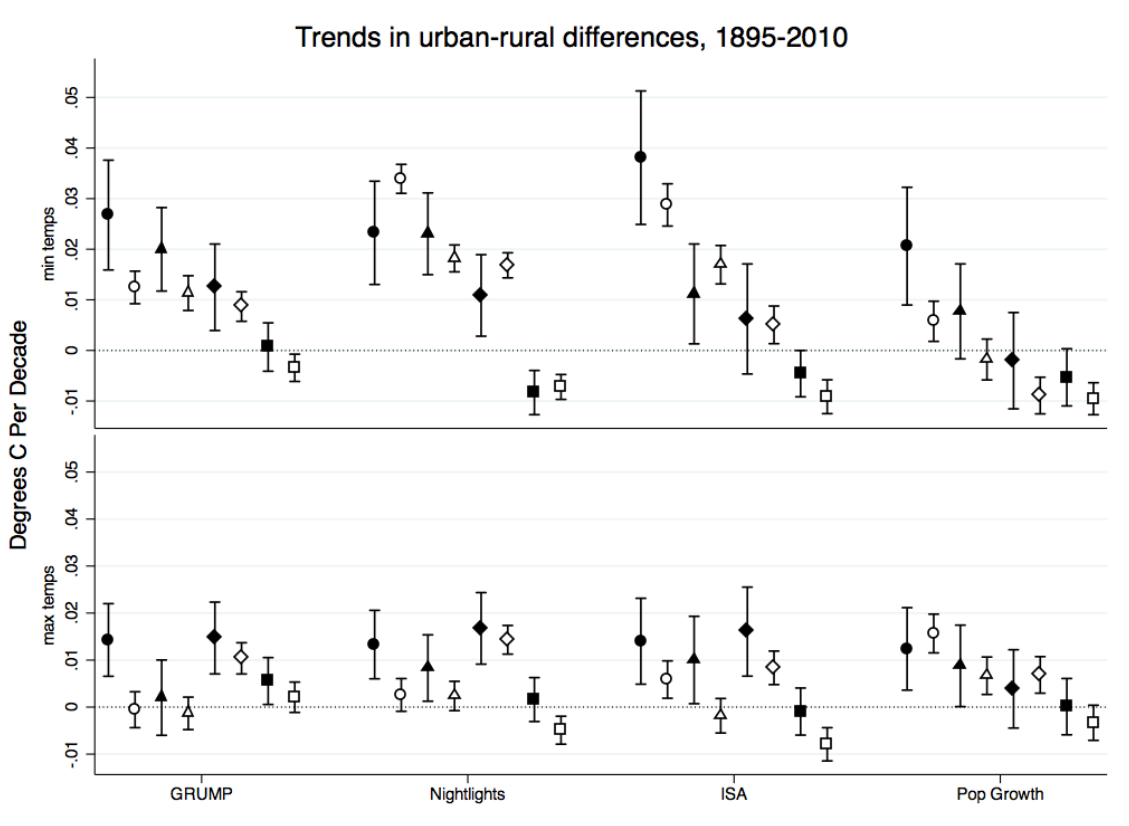


Figure 1: *1895-2010 trends and 95% confidence intervals in urban-rural differences by proxy type. Circles represent TOB adjusted data, Triangles represent version 2.0 data adjusted using rural neighbors only (v2-rural neigh), Diamonds represent version 2.0 homogenized data (v2), and Squares represent version 2.0 homogenized data with additional corrections using GISS's Step 2 method (Step 2). Solid shapes show results from the station pair method, and hollow shapes show results from the spatial gridding method.*

As expected, the urban signal is larger in minimum temperatures than in maximum temperatures. Urban-rural difference trends in minimum temperature range between 0.05 and 0.5 °C per century in minimum temperatures for the 1895-2010 period for the unhomogenized data depending on the classification and comparison method (e.g. station pairing or spatial gridding).

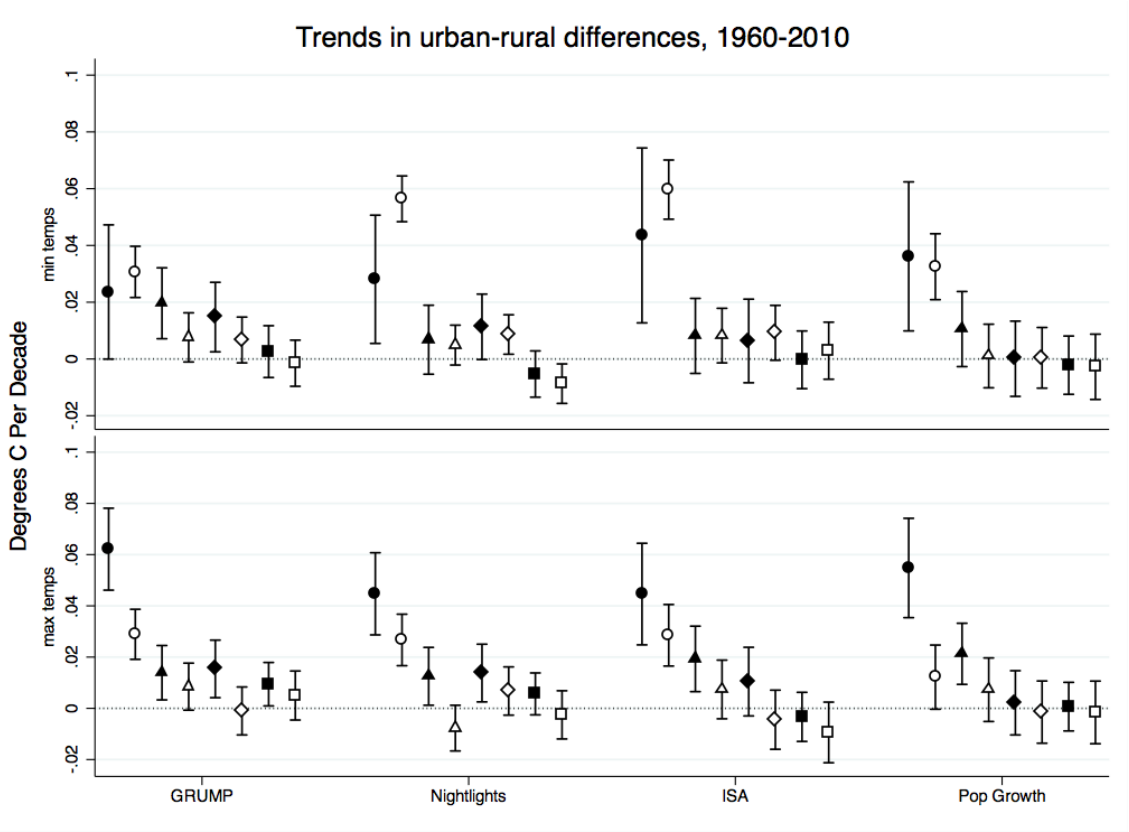


Figure 2: *1960-2010 trends and 95% confidence intervals in urban-rural differences by proxy type. Circles represent TOB adjusted data, Triangles represent version 2.0 data adjusted using rural neighbors only (v2-rural neigh), Diamonds represent version 2.0 homogenized data (v2), and Squares represent version 2.0 homogenized data with additional corrections using GISS's Step 2 method (Step 2). Solid shapes show results from the station pair method, and hollow shapes show results from the spatial gridding method.*

As shown in Figure 2, there is also evidence of a significant urban signal in the unhomogenized data during the past 50 years, with urban-rural difference trends of between 0.2 and 0.6 °C per century across all urbanity proxies for the period 1960-2010. This large urban warming signal does not appear to be a result of any correlation between instrument changes and urban form because it occurs with a similar magnitude in both the station pairing method (which controls for instrument type) and the spatial gridding method (which does not).

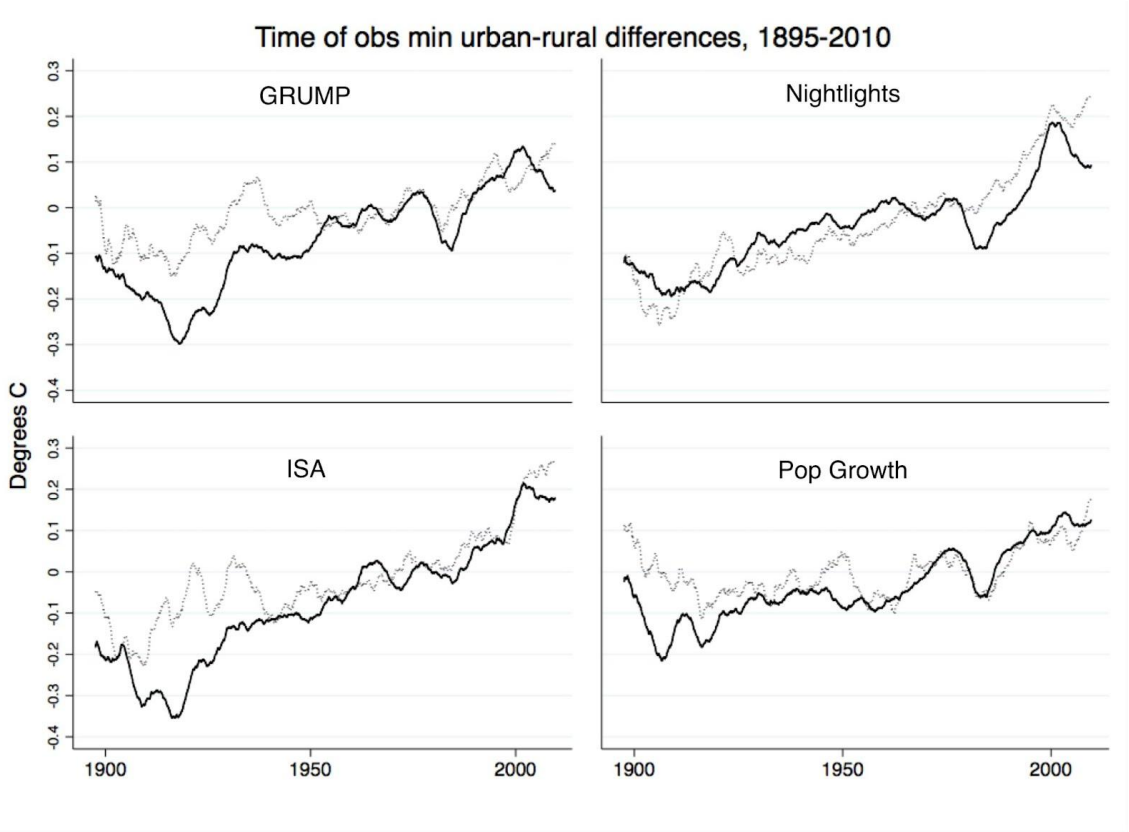


Figure 3: *Running 5-year mean of urban and rural differences for time of observation-adjusted min USHCN station data from 1895 to 2010, using both station-pair (solid line) and spatial gridding (dashed line) methods for GRUMP, Nightlight, ISA (10%), and Population Growth urbanity proxies.*

For minimum temperatures, the urban warming signal over both century and half-century timeframes is larger in the more restrictive urban classification—ISA—that contains relatively few urban stations, and are smaller in the classifications—GRUMP, Nightlights, and Population Growth—that contain a more even split between urban and rural designations. The station pairing method often shows significantly larger urban warming than the spatial gridding method; however, the pairing method does not account for the potential biases related to the spatial distribution of the station pairs. As shown in Figure 3, the divergences between station pairing and spatial gridding methods are particularly pronounced prior to 1950, which may be indicative of a larger geographic bias to the station pairs during that period. On the other hand, both methods produce similar results for periods after 1950.

As supplementary Figure SI.1 shows, the rural-urban differences are even larger in the raw minimum temperatures than in the TOB-adjusted data especially for the period since 1950 when

time-of-observation changes were prevalent. However, as mentioned above, this difference is not likely driven by any physical phenomena related to UHI. Rather it likely reflects a higher frequency of time of observation changes at non-urban stations.

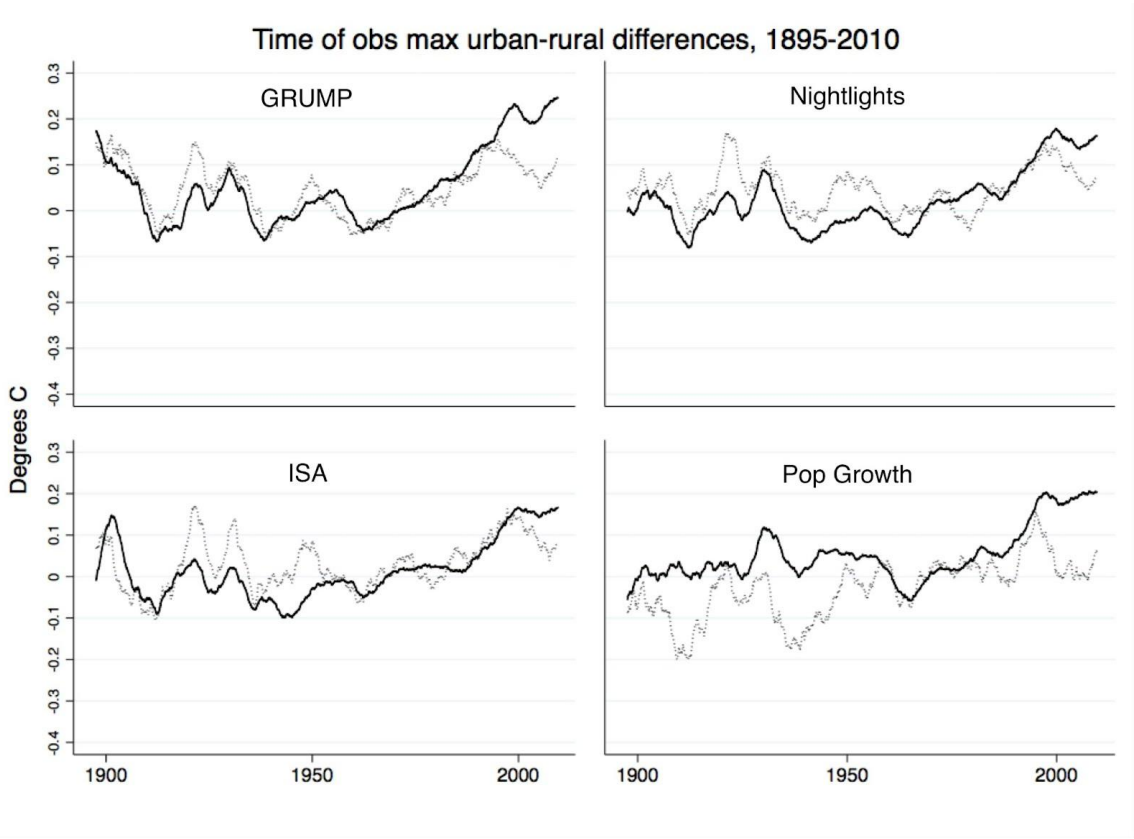


Figure 4: *Running 5-year mean of urban and rural differences for time of observation-adjusted max USHCN station data from 1895 to 2010, using both station-pair (solid line) and spatial gridding (dashed line) methods for GRUMP, Nightlight, ISA (10%), and Population Growth urbanity proxies.*

Maximum temperature urban-minus-rural trends in the unhomogenized (TOB) data are also significantly larger than zero over the period 1895 to 2010 for most urban classifications, but are smaller than the trends in minimum temperature urban-rural differences. They also show considerably less variation across urbanity proxy, with urban warming trends of around 0.08 to 0.22 °C per century for the station pairing method and -0.04 to 0.2 °C per century for the spatial gridding method. However, maximum temperature urban-rural difference trends are larger over the period 1960 to 2010, particularly in the GRUMP and Population Growth proxies where they exceed minimum urban minus rural trends. In this case, there is also a greater divergence between analysis methods, with the station pairing method showing much larger urban warming

than the spatial gridding method, which again, likely reflects a spatial bias caused by the non-uniform distribution of station pairs.

By comparing the trends of rural stations to those of all USHCN stations, we can use the spatial gridding method to get an estimate of the extent to which overall CONUS minimum temperature trends over the past century may have been driven by the urban warming signal (see Table SI.1). By this estimate, the unhomogenized minimum temperature data from rural USHCN stations yields trends that are between 14 and 21 percent smaller on average over the period 1895-2010 period than the trends from the full USHCN network. This difference decreases to between about 6 and 9 percent during the last 50 years.

Homogenized Version 2 Data (v2)

The pairwise homogenization algorithm (PHA) significantly reduces the difference between urban and rural minimum temperature trends according to all analysis methods and station classifications. This is particularly true over the 1960-2010 period, where a the vast majority of the urbanity proxies and methods indicate no significant urban warming in the minimum data. Maximum temperatures are a bit more mixed, though most proxies and methods show no significant urban warming in the maximum data over the period. As shown in Figure 5, there is still a small but significant minimum urban warming prior to 1960 in all urbanity proxies except for Population Growth. The station pairing method suggests some residual urban signal before 1960, but this residual signal is small in the spatial gridding method for all proxies after 1930.

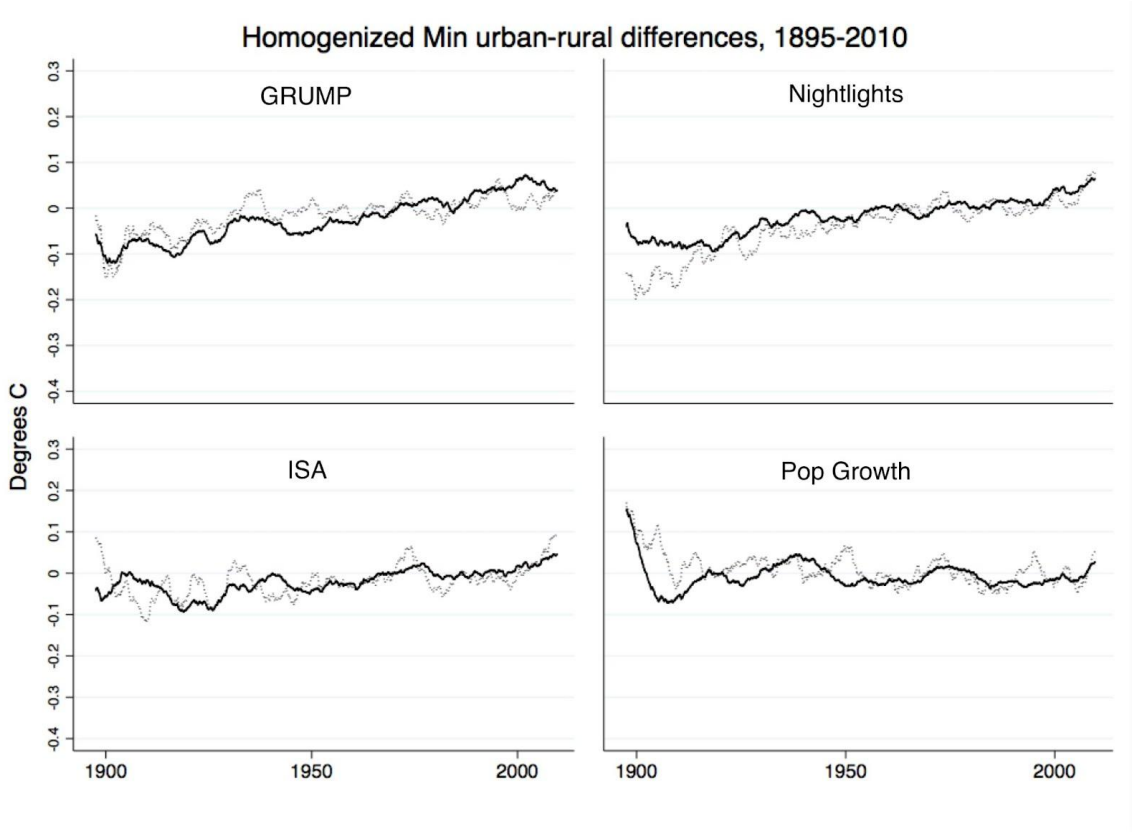


Figure 5: *Running 5-year mean of urban and rural differences for v2 homogenized minimum temperature USHCN station data from 1895 to 2010, using both station-pair (solid line) and spatial gridding (dashed line) methods for GRUMP, Nightlight, ISA (10%), and Population Growth urbanity proxies.*

The effect of homogenization is most pronounced in the more restrictive urbanity proxies like ISA that contain relatively few urban stations and show larger urban warming trends prior to homogenization. The divergences between urban and rural temperatures that remain prior to 1930 even after homogenization are likely in part due to the combination or poorer metadata for that time period and fewer coop station records that can be used as neighbors.

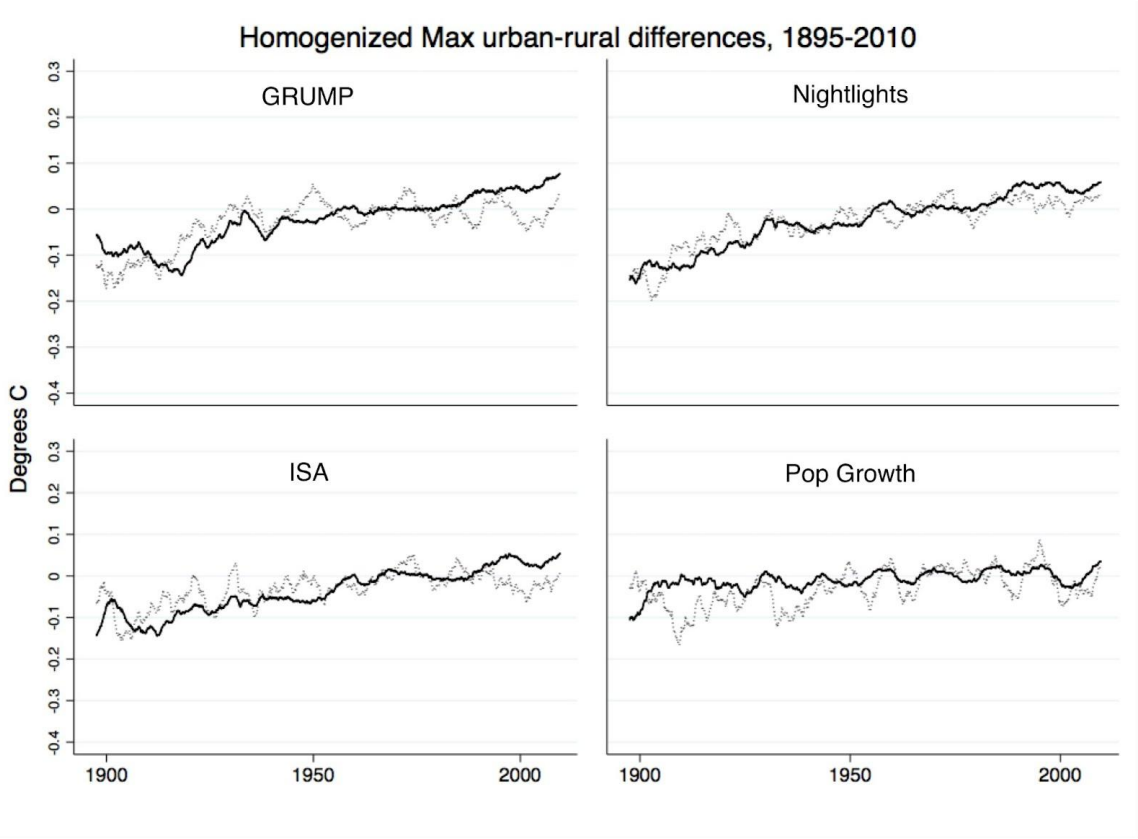


Figure 6: *Running 5-year mean of urban and rural differences for v2 homogenized max USHCN station data from 1895 to 2010, using both station-pair (solid line) and spatial gridding (dashed line) methods for GRUMP, Nightlight, ISA (10%), and Population Growth urbanity proxies.*

Notably, urban-rural differences in maximum temperatures over the century timeframe are in most cases not significantly reduced by homogenization, as shown in Figure 6.

Comparing homogenized rural HCN stations to all HCN stations, we find that rural stations show between 3 and 13 percent less average temperature (tave) warming over the 1895-2010 period, and a slight but not significantly different from zero reduction in warming over the 1960-2010 period (see Table SI.1). Thus, residual urban signals not removed by data homogenization appear to be significant only for the period prior to 1960 and effectively only prior to about 1930. In summary, pairwise homogenization effectively removes the urban signal present in minimum temperature data from the last 50 to 80 years, and reduces it by around 50% or more for the period prior to 1930 (as can be seen when comparing Figure 3 and Figure 5).

Homogenized version 2 data with added GISTEMP correction (v2+step 2)

As reported in *Hansen et al.* [2010], applying the GISTEMP Step 2 UHI correction to the USHCN version 2 data has the impact of reducing the mean CONUS temperature trend from 0.73°C to 0.65°C over the period 1900-2009. As shown in Fig. SI.1, this reduction appears to result almost entirely from trend adjustments in the data for years prior to 1930. After 1930, the version 2.0 (52d) and version 2.5 (52i) data are not significantly impacted by the Step 2 adjustment. Moreover, this trend reduction is required only because of an urban signal in the early minimum temperature data, which get reduced by about 0.0113°C/decade by the Step 2 adjustment. The impact on maximum temperature is only 0.00288°C/decade. The average of these impacts is equivalent to the impact reported by *Hansen et al.* [2010]. As shown in Figures 7 and 8, the GISS Step 2 adjustment is effectively removing the residual urban signal in both minimum and maximum temperatures across all proxies without any significant over adjustment, even for the most restrictive definitions of urbanity.

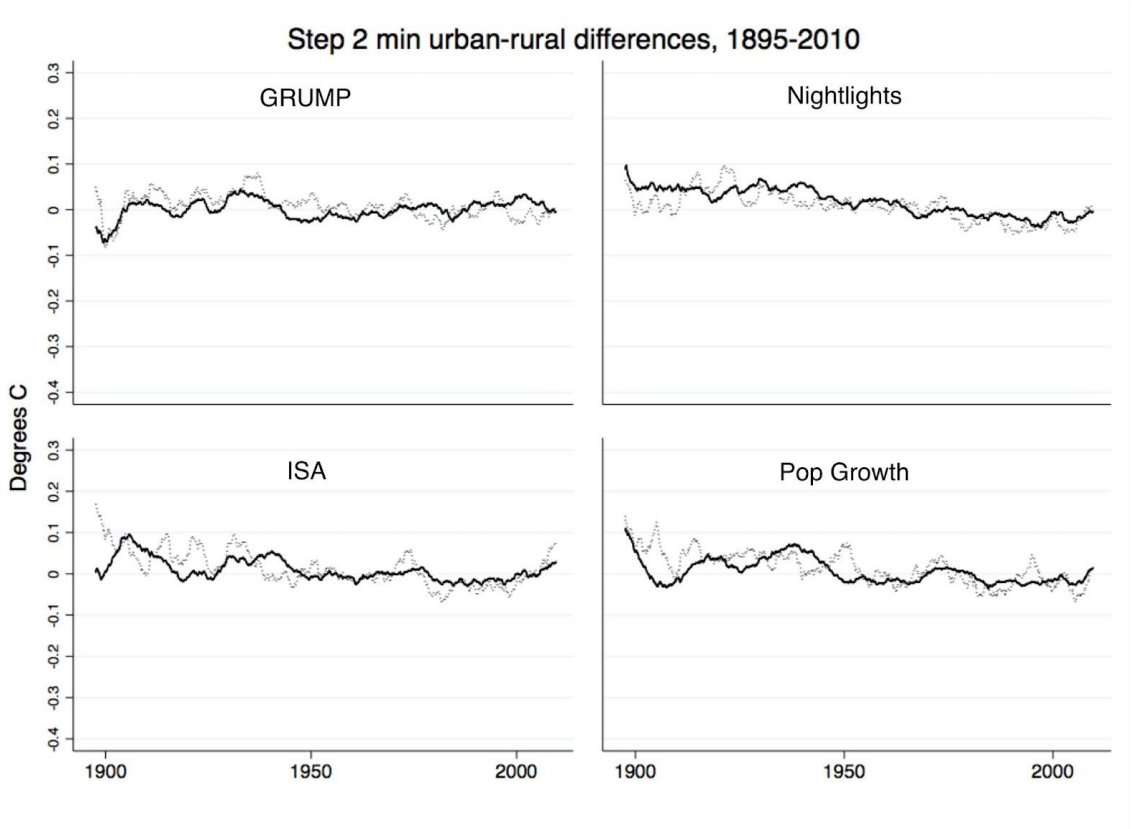


Figure 7: *Running 5-year mean of urban and rural differences for Step 2 min USHCN station data from 1895 to 2010, using both station-pair (solid line) and spatial gridding (dashed line) methods for GRUMP, Nightlight, ISA (10%), and Population Growth urbanity proxies.*

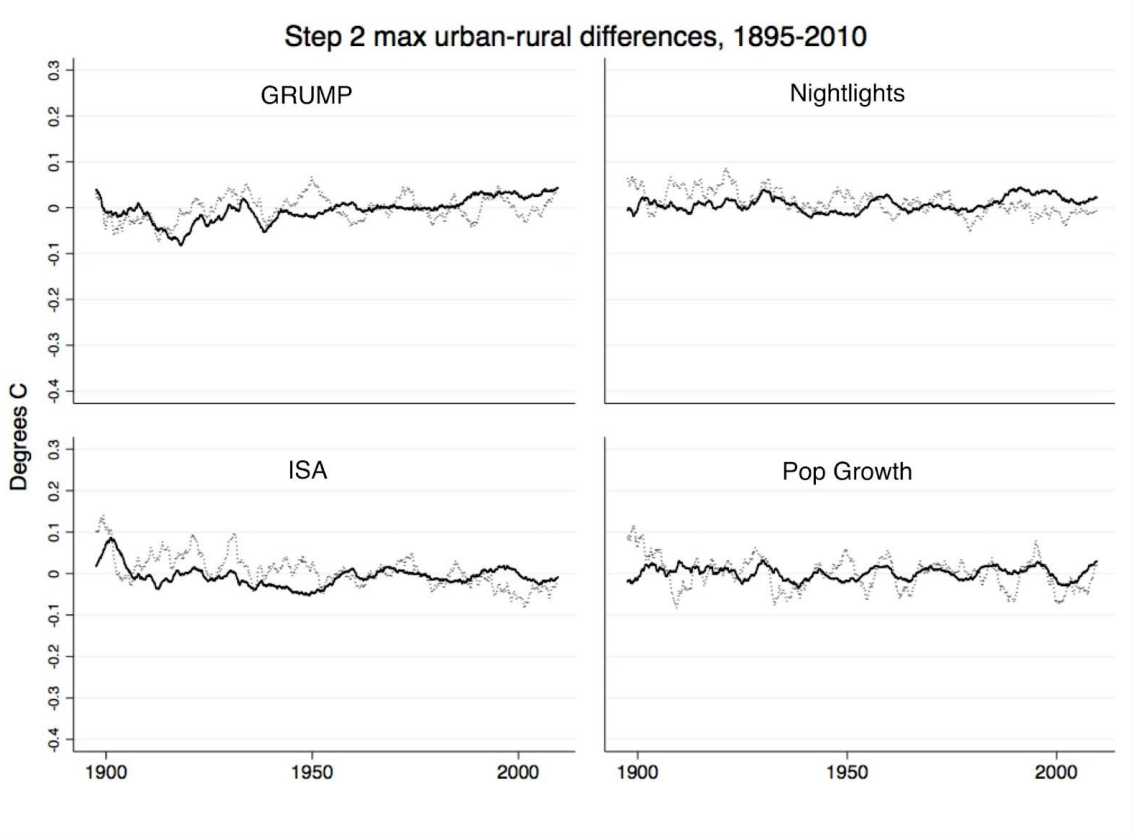


Figure 8: *Running 5-year mean of urban and rural differences for Step 2 max USHCN station data from 1895 to 2010, using both station-pair (solid line) and spatial gridding (dashed line) methods for GRUMP, Nightlight, ISA (10%), and Population Growth urbanity proxies.*

Homogenized version 2 data using only Coop neighbors classified as rural (v2-rural neigh)

In all of the urbanity proxies and analysis methods, the differences between urban and rural station minimum temperature trends are smaller in the homogenized data than in the unhomogenized data, which suggests that homogenization can remove much and perhaps nearly all (since 1930) of the urban signal without requiring a specific UHI correction. However, the trends in rural station minimum temperatures are slightly higher in the homogenized minimum temperature data than in the TOB-only adjusted data. One possible reason for this is that the PHA is appropriately removing inhomogenities caused by station moves or other changes to rural stations that have had a net negative impact on the CONUS average bias (e.g., many stations now classified as rural were less rural in the past since they moved from city

centers to airports or waste water treatment plants). Another possibility is that homogenization is causing nearby UHI-affected stations to "correct" some rural station series in a way that transfers some of the urban warming bias to the temperature records from rural stations. In such a case, a comparison of the homogenized data between rural and urban stations would then show a decreased difference between the two by removing the *appearance* of an urbanization bias without actually removing the bias itself.

To help determine the relative merits of these two explanations, the PHA was run separately allowing only rural- and only urban-classified Coop stations to be used as neighbors in calculating the PHA corrections for USHCN stations. In Figure 9, the spatially averaged U.S minimum temperature anomalies for rural stations are shown for the four different datasets: the unhomogenized (TOB-adjusted only); the version 2 (all-Coop-adjusted; v2) data; the homogenized dataset adjusted using only coop stations classified as rural; and the homogenized dataset adjusted using only urban coop stations.

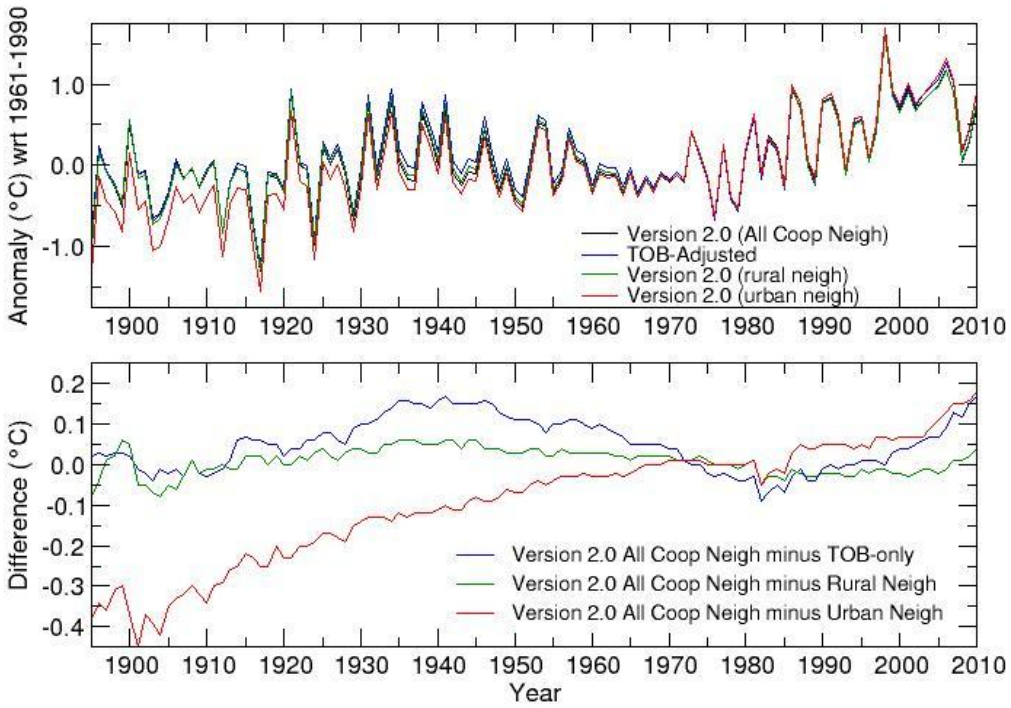


Figure 9: *Comparison of spatially gridded minimum temperatures for the TOB-only adjusted USHCN data, v2 USHCN data (homogenized using all Coop station series as reference series), USHCN data homogenized using series from Coop stations only classified as rural according to the impervious surface method, and USHCN data homogenized using series from Coop stations only classified as urban (according to the impervious surface method). Top Panel: CONUS average anomalies for the four versions of the USHCN data. Bottom Panel: the differences between the USHCN v2 data homogenized with all Coop station series and: data adjusted only for the tob-bias (blue); data homogenized using only rural station series (green); and, data homogenized using only urban station series (red).*

The large difference in the trends between the urban-only adjusted and the rural-only adjusted datasets suggests that when urban coop station series are used exclusively as reference series for the USHCN some of their urban-related biases can be transferred to USHCN station series during homogenization. However, the fact that the homogenized all-coop-adjusted minimum temperatures are much closer to the rural-station-only adjustments than the urban-only adjustments suggests that the bleeding effect from the ISA classified urban stations is likely small in the USHCN version 2 dataset. This is presumably because there are a sufficient number of rural stations available for use as reference neighbors in the Coop network to allow for the identification and removal of UHI-related impacts on the USHCN temperature series. Furthermore, as the ISA classification shows the largest urban-rural difference in the TOB data, it is likely that greater differences between rural-station-only-adjusted and all-coop-adjusted series using stricter rural definitions result from fewer identified breakpoints due to less network coverage, and not UHI-related aliasing. Nevertheless, it is instructive to further examine the rural-only and urban-only adjustments to assess the consequences of using these two subsets of stations as neighbors in the PHA.

Figure S.I.2 shows the cumulative impact of the adjustments using the rural-only and urban-only stations as neighbors to the USHCN. In this example, the impermeable surface extent was used to classify the stations. The cumulative impacts are shown separately for adjustments that are common between the two runs (i.e., adjustments that the PHA identified for the same stations and dates) versus those that are unique to the two separate urban-only and rural-only reference series runs. In the case of both the common and unique adjustments, the urban-only neighbor PHA run produces adjustments that are systematically larger (more positive) than the rural-only neighbor run. The magnitude of the resultant systematic bias for the adjustments common to both algorithm versions is shown in black. The reason for the systematic

differences is likely that UHI trends or undetected positive step changes pervasive in the urban-only set of neighboring station series are being aliased onto the estimates of the necessary adjustments at USHCN stations. This aliasing from undetected urban biases becomes much more likely when all or most neighbors are characterized by such systematic errors.

Figure S.I.3. provides a similar comparison of the rural-only neighbor PHA run and the all-Coop (v2) neighbor run. In this case, the adjustments that are common to both the rural-only and the all-Coop neighbor runs have cumulative impacts that are nearly identical. This is evidence that, in most cases, the Coop neighbors that surround USHCN stations are sufficiently "rural" to prevent a transfer of undetected urban bias from the neighbors to the USHCN station series during the homogenization procedure. In the case of the adjustments that are unique to the separate runs, the cumulative impacts suggest that the less dense rural-only neighbors are missing some of the negative biases that occurred during the 1930 to 1950 period, which highlights the disadvantage of using a less dense station network. In fact, the all-Coop neighbor v2 dataset has about 30% more adjustments than the rural-only neighbor PHA run produces. Results using the other three station classification approaches are similar and are provided as Figures S.I.3 – S.I.8.

CONCLUSIONS

According to all four proxy measures used to identify station environments that are currently urban, there is consistent evidence that urban stations have a systematic bias relative to rural stations throughout the USHCN period of record. This bias has led to an apparent urban warming signal in the unhomogenized data that accounts for approximately 14 to 21 percent of total rise in USHCN minimum temperatures averaged over the CONUS for the period since 1895, and 6 to 9 percent of the rise over the past 50 years. Homogenization of the monthly temperature data via NCDC's Pairwise Homogenization Algorithm (PHA) removes the majority of this apparent urban bias, especially over the last 50 to 80 years. Moreover, results from the PHA using the full set of Coop station series as reference series and using only those series from stations currently classified as rural are broadly consistent, which provides strong evidence that the reduction of the urban warming signal by homogenization is a consequence of the real elimination of an urban warming bias present in the raw data rather than a consequence of simply forcing agreement between urban and rural station trends through a spreading of the urban signal to series from nearby stations.

As noted in the introduction, one of the challenges in quantifying the UHI signal in land surface air temperature records is that changes affecting urban stations can occur at both the micro and meso-scales. Changes at the micro-scale (e.g., small station moves, growth of a tree) are not necessarily of interest in evaluations of the UHI signal because they are highly localized and may have no relevance to the broader land use changes associated with urbanization that can affect the mesoscale temperature signal. For this reason, micro-scale changes reasonably can be included in the list of inhomogeneities that should be corrected for via homogenization (along with instrument changes and time of observation changes). In contrast, it may be desirable to preserve changes in the in the meso-scale signal because these changes encompass a broader footprint and are arguably more likely to be related to larger-scale land use changes. Unfortunately, it may not be possible to distinguish (at least automatically) changes occurring at the micro-scale from changes at the meso-scale, especially if only one station record is available to sample the meso-scale signal. Whatever the cause, when any station series exhibits a sustained change relative to highly correlated surrounding stations, the change is likely to be identified by the PHA as uniquely local, and its impact on that stations temperature trend will be removed with a bias adjustment. This happens whether the USHCN station is from a rural or urban environment, which means that the same challenge that exists for identifying UHI impacts also exists for identifying the impacts of other types of (non-urban) land use changes.

Nevertheless, the pairing of urban and rural stations in a manner that controls for instrument type and time of observation changes reveals larger trends at urban stations, which is consistent with the understanding that land use changes associated with urbanization lead to larger historic temperature trends at urban stations. However, that this larger trend signal is effectively removed through homogenization suggests that the urban environments characterized by larger trends do not have large spatial scales that allow them to be sampled by a number of Coop stations (or that the urban temperature signal is heterogeneous) and thus the local urban signal is being effectively removed via homogenization.

Because homogenization is largely successful in removing urban bias in the USHCN temperature data, it appears that only about 5% of the period-of-record USHCN version 2 minimum temperature trends across the CONUS can be attributed to local urban influences and, further, that most of this contribution is coming from data for years prior to 1930. This residual urban bias for the earlier years in the record may be a consequence of the reduced station density of the Coop network in the early part of the twentieth century, which limits the

number of pairs available for detecting inhomogenities some of which may be related to urbanization.

NASA GISS's (GISTEMP) "Step 2" nightlight-based UHI adjustments effectively remove the remaining urban-rural differences during this early period, suggesting that the additional UHI-specific adjustment is achieving the goal of forcing agreement between urban and rural temperature trends. Nevertheless, the recently released USHCN version v2.5 data (homogenized with the PHA algorithm version "52i as shown in figure S.I.1) improves the pre-1930 period considerably vis-à-vis v2.0 (except in the case of GRUMP), which may also mean that homogenization procedures may be able to more fully account for urban-related biases in the future, at least in areas with sufficient station density. In any case, at present, the net effect of urban-correlated biases on the version 2.5 adjusted data is evidently small, accounting for less than 5% of the trend since 1895 (and between 0 and 2% since 1960). While it would likely be worthwhile to further characterize the uncertainty in UHI-related warming in datasets like the USHCN (e.g., by exploring a range of cutoffs for classifying a station as urban with the various proxies or by quantifying more site-specific aspects of a stations environment), UHI does not appear to represent a significant contributing factor in the CONUS-average temperature signal over the past 50-80 years.

ACKNOWLEDGMENTS

The authors thank Peter Thorne, Russ Vose, Jay Lawrimore, Tom Peterson and three anonymous reviewers for helpful comments on this manuscript. The authors also thank Steven Mosher for useful feedback and help in obtaining Impermeable Surface Area values for U.S. stations.

REFERENCES

Arnfield A.J. (2003), Two decades of urban climate research: a review of turbulence, exchanges of energy and water, and the urban heat island. *Int J Climatol.*, 23:1–26. DOI:10.1002/joc.859.

Arnold, C.L. and Gibbons, C.J. (1996), Impervious surface coverage. *American Planning Association Journal*, *62,* 243-258.

Balk, D., F. Pozzi, , G. Yetman, , U. Deichmann, and A. Nelson (2004), The distribution of people and the dimension of place: methodologies to improve the global estimation of urban extents. Working Paper, CIESIN, Columbia University. Palisades, NY, pp 31.

Barnes, N., and D. Jones (2011), Clear Climate Code: Rewriting legacy science code for clarity, *IEEE*, 28, 36-42.

Baugh, K., C. Elvidge, T. Ghosh, and D. Ziskin (2010), Development of a 2009 Stable Lights Product using DMSP-OLS data, Proceedings of the 30th Asia-Pacific Advanced Network Meeting, 114-130.

Boehm, R., (1998), Urban bias in temperature series—A case study for the city of Vienna, *Climate Change*, 38, 113–128.

Center for International Earth Science Information Network-CIESIN, Columbia University; International Food Policy Research Institute-IFPRI; The World Bank; and Centro Internacional de Agricultura Tropical-CIAT (2004), Global Rural-Urban Mapping Project (GRUMP), Alpha Version. Palisades, NY: Socioeconomic Data and Applications Center (SEDAC), Columbia University. Available at http://sedac.ciesin.columbia.edu/gpw.

DeGaetano, A.T. (2006), Attributes of several methods for detecting discontinuities in mean temperature series. *J. Climate,* **19,** 838–853.

Elvidge, C.D.,B.T. Tuttle, ,P.C. Sutton, , K.E. Baugh, A.T. Howard, C. Milesi, B.L. Bhaduri, R. Nemani (2007), Global Distribution and Density of Constructed Impervious Surfaces, *Sensors*, 7, 1962-1979.

Gallo, K.P., T.W. Owen, D.R. Easterling, and P.F. Jamason (1999), Temperature trends of the U.S. Historical Climatology Network based on satellite designated land use/land cover*, J. Climate*, 12, 1344-1348, doi: 10.1175/1520-0442(1999)012<1344:TTOTUS>2.0.CO;2.

Gallo, K.P. and T.W. Owen (2002), A sampling strategy for satellite sensor-based assessments of the urban heat-island bias*, Int. J. Remote Sensing*, 23, 1935-1939, doi: 10.1080/01431160110097259.

Karl, T.R, C. N. Williams Jr., P. J. Young, and W. M. Wendland (1986), A model to estimate the time of observation bias associated with monthly mean maximum, minimum, and mean temperature for the United States. *J. Climate Appl. Meteor.,* **25,** 145–160.

Karl, T.R., H.F. Diaz, and G. Kukla (1988), Urbanization: its detection and effect in the United States climate record, *J. Climate*, 1, 1099-1123.

Kukla, G., J. Gavin, and T.R. Karl, (1986), Urban Warming, *J. Clim. Applied. Meteor.*, 25, 1265-1270.

Hansen, J., R. Ruedy, J. Glascoe, and M. Sato (1999), GISS analysis of surface temperature change. *J. Geophys. Res.*, **104**, 30997-31022.

Hansen, J.E., R. Ruedy, M. Sato, M. Imhoff, W. Lawrence, D. Easterling, T. Peterson, and T. Karl (2001), A closer look at United States and global surface temperature change. *J. Geophys. Res.*, **106**, 23947-23963.

Hansen, J.E., R. Ruedy, M. Sato, and K. Lo (2010), Global Surface Temperature Change, *Rev. Geophys.*, 48, RG4004, doi:10.1029/2010RG000345.

Imhoff, M. L., W. T. Lawrence, D. C. Stutzer, and C. D. Elvidge (1997), A technique for using composite DMSP‐OLS “city lights” satellite data to map urban area, *Remote Sens. Environ.*, 61, 361–370, doi:10.1016/S0034-4257(97)00046-1.

Jones, P.D., D.H. Lister, and Q. Li (2008), Urbanization effects in large-scale temperature records, with an emphasis on China, J. Geophys. Res., 113, D16122, DOI:10.1029/2008JD009916.

Jones, P.D. and D.H. Lister (2010), The Urban Heat Island in Central London and urban-related warming trends in Central London since 1900. *Weather*, 64, 323-327, DOI: 10.1002/wea.432.

Lawrimore, J.H., M.J. Menne, B.E. Gleason, C.N. Williams, D.B. Wuertz, R.S. Vose, and J. Rennie, 2011: An overview of the Global Historical Climatology Network Monthly Mean Temperature Dataset, Version 3. *Journal of Geophysical Research-Atmospheres*, 116, D19121, doi:10.1029/2011JD016187.

McCarthy, M.P., M. J. Best, and R. A. Betts (2010), Climate change in cities due to global warming and urban effects. *Geophys. Res. Lett.*, 37, L09705, doi:10.1029/2010GL042845, 2010.

Menne, M.J., and C.N. Williams (2009), Homogenization of temperature series via pairwise comparisons, *Journal of Climate*, 22, 1700-1717.

Menne, M.J., C.N. Williams Jr., and R.S. Vose (2009), The United States Historical Climatology Network monthly temperature data–Version 2. *Bulletin of the American Meteorological Society*, 90, 993-1007.

Menne, M. J., C. N. Williams, Jr., and M. A. Palecki (2010), On the reliability of the U.S. surface temperature record. *Journal of Geophysical Research*, 115, D11108, doi:10.1029/2009JD013094.

Mitchell, J. M. (1953), On the causes of instrumentally observed secular temperature trends*, J. Meteorol.*, 10, 244–261.

National Geophysical Data Center/NESDIS/NOAA (2002), Land Use Change Project Data: 2000 U.S. Census Population Grids (http://dmsp.ngdc.noaa.gov/html/download.html).

National Climatic Data Center/NOAA (2012), Historical Observing Metadata Repository (http://www.ncdc.noaa.gov/land-based-station-data/station-metadata)

Oke, T.R., (1982), The energetic basis of the urban heat island, *Quarterly Journal of the Royal Meteorological Society*, 108, 1–24.

Parker, D.E., (2006): A demonstration that large-scale warming is not urban. *J. Clim.*, 19, 4179-4197, DOI:10.1175/JCLI3730.1.

Parker, D. E. (2010), Urban heat island effects on estimates of observed climate change, Wiley Interdiscip. Rev. Clim. Change, 1, 123–133, doi:10.1002/wcc.21.

Peterson, T.C., K.P. Gallo, J.H. Lawrimore, A. Huang, D.A. McKittrick (1999), Global rural temperature trends. Geophys. Res. Lett., 26, 329-332, DOI: 10.1029/1998GL900322.

Peterson, T.C. (2003), Assessment of Urban Versus Rural In Situ Surface Temperatures in the Contiguous United Stats: No Difference Found. *J. Climate*, 16:18, 2941-2959.

Peterson, T. C., and T.W. Owen (2005), Urban Heat Island Assessment: Metadata Are Important. J. Climate, 18:14, 2637-2646.

Potere, D., A. Schneider, S. Angel, and D.L. Civco, D.L (2009), Mapping urban areas on a global scale. *International Journal of Remote Sensing,* 30(24),6531-6558.

Quayle, R. G., D. R. Easterling, T. R. Karl, and P. Y. Hughes (1991), Effects of recent thermometer changes in the Cooperative Station Network, *Bull. Amer. Meteorol. Soc.*, 72, 1718–1723, doi:10.1175/1520-0477(1991)072<1718:EORTCI>2.0.CO;2.

Quinlan, F. T., T. R. Karl, and C. N. Williams Jr. (1987), United States Historical Climatology Network (HCN) serial temperature and precipitation data. NDP-019, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, TN.

Schueler, T.R., (1994), The importance of imperviousness. *Watershed Protection Techniques*, *1(3),* 100-111.

Schaal, L. A., and R. F. Dale (1977), Time of observation bias and "climatic change," *J. Appl. Meteorol.*, 16, 215–222, doi:10.1175/1520-0450(1977)016<0215:TOOTBA>2.0.CO;2.

Spronken-Smith, R. A., and T. R. Oke (1998), The thermal regime of urban parks in two cities with different summer climates. *Int. J. Remote Sens.,* **19,** 2085–2104.

U.S. Census Bureau, 2002: Population of States and Counties of the United States: 1790-2000, Washington, D.C., 226 pp.

Venema, V. K. C., et al. (2012), Benchmarking monthly homogenization algorithms, *Clim. Past Discuss.*, 7, 2655–2718.

Vose, R.S., C.N. Williams, T.C. Peterson, T.R. Karl, and D.R. Easterling, 2003: An evaluation of the time of observation bias adjustment in the US Historical Climatology Network. *Geophysical Research Letters*, 30 (20), 2046, doi:10.1029/2003GL018111.

Vose, R.S., D. Arndt, V. F. Banzon; D.R. Easterling; B.E. Gleason; B. Huang; E. Kearns; J.H. Lawrimore; M.J. Menne; T.C. Peterson; R.W. Reynolds; T.M. Smith; C.N. Williams; D.L. Wuertz, 2012: NOAA's Merged Land-Ocean Surface Temperature Analysis. *Bulletin American Meteorological Society*, in press.

Williams, C.N., M.J. Menne, and P.W. Thorne (2012a), Benchmarking the performance of pairwise homogenization of surface temperatures in the United States. *J. Geophys. Res.*, 117, D5, doi:10.1029/2011JD016761.

Williams, C.N, M.J. Menne and J.H Lawrimore (2012b), Modifications to Pairwise Homogeneity Adjustment software to address coding errors and improve run-time efficiency. NOAA

Technical Report NCDC No. GHCNM-12-02 (available at http://www1.ncdc.noaa.gov/pub/data/ghcn/v3/techreports/Technical%20Report%20NCDC%20No12-02-3.2.0-29Aug12.pdf).

SUPPLEMENTARY INFORMATION

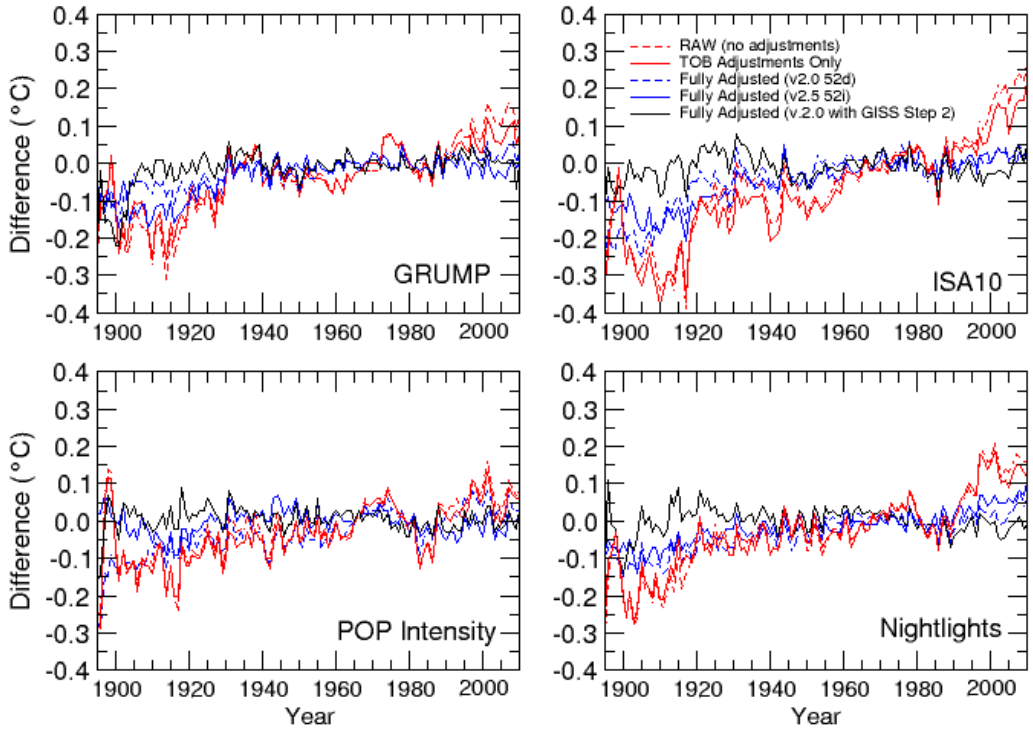


Figure S.I. 1. *CONUS average urban and rural minimum temperature differences for five different versions of the USHCN station data and four different station classifications. The five different dataset versions are "Raw" (no bias adjustments-dashed red); TOB-only adjusted (solid red); v2 fully adjusted data homogenized using the pairwise algorithm version "52d" (dashed blue); fully adjusted v2 data homogenized using the pairwise algorithm version "52i" (solid blue); and fully adjusted v2 data homogenized using algorithm version "52d" with the NASA GISS*

(GISTEMP) "Step 2" UHI correction. CONUS averages were computed as described in Menne et al (2009, 2010).

Table SI.1. *Minimum and maximum trends in CONUS temperatures for specified sets of stations using the Menne et al [2009] spatial gridding method for time of observation-adjusted (TOB), homogenized v2.0 (52d), and homogenized v2.5 (52i) series.*

Stations	Dates	Series	TMIN Trend	TMAX Trend
All Stations	1895-2010	TOB	0.074	0.028
All Stations	1895-2010	52d	0.075	0.061
All Stations	1895-2010	52i	0.070	0.056
GRUMP Rural	1895-2010	TOB	0.060	0.027
GRUMP Rural	1895-2010	52d	0.068	0.053
GRUMP Rural	1895-2010	52i	0.068	0.056
ISA Rural	1895-2010	TOB	0.064	0.026
ISA Rural	1895-2010	52d	0.072	0.060
ISA Rural	1895-2010	52i	0.070	0.060
Nightlight Rural	1895-2010	TOB	0.062	0.025
Nightlight Rural	1895-2010	52d	0.069	0.056
Nightlight Rural	1895-2010	52i	0.068	0.056
Pop Growth Rural	1895-2010	TOB	0.064	0.025
Pop Growth Rural	1895-2010	52d	0.076	0.058
Pop Growth Rural	1895-2010	52i	0.071	0.059

All Stations	1960-2010	TOB	0.255	0.127
All Stations	1960-2010	52d	0.248	0.189
All Stations	1960-2010	52i	0.236	0.196
GRUMP Rural	1960-2010	TOB	0.234	0.106
GRUMP Rural	1960-2010	52d	0.242	0.184
GRUMP Rural	1960-2010	52i	0.237	0.190
ISA Rural	1960-2010	TOB	0.240	0.119
ISA Rural	1960-2010	52d	0.247	0.193
ISA Rural	1960-2010	52i	0.234	0.198
Nightlight Rural	1960-2010	TOB	0.233	0.113
Nightlight Rural	1960-2010	52d	0.244	0.185
Nightlight Rural	1960-2010	52i	0.234	0.194
Pop Growth Rural	1960-2010	TOB	0.236	0.120
Pop Growth Rural	1960-2010	52d	0.248	0.190
Pop Growth Rural	1960-2010	52i	0.236	0.193

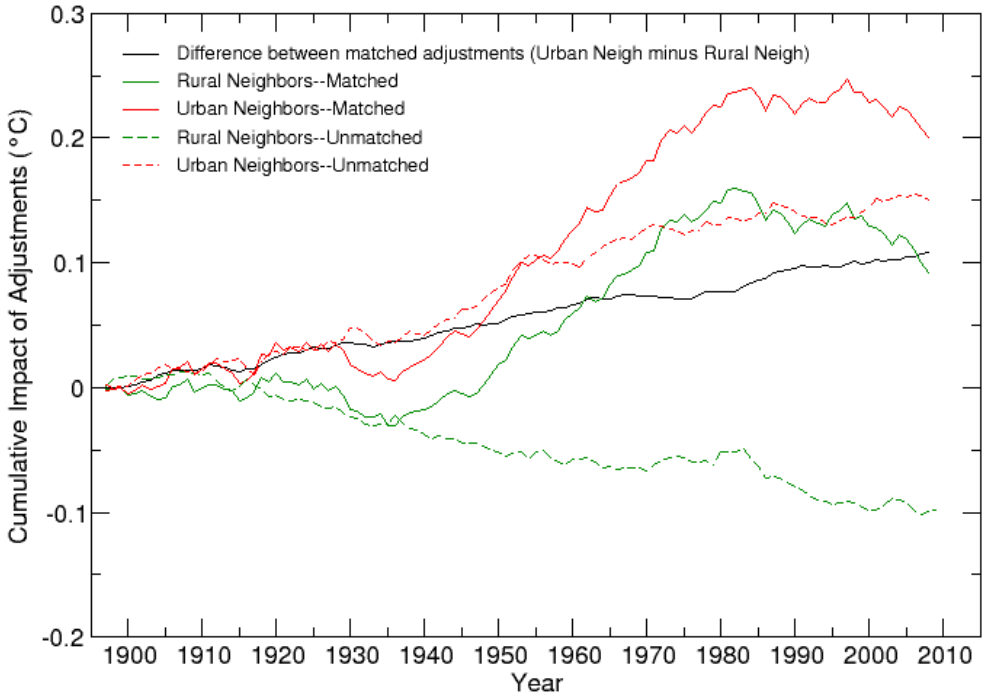


Figure S.I. 2. *Cumulative average of PHA-derived minimum temperature adjustments using Coop station reference series classified as urban only (red lines) and as rural only (green lines) according to the impermeable surface area (ISA10) classification method. The cumulative average of the adjustments that are common to both datasets are shown as solid lines and those that are unique are shown as dashed lines.*

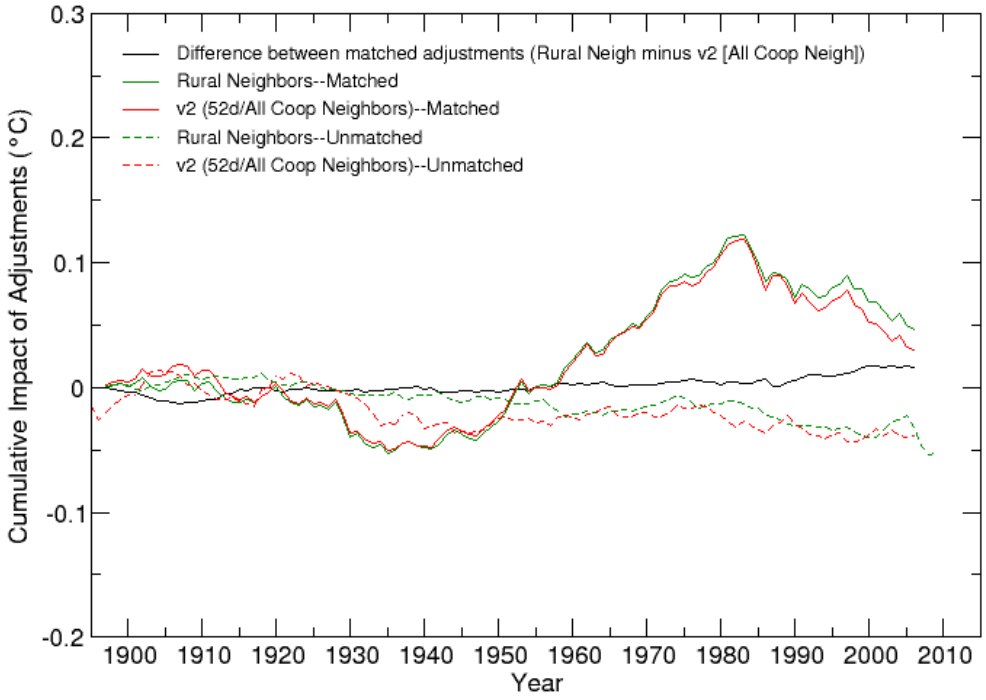


Figure S.I. 3. *Cumulative average of PHA-derived minimum temperature adjustments using all Coop station series as reference series-v2-"52d" (red lines) and classified as rural only (green lines) according to the impermeable surface classification method. The cumulative average minimum temperature adjustments that are common to both datasets are shown as solid lines and those that are unique are shown as dashed lines.*

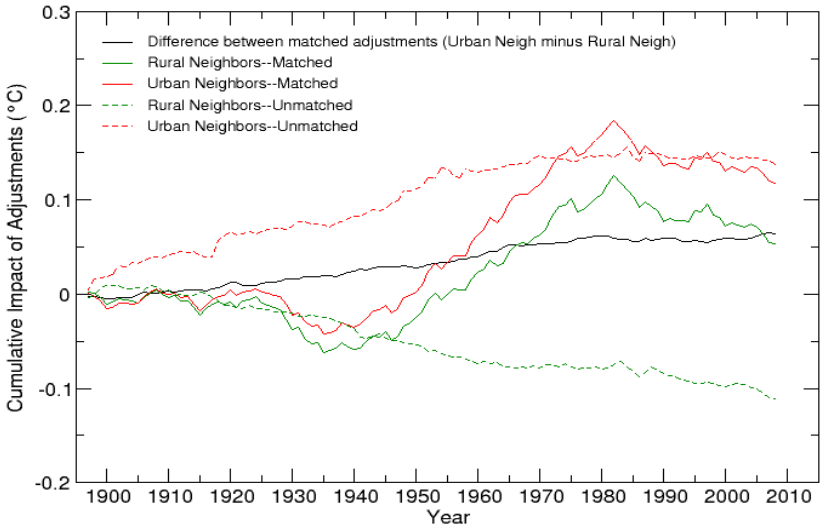


Figure S.I. 4. *As in Fig. S.I.2 but from stations classified using GRUMP.*

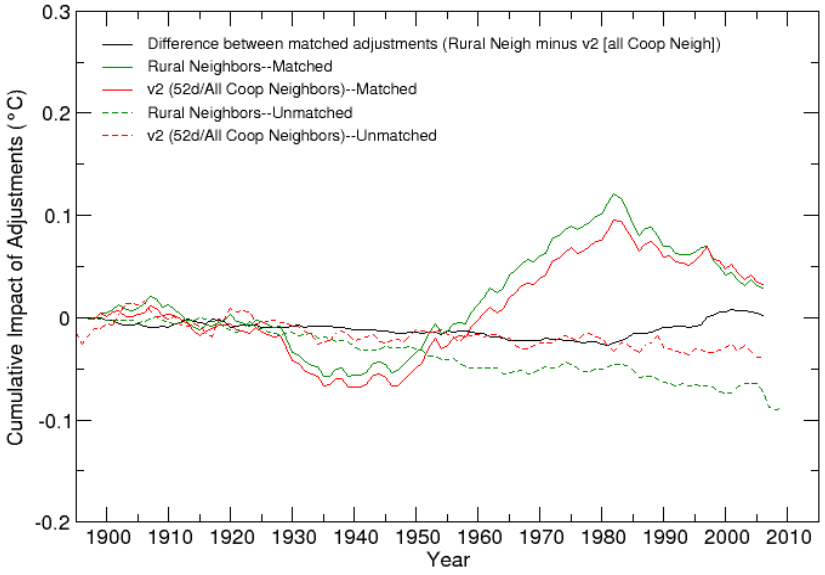


Figure S.I. 5. *As in Fig. S.I.3 but from stations classified using GRUMP.*

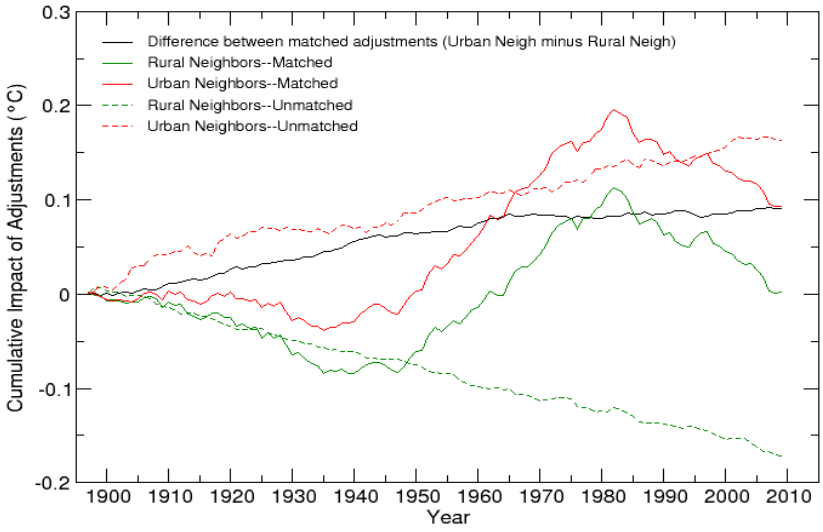


Figure S.I. 6. *As in Fig. S.I.2 but from stations classified using Nightlights.*

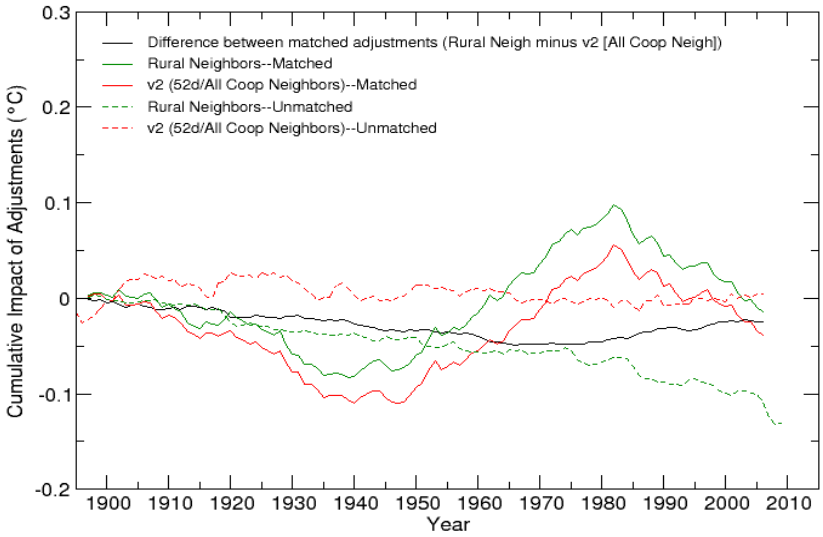


Figure S.I. 7. *As in Fig. S.I.3 but from stations classified using Nightlights.*

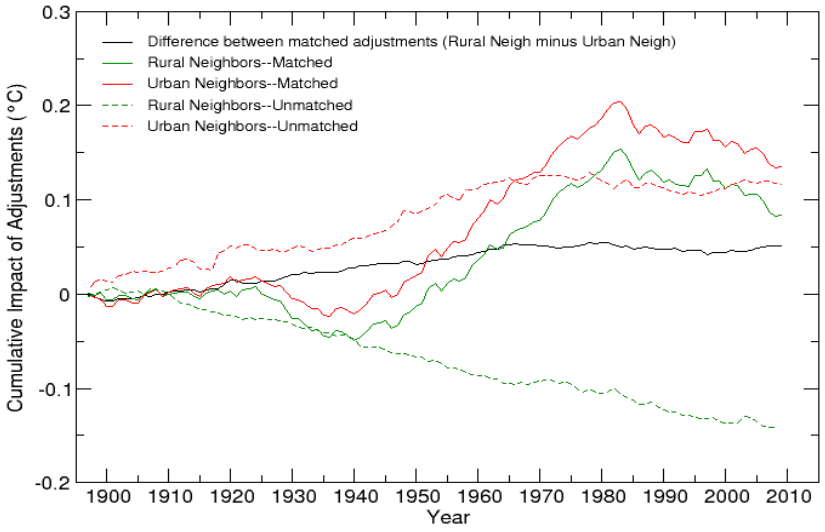


Figure S.I. 8. *As in Fig. S.I.2 but from stations classified using population growth.*

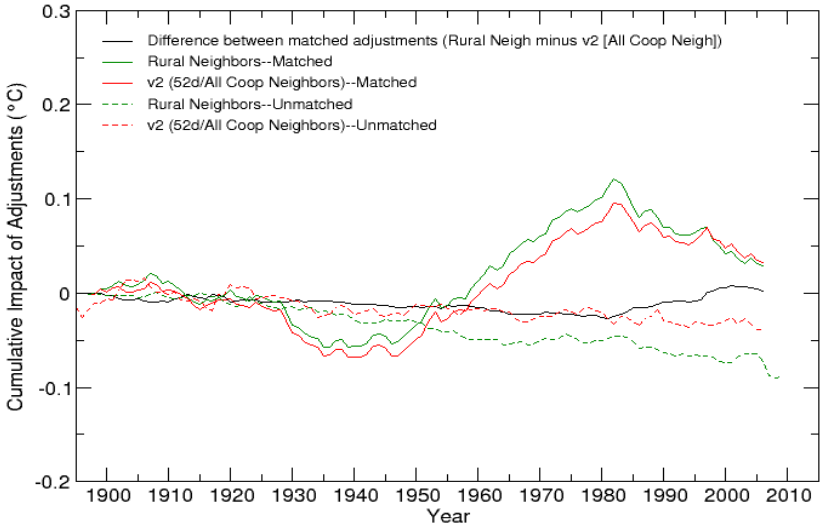


Figure S.I. 9. *As in Fig. S.I.3 but from stations classified using population growth.*

DATA SOURCES

Urbanity Proxies

For the satellite nightlights, the Global Radiance Calibrated Nighttime Lights data-set year 2006, satellite F16 was used: http://www.ngdc.noaa.gov/dmsp/download_radcal.html

http://www.google.com/url?q=http%3A%2F%2Fwww.ngdc.noaa.gov%2Fdmsp%2Fdownload_radcal.html&sa=D&sntz=1&usg=AFQjCNEErK-ouoX_mn-v7gQJJuqlmTH8dA

For population growth and low population proxies, Gridded 1 km Population Estimates for the Conterminous U.S., 1930-2000 were used: http://www.ncdc.noaa.gov/oa/climate/research/population/popdata.html

http://www.google.com/url?q=http%3A%2F%2Fwww.ncdc.noaa.gov%2Foa%2Fclimate%2Fresearch%2Fpopulation%2Fpopdata.html&sa=D&sntz=1&usg=AFQjCNEWRo1mmvTpvi278n0ZXRV2tjhlig

For impermeable surfaces, Global Distribution and Density of Constructed Impervious Surfaces was used: http://www.ngdc.noaa.gov/dmsp/download_global_isa.html

For urban boundaries, Global Rural-Urban Mapping Project (GRUMP) data was used: http://sedac.ciesin.columbia.edu/gpw/documentation.jsp

Station and Temperature Data

USHCN TOB min data:

ftp://ftp.ncdc.noaa.gov/pub/data/ushcn/v2/monthly/9641C_201105_tob.min.gz

ftp://ftp.ncdc.noaa.gov/pub/data/ushcn/v2/monthly/9641C_201011_F52.min.gz

USHCN TOB max data:

ftp://ftp.ncdc.noaa.gov/pub/data/ushcn/v2/monthly/9641C_201105_tob.max.gz

ftp://ftp.ncdc.noaa.gov/pub/data/ushcn/v2/monthly/9641C_201011_F52.min.gz

USHCN v2 homogenized min data:
ftp://ftp.ncdc.noaa.gov/pub/data/ushcn/v2/monthly/9641C_201105_F52.min.gz

USHCN v2 homogenized max data:
ftp://ftp.ncdc.noaa.gov/pub/data/ushcn/v2/monthly/9641C_201105_F52.max.gz

USHCN v2 rural-neighbor homogenized min data:
ftp://ftp.ncdc.noaa.gov/pub/data/ushcn/v2/monthly/uhi/

USHCN v2 rural-neighbor homogenized max data:

ftp://ftp.ncdc.noaa.gov/pub/data/ushcn/v2/monthly/uhi/

Station instrument types:

ftp://ftp.ncdc.noaa.gov/pub/data/ushcn/v2/monthly/uhi/

MMTS transition dates:

ftp://ftp.ncdc.noaa.gov/pub/data/ushcn/v2/monthly/uhi/

SOURCE CODE

Code used for both the station pairing and spatial gridding analysis is available for the statistical software STATA here: ftp://ftp.ncdc.noaa.gov/pub/data/ushcn/v2/monthly/uhi/code/stata

A Java version of the code is also available:
ftp://ftp.ncdc.noaa.gov/pub/data/ushcn/v2/monthly/uhi/code/java

Code used in the gridding method described in *Menne et al.* [2009; 2010] is available here:

Code used to produce both fully homogenized and rural-homogenized data via the Pairwise Homogenization Algorithm in *Menne and Williams* [2009] is available here:
ftp://ftp.ncdc.noaa.gov/pub/data/ushcn/v2/monthly/uhi/code/pha

Code used to apply the NASA GISS Step 2 adjustment is available here:
http://code.google.com/p/ccc-gistemp/

Numerical values for trends and confidence intervals for Figures 1 and 2 are available here: ftp://ftp.ncdc.noaa.gov/pub/data/ushcn/v2/monthly/uhi/trends

5. PAPER 2: EVALUATING THE IMPACT OF U.S. HISTORICAL CLIMATOLOGY NETWORK HOMOGENIZATION USING THE U.S. CLIMATE REFERENCE NETWORK

Zeke Hausfather[1,2], Kevin Cowtan[3], Matthew J. Menne[4], and Claude N. Williams, Jr[4].

[1]Energy and Resources Group, University of California Berkeley, Berkeley CA 94720.
[2]Berkeley Earth, Berkeley CA 94720.
[3]Department of Chemistry, University of York, Heslington, York, YO10 5DD, United Kingdom.
[4]National Centers for Environmental Information (NCEI), NOAA, Asheville NC.

Geophysical Research Letters 43 (4), 1695-1701, 2016.

ABSTRACT

Numerous inhomogeneities including station moves, instrument changes, and time of observation changes in the U.S. Historical Climatological Network (USHCN) complicate the assessment of long-term temperature trends. Detection and correction of inhomogeneities in raw temperature records have been undertaken by NOAA and other groups using automated pairwise neighbor-comparison approaches, but these have proven controversial due to the large trend impact of homogenization in the United States. The new U.S. Climate Reference Network (USCRN) provides a homogenous set of surface temperature observations that can serve as an effective empirical test of adjustments to raw USHCN stations. By comparing nearby pairs of USHCN and USCRN stations, we find that adjustments make both trends and monthly anomalies from USHCN stations much more similar to those of neighboring USCRN stations for the period from 2004-2015 when the networks overlap. These results improve our confidence in the reliability of homogenized surface temperature records.

INTRODUCTION

The U.S. Historical Climatological Network (USHCN) is a group of 1,218 stations selected from the larger U.S. Cooperative Observer Program to provide a spatially-representative estimate of contiguous U.S. temperatures (CONUS) from 1895 through the present [Fiebrich 2009]. These

stations were selected based on long, continuous temperature records, rural or small-town locations, and other factors intended to produce as unbiased an estimate as possible of long-term climate changes [Quinlan et al., 1987; Menne et al., 2009]. Despite these selection criteria, significant systemic inhomogeneities plague the USHCN. These include time of observation changes [Karl et al., 1986; Vose et al., 2003], instrument changes [Quayle et al., 1991; Doesken, 2005; Hubbard and Lin, 2006], station location changes [Changnon and Kunkel 2006], changes in broader urban form surrounding station locations [Karl et al., 1988; Peterson and Owen, 2005; Hausfather et al., 2013], and changes in localized station site characteristics [Fall et al., 2011; Menne et al., 2010; Muller et al., 2013]. Most stations in the USHCN have been subject to three or more of these inhomogeneities during the past century, and few if any have completely homogenous records [Menne et al., 2009]. These inhomogeneities can have large non-symmetric effects on estimates of U.S. temperature trends. The two largest trend effects are due to correcting time-of-observation changes and instrument changes from liquid-in-glass (LiG) to minimum-maximum temperature systems (MMTS).

Time-of-observation changes introduced a large cooling bias due to widespread observation time changes from afternoon to morning between 1950 and present. This results in a shift from minimum-maximum thermometers occasional double-counting of maximums to a double-counting of minimums, with a net U.S. average negative bias of about 0.25 °C [Vose et al., 2003]. The widespread transition from LiG to MMTS instruments between 1980 and 2000 also resulted in a cooling bias; MMTS instruments tend to measure maximum temperatures about 0.5 °C lower and minimum temperatures about 0.35 °C higher than LiG instruments, resulting a net negative trend bias of around 0.15 °C [Hubbard and Lin, 2006].

The raw USHCN temperature records are adjusted (homogenized) to attempt to remove biases introduced by these inhomogeneities. Two distinct adjustments are performed on USHCN data: a correction for time of observation [Karl et al., 1986], and a Pairwise Homogenization Algorithm (PHA) to detect and remove all other biases [Menne and Williams, 2009]. The adjustments to USHCN records have been evaluated extensively using synthetic data [Williams et al., 2012; Venema et al., 2012], and they generally perform well in removing both regional and local biases independent of the sign of the bias. Adjusted USHCN trends are also quite similar to results from independent reanalysis datasets, while raw USHCN trends are significantly lower [Vose et al., 2012]. Other independent groups have also found similar results to NOAA using differing automated adjustment approaches [Rohde et al 2013]. However, the net effect of

adjustments on the USHCN is quite large, effectively doubling the mean temperature trend over the past century compared to the raw observational data [Menne et al 2009]. This has resulted in a controversy in the public and political realm over the implications of much of the observed U.S. warming apparently resulting from temperature adjustments.

In part as a response to criticisms of the quality of the USHCN, NOAA began setting up a U.S. Climate Reference Network (USCRN) in 2001. The USCRN stations are sited in pristine environments in rural areas away from any potential direct urban influence. Stations include three NIST-calibrated redundant temperature sensors that make measurements every 2 seconds and automatically report the data to a centralized server via satellite uplink. Stations are actively monitored and regularly maintained by NOAA employees. The USCRN is currently comprised of 114 conterminous U.S. stations and has had sufficient station density and distribution to provide relatively good spatial coverage of the U.S. since the start of 2004 [Diamond et al., 2013].

The period of overlap between the records is now sufficiently long to effectively assess the impact of temperature adjustments to USHCN stations using the USCRN as an unbiased reference. The USCRN has been used to evaluate other observational networks before; for example, Otkin et al. [2005] used the USCRN to validate insolation estimates, Gallo [2005] examined proximate USCRN station pairs to assess the impact of microclimate influences, and Leeper [2015] examined absolute temperature and precipitation differences between proximate U.S. Cooperative Observer Program and USCRN stations.

METHODS

The USCRN record is homogeneous by design, while the USHCN has large known inhomogeneities. This means that an effective homogenization algorithm would tend to make the USHCN network trends and anomalies very similar to those of the USCRN network, and we can use this fact to empirically assess the effectiveness of homogenization during the period of overlap between the networks. To evaluate the efficacy of USHCN homogenization with respect to the USCRN, we focus on the period between January 2004 and August 2015 where both USHCN and USCRN networks have reasonably comprehensive spatial coverage of the U.S. We look at CONUS spatially-weighted average temperatures for USCRN and both USHCN raw and adjusted series. We also examine individual proximate pairs of USHCN and USCRN

stations. In all cases we separately perform the analysis for minimum (tmin), maximum (tmax), and average (tavg) monthly temperatures.

The adjusted (version 52j) USHCN series contains the same 1,218 temperature stations as the raw USHCN series, but uses the full set of around 10,000 temperature stations available in the U.S. for the detection and removal of inhomogeneities. Included in those 10,000 are the 114 USCRN stations, which raises the possibility that the adjusted USHCN data and USCRN data may not be completely independent. To ensure that the USCRN stations can provide an independent empirical test, we generated a variant of adjusted USHCN series that excluded all USCRN series from the full station population prior to any homogenization. This had relatively little effect for most stations, as the PHA requires the agreement of the preponderance of neighboring stations to flag inhomogeneities. A figure showing the difference between this new without-USCRN adjusted USHCN series and the standard with-USCRN USHCN adjusted series is available in the supplementary materials (Figure SM1).

To calculate CONUS temperature anomalies we follow a standard approach of assigning each station to a 2.5 by 3.5 latitude/longitude grid cell, transforming monthly values for each station into anomalies by subtracting the average for each month over a baseline period (in this case, 2004 through the end of 2014 to reflect the period of network overlap), average the anomalies from all stations within each gridcell, and creating a weighted average of all gridcells based on the respective land area of each grid cell (EPA 2013). We further exclude any gridcell-months prior to averaging that do not contain at least one USHCN-raw, USHCN-adjusted, and USCRN record to ensure that spatial coverage is comparable between the resulting records. Trend confidence intervals for the resulting CONUS records are calculated using an ARMA[1,1] model to account for autocorrelation in the data.

To evaluate proximate pairs of USHCN/USCRN stations, we examine all possible permutations of USHCN and USCRN station pairs that are within a given distance of each other. We examine distances of 50 miles (80 km), 100 miles (161 km), and 150 miles (241 km), though most of the figures presented herein focus on the 100-mile (161 km) case (the others are available in the supplementary materials). We further limit valid station pairs to those whose record begins prior to January 2006 and ends no earlier than July 2014, and exclude them from the analysis if they do not have at least 8 years (96 months) of data, ensuring all resulting station pair trends will be

calculated over a period of at least 8 years. These values are chosen in an attempt to maximize both the overlapping period and the number of station pairs available to evaluate.

These selection criteria result in 191 USHCN/USCRN station pairs (with 68 unique USCRN stations) at a distance cutoff of 50 miles (80 km), 651 station pairs (75 unique USCRN) at 100 miles (161 km), and 1393 station pairs (76 unique USCRN) at 150 miles (241 km). Distances are calculated via the spherical law of cosines formula. Each station pair record is trimmed to include only months where USCRN, USHCN-raw, and USHCN-adjusted readings are all available, to remove any impact of USHCN-adjusted in-filled values when USHCN-raw data is not available. Temperature readings for each station are converted into monthly temperature anomalies over the full period of overlap between the paired stations. A difference series is calculated by subtracting USCRN anomalies from USHCN anomalies for each month:

$$Diff_m = HCN_m - CRN_m$$

The trends in these pair difference series are calculated using a simple OLS regression. Mean squared differences between pair anomalies are also calculated to provide an additional metric of variation. The station pair difference time-series exhibit some residual autocorrelation (as verified by examining Durbin's alternative test for autocorrelation for station pair difference series), with more than half of the pair-differences having significant autocorrelation ($p < 0.05$) when differencing raw USHCN stations from their USCRN pair. However, because the measure of interest is the distribution of difference trends between all pairs pre- and post-adjustment rather than the uncertainty in difference trends for individual station pairs, the use of a simple OLS trend calculation rather than a more computationally-intensive approach that explicitly accounts for autocorrelation should have no meaningful effect on the results.

Additionally, we look at pairs of USCRN/USCRN and USHCN/USHCN stations to determine the variation of anomalies and trends as a function of distance within each network, similar to the approach taken in Gallo [2005]. The analysis undertaken for these in-network pairs is the same as for between-network pairs, though for distances up to 2,000 miles a random subset of 10,000 USHCN/USHCN station pairs are selected to make the calculations more tractable. Code used in performing these analyses is available in the supplementary materials.

RESULTS

Over the past 10 years there is relatively little difference between the raw and adjusted USHCN temperature series in the overall CONUS temperature record. The impact of adjustments over this period is largely trend-neutral due to a lack of detected systemic trend-biasing inhomogenities. Accordingly, at the CONUS-level the USCRN record does not allow for an effective differentiation between raw and adjusted USHCN series, as shown in Figure 1.

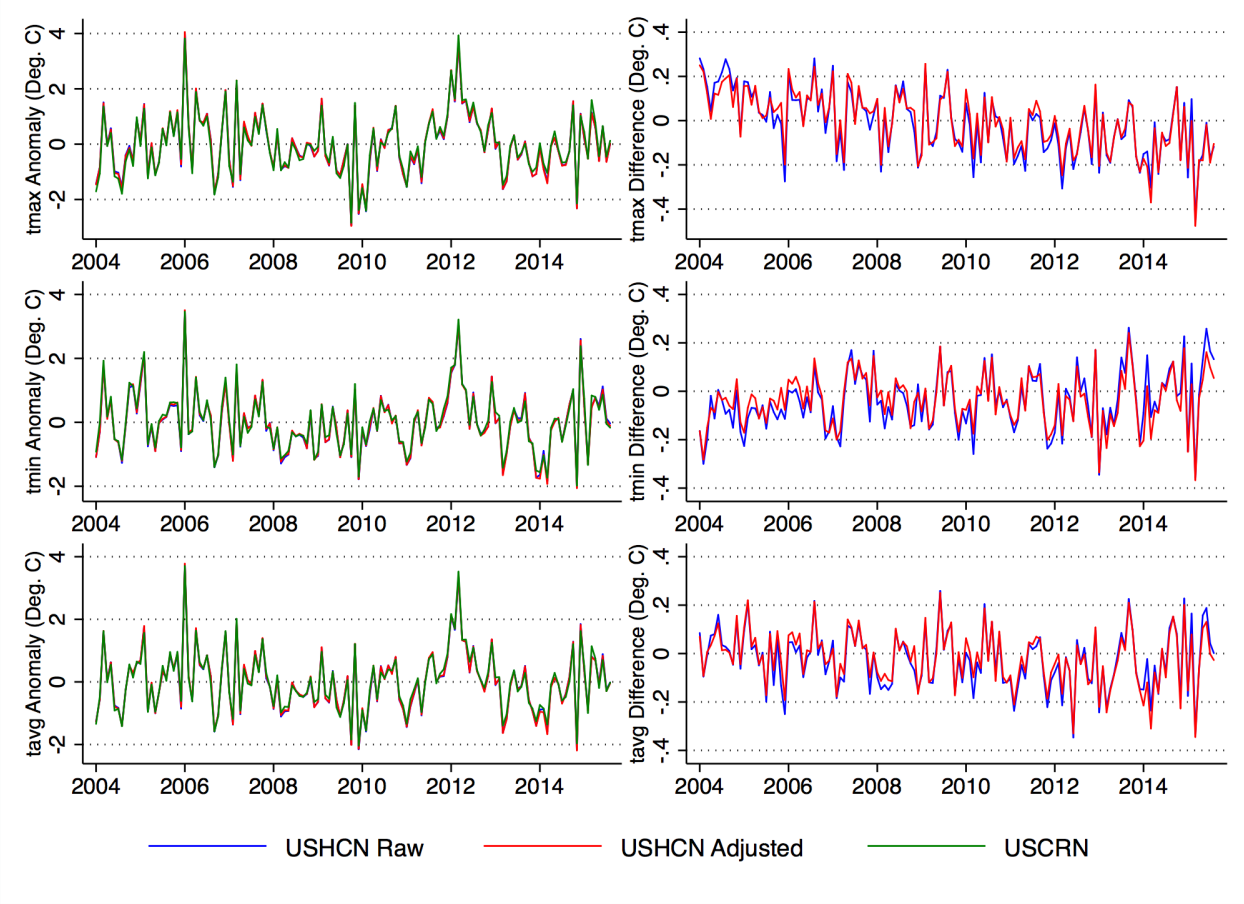


Figure 1: *Maximum (Tmax), minimum (Tmin), and mean (Tavg) CONUS values for USCRN, USHCN raw, and USHCN adjusted data. Left column: CONUS temperature anomalies for each series. Right column: USHCN raw minus USCRN (in blue) and USHCN adjusted minus USCRN (in red). CONUS reconstructions are spatially-limited to grid cells where values for all three datasets are present for any given month. For detailed statistics of the data shown, see supplementary materials Table SM1.*

The CONUS-averaged USCRN and both USHCN series are largely indistinguishable for both minimum and mean temperatures. These results are similar to those of Menne et al [2010] and

Diamond et al [2013], who also found little distinguishable differences between average USCRN and USHCN temperatures. However, significant differences ($p < 0.05$) exist for maximum temperatures, where both the raw and adjusted USHCN series have a lower temperature trend over the 2004-2014 period than the USCRN series.

While CONUS-averaged temperatures show little difference between USHCN adjusted, USHCN raw, and USCRN series, the same is not true when we look at individual pairs of proximate USCRN/USHCN stations within 100 miles (161 km) of each other (Figure 2). Here the effect of adjustments is to bring raw USHCN station trends much closer to their USCRN counterparts for maximum, minimum, and average temperatures. The effect of adjustments is particularly pronounced for more divergent trends. These results hold across pair-distances cutoffs of 50 and 150 miles (80 and 241 km) as well (see Figures SM5 and SM6 in the supplementary materials).

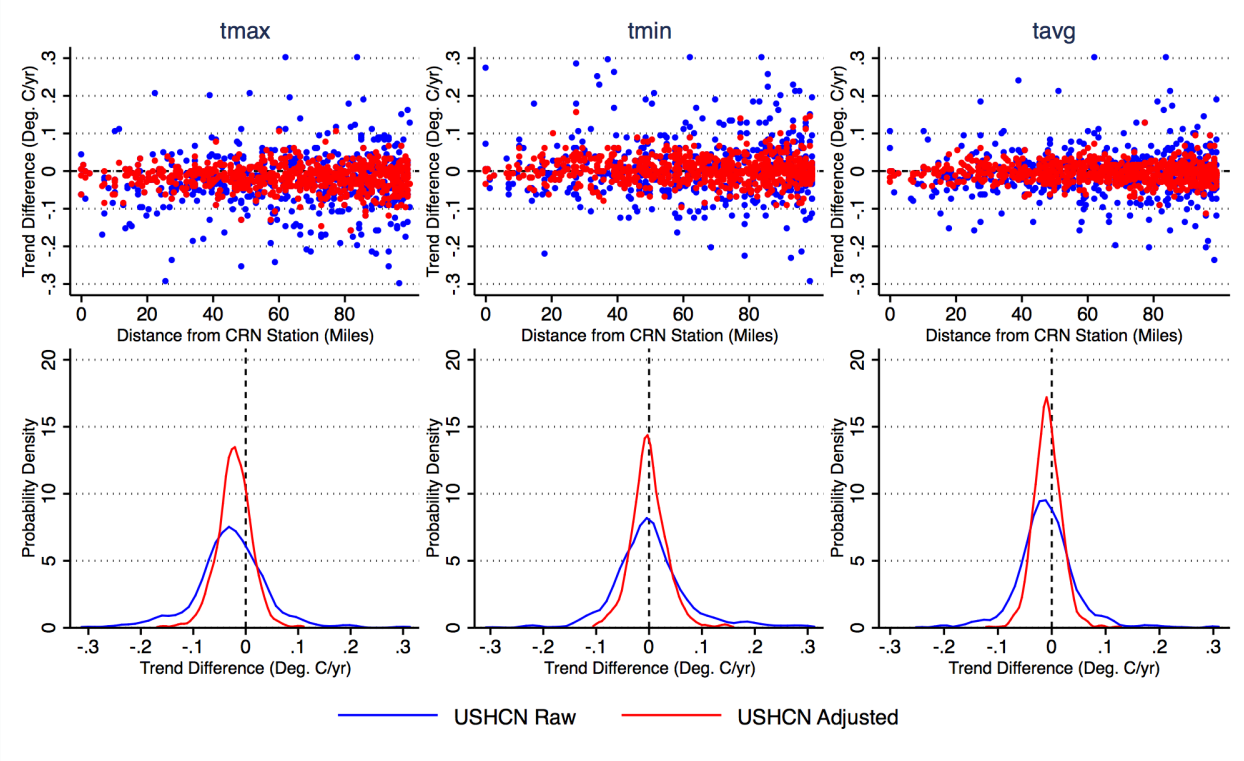


Figure 2: *Maximum (tmax), minimum (tmin), and mean (tavg) trend differences from USHCN-USCRN station pairs within 100 miles (161 km) of each other for both raw and adjusted USHCN data. The top panel shows a scatter plot of trend differences (in degrees °C per year) as a function of distance between station pairs; the bottom panel shows the probability density*

function of station pair trends with kernel density displayed on the y-axis. For detailed statistics of the data shown, see supplementary materials Table SM2.

If adjustments to USHCN data removed all inhomogenities present in the data, we would expect the trend differences between USHCN and USCRN stations to constitute a mean-zero normal distribution, with some variation of trends differences as a function of distance. The probability density functions in Figure 2 show a clear narrowing of the distribution around zero trend differences, particularly for minimum and mean temperatures. For maximum temperatures the distribution is narrower, but has a slight negative mean. This means that adjusted (and raw) USHCN stations generally have a lower maximum temperature trend than their nearby USCRN pairs, similar to the results from the CONUS-wide analysis. Adjustments move the trend difference slightly closer to zero, but a statistically significant ($p < 0.01$, via a two-sample t-test) gap remains.

This maximum temperature trend difference appears to be widespread among USCRN/USHCN pairs, and is not a result of any distinct subset of outliers, perhaps suggesting that the differences might be instrumental in origin rather than a result of station moves, microsite changes, or other inhomogeneities that would only affect a subset of USHCN stations during the 2004-2015 period. USCRN stations used platinum resistance thermometers in fan-aspirated solar shields, while USHCN stations primarily use MMTS instruments with no fan aspiration. Interestingly, the max temperature trend bias between USCRN and USHCN stations has the opposite sign as the absolute max temperature bias; Leeper et al. [2015] find that fan-aspirated USCRN stations read maximum temperatures as 0.48 °C colder than proximate USHCN stations, and minimum temperatures 0.36 °C warmer.

There is also a possibility that the PHA is less effective in detecting (and removing) inhomogenities near the end of the record, as post-breakpoint records will be too short to allow reliable detection [Menne and Williams 2009]. However, the difference between USHCN and USCRN maximum temperatures increases fairly monotonically between 2004 and 2015 (figure SM4), suggesting that ‘end effects’ are not responsible for the failure of homogenization to remove this difference. We also examine how these USCRN/USHCN maximum temperature differences vary regionally (figure SM5 in the supplementary materials), and find that the effect is easily noticeable in the Eastern and Central U.S. but somewhat smaller in the Western U.S.

The variation in trend differences over distances between USHCN-adjusted/USCRN pairs is considerably smaller than that of USHCN-raw/USCRN pairs. There is some variation expected with distance, so to test whether or not adjustments are producing a realistic distribution of trend differences over distance we compare them to the distribution of trend differences between pairs of similarly-proximate USCRN stations, as shown in Figure 3. Here pairs of stations within 150 miles (241 km) are used due to the limited number of CRN stations in close proximity.

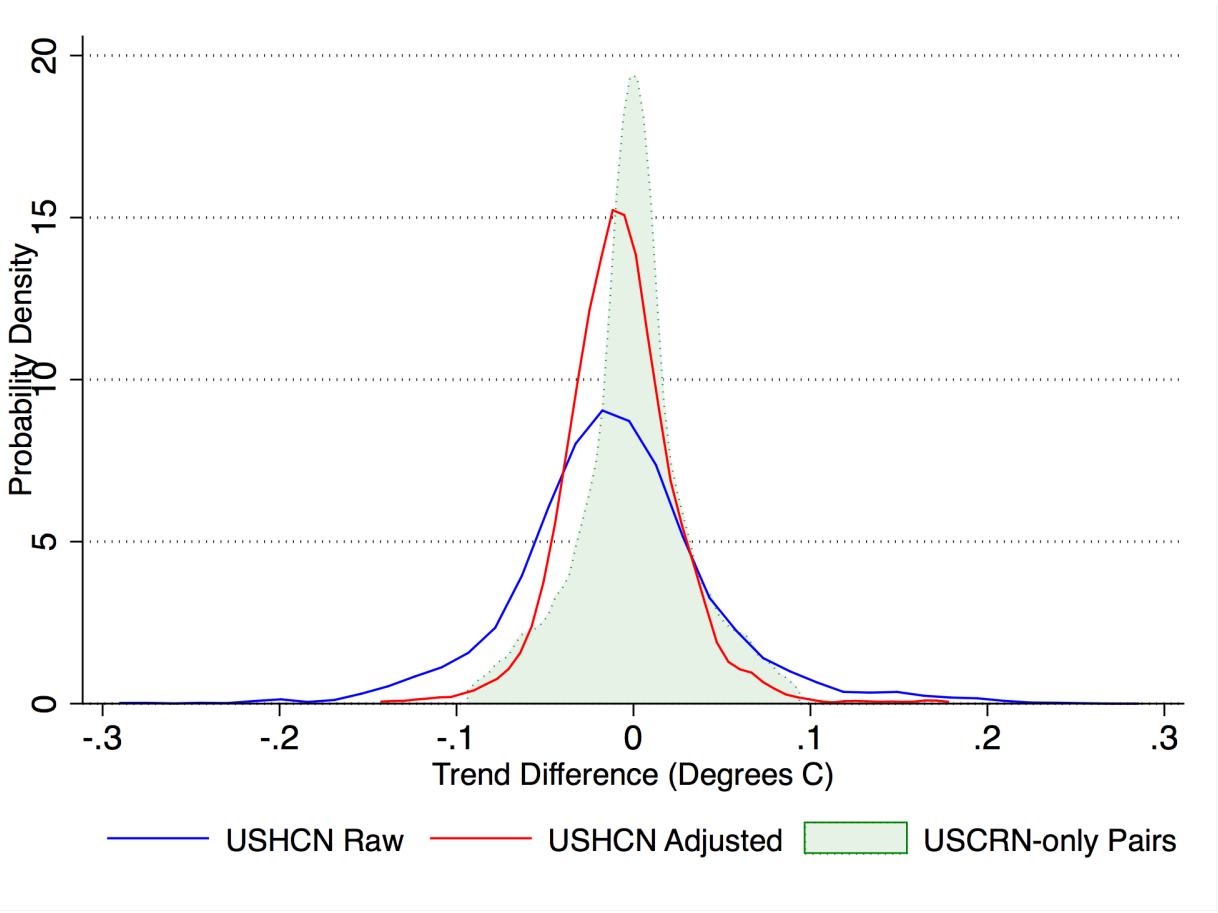


Figure 3: *Probability density function of tavg trend differences between USCRN and USHCN pairs within 150 miles (241 km), with a range of expected trend variation (green shading) based on pairs of USCRN-only stations within 150 miles (241 km) of each other, with kernel density displayed on the y-axis.*

The adjustments to USHCN stations create a spatial structure of trends more similar to the USCRN stations over longer distances as well. Figure 4 shows the standard deviation of trend differences between within-network station pairs (USCRN to USCRN; raw USHCN to raw USHCN; adjusted USHCN to adjusted USHCN) as a function of distance for the period from January 2004 to October 2015. Raw USHCN stations have much greater variation in trends

between station pairs across all distances; the adjustments consistently reduce this variation to the level seen in the homogenous USCRN stations.

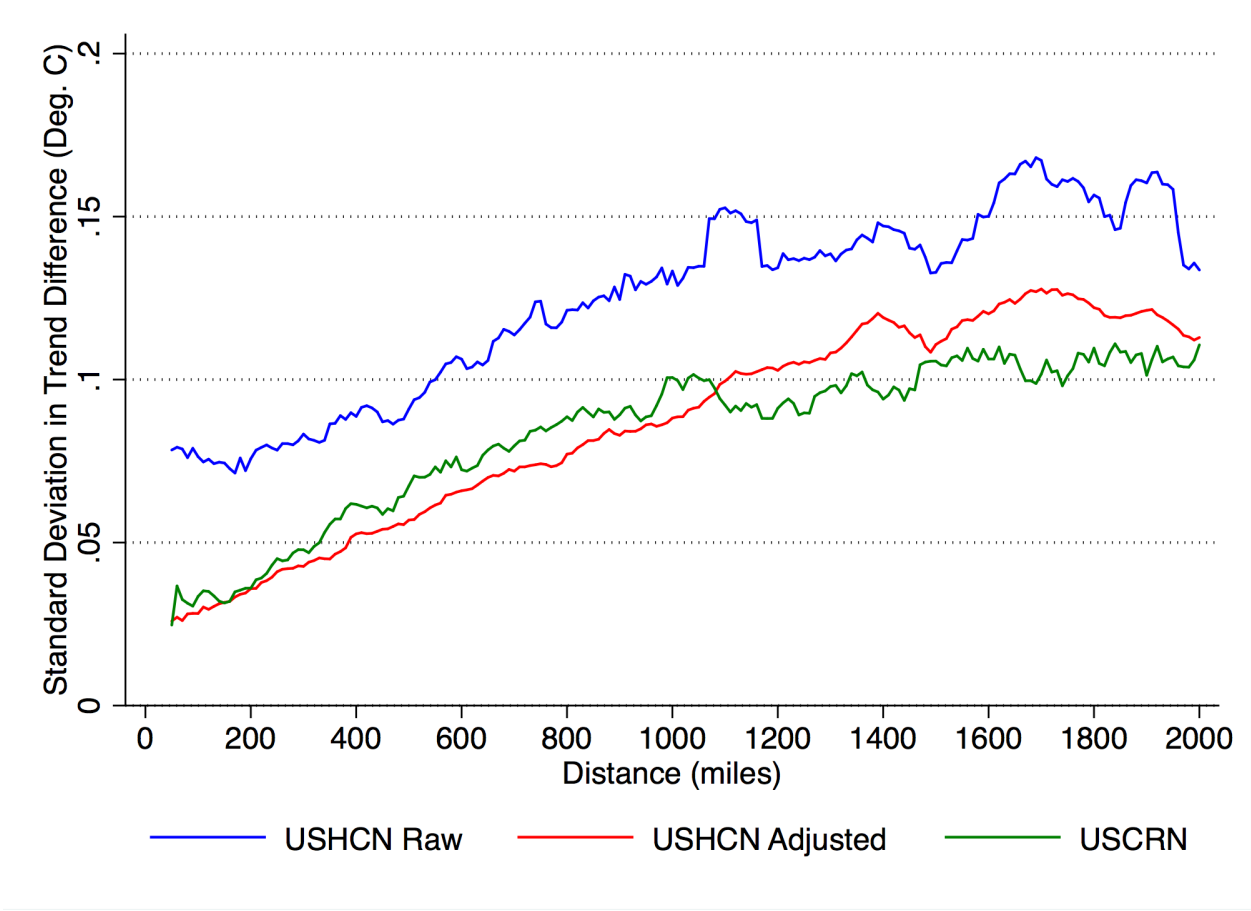


Figure 4: *Standard deviation of trend differences between in-network station pairs as a function of distance.*

Mean squared differences between USHCN/USCRN station pair anomalies are also calculated (shown in supplementary materials figures SM6 and SM7). These provide a measure of the difference in anomalies for individual stations somewhat independent of trend impacts. For minimum, maximum, and mean temperature series of station pairs within 100 miles (161 km) the mean squared difference of the adjusted data is statistically significantly smaller ($p < 0.01$) than that of the raw data, indicating that adjustments are making anomalies of USHCN stations more similar to USCRN stations.

CONCLUSIONS

During the period of overlap between the USHCN and USCRN networks, we can confidently conclude that the adjustments to the USHCN station records made them more similar to proximate homogenous USCRN station records, both in terms of trends and anomalies. There are no systematic trend biases introduced by adjustments during this period; if anything adjusted USHCN stations still underestimate maximum (and mean) temperature trends relative to USCRN stations. This residual maximum temperature bias warrants additional research to determine the exact cause.

While this analysis can only directly examine the period of overlap, the effectiveness of adjustments during this period is at least suggestive that the PHA will perform well in periods prior to the introduction of the USCRN, though this conclusion is somewhat tempered by the potential changing nature of inhomogeneities over time. This work provides an important empirical test of the effectiveness of temperature adjustments similar to Vose et al. [2012], and lends support prior work by Williams et al [2012] and Venema et al [2012] that used synthetic datasets to find that NOAA's pairwise homogenization algorithm is effectively removing localized inhomogeneities in the temperature record without introducing detectable spurious trend biases.

ACKNOWLEDGMENTS

The USHCN data is available from ftp.ncdc.noaa.gov/pub/data/ushcn/v2.5/
The USCRN data is available from ftp.ncdc.noaa.gov/pub/data/uscrn/products/monthly01/
Computer code is available from http://www.ysbl.york.ac.uk/~cowtan/crn2016/
ZH is funded by Berkeley Earth. MM and CW are funded by NOAA. No specific funding or grants supported this project.

REFERENCES

Changnon, S. A., and K. E. Kunkel (2006), Changes in Instruments and Sites Affecting Historical Weather Records: A Case Study. Atmos. Oceanic Technol., 23, 825–828.

Diamond, H. J., T. R. Karl, M. A. Palecki, C. B. Baker, J. E. Bell, R. D. Leeper, D. R. Easterling, J. H. Lawrimore, T. P. Meyers, M. R. Helfert, G. Goodge, and P. Thorne (2013), U.S. Climate Reference Network after One Decade of Operations: Status and Assessment. Bulletin of the American Meteorological Society, 485-498.

Doesken, N. (2005), The National Weather Service MMTS (Maximum Minimum Temperature System) - 20 Years After. 15th Conference on Applied Climatology; 13th Symposium on Meteorological Observations and Instrumentation. JP1.26.

U.S. Environmental Protection Agency (EPA) (2013), Technical Documentation: U.S. and Global Temperature. Available: http://www3.epa.gov/climatechange/pdfs/temperature_documentation-2013.pdf

Fall, S., A. Watts, J. Nielsen-Gammon, E. Jones, D. Niyogi, J. Christy, and R.A. Pielke Sr. (2011), Analysis of the impacts of station exposure on the U.S. Historical Climatology Network temperatures and temperature trends. J. Geophys. Res., 116.

Fiebrich, C. A. (2009), History of surface weather observations in the United States. Earth-Science Reviews, 93(3-4), 77-84.

Gallo, K. (2005), Evaluation of Temperature Differences for Paired Stations of the U.S. Climate Reference Network. Journal of Climate, 18, 1629-1636.

Hausfather, Z., M. J. Menne, C. N. Williams, Jr., T. Masters, R. Broberg, and D. Jones (2013), Quantifying the Effect of Urbanization on U.S. Historical Climatology Network Temperature Records. Journal of Geophysical Research, 118-2, 481-494.

Hubbard, K. G., and X. Lin (2006), Reexamination of instrument change effects in the U.S. Historical Climatological Network. Geophysical Research Letters, 33.

Karl, T. R., C. N. Williams Jr., P. J. Young, and W. M. Wendland (1986), A model to estimate the time of observation bias associated with monthly mean maximum, minimum, and mean temperature for the United States. J. Clim. Appl. Meteorol., 25, 145–160.

Karl, T. R., H. F. Diaz, and G. Kukla (1988), Urbanization: Its detection and effect in the United States climate record, J. Climate, 1, 1099–1123.

Leeper, R. D., J. Rennie, and M.A. Palecki (2015), Observational Perspectives from U.S. Climate Reference Network (USCRN) and Cooperative Observer Program (COOP) Network: Temperature and Precipitation Comparison. J. Atmos. Oceanic Technol., 32, 703–721.

Menne, M. J., and C. N. Williams (2009), Homogenization of temperature series via pairwise comparisons, J. Climate, 22, 1700–1717.

Menne, M. J., C. N. Williams, Jr., and R. S. Vose (2009), The United States Historical Climatology Network monthly temperature data—Version 2. Bull. Am. Meteorol. Soc., 90, 993–1007.

Menne, M. J., C. N. Williams, Jr., and M. A. Palecki (2010), On the reliability of the U.S. surface temperature record. J. Geophys. Res., 115, D11108, doi:10.1029/2009JD013094.

Muller, R. A., J. Wurtele, R. Rohde, R. Jacobsen, S. Perlmutter, A. Rosenfeld, J. Curry, D. Groom, C. Wickham, and S. Mosher (2013). Earth Atmospheric Land Surface Temperature and Station Quality in the Contiguous United States. Geoinfor Geostat: An Overview 1:3

Otkin, J., M. C. Anderson, J. R. Mecikalski, and G. R. Diak (2005), Validation of GOES-based insolation estimates using data from the U.S. Climate Reference Network. Journal of Hydrometeorology, 6, 460-475.

Peterson, T. C., and T. W. Owen (2005), Urban heat island assessment: Metadata are important. J. Climate, 18(14), 2637–2646.

Quayle, R. G., D. R. Easterling, T. R. Karl, and P. Y. Hughes (1991), Effects of recent thermometer changes in the Cooperative Station Net- work, Bull. Am. Meteorol. Soc., 72, 1718–1723, doi:10.1175/1520-0477 (1991)072<1718:EORTCI>2.0.CO;2.

Quinlan, F. T., T. R. Karl, and C. N. Williams, Jr. (1987), United States Histor- ical Climatology Network (USHCN) serial temperature and precipitation data. NDP-019, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, TN.

Rohde, R., R. A. Muller, R. Jacobsen, E. Muller, S. Perlmutter, A. Rosenfeld, J. Wurtele, D. Groom, and C. Wickham (2013), A New Estimate of the Average Earth Surface Land Temperature Spanning 1753 to 2011. Geoinfor Geostat: An Overview 1:1

Venema, V. K. C. et al. (2012), Benchmarking monthly homogenization algorithms for monthly data, Clim. Past, 8, 89–115, doi:10.5194/cp-8- 89-2012.

Vose, R. S., C. N. Williams, T. C. Peterson, T. R. Karl, and D. R. Easterling (2003), An evaluation of the time of observation bias adjustment in the US Historical Climatology Network. Geophys. Res. Lett., 30(20), 2046, doi:10.1029/2003GL018111

Vose, R. S., S. Applequist, M. J. Menne, C. N. Williams Jr. and P. Thorne. An intercomparison of temperature trends in the U.S. Historical Climatology Network and recent atmospheric reanalyses. Geophys. Res. Lett., 39, L10703, doi:10.1029/2012GL051387

Williams, C. N., M. J. Menne, and P. W. Thorne (2012), Benchmarking the performance of pairwise homogenization of surface temperatures in the United States. J. Geophys. Res., 117, D5, doi:10.1029/2011JD016761.

SUPPLEMENTARY MATERIALS

These supplementary materials provide links to data, code, and a number of figures and tables referenced in the article text. Each figure includes a caption describing its contents and (when relevant) how it was calculated.

DATA AND CODE

U.S. HCN v2.5 raw data: ftp.ncdc.noaa.gov/pub/data/ushcn/v2.5/ (.raw files)
U.S. HCN v2.5 adjusted data: ftp.ncdc.noaa.gov/pub/data/ushcn/v2.5/ (.FLs.52j files)
U.S. CRN data: ftp.ncdc.noaa.gov/pub/data/uscrn/products/monthly01/

Annotated Code: http://www-users.york.ac.uk/~kdc3/papers/crn2016/

SUPPLEMENTARY FIGURES

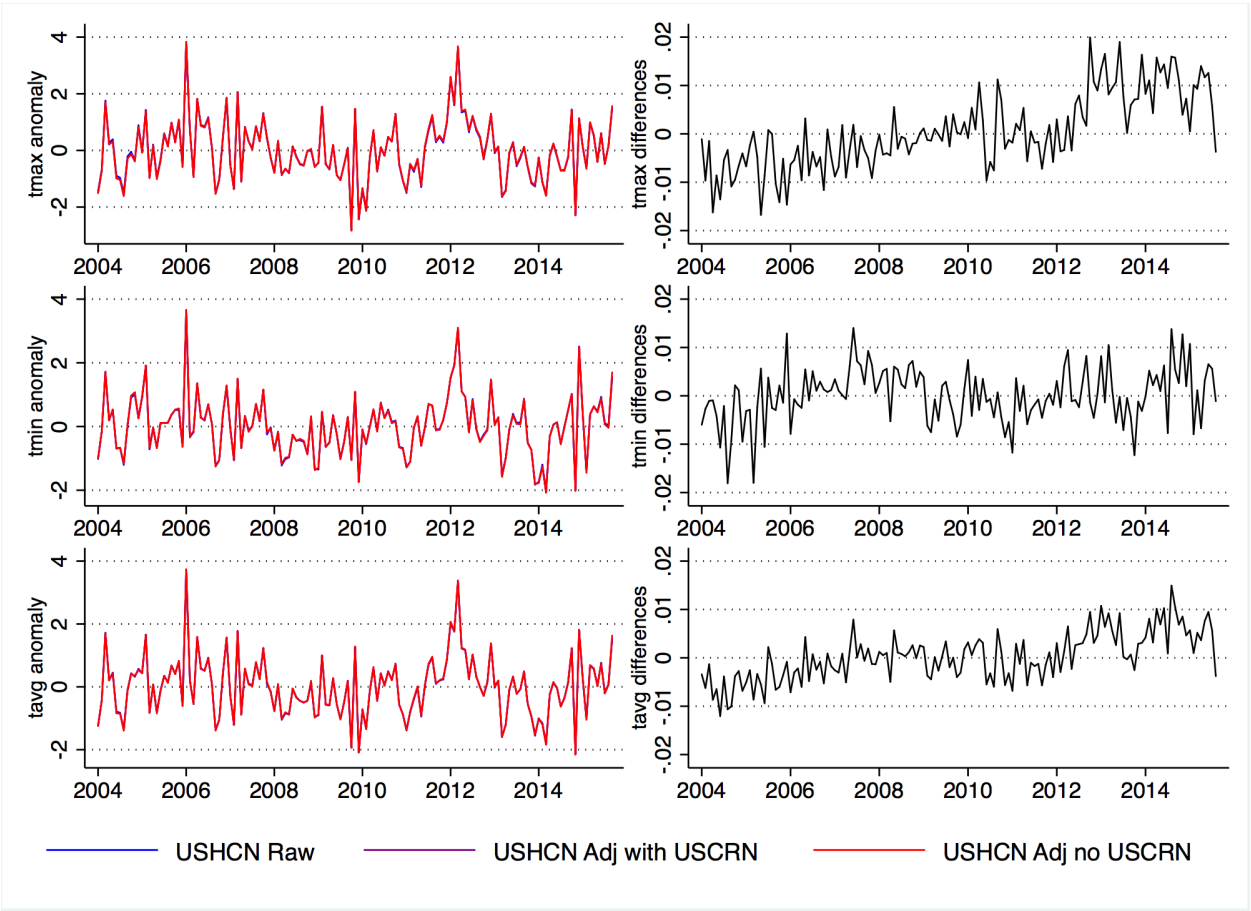


Figure SM1: *Maximum (Tmax), minimum (Tmin), and mean (Tavg) CONUS values for HCN raw, and HCN adjusted with CRN, and HCN adjusted without CRN data. Left column: CONUS temperature anomalies for each series. Right column: HCN adjusted with CRN minus HCN adjusted without CRN (in black). CONUS reconstructions are spatially-limited to grid cells where values for all three datasets are present for any given month. The inclusion of CRN in the homogenization process slightly increases Tmax trends (and Tavg trends) but not Tmin trends, consistent with the higher Tmax trends seen in CRN stations vis-a-vis nearby HCN adjusted stations.*

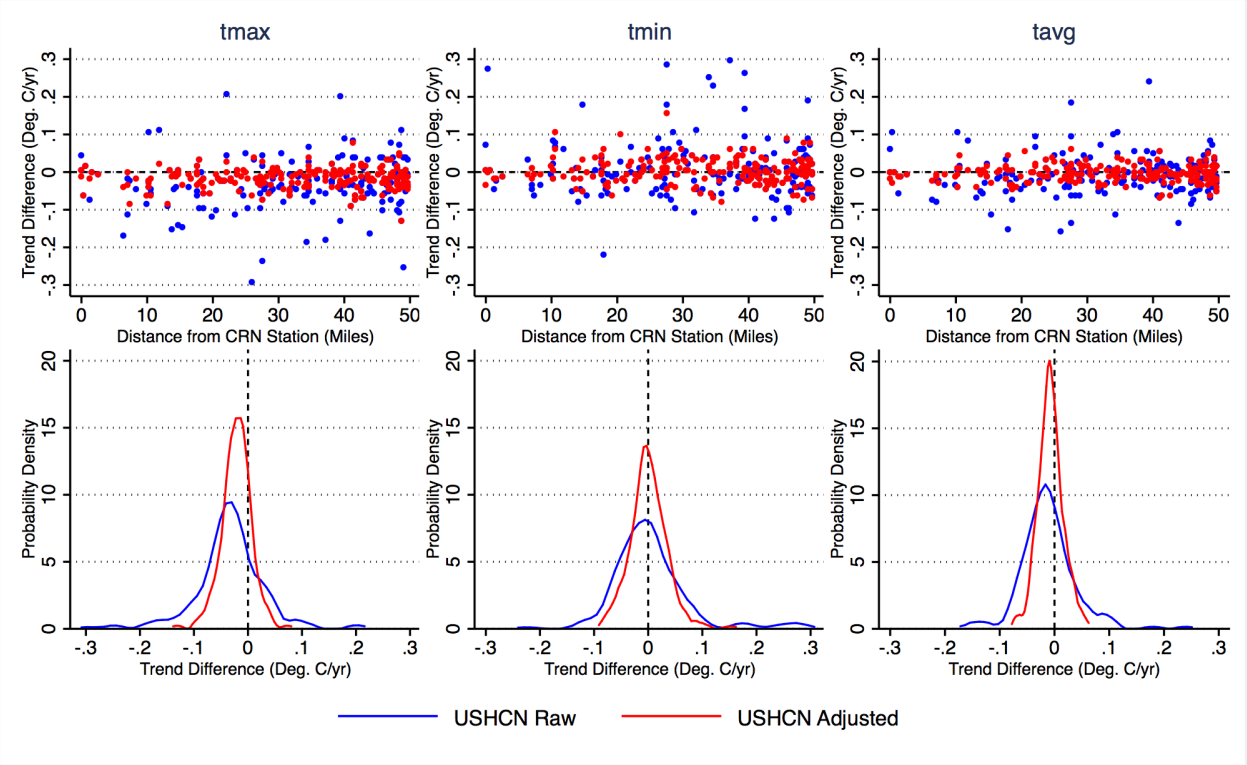


Figure SM2. *Maximum (tmax), minimum (tmin), and mean (tavg) trend differences from HCN-CRN station pairs within 50 miles of each other for both raw and adjusted HCN data. The top panel shows a scatter plot of trend differences (in degrees °C per year) as a function of distance between station pairs; the bottom panel shows the probability density function of station pair trends.*

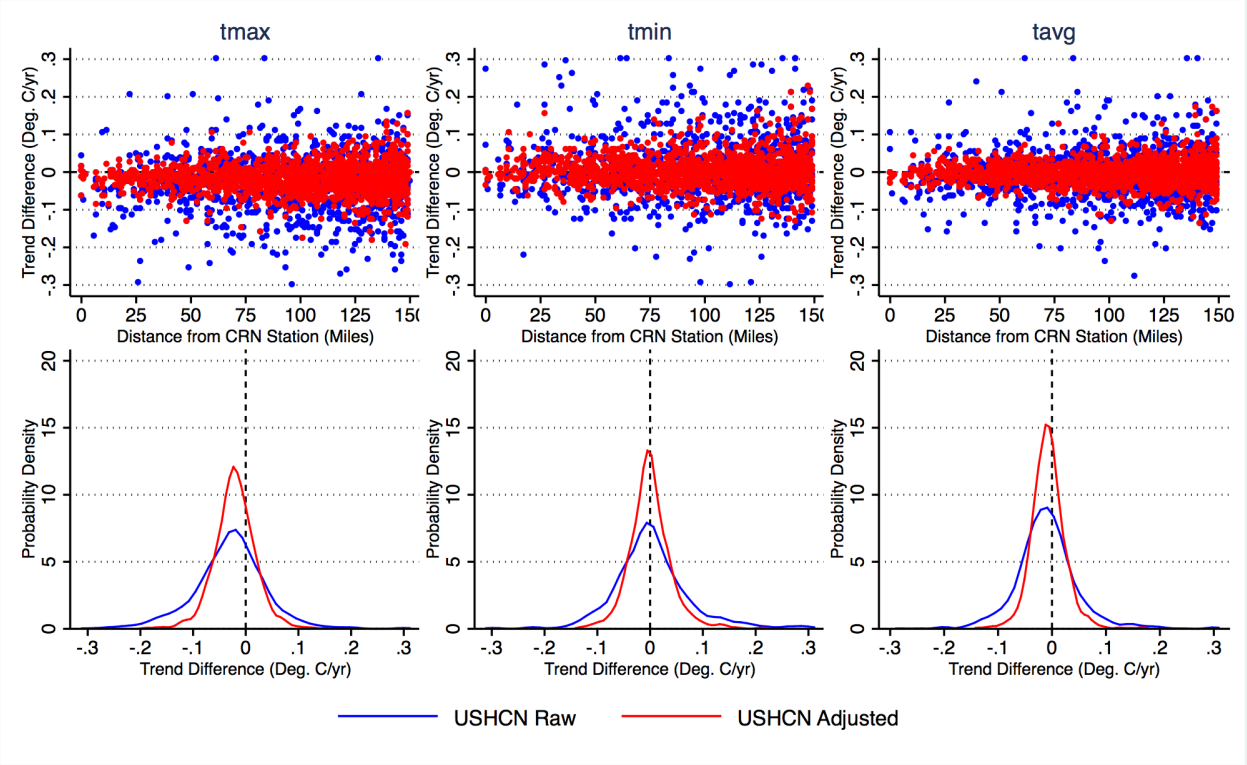


Figure SM3. *Maximum (tmax), minimum (tmin), and mean (tavg) trend differences from HCN-CRN station pairs within 150 miles of each other for both raw and adjusted HCN data. The top panel shows a scatter plot of trend differences (in degrees °C per year) as a function of distance between station pairs; the bottom panel shows the probability density function of station pair trends.*

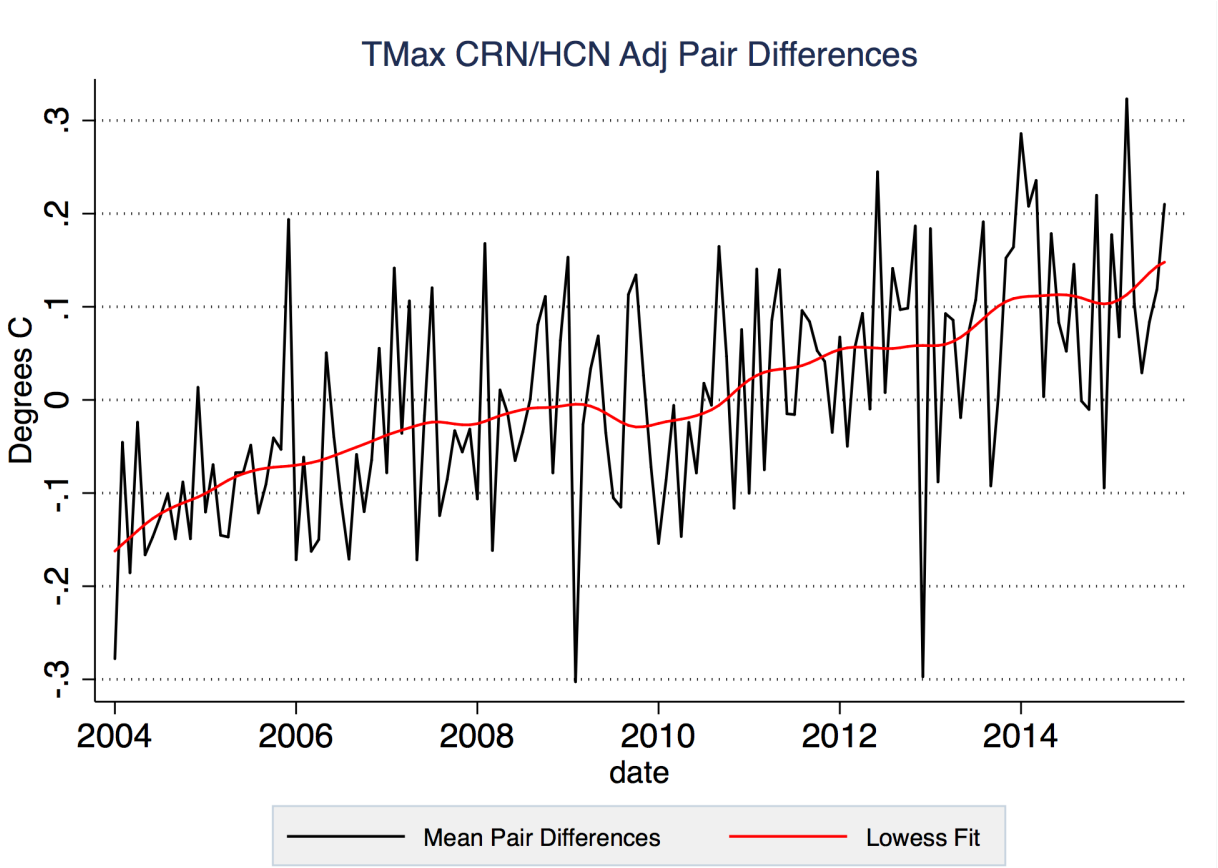


Figure SM4: *TMax mean difference between CRN and HCN-adjusted station pairs, with the mean calculated by weighting each unique CRN station by inverse of the number of unique HCN stations that it is paired with for each month in order to avoid overweighting CRN stations with more HCN pairs in the analysis. A lowess fit (bandwidth 0.2) is also shown for reference.*

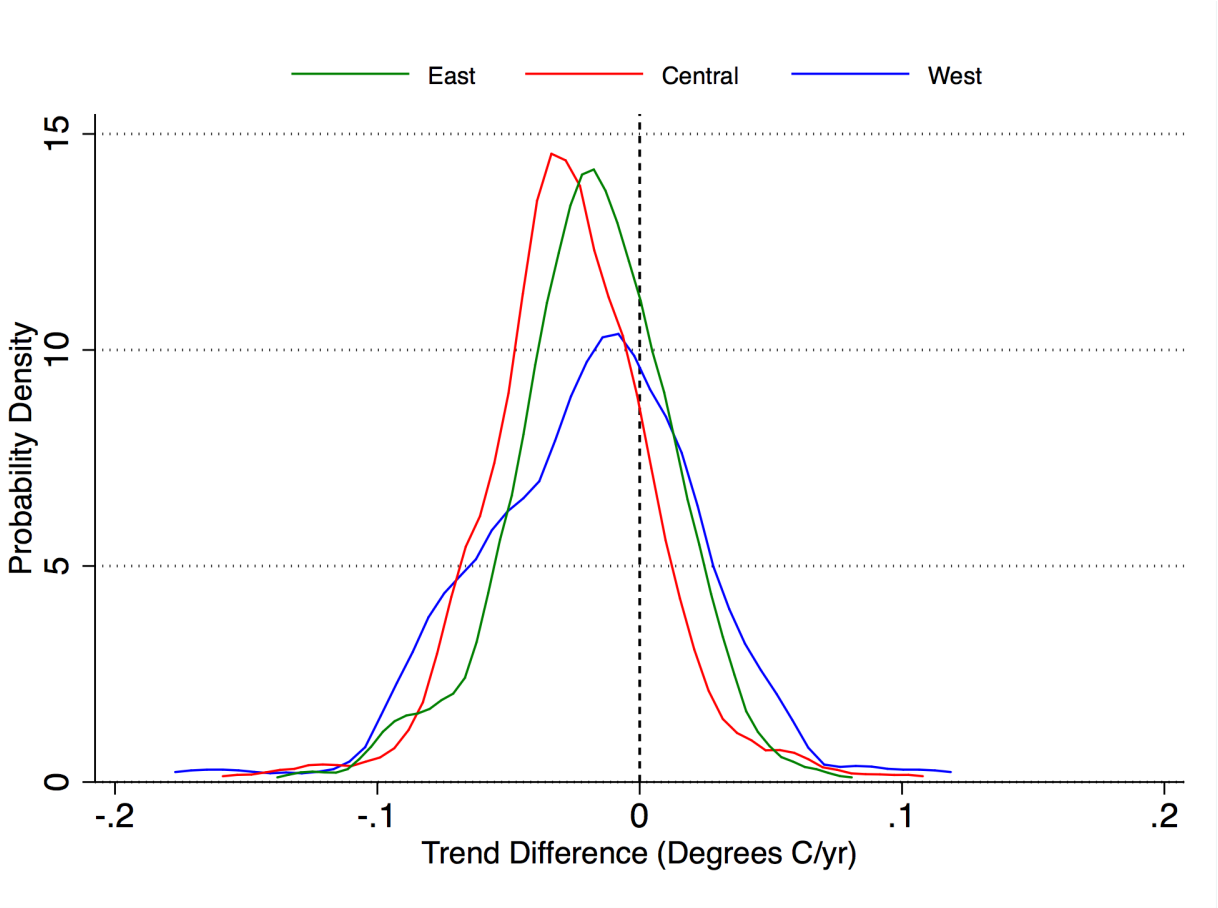


Figure SM5: *Probability density function of tmax trend differences between CRN and HCN adjusted pairs within 100 miles by geographic region. East is > -90 latitude; West is < -110 latitude; Central is between the two. The East region contains 320 station pairs, the Central region contains 255 station pairs, and the West region contains only 76 station pairs.*

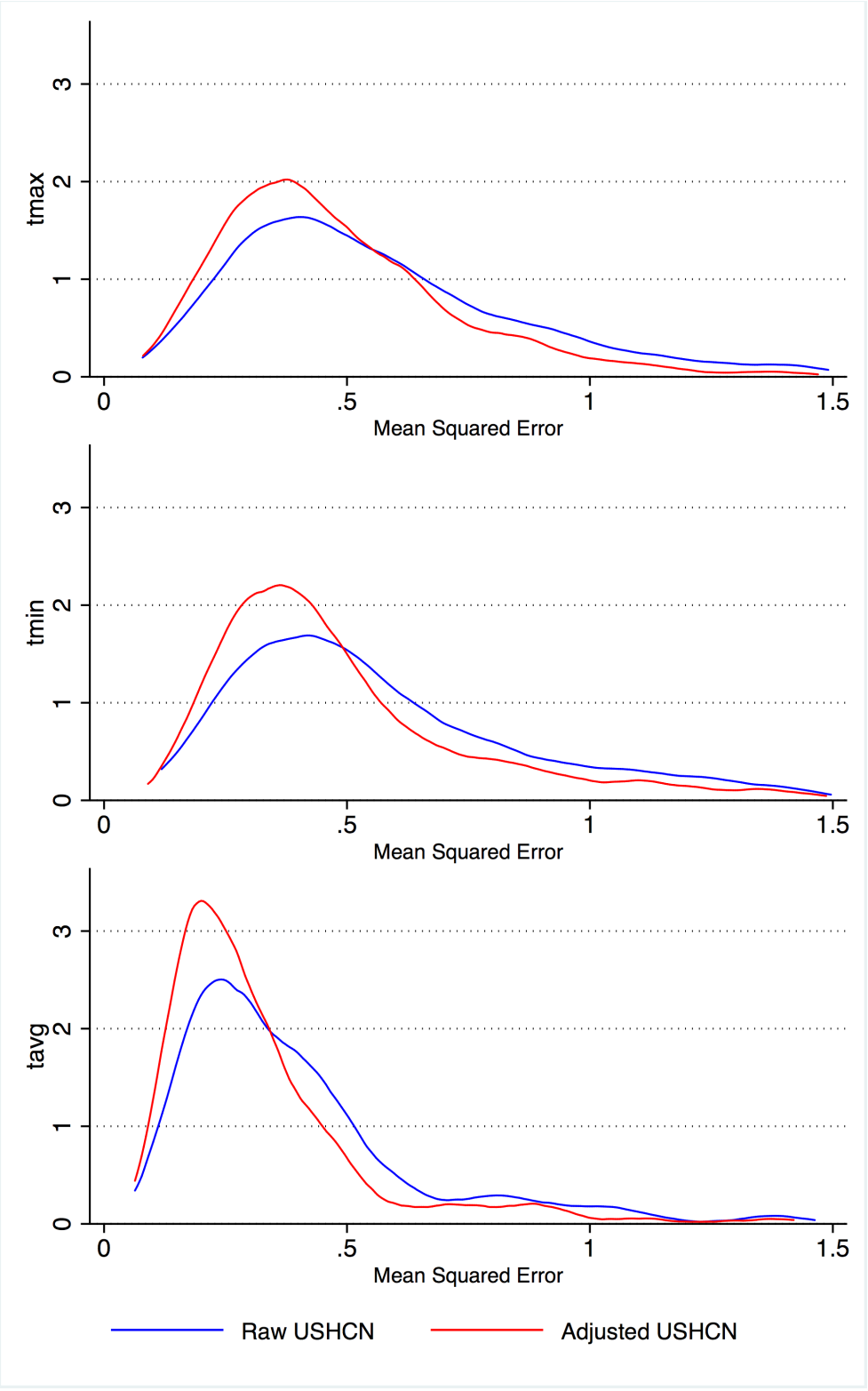


Figure SM6: *Probability density of maximum (tmax), minimum (tmin), and mean (tavg) mean square errors from USHCN-USCRN station pairs within 100 miles of each other for both raw and adjusted USHCN data. Values on y-axis are in degrees C.*

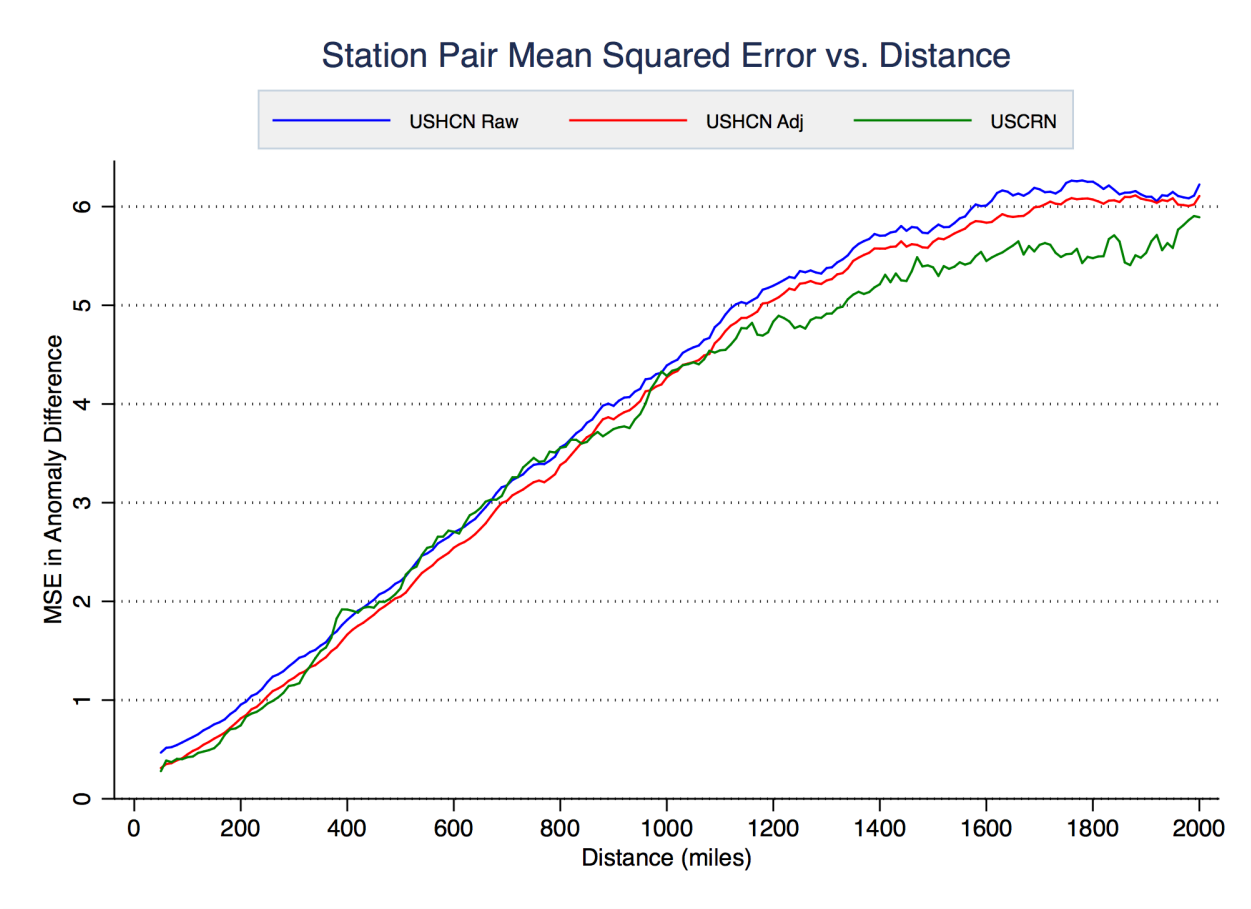


Figure SM7: *Mean squared errors in differences between in-network station pairs as a function of distance out to 2000 miles.*

Table SM.1: Statistics on CONUS raw/adjusted USHCN and USCRN trends			
Measurement	**Series**	**Trend**	**Confidence Interval**
Tmax	USHCN Raw	-0.058	-1.150 to 1.034
Tmax	USHCN Adjusted	-0.055	-1.165 to 1.055
Tmax	USCRN	0.179	-0.916 to 1.276
Tmin	USHCN Raw	-0.086	-0.717 to 0.546
Tmin	USHCN Adjusted	-0.159	-0.793 to 0.475
Tmin	USCRN	-0.176	-0.821 to 0.47
Tavg	USHCN Raw	-0.074	-0.897 to 0.749
Tavg	USHCN Adjusted	-0.108	-0.938 to 0.722
Tavg	USCRN	-0.028	-0.872 to 0.817

Tmax	USHCN Raw minus USCRN	-0.238	-0.300 to -0.176
Tmax	USHCN Adjusted minus USCRN	-0.234	-0.244 to -0.225
Tmax	USHCN Adj minus USHCN Raw	0.003	-0.042 to 0.048
Tmin	USHCN Raw minus USCRN	0.09	0.011 to 0.169
Tmin	USHCN Adjusted minus USCRN	0.017	-0.057 to 0.091
Tmin	USHCN Adj minus USHCN Raw	-0.073	-0.097 to -0.049
Tavg	USHCN Raw minus USCRN	-0.046	-0.102 to 0.009
Tavg	USHCN Adjusted minus USCRN	-0.080	-0.134 to -0.026
Tavg	USHCN Adj minus USHCN Raw	-0.034	-0.063 to -0.004

Table SM1: *Mean and 95% confidence intervals in trends and trend differences from data shown in Figure 1. Trend differences confidence intervals are in degrees °C per decade from January 2004 through October 2015. Confidence intervals are calculated using an ARMA[1,1] approach to account for autocorrelation.*

Table SM.2: Statistics on raw/adjusted USHCN/USCRN pair differences				
Measurement	**Pair Distance**	**Pair Type**	**Trend Diff Mean**	**Trend Diff StDev**
Tmax	50	Raw	-0.033	0.063
Tmax	50	Adjusted	-0.021	0.027
Tmax	100	Raw	-0.029	0.071
Tmax	100	Adjusted	-0.022	0.033
Tmax	150	Raw	-0.030	0.070
Tmax	150	Adjusted	-0.021	0.039
Tmin	50	Raw	0.004	0.071
Tmin	50	Adjusted	0.001	0.034
Tmin	100	Raw	0.003	0.073
Tmin	100	Adjusted	0.000	0.034
Tmin	150	Raw	0.005	0.079
Tmin	150	Adjusted	0.000	0.040
Tavg	50	Raw	-0.012	0.050
Tavg	50	Adjusted	-0.007	0.023
Tavg	100	Raw	-0.010	0.058
Tavg	100	Adjusted	-0.008	0.026

Tavg	150	Raw	-0.010	0.060
Tavg	150	Adjusted	-0.007	0.032

Table SM2: *Mean and standard deviation in trend differences from data shown in Figure 2, Figure SM2, and Figure SM3. Pair distance is in miles; trend differences and standard deviations are in degrees °C per year (following the convention in Figure 2) from January 2004 through October 2015.*

III. OCEAN TEMPERATURES

Oceans contribute the bulk of global surface temperature measurements, as they cover slightly over 70% of Earth's surface. Measuring the temperature of the oceans is subject to large uncertainties. While correlation lengths of temperature changes are historically higher than over land, sampling of ocean temperatures has been much sparser, particularly for variables like ocean heat content. Correcting for changes in measurement techniques over time has posed a particular challenge, as the non-stationary nature of most ocean temperature observation platforms makes the neighbor-comparison approaches used for land temperatures difficult to apply.

Sea surface temperatures (SST) are the single largest contributor to uncertainty in overall GMST records, with the magnitude of adjustments in response to past changes in measuring technique dwarfing any changes made to the land record on a global scale.

Major inhomogeneities have arisen over time in SST records associated with the transition from measuring temperature from wooden (and later canvas) buckets thrown over the side of ships to temperatures measured in engine room intake valves.[48] Bucket-based measurements were systematically cooler, due both to the higher ambient warmth of ship engine rooms and (more importantly) evaporative cooling that occurs during the period between when the bucket is pulled out of the water, hauled onto the ship deck, and finally has its temperature measured. Similarly, the transition between engine room intake measurements and those taken by automated buoys over the past two decades has introduced a bias.

The ocean temperature research presented in this dissertation has focused on three areas:

1) Understanding and reconciling differences between SST records in the last two decades (which were the key issue of contention in the controversy surrounding the Karl et al paper).[23]
2) Better understanding the ocean temperature record in the period around WW2 using island and coastal land stations as a more homogenous reference.
3) Evaluating recent changes in ocean heat content measurements and comparing them to climate model projections.

1. RECONCILING RECENT DIVERGENCES IN SST RECORDS

In 2015 NOAA updated its sea surface temperature record from ERSST version 3b to version 4. ERSSTv3b has long been something of an outlier with respect to other SST records over the past decade, showing noticeably less warming. This was due to the combination of warmer engine room intake data from ships with cooler buoy data (which began providing SST measurements in earnest in mid-1990s and today provide upwards of 90% of all SST measurements). Measurements from buoys are cooler because the instrument sits directly in the water, compared to the warmer environment of ship engine rooms after water has been pulled through the hull. Multiple studies of collocated ships and buoys have found an offset between the two of around 0.1C,[13,14] though there is some evidence this might have changed a bit in recent years.

However, instead of just bringing their record up to match those of Hadley[13] and the Japanese Meteorological Agency (JMA),[49] the new NOAA data actually showed considerably more warming, particularly after 2004. To investigate which of these records provided an accurate depiction of SSTs in recent years and to better understand why they differed, we embarked on a project to analyze many different sources of SST data.

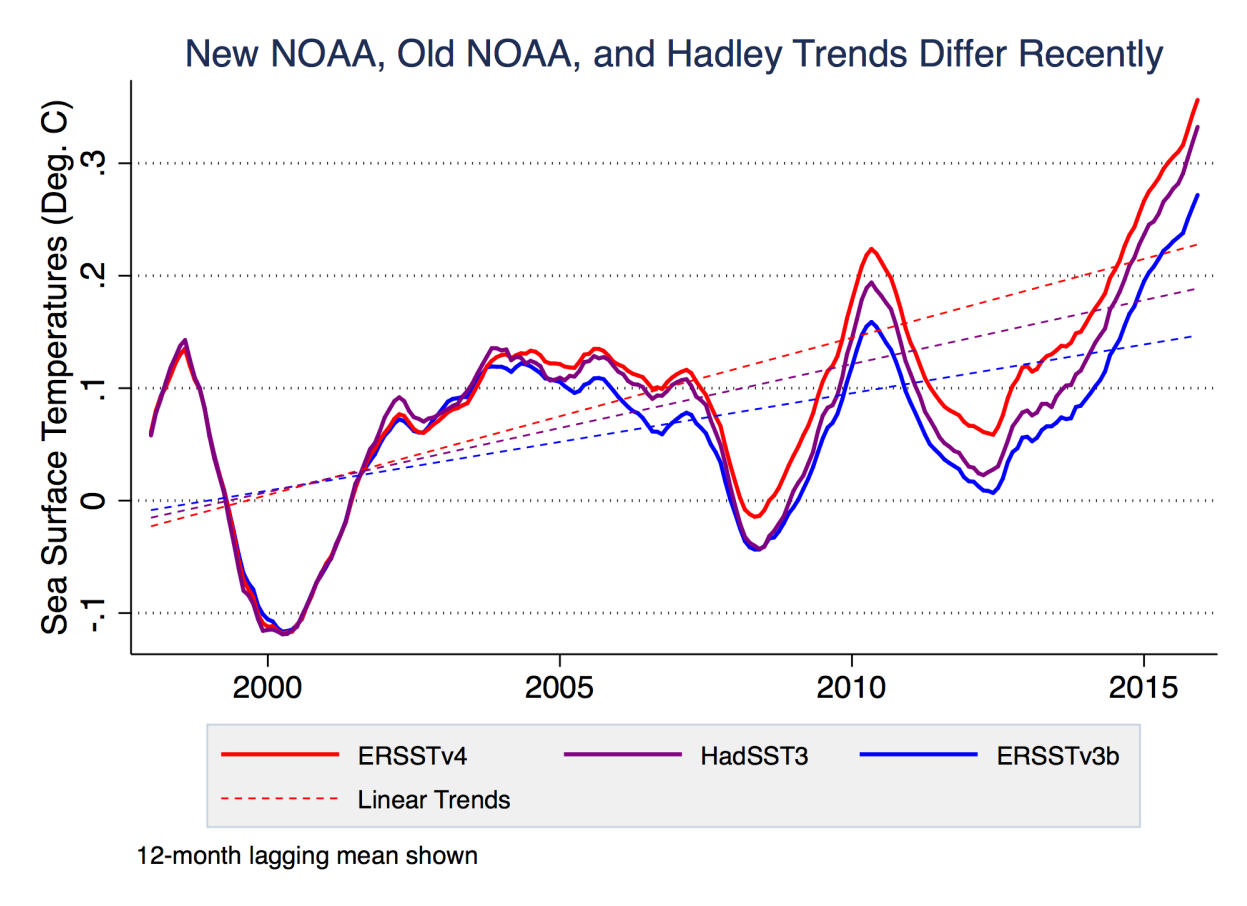


Figure III.1. *Different composite SST records from ERSSTv4, v3b, and HadSST3 show notably different trends in recent years. Figure based on data in Hausfather et al 2017.*[50]

Our insight was to create multiple different SST records, each from a different type of instrument that would require little or no adjustments to the raw readings. These "instrumentally homogenous" SST records (IHSSTs) would each provide an independent assessment of what had happened in recent decades. Because the global scientific community has invested so much time into monitoring SSTs in different ways over the last 20 years, it was possible to create separate SST records from buoys, satellite radiometers, and Argo floats. We then compared these new instrumentally homogenous records to the composite records published by NOAA, Hadley, and JMA. As most of the uncertainties associated with major SST records involve how to adjust for changes in instrumentation, we avoided this with our approach at the expense of having more limited temporal series.

The figure below shows these comparisons for buoys and satellite radiometer data IHSSTs. Satellite radiometer data uses measurements from the along-track side radiometer instruments

(ATSR) through 2014 combined with the Advanced Very High Resolution Radiometer (AVHRR) that extends through present.

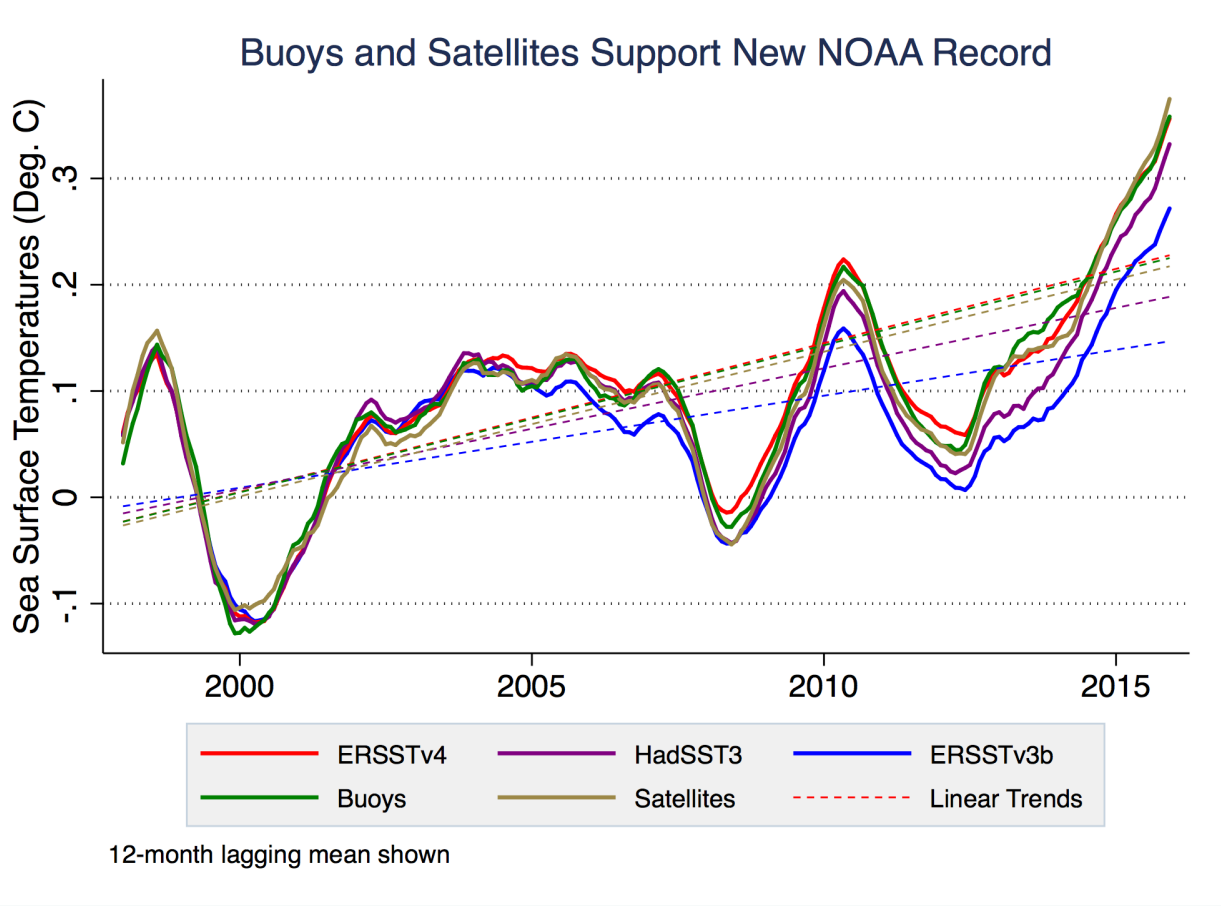


Figure III.2. *Instrumentally homogenous sea surface temperature records from buoys and satellite radiometers [51,52] agree well with NOAA's ERSSTv4 record, and show notably more warming than ERSSTv3b and HadSST3 records.*

We found that instrumentally homogenous records from buoys, satellite radiometers, and Argo floats all showed warming in similar years quite similar to NOAA's ERSSTv4 (and new v5) record, and showed more warming than ERSSTv3 as well as the commonly used HadSST3 and COBE-SST datasets. These results were robust across an ensemble of reconstructions for each dataset, and were considerably larger than the respective trend uncertainties of each.

While both HadSST3 and COBE-SST2 include explicit corrections for the ship-to-buoy measurement transition, they still show less warming than the IHSSTs we examined. These results show that the most commonly used global SST dataset, HadSST3, suffers from a cool bias in recent years that contributed to the appearance of a "hiatus" in global surface

temperatures in the early 21st century. This appears to be due to those datasets assuming a fixed offset of around 0.1°C between ship and buoy measurements over time; in reality, changes in the composition of the shipping fleet in recent decades have led to a systematic cooling bias when compared to co-located homogenous buoy measurements.

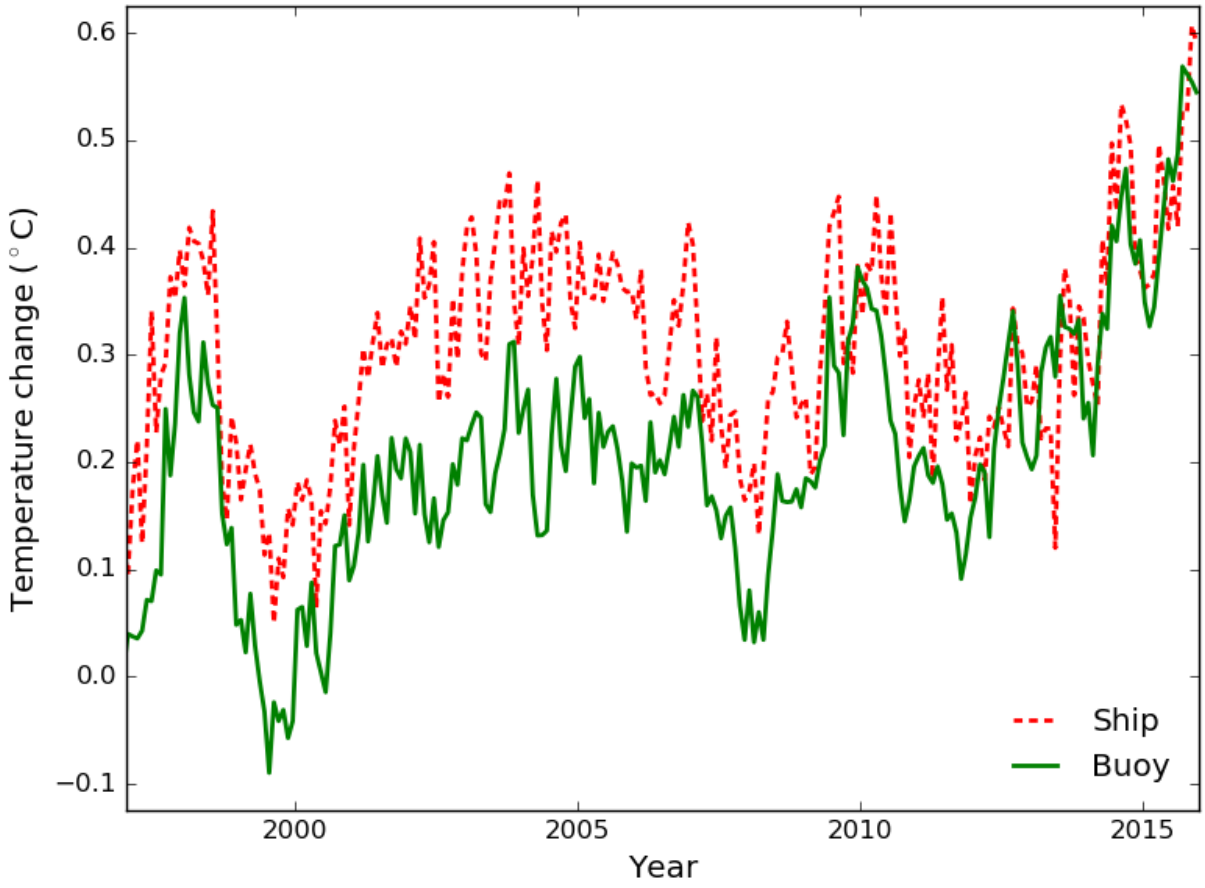


Figure III.3 Buoy-only and ship-only temperature anomalies from January 1997 through December 2015. Figure from Hausfather et al 2017.[50]

While buoys all use the same instrument type and are largely unchanged over the past two decades (apart from more buoys being deployed), the same is not true for ship-based measurements. The depth of the hull, the speed of ships, and the type of ships have all changed over the past two decades. The number of ship measurements have also fallen by about a third over this period. These changes appear to have introduced a spurious cool bias in the ship records that impacts Hadley and COBE SST records, but that the new NOAA record

mostly avoids by putting more weight on the higher quality (and more homogenous) buoy record.

This residual ship bias has been corrected in the new version of the Hadley SST dataset – HadSST4 – that will go into operational use in early 2020. In response to our paper, Hadley has included an explicit comparison between their new HadSST4 record and instrumentally homogenous records from buoys, satellites (ARC), and Argo floats.

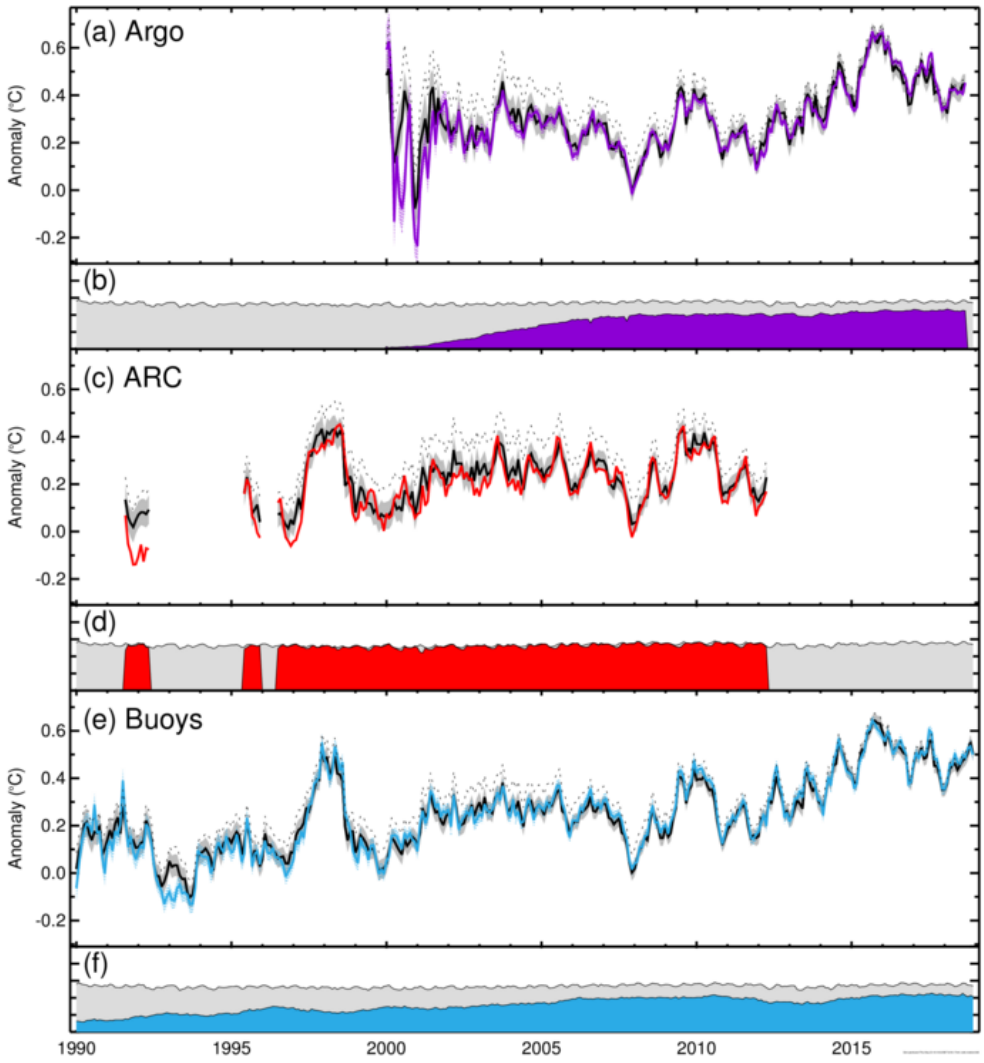


Figure III.4. Comparison of the new HadSST4 dataset to IHSSTs highlighted in Hausfather et al 2017.[50] Figure 11 from Kennedy et al 2019.[53]

2. USING ISLAND AND COASTAL STATIONS TO IMPROVE WW2-ERA RECORDS

One of the main uncertainties in sea surface temperature records occurs during the period before and after WW2. Large differences can be seen between NOAA's ERSSTv4/v5 and Hadley's HadSST3 during this period, due to different approaches used to identify and correct for changes in measurement techniques. The wartime years in particular show up quite differently, with NOAA's record exhibiting a large spike in temperatures while Hadley finds a much more gradual ramp-up.

To help better understand sea surface temperatures during this period, we developed an entirely new global sea surface temperature record from 1850 through present that used island and coastal land stations to homogenize ocean temperatures.[54] These island and coastal stations were much less affected by WW2 than ship-based ocean measurements, providing a more stable and continuous set of measurements.

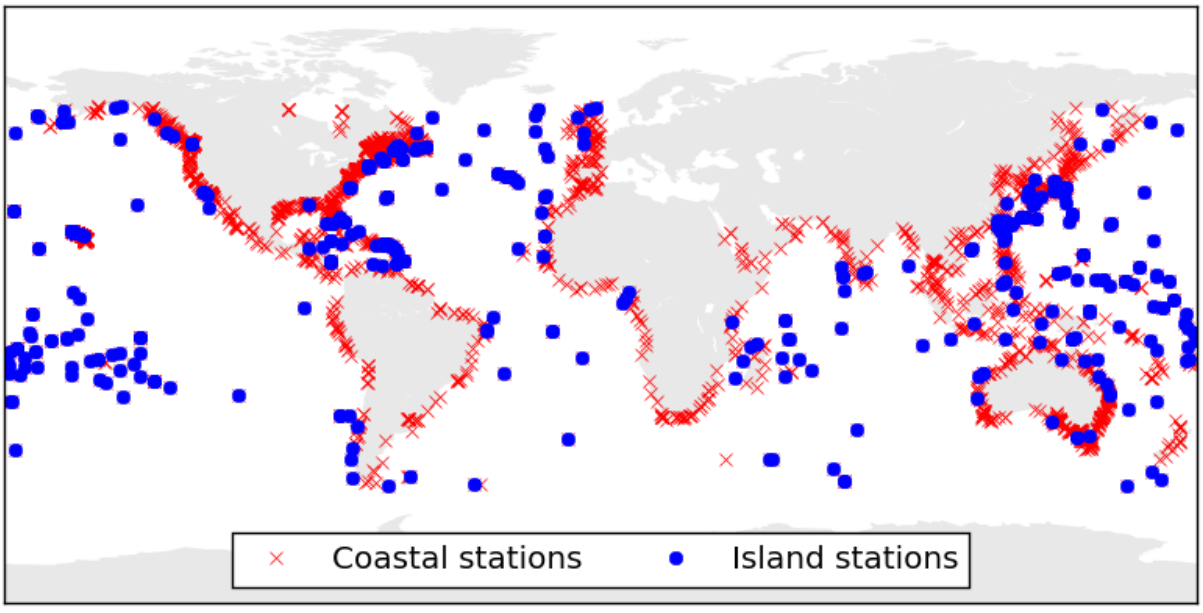


Figure III.5. *Island and coastal stations used for the hybrid SST reconstruction in Cowtan et al 2017.*[54]

These island and coastal stations are reasonably well distributed around the world and can catch large divergences between SST and nearby surface temperature measurements. We used high-resolution climate model runs to determine the necessary corrections for the difference in temperature changes over various time period in coastal vs. ocean locations, and a generalized least squares approach was employed to interpolate island and coastal records to create a spatially complete field.

The results are shown in the figure below, which compares HadSST3, ERSSTv5, and our new "Hybrid SST" series.

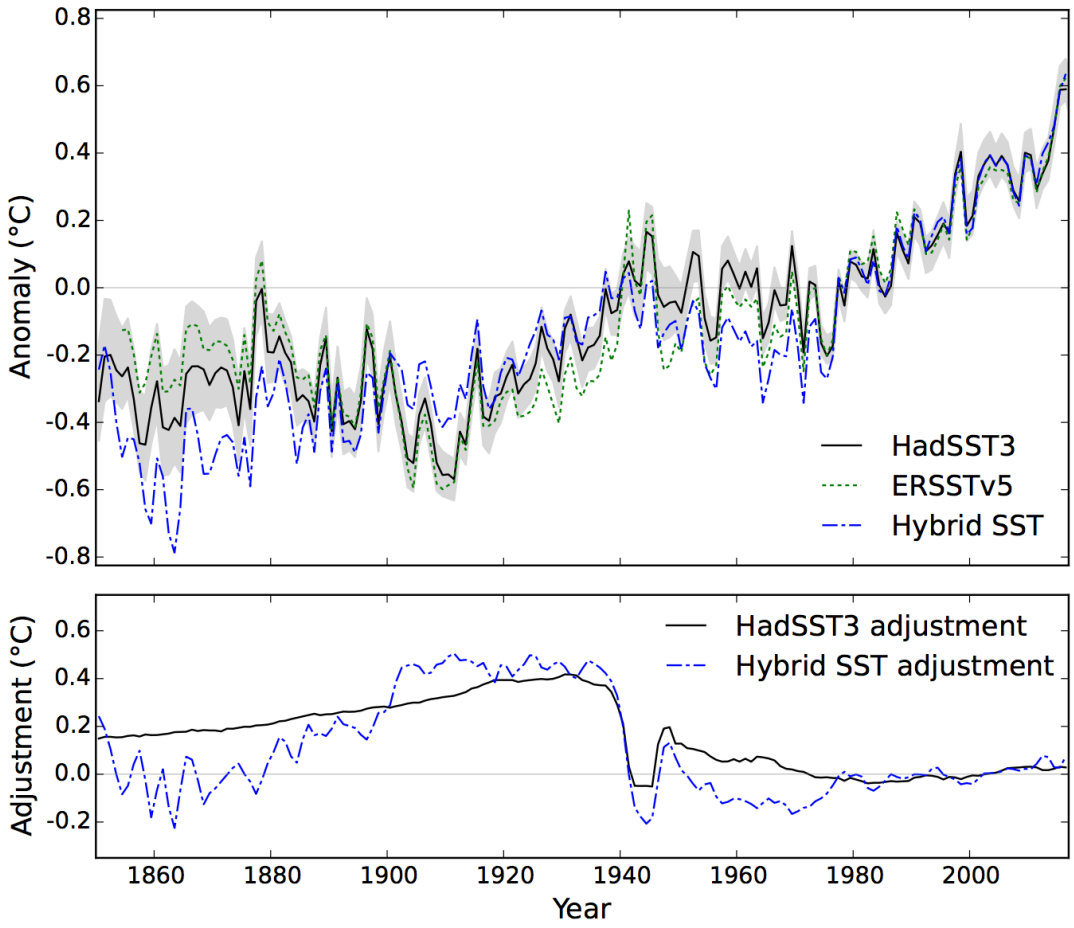


Figure III.6. *Comparison of the hybrid temperature reconstruction (using all coastal and island) to co-located data from HadSST3 and ERSSTv5 for the period 1850-2016. The shaded region is the 95% confidence region for the HadSST3 anomalies including combined bias adjustment and measurement and sampling errors. The lower panel shows the adjustment applied to the raw data in the HadSST3 and coastal hybrid records.*

In the years after 1975 the new hybrid SST record agrees quite well with both HadSST3 and ERSSTv5, though it shows more warming than HadSST3 and is quite similar to ERSSTv5 in the post-2000 period, adding further support to our prior instrumentally homogenous SST paper. The record shows quite good agreement with HadSST3 during the WW2 years, and rejects the "spike" that shows up in ERSST records. The hybrid record is cooler after WW2, similar to ERSSTv5 during that period, and also shows no cold excursion during the 1900-1910 period. The new record shows much better agreement with CMIP5 models during the 20th century than other existing SST records.

This new record is not intended to be a replacement for more conventional metadata-based approaches. Rather, it's a new independent approach that can help settle disputes between existing records (e.g. the WW2 "spike") and pinpoint periods that may deserve more scrutiny (the 1900-1910 cool excursion). The paper was published in late 2017 in the Quarterly Journal of the Royal Meteorological Society.[54]

The recently released HadSST4 record makes a number of changes that bring it more in-line with our hybrid SST record, providing a nice independent validation of our approach. The new version is significantly cooler for the period from 1950-1975, and also shows more temperature change on a centennial timescale. When compared to the coastal hybrid temperature record in the figure below, the big step around 1975 has been completely eliminated.

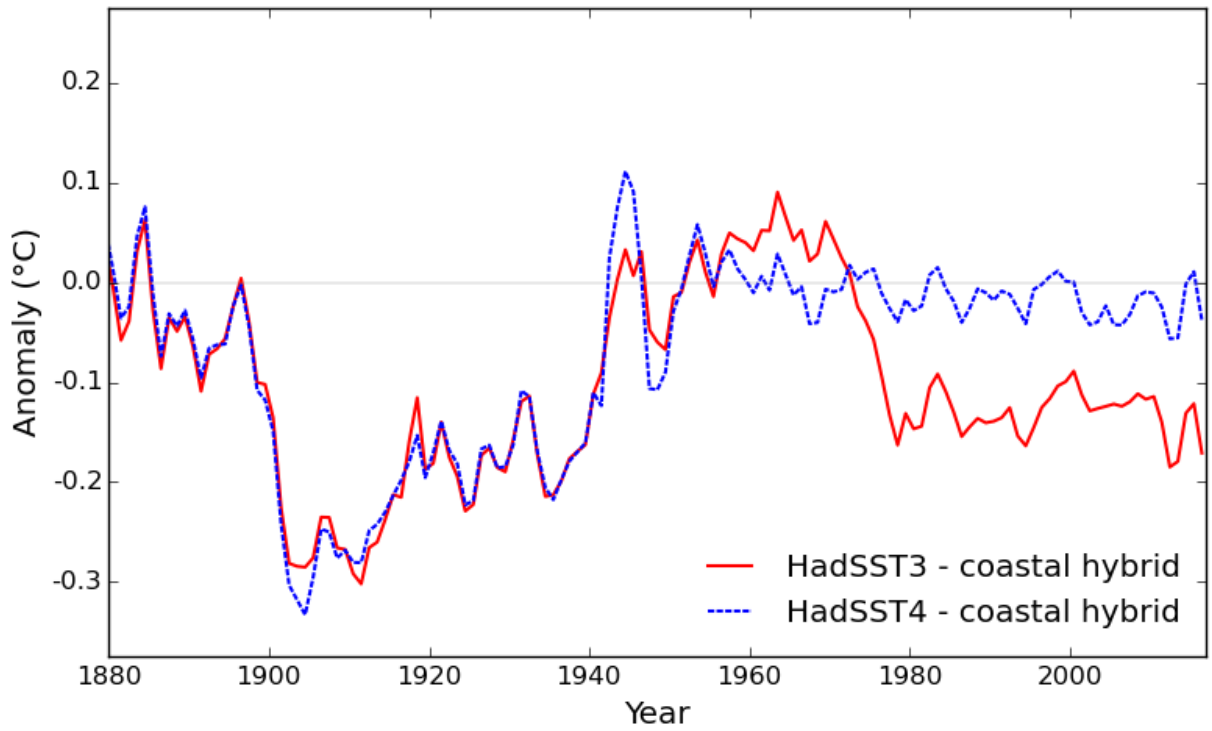


Figure III.7. *HadSST3 [13] and HadSST4 [53] records differenced from the Cowtan et al 2017 [54] hybrid SST record.*

3. REEVALUATING OCEAN HEAT CONTENT CHANGES

Ocean heat content (OHC) is one of the main measures of climate change, with around 93% of all heat trapped by greenhouse gases in the atmosphere accumulating in Earth's oceans. The "fingerprint" of human influence on the climate is much easier to detect in the oceans, as it is much less affected by year-to-year natural variability than more commonly used surface temperature records.[55]

Back in 2013, the IPCC Fifth Assessment Report (AR5) featured five different OHC estimates that generally showed oceans warming more slowly than most models projected.[56] However, over the past five years the research community has made substantial progress in improving

long-term OHC records and has identified several problems with prior OHC estimates. Improvements include properly accounting for limitations in some older OHC instruments and taking advantage of better methods of accounting for gaps in the coverage and completeness of ocean temperature measurements.[57]

OHC is a record of the heat content of Earth's oceans. It is generally given for a range of depths from 0-700 meters or 0-2000 meters, as historical measurements below 2000 meters are extremely sparse.[58] The majority of the heat trapped by greenhouse gases in the atmosphere over the past century – over 65% – has accumulated in the top 700 meters of the ocean, with most of the remainder in the top 2000 meters. OHC is a very different metric from sea surface temperatures, which only measure the top meter or so of the ocean and more closely match changes in air temperatures.

OHC is challenging to measure. However, since the mid-2000s, scientists have had the benefits of the Argo network – thousands of autonomous robots that dive down to depths of 2000 meters or so and measure temperature, salinity, pH and other ocean characteristics as they slowly ascend. Once the Argo floats surface, the data they have collected is relayed to a central data repository by satellite.

Prior to the mid-2000s, measurements were much less frequent and used devices called "expendable bathythermographs" (XBTs) – a temperature probe connected by wire to a ship, which sinks down into the ocean until the wire runs out and the probe is lost – that required extensive calibration to produce a consistent record.

XBTs are problematic in that they do not measure the depth at which they take temperatures. Using them for OHC estimates requires assumptions around the rate of both horizontal and vertical movement in the ocean. Differing corrections to XBT measurements, as well as different approaches to infilling regions where data is sparse, have led to large differences in OHC estimates prior to the Argo era.

In our paper in Science,[57] we analyzed a number of OHC records published by different groups. We examined the five records included in AR5, as well as four new or updated records that have been published over the past few years.

The figure below shows the rate of OHC warming in these records, as well as the rate projected by the CMIP5 climate models featured in AR5. The rate of warming is shown for the 1971-2010 period highlighted in AR5; estimates in blue were included in AR5, while those in purple were

published more recently. The range of OHC warming across all the CMIP5 models is shown in grey.

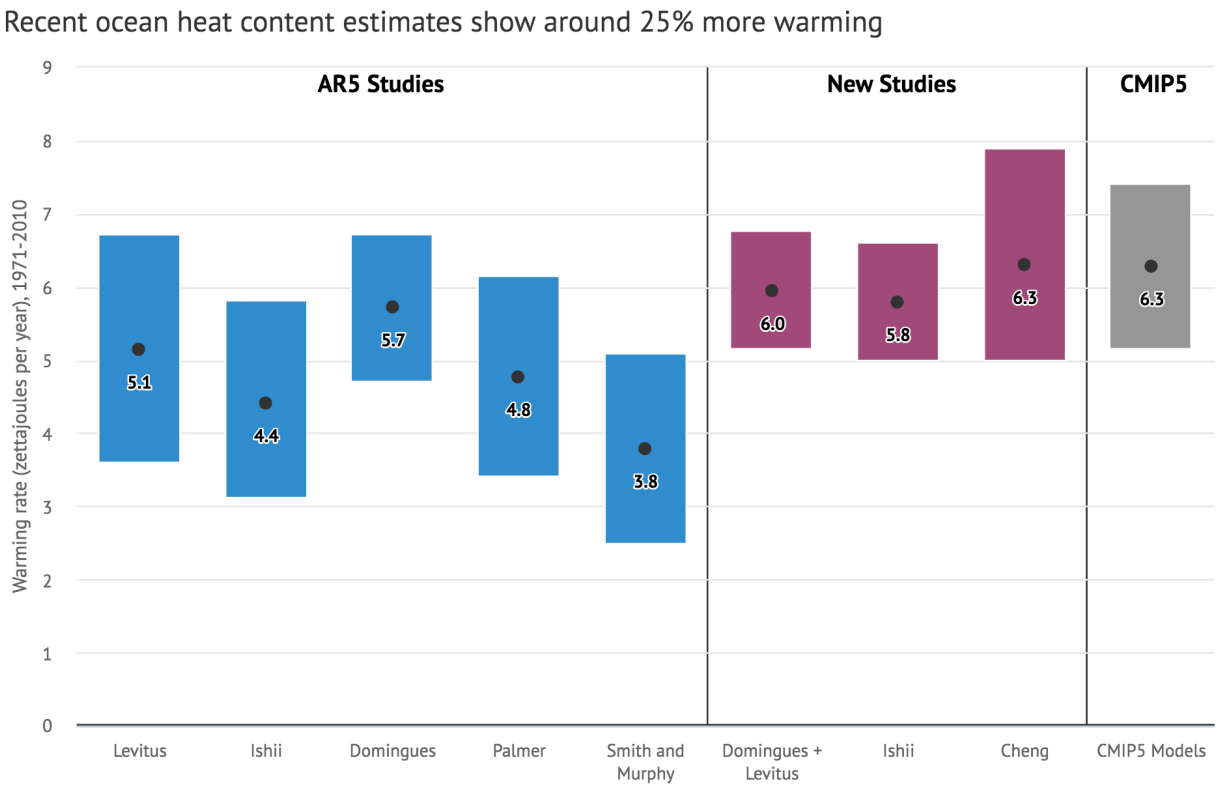


Figure III.8. *Rate of ocean heat content warming for the top 2000 meters of the ocean, from 1971-2010, in zetajoules (10^21 joules) per year. For each study the bar shows the 90% trend uncertainty, with the mean estimate shown by the black dot in the middle. For CMIP5 models, the bar spans the 90% range of models, with the black dot showing the multi-model mean*

OHC records published in recent years show, on average, about 25% – with a range of 6% to 60% – more warming than the OHC records featured in AR5. These three new records have been corrected for issues that were identified in data collected from XBTs. They also employ better statistical methods to account for limited coverage of data sampling for OHC, particularly prior to the Argo era.

The figure below shows the change in OHC over time in all four new studies, and compares them with the range of OHC changes across the CMIP5 models (grey band). The average of all the models is shown by the black line. The climate models and observations are plotted with respect to a 1981-2010 baselines, which results in model uncertainties that expand both before and after the baseline period.

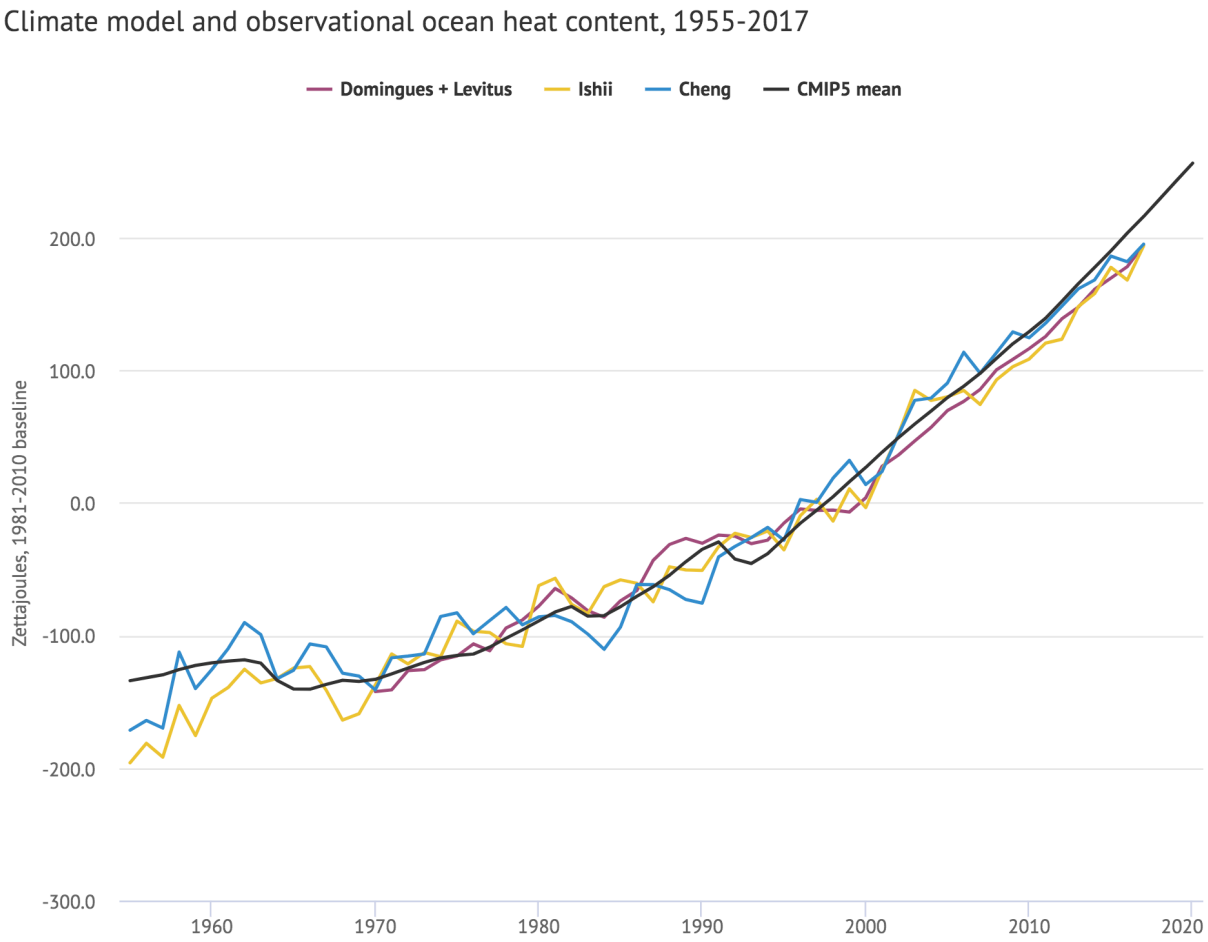


Figure III.9. *Change in ocean heat content – in zetajoules – for the top 2000 meters of the ocean with respect to a 1981-2010 baseline period in CMIP5 climate models (black line is the multimodel mean, grey area represents the 95th percentile range of model runs; RCP8.5 runs are used after historical runs end in 2005) and recent observational records (coloured lines). Data available here.*

As with the earlier chart, these new records agree well with climate model projections over the past few decades, resolving the apparent discrepancy between CMIP5 models and the OHC observations featured in the IPCC AR5.

That four independent groups of researchers have all found similar results for OHC in recent years makes us more confident that these results are accurate – and that prior OHC records suffered from problems that led them to systematically underestimate OHC changes. The agreement between climate model projections and OHC observations over the past few decades also gives us confidence that climate models are able to reliably project OHC changes into the future.

4. PAPER 3: ASSESSING RECENT WARMING USING INSTRUMENTALLY HOMOGENEOUS SEA SURFACE TEMPERATURE RECORDS

Zeke Hausfather[1,2], Kevin Cowtan[3], David C. Clarke[4], Peter Jacobs[5], Mark Richardson[6], and Robert Rohde[2]

[1]Energy and Resources Group, University of California, Berkeley, California, USA, [2]Berkeley Earth, Berkeley, California, USA, [3]Department of Chemistry, University of York, York, UK, [4]Independent Researcher, Montreal, Quebec, Canada, [5]Department of Environmental Science and Policy, George Mason University, Fairfax, Virginia, USA [6]NASA Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA USA.

Science Advances 3 (1), 2019

ABSTRACT

Sea surface temperature (SST) records are subject to potential biases due to changing instrumentation and measurement practices[1]. Significant differences exist between commonly-used composite sea surface temperature reconstructions from NOAA's Extended Reconstruction Sea Surface Temperature (ERSST)[2], the Hadley Centre SST data set (HadSST3)[3], and the Japanese Meteorological Agency's Centennial Observation-Based Estimates of SSTs (COBE-SST)[4] from 2003 to present. The update from ERSST version 3b to version 4 resulted in an increase in the operational SST trend estimate during the last 18 years from 0.07°C/decade to 0.12°C/decade, indicating a higher rate of warming in recent years[5]. Here we show that ERSST version 4 trends generally agree with largely-independent, near-global and instrumentally-homogeneous SST measurements from floating buoys, Argo floats, and radiometer-based satellite measurements that have been developed and deployed during the past two decades. We find a large cooling bias in ERSSTv3b and smaller but significant cooling biases in HadSST3 and COBE-SST from 2003 to present with respect to most series examined. These results suggest that reported rates of SST warming in recent years have been underestimated in these three datasets.

One Sentence Summary:

Instrumentally homogenous SST records show a cooling bias in composite SST products and validate NOAA's recent record revision.

INTRODUCTION

Accurate SST data are necessary for a wide range of applications, from providing boundary conditions for numerical weather prediction, to assessing the performance of climate modeling, to understanding drivers of marine ecosystem changes. However in recent years SST records are subject to large inhomogeneities due to a dramatic increase in the use of buoy-based measurements and changing characteristics of ships taking measurements[3,6]. Prior to the past two decades, a large majority of SST measurements were taken by ships, first with buckets thrown over the side and increasingly through engine room intakes (ERI) after 1940. Since 1990, the number of SST measurements coming from buoys has increased around 25-fold, while the number of observations from ships has fallen by around 25 percent[1,7]. The observations have gone from 80% ship-based in 1990 to 80% buoy-based in 2015. Modern ship-based measurements (primarily ERI, though hull contact sensors and other devices are also used) tend to be biased warm by around 0.12°C relative to buoys, whose sensors are directly in contact with the ocean's surface[2,3,8]. As the number of ships actively taking measurements available in the International Comprehensive Ocean-Atmosphere Data Set (ICOADS) database[7] has fallen, a growing portion of ships are also using non-ERI systems that may introduce further changes in the combined record[3]. While buoy records are widely considered to be more accurate than ship-based measurements, their integration with ship records into longer SST series poses a number of challenges[1].

NOAA's Extended Reconstruction Sea Surface Temperature (ERSST)[2], the Hadley Centre SST data set (HadSST3)[3], and the Japanese Meteorological Agency's Centennial Observation-Based Estimates of SSTs (COBE-SST)[4] are composite SST series that assimilate data from multiple different instrument platforms (ships and buoys from ICOADS, and some satellite data in the case of COBE-SST) and measurement methods (wood buckets, canvas buckets, engine intake valves, etc.) to create consistent long-term records. These three composite ocean SST series are used by the primary groups reporting global temperature records: NASA's

GISTEMP[9], Met Office Hadley Centre's and University of East Anglia's Climatic Research Unit's HadCRUT[10], NOAA's GlobalTemp[11,12], the Japan Meteorological Agency[13], Berkeley Earth[14], and Cowtan and Way[15]. As the oceans cover 71 percent of the Earth's surface, changes to SST series have large impacts on the resulting global temperature records.

ERSST was recently updated from version 3b to version 4 (ERSSTv4), adding corrections to account for the increasing use of buoy measurements and incorporating adjustments to ship-based measurements based on Nighttime Marine Air Temperature (NMAT) data from the Met Office Hadley Centre and National Oceanography Centre's HadNMAT2[2,16-18]. ERSSTv3b did not include any SST bias adjustments after 1941, while ERSSTv4 continues these adjustments through present. Although the largest changes to the ERSST record occurred in the WW2-era, v4 also resulted in a higher rate of warming after 2003. This led Karl et al.[5] to conclude that the central estimate of the rate of global mean surface temperature change during the 1998-2012 period was comparable to that of the 1951-2012 period, in contrast to the IPCC characterization of the recent period as a 'hiatus'[19]. These updates also created a notable divergence between ERSSTv4, HadSST3, and COBE-SST in the period from 2003 to present, and raise the question of which composite SST series provides the most accurate record in recent years.

Over the past two decades, reasonably spatially-complete, instrumentally-homogeneous sea surface temperature (IHSST) measurements are available from drifting buoys, Argo floats[37], and satellites (see Methods for the details of each IHSST series). To assess how well the composite SST records correct for biases due to changing instrumentation, we compare each of them in turn to IHSST series created using only drifting buoys, only Argo floats, and only satellite infrared radiometer data. Because these IHSST series are created from relatively homogeneous measurements from a single type of instrument, they should be less subject to bias due to changing measurement methods, though other factors such as differences in spatial coverage or instrumental drift (in the case of satellites) need to be carefully accounted for.

Each of the three IHSST series (buoys, Argo floats, and satellites) span a different period of time. Buoy data have reasonably complete spatial coverage of the oceans from the late 1990s through present. Argo floats achieve sufficient coverage for analysis from January 2005, while reliable satellite data spans 1996 to present. Two sources of infrared radiometer-based satellite sea skin temperature are considered: the ARC SST product[23] from Along Track Scanning Radiometers (ATSR) data, which only provide data through the end of 2011, and the European

Space Agency Climate Change Initiative experimental record (hereafter CCI)[29] which combines ATSR and Advanced Very High Resolution Radiometer (AVHRR) data to obtain a continuous record for the whole period. The experimental version of the CCI record is not strictly instrumentally homogenous and is not fully independent from in situ SST observations, but closely matches the independent ARC SST record during the period of overlap; the next official release of the CCI containing AVHRR and ATSR data should be fully independent of in situ observations. Three different Argo-based near-surface temperature datasets are examined from the Asia-Pacific Data Research Center (APDRC)[32], the Japan Agency for Marine-Earth Science and Technology (hereafter H2008)[33,34], and Roemmich and Gilson (hereafter RG2009)[22], with a number of different datasets chosen to reflect the uncertainty introduced by attempting to reconstruct near-SSTs using Argo data.

RESULTS

From 1997 to present, ERSSTv3b has the lowest central trend estimate of the operational versions of the four composite SST series assessed, at 0.07°C per decade. HadSST3 is modestly higher at 0.09°C per decade, COBE-SST is 0.08°C per decade, while ERSSTv4 shows a trend of 0.12°C per decade over the region of common coverage for all four series. We find that ERSSTv3b shows significantly less warming than the buoy-only record and satellite-based IHSSTs over the periods of overlap ($p < 0.01$, using an ARMA[1,1] model to correct for autocorrelation), as shown in Figure 1. While v3b is comparable to v4 and the buoy and satellite records prior to 2003, notable divergences are apparent thereafter.

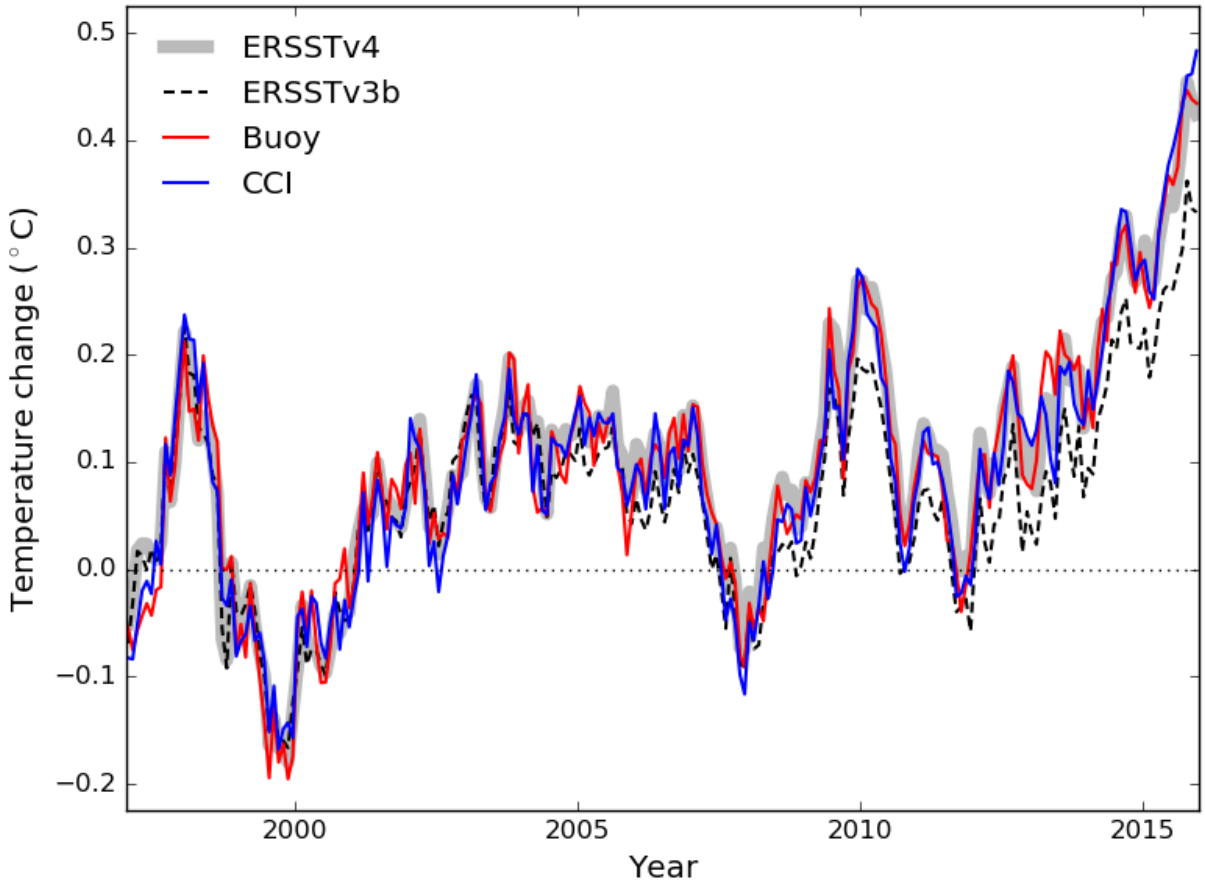


Figure 1: *Comparison of the different ERSSTv3b, v4, buoy-only, and CCI SST monthly anomalies during the period from January 1997 through December 2015, restricting all series to common coverage. ERSSTv4 is shown as a broad band for visualisation purposes; this band does not represent an uncertainty range. The series are aligned on the period 1997-2001 for comparison purposes. Spatial trend maps are also available in Figure S1, and a similar comparison with Argo data is shown in Figure S2.*

Both the buoy-only and CCI series are very similar to ERSSTv4 during their respective periods of overlap; trends in differences are insignificant in all cases. This strongly suggests that the improvements implemented in ERSSTv4 removed a cooling bias in ERSSTv3b. The ERSSTv4 record is expected to show good agreement with the collocated buoy record, because of new ship-buoy bias corrections and the increased weight attached to buoy observations in ERSSTv4. Thus this agreement represents a replication of the ERSSTv4 result from the same data using a substantially different methodology. The CCI data are not used in the ERSSTv4 record, and therefore represent an independent validation of the ERSSTv4 record.

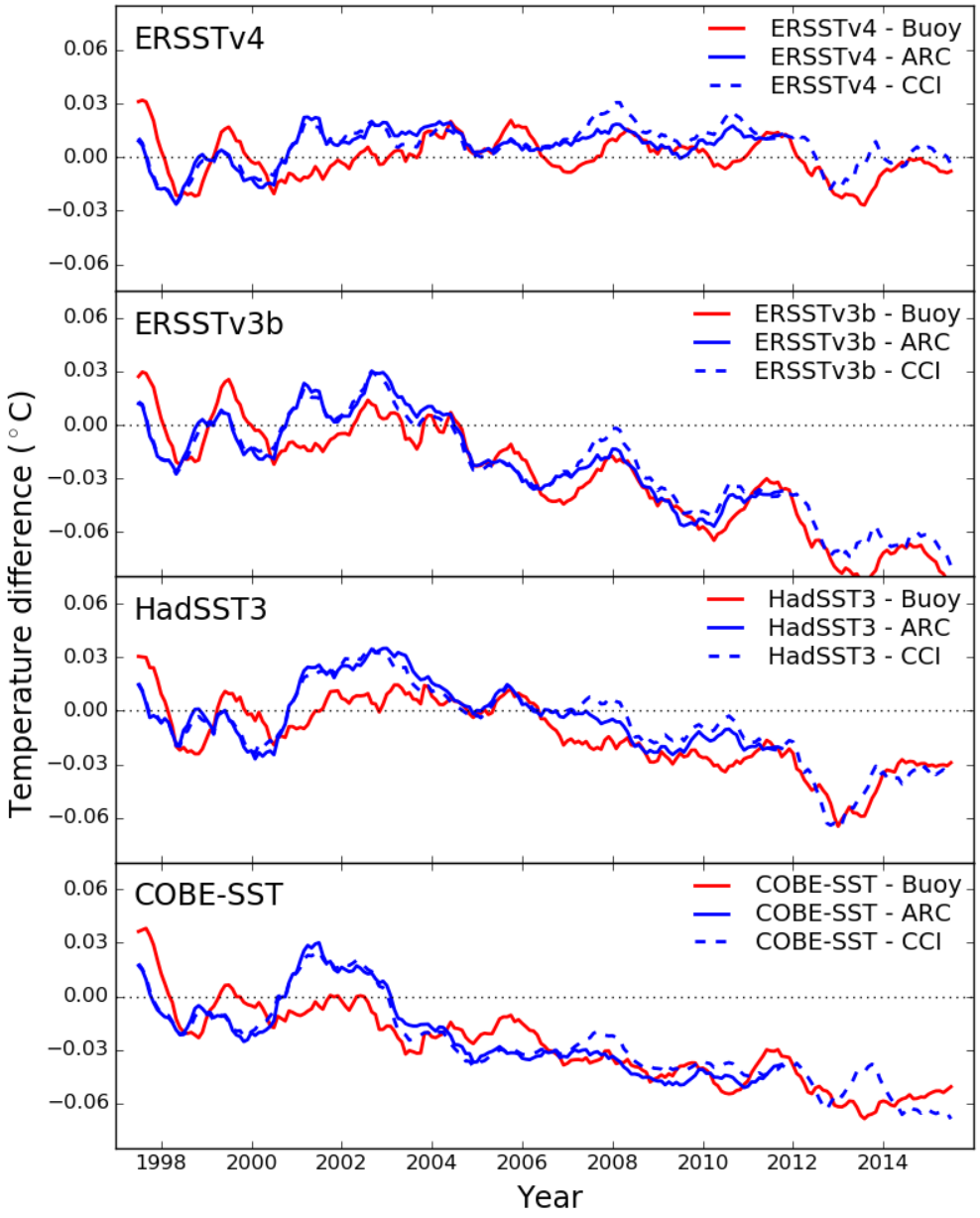


Figure 2: *12-month centered moving average of temperature difference series between composite and buoy-only, CCI, and ARC SST anomalies. Values below zero indicate that the composite series has a cool bias relative to the IHSST record.*

In addition to ERSST, we also examine how the other two commonly-used composite sea surface records, HadSST3 and COBE-SST, compare with the buoy-only and satellite-based IHSST records. Both show significant cool biases in the period from 2003-present relative to the buoy-only record, though the magnitude of this cool bias is smaller than that found in

ERSSTv3b. Difference series between all four composite records and the buoy-only and satellite-based IHSST records are shown in Figure 2. Each difference series is constructed by restricting all four composite SST series to common grid cells for each month, and comparing all grid cells where the composite records and the IHSST in question have data available. Our conclusions are similar when we consider all-product common coverage or interpolating products to global coverage; details of the spatial coverage approach and uncertainty calculations can be found in the Methods.

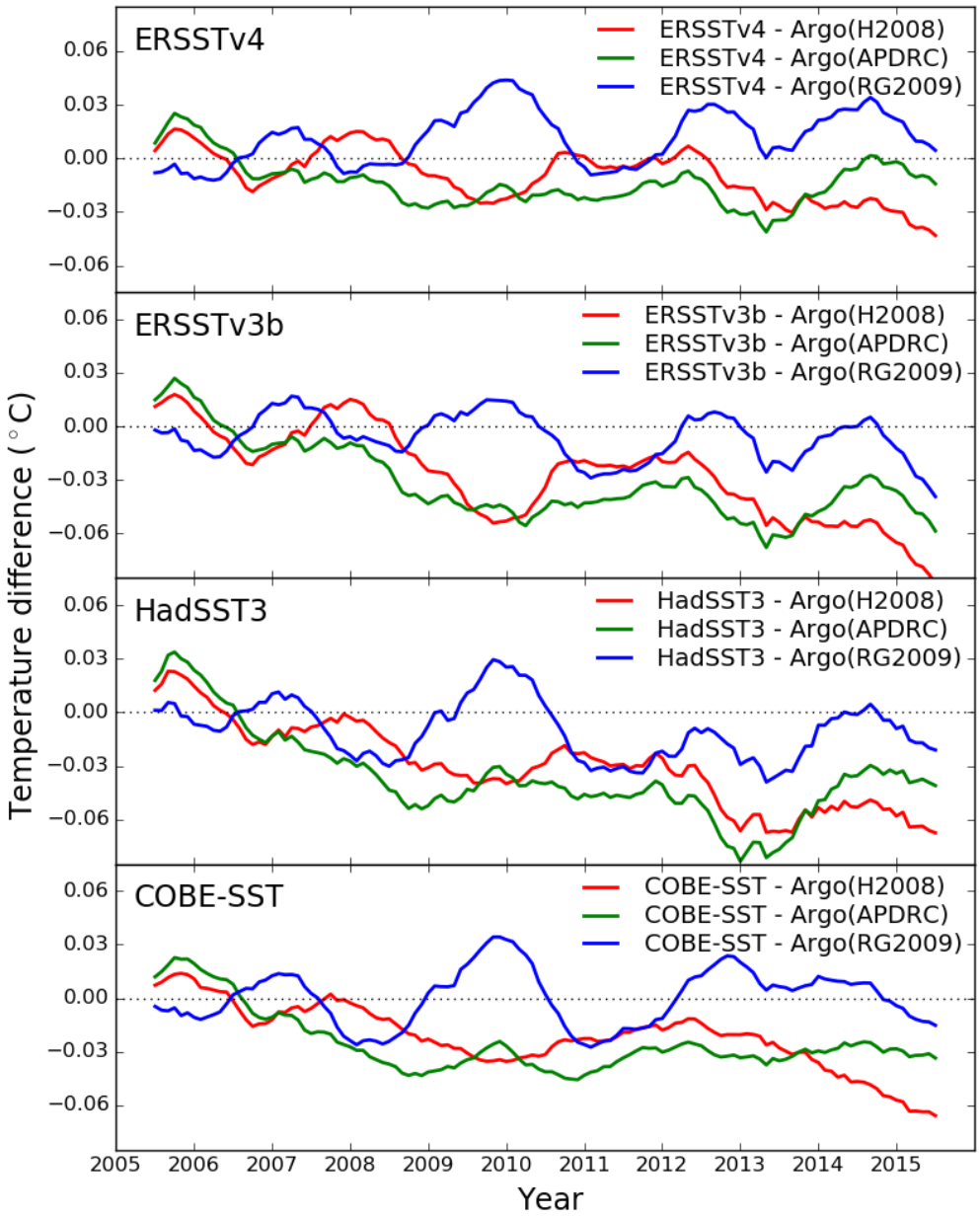


Figure 3: *12-month centered moving average of temperature difference series between composite and Argo near-SST anomalies.*

Two of the three Argo near-SST records assessed, APDRC and H2008, agree well with the buoy-only and satellite-based records and suggest a cool bias in ERSSTv3b during the 2005-2015 period when sufficient Argo data are available (Figure 3). The RG2009 series is more ambiguous, with trends that are not significantly different ($p > 0.05$) from either ERSSTv3b or v4. Similarly, while APDRC and H2008 both suggest cool biases in HadSST3 and COBE-SST, RG2009 does not show a significant trend in the difference series with the any of the composite temperature records (see Figure 4). Differences between the Argo series emerge through different interpolation techniques and additional data incorporation; APDRC use Aviso satellite altimetry for sea surface height estimates, H2008 use a small amount of data from the Triangle Trans-Ocean Buoy Network and conductivity-temperature depth profilers (mostly prior to 2005)[34], while RG2009 relies solely on Argo data.

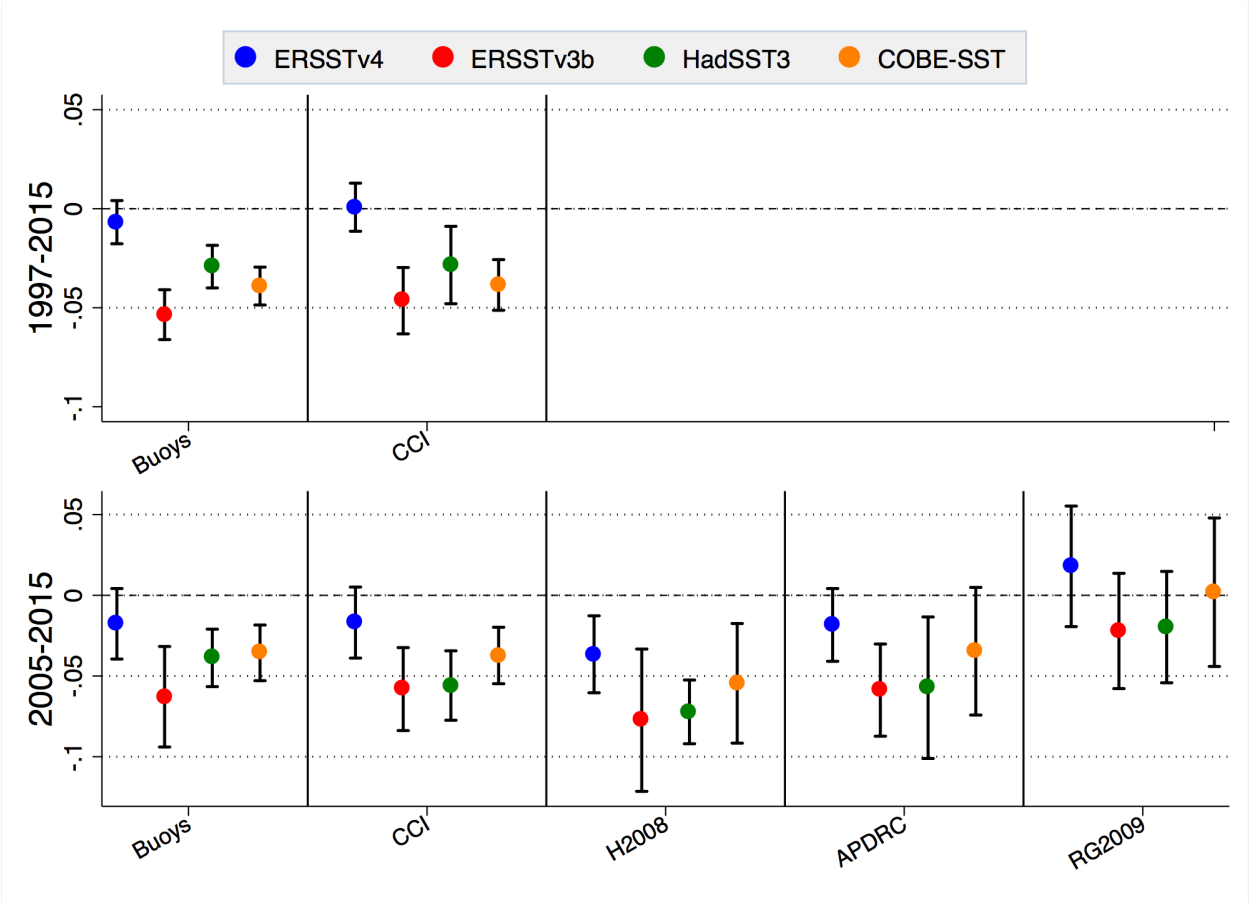


Figure 4: *Trends and 95% confidence intervals (degrees °C per decade) in difference series for each IHSST and composite SST series, masked to common composite SST coverage. Each*

difference series represents a composite series minus a IHSST series. Confidence intervals for trends are calculated using an ARMA[1,1] autocorrelation model. Values below zero indicate that the composite series has a lower trend than the IHSST series over the period examined. The two trend periods examined are Jan. 1997 through Dec. 2015 and Jan. 2005 through Dec. 2015.

To assess the significance of differences between composite series and IHSSTs, we examined whether trends in differences between the datasets were statistically different from zero (i.e. $p < 0.05$), as shown in Figure 4. We looked at two periods: 1997-2015 (where buoys, CCI, and the four composite series have records), and 2005-2015 (buoys, CCI, three Argo series, and four composite series). When comparing ERSSTv4 to all six IHSSTs during both periods, there are no significant trends in differences between the datasets except in the case of H2008, which showed slightly greater warming over the 2005-2015 period. ERSSTv3b, HadSST3 and COBE-SST all show a significantly lower warming trend over the period since 1997 compared to the buoy-only and CCI records (ARC SST shows nearly identical trends to CCI during its period of coverage from 1997-2012, as shown in Figure S3). During 2005-2015, ERSSTv3b, HadSST3, and COBE-SST have significantly lower trends than the H2008 Argo record, and ERSSTv3b and HadSST3 have significantly lower trends than the APDRC Argo record. For the RG2009 Argo record, no significant trend difference can be found for any of the composite temperature series during 2005-2015.

Both ERSSTv4[16] and HadSST3[3] incorporate detailed assessments of fully correlated (parametric) and partially correlated (sampling and measurement) uncertainties into their respective composite SST series. ERSSTv4 assesses these combined “bias” uncertainties via an ensemble of SST reconstructions incorporating a range of parametric setting combinations, most recently in an expanded 1000-member ensemble[17]. HadSST3 provides a 100-member ensemble to assess parametric uncertainty, but treats sampling and measurement uncertainty separately. We derived a 1000-member ensemble from the HadSST3 ensemble, with each member expanded to 10 members by adding an AR1 time-series with standard deviation and autocorrelation scaled to match the missing partially correlated uncertainty. We repeat the buoy-only and CCI IHSST comparisons on each of the realizations masked to common coverage (Figure 5).

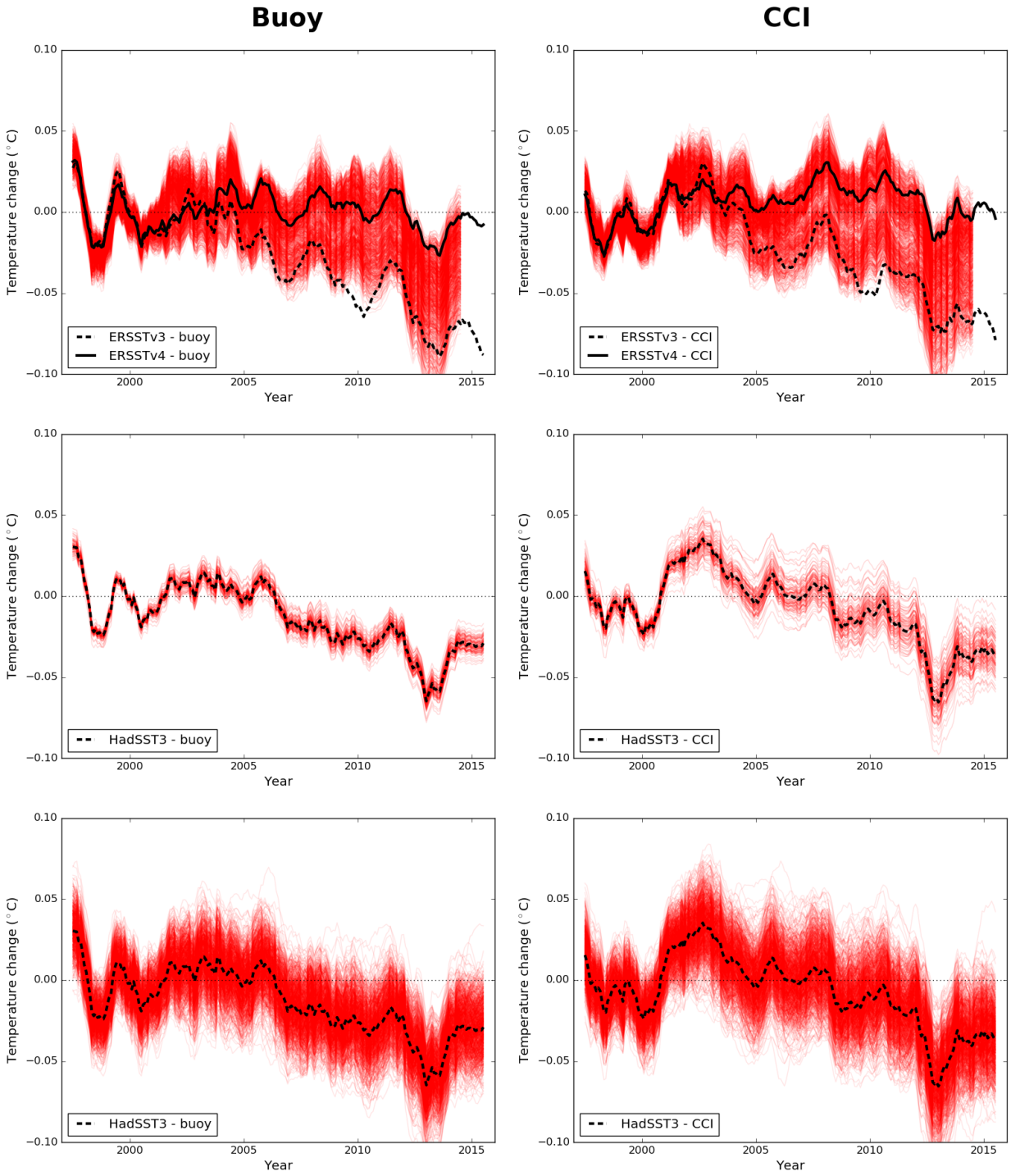


Figure 5: *12-month centered moving average of temperature difference series between collocated ERSSTv4/HadSST3 ensemble realizations and IHSST anomalies. The left column contains the difference series with the buoy-only record. The right column contains the difference series with the CCI record. The top row shows 1000 ERSSTv4 ensemble members, with operational versions of ERSSTv3b and v4 highlighted in black (note that the ERSST*

ensemble runs only go through 2014). The middle row shows the 100 published HadSST3 ensemble members, with the operational version in black. The bottom row displays the 1000 expanded ensemble members as discussed in the text.

The ERSSTv4 ensemble is not symmetric around the operational "best" estimate, which is based on the most empirically justified combination of parameter settings[2]; the majority of realizations have lower trends, with the lower bound of the ensemble encompassing ERSSTv3b. Only 16 of the 1000 ERSSTv4 realizations have a trend greater than that of the buoy-only IHSST record. The HadSST3 ensemble, in contrast, is largely symmetric around the operational estimate, which is based on the median of the ensemble. All of the 100-member and 1000-member HadSST3 ensemble realizations have lower trends than the buoy-only record. The increased spread of the difference between the HadSST ensemble members and CCI compared to the corresponding differences with the buoy record may arise from the interaction of the greater regional variability in the difference between HadSST and CCI coupled with the time varying coverage of HadSST.

The structural uncertainty in the buoy record can be estimated by comparing two subsets of the buoy data, and is about 0.05°C in 1997, dropping to 0.027°C for the period 2005-2015 (Figure S4) as the number of observations increases. The structural uncertainties estimated using equation (8) (in Methods) from an intercomparison of the IHSST records are 0.024°C, 0.020°C and 0.012°C for the buoy, Argo-H2008 and CCI records respectively. The structural uncertainties in the *trends* over the period 2005-2015 using equation (10) are 0.012°C per decade, 0.014°C per decade and 0.009°C per decade for the buoy, Argo-H2008 and CCI records. If the Argo-RG2009 data are used in place of the Argo-H2008 data, the trend uncertainties are 0.014°C per decade, 0.020°C per decade and 0.012°C respectively, representing a small increase in the uncertainties for the buoy and CCI records and a larger increase in the uncertainty for the Argo data.

The trend uncertainties estimated from equation (8) are very similar to the uncertainty of 0.013°C per decade estimated from the ERSSTv4 1000 member ensemble. This represents a useful validation of the ERSST ensemble, because the methods are independent: the ERSST ensemble relies on a bottom-up estimation of uncertainty from the different uncertainties in the methodology, while equation (8) yields a top down estimate based on the differences between independent data sources. The trend uncertainties estimated from equation (8) are 10-20% of

the linear trend uncertainties in the corresponding temperature trends, which include the effect of internal variability. The uncertainties are based on the region of common coverage, and inclusion of poorly sampled regions will increase the structural uncertainty. The limited time span means that uncertainties are somewhat determined by a few outliers in each temperature series, however the results show that linear trend uncertainty should not be used as an estimate of the structural uncertainty in the trend.

The resulting difference series and trends in the all of the Figures will differ modestly based on how spatial coverage is handled. For each IHSST difference series, we restrict coverage for each month to that common between the IHSST series in question and the four composite records. This serves to maximize the spatial overlap between the datasets and provide a more accurate global estimate of differences for each individual IHSST, but also results in difference series and trends that are not strictly comparable between IHSSTs due to coverage differences. This is particularly pronounced in the 1997-2005 period, when the buoy-only record has less coverage than the more spatially-complete ARC and CCI satellite radiometer-based records. Some coverage differences also arise in the 2005-2015 period between Argo-based records and buoy/CCI records, as Argo data are largely unavailable north of 60°N, south of 60°S, or in the Malay Archipelago.

To ensure that our results are robust to the choice of how spatial coverage is handled, we performed two additional tests to account for both spatial and temporal-spatial consistency across series. In the first test, we restricted all series examined for the two time periods in question (1997-2015 and 2005-2015) to only 1x1 lat/lon grid cells containing records from all series examined over that timeframe. During the 1997-2015 period we only looked at grid cells with common coverage across the four composite series, buoys, and CCI, while during the 2005-2015 period we examined only grid cells with common coverage between the composite series, buoys, CCI, and all three Argo-based series. This results in a record that is less spatially complete for any given IHSST-composite series comparison, but is strictly comparable between IHSSTs. Difference series and trends for this common coverage approach are shown in Figures S5-S7. Results are largely comparable to those in the main paper, with a slightly higher trend in CCI difference series during the 1997-2015 period and lower CCI trend during the 2005-2015 period as the only notable difference.

In the second coverage test, we applied a kriging spatial interpolation approach the two series (buoys and HadSST3) that contain large gaps in spatial coverage for all months to create fully spatially and temporally-complete records (the three Argo series and the other three composite series have their own interpolation provided, while satellite records are largely spatially complete apart from high latitudes). We then restricted all series to common coverage over the 1997-2015 and 2005-2015 periods following the approach of the common coverage test. This introduces some additional uncertainty due to the kriging, but ensures that the spatial coverage represented by the difference series and trends does not change from month to month, and that all series have nearly complete coverage over the period of overlap. The results for the kriged series are shown in Figures S8-S10. Here the cooling bias in ERSSTv3b, COBE-SST, and HadSST3 is more pronounced with respect to the buoy-only and CCI records, though the overall results are comparable. Interpretation of the Argo records is largely unchanged for any of the spatial coverage approaches examined.

In addition, the collocated buoy and CCI records show a spatial disagreement which is not apparent in Figures 2 and 4, which is only apparent when the CCI coverage is reduced to match the buoy coverage (see Figures S11 and S12). This arises from regional differences between the CCI record and other records, particularly before 2001. CCI shows greater warming than ERSSTv4 in the Southern Ocean, but less in the northern mid-latitudes. The Southern Ocean is consistently cloud-covered, so CCI might be expected to be less accurate in these regions. Winds can also affect skin temperature retrievals relative to those at depth. In situ observations are prevalent in the northern hemisphere, and so may be more reliable. In the Southern Ocean in-situ observations are sparse and so temperature trends remain uncertain. The regional deviations from the in-situ records and their impact on trends mean that comparisons with CCI should be treated with care.

Coverage biases may also be are also impacted by the choice of baseline for geographical map series. The results presented use a nineteen year 1997-2015 baseline for both the ERSSTv3 data to which the other series are then matched, and for the high resolution climatology used in constructing the buoy record. Changing either of these to a 30 year 1986-2015 baseline has no perceptible effect on the results.

DISCUSSION

Trends in IHSSTs constructed from buoy and satellite data agree with ERSSTv4 over the period from 1997-2015 but are significantly higher ($p<0.01$) than the ERSSTv3b trend, supporting the conclusions of Karl et al.[5]. Both buoys and satellites also suggest a significant ($p<0.05$) cooling bias in HadSST3 and COBE-SST. Over 2005-2015, four of five IHSST series agree with ERSSTv4 or suggest that it might be slightly cold-biased. By contrast, four of five suggest cool biases in both ERSSTv3b and HadSST3, while three of five suggest a cool bias in COBE-SST. One of the three Argo series (RG2009) is statistically indistinguishable from all four of the composite SST products during the 2005-2015 period.

The difference in IHSST records relative to HadSST3 are particularly noteworthy, as HadSST3 includes explicit buoy-ship offset adjustments comparable to those used by ERSSTv4 and continues ship SST corrections through present[3]. The source of the apparent cooling bias in recent years in HadSST3 is unclear, though it is likely related to biases in ship records introduced by the changing composition of shipping fleets and a general decline in the number of available ship-based SST measurements[7]. When comparing IHSSTs to a ship-only SST record (restricting to common coverage), we identify a strong cool bias in the ship record, particularly since 2010. Not only are ship temperatures lower than the buoys at the start of the period of study (due to a ~0.1 °C offset), but the ship record substantially underestimates the rate of warming over the later part of the period (Figure S13). This result is supported by the satellite observations of skin temperature, the buoy measurements in the top meter of the ocean, and Argo observations from three different methodologies over depths spanning 2.5m to 20m (Figure S14). ERSSTv4 mostly avoids this potential bias in ship records by assigning an increased weight to buoys in recent years,[2] though the slightly higher trends in buoys, CCI, and two of the three Argo series vis-a-vis ERSSTv4 over 2005-2015 (Figure 4, lower panel) might be driven by some residual ship-related bias.

The difference in trend between ERSSTv3b and ERSSTv4 is smaller than the difference in trend between the buoy and ship records, as ERSSTv3b also incorporates data from buoys but does not account for the offset between the ship and buoy temperatures or assign the buoys more weight than ship-based measurements. HadSST3 falls between the two versions, incorporating an offset adjustment between ships and buoys and some corrections to the ship observations but weighting ships and buoys equally. Nighttime Marine Air Temperatures (HadNMAT2), which are used as part of the ERSSTv4 homogenization, also appear to show a

cool bias comparable to if not larger than that of HadSST3 relative to the IHSSTs in the period after 2003 (Figure S15), possibly due to residual inhomogeneities in NMAT records. While COBE-SST is also significantly cooler in recent years than the buoy-only record and CCI, a new version (COBE-SST2) incorporates buoy adjustments and shows better agreement with the IHSST records, but does not extend up to present and is not yet in operational use in the JMA global land/ocean temperature product (Figure S16)[20].

Interpreting the Argo Results

The Argo records cover a shorter period (11 years rather than 19), and their results are less clearcut than the buoy and CCI IHSSTs. The H2008 and APDRC records support ERSSTv4 (and even suggest that it might be a bit too cool), though APDRC results are somewhat sensitive to the choice of start year (Figure S17). RG2009 falls between ERSSTv3 and ERSSTv4 in trend and does not reject either. Similarly, while H2008 and APDRC suggest a cool bias in HadSST3 and (to a lesser extent) in COBE-SST over the 2005-2015 period, the results of RG2009 are ambiguous and do not allow any differentiation between composite record trends.

The shortness of the Argo records and their divergence limits the weight which can be placed on them. If the faster warming H2008 and APDRC records are accurate, than all of the IHSSTs (buoys, satellites, and Argo floats) are in basic agreement in rejecting the slower warming ERSSTv3 record. However if the slower warming RG2009 record is correct, this would imply that either the buoy and CCI IHSSTs are too warm over 2005-2015 or that there may be a variation in temperature trend with depth: the skin record and the top meter show faster warming, while the deeper ship and Argo records show slower warming. Different observational platforms sample sea "surface" (or near-surface) temperature at different depths in the mixed layer, with satellites, buoys, ships and Argo floats observing the temperature at increasing depths. If H2008 or APDRC records are more accurate, it seems unlikely that depth plays a role in the differences between temperature trends, since the slower warming ship record is bracketed in depth by the satellite/buoy records and the Argo records. This would also suggest that measurement depth does not explain any part of the slower warming found in the ship record. If the RG2009 record is correct, however, it may suggest that the slower warming ship record arises from a combination of both depth and the bias in the ship record (since the ship record exhibits less warming than even RG2009, as shown in Figure S14).

Argo instruments have temperature profiles at depths throughout the mixed layer (and below), with the shallowest observations in any of the Argo products in the 2.5-7.5 meter range. While the Argo records show no discernable reduction in trends between 5 meter, 10 meter, and 20 meter depths (Figure S18), the Argo record cannot exclude a difference with the top meter measured by the buoys. If there is a significant difference in temperature trend between the top meter and the remainder of the mixed layer, this would present a problem in the construction of a homogeneous sea surface temperature product from the combination of ship and buoy records. Similarly, the majority of CMIP5 climate models have a top layer spanning 0-10 meters, and so may not resolve the top meter of the ocean. This could present a challenge in both testing for the depth effect in models, and in the comparison of models to observations. However, as two of the three Argo-based records analyzed show no significant difference with buoy and CCI surface records and the Argo series is rather short, any conclusions about depth-related effects appear to be premature.

CONCLUDING REMARKS

Adjustments to correct for inhomogeneities in SSTs in recent years have a large impact on the resulting decadal-scale global temperature trends. Assessing the effectiveness of these adjustments is critical to improving our understanding of the structure of modern climate changes and the extent to which trends in recent periods may have been anomalous with respect to longer-term warming. Using independent instrumentally-homogeneous SST series, we find that NOAA's new ERSSTv4 effectively corrects a significant cooling bias present in ERSSTv3b during the past two decades without introducing any detectable residual trend bias. We also conclude that two other widely-used composite SST series, HadSST3 and COBE-SST, likely suffer from spurious cooling biases in recent years present in ship-based records.

Some uncertainty remains, particularly in Argo-based near-SST reconstructions. While two of the three Argo reconstructions examined agree well globally with the buoy and radiometer-based IHSSTs, the third does not allow for any effective differentiation between composite SST series. Similarly, while CCI and ARC-SST radiometer-based estimates agree quite well with the buoy-only record globally, there are significant zonal differences. The time period considered is relatively short, with most of the divergence between composite SST records occurring after

2003, and sufficient Argo data is only available subsequent to 2005. Nonetheless, SST time series from drifting buoys and independent satellite radiometers support ERSSTv4 and suggest a cool bias in other series such as ERSSTv3b, while two of three Argo series agree with ERSSTv4 and one suggests that it is too cool. Overall these results suggest that the new ERSSTv4 record represents the most accurate composite estimate of global sea surface temperature trends during the past two decades, and thus support findings[15] that previously reported rates of surface warming in recent years have been underestimated.

METHODS

We compare composite SST records including ERSSTv3b, ERSSTv4, HadSST3, and COBE-SST to three separate IHSST records constructed from ICOADS-reporting buoys, near-surface measurements from Argo floats, and radiometer-based satellite SST records. We obtain existing spatially gridded fields for each SST series (and create novel ones in the case of buoy-only and ship-only records), convert each to standardized 1 degree latitude by 1 degree longitude uniform grid (hereafter 1x1 degree grid).

Temperature averaging in the presence of varying geographical coverage requires that all of the temperature series be aligned on a common baseline. It is common practice to apply an offset to each cell and month of the year to bring the mean of that cell and month to zero over a 30 year baseline period; however, this is impractical for the short buoy record. Fixing the baseline for an incomplete record is problematic in the case where the months for which observations are present are unusually hot or cold, however the problem may be addressed by aligning the data against a more complete record containing the same weather signal. The spatially complete ERSSTv3b record was therefore aligned to zero on the period 1997-2015, and then the other datasets are aligned to the normalised ERSSTv3b map series. This method is a conservative choice in attempting to detect a bias in the ERSSTv3b record, as it may bias the compared series slightly towards it.

Data series are carefully aligned to ensure accurate intercomparisons of SST series. The process is as follows: OISST is used to construct a high resolution daily climatology on the baseline period (1997-2015) - yielding 365 fields, one for each day (leap days are also treated). The buoy series is then calculated using this high resolution daily climatology, yielding 228

monthly fields (19 years * 12 months). ERSSTv3b is also aligned to the 1997-2015 baseline. All of the composite series and IHSSTs (including the buoy series) are then aligned to the baselined ERSSTv3b based on whatever months are available for each grid cell. These are then masked to common coverage and plotted in Figure 1. This makes use of the spatial completeness of ERSSTv3b to avoid artifacts due baselining temporally incomplete cells on an incomplete baseline period; we use ERSSTv3b for this purpose to avoid biasing our results towards ERSSTv4. Pairwise difference map series are calculated between the aligned maps. The study is restricted to the period 1997-2015, with the start date determined by buoy coverage and a data break in the ATSR-based SST data. Details of how each dataset was obtained and processed are provided below.

ERSST, HadSST and COBE-SST

Both ERSST v3[11] and v4[12] are produced on a 2x2 degree grid, with sea ice cells recorded as -1.8°C. The ice cells were set to missing, and then the data were expanded to a 1x1 degree grid, repeating each value from the original grid to the 4 corresponding cells in the finer grid. HadSST3[3] is produced on a 5x5 degree grid with no values for sea ice cells, and is expanded to the 1x1 degree grid by repeating each value from the original grid to the 25 corresponding cells in the finer grid. COBE-SST[4] and COBE-SST2[20] are distributed as a 1x1 gridded product; cells with sea ice are recorded as -1.8°C similar to ERSST, and were set to missing. As both HadSST3 and ERSSTv4 include ensembles of realizations with different parameterizations, for the main analysis in the paper (e.g. Figures 1-4) the operational version of each series was used. This is the ensemble median in the case of HadSST3, while ERSSTv4 provides a preferred realization.

Different approaches are used in the construction of the gridded SST products. In the HadSST record, observations only contribute to the grid cell and month in which they occur, leading to some cells for which no temperature estimate is available. In the COBE-SST records, optimal interpolation is used in both space and time to create a spatially complete field from the available data. The ERSST and COBE-SST2 datasets combine a low resolution reconstruction with the fitting of empirical orthogonal teleconnections (EOT) to the observations to produce a spatially complete field in which local temperatures can be inferred from distant observations (up to a specified distance) through teleconnections. All the records include data from ICOADS (albeit some from different releases of the database), however in addition to differences in the

processing methods, ERSSTv4 attaches an increased weight to buoy observations on the basis of their lower estimated uncertainty.

Because some of the composite SST series include interpolation of observations into proximate grid cells with missing data, all composite SST series were restricted to grid cells common to the HadSST3, ERSSTv4, and COBE-SST datasets for any given month. Since HadSST3 includes no explicit interpolation (apart from that implicit in its use of relatively large 5x5 degree grid cells), this should remove any differences between series due to interpolation. Failing to account for interpolation could lead to difficulty in cross-comparison of difference series between IHSST and different composite SST records.

Buoys

The buoy data are from the ICOADS release 2.5 data[7]. Drifting buoys are selected by the World Meteorological Organization buoy identifier and the presence of a value in the SST field (thus excluding Argo buoys with WMO identifiers). Moored buoys were excluded from the analysis due to an offset in temperatures between drifting and moored buoys (perhaps due to measurement depth; see Figure S19), which would introduce a bias as the proportion of moored and drifting buoys changes over the period of interest. A large majority of measurements in recent years come from drifting rather than moored buoys, and the use of drifting buoys only has no major impact on the results. The temperature field is determined by averaging buoy observations over the span of a month for each cell in a global grid. The grid consists of cells of equal area, with equatorial cells spanning a range of 5 degrees in both longitude and latitude. At higher latitudes the longitudinal width of a cell in degrees is increased by calculating the area of the latitude band, dividing by the area of a 5x5 cell at the equator, and using that many cells in the latitude band to maintain a constant area.

The data are processed one month at a time. For each buoy, data are divided into days. The (typically hourly) temperature, latitude and longitude data for that day are averaged. Buoys which show temperature variations with a standard deviation exceeding 1°C or positional variation with a standard deviation exceeding 0.5° of latitude or longitude during a single day are excluded for the whole month: this can occur if a buoy is beached or picked up by a ship. The temperature is then converted to an anomaly using a climatology calculated from OISSTv2[21] for that day of the year and for the corresponding latitude and longitude on a finer ½ degree grid.

This mitigates the biasing effects of temperature observations at the beginning or end of a month or the northern or southern edges of a 5 degree latitude band. The daily mean temperature anomaly for the buoy is then added to a list for the corresponding grid cell. Once all buoy records have been processed, all temperature anomalies for a given cell are averaged to produce a final anomaly value for that cell.

This method for constructing the buoy-only temperature record was chosen for simplicity, with the aim of reducing the possibility of methodological artefacts such as infilling distorting the result: a consequence of this is that the resulting temperature reconstruction is limited to regions where observations are available. However simplicity does not in itself preclude bias: an overly simple method might for example fail to detect some faulty observations. This possibility will be addressed through internal consistency checks on the buoy data.

Another possible source of bias is mis-calibration of the temperature sensor, leading to systematically lower or higher readings. Normally these would contribute noise rather than a bias in the trends as the mis-calibrated buoy moves into more or less sampled regions and so receives a different weight in the temperature calculation. However if new buoys are introduced which are systematically different in calibration relative to older buoys, a bias in the trends could result. There is no sign of such a bias in the comparisons between different IHSSTs, and the cross-validated uncertainties are lowest for the recent period where the composite records show most difference.

Additional inter-buoy comparisons were performed to address this possibility. For each grid cell and month where at least 3 buoys contributed observations, a bias estimate was calculated from the difference between the mean anomaly for the buoy and the mean of the anomalies for all the remaining buoys in that cell. All the bias estimates for a buoy were collected, and buoys for which the magnitude of the mean bias or standard deviation of the bias estimates exceeded 1°C were eliminated, reducing the total number of buoys by about 10%. In a further test, the temperature record was recalculated applying the resulting bias adjustment to the readings from each buoy in turn.

Four versions of the buoy record were prepared to evaluate the potential impact of buoy biases, as follows:

(a) Using all of the data, omitting the test for daily variability..

(b) Filtering on the basis of daily variability only (the default per-buoy filter, described at the start of this section).
(c) Filtering on the basis of daily variability and inter-buoy variability (i.e. the additional filter described in the previous paragraph).
(d) Filtering on the basis of monthly and inter-buoy variability and application of the bias correction (as (c) but then recalculating the buoy record on the after applying a correction to each buoy on the basis of its mean difference with passing buoys).

The resulting temperature series are shown in Figure S20, along with the differences of the other methods from the default method. The largest difference arises from using all of the data without filtering for daily variability. Inter-buoy variability and bias correction make a rather smaller difference. The differences between the methods are small compared to the differences between the composite records. The default method using a per-buoy filter shows the lowest trend over 1997-2015, and is therefore a conservative choice.

The buoy coverage is limited, particularly in the 1990s, and comparisons to other datasets may be impacted by coverage bias. In order to produce an unbiased comparison to other datasets, all the datasets are expanded onto a 1x1 degree grid. Comparisons are made using only the cells for which the datasets being compared have values. The area weighted mean temperature is then calculated for each record using the common coverage cells. The percent of global ocean covered by buoy measurements varies from around 40% in the mid 1990s to around 70% in recent years.

Ships

The ship record is constructed in the same way as the buoy record, with one exception: many ships only report once per day, and from 2007 some ship identifiers are masked for security reasons (though this has been improved in Release 3 of ICOADS). The test to detect excessive motion or variation within a single day is therefore omitted. The only quality control applied to the ship record therefore arises from the calculation of the global mean of the sea surface temperature field, which excludes observations which fall in land areas. The ship observations are subject to significant quality issues, and the limited quality control implemented in this record therefore provides no more than a general indication of the presence and scale of any bias in the ship record.

Argo floats

Three different gridded Argo data provided online by the International Pacific Research Center (IPRC) Asia-Pacific Data Research Center (APDRC)[32], the Japan Agency for Marine-Earth Science and Technology (H2008)[33,34], and Roemmich and Gilson (RG2009)[22] were used. These data are produced on a monthly 1x1 degree grid, and have been smoothed and infilled by the data provider using a variational analysis technique to provide global coverage over all cells unaffected by seasonal sea ice. Sea surface height was used as part of the interpolation process in APDRC, while Cells containing sea ice were represented by missing data. The data did not require regridding, and were aligned to the ERSSTv3b data as described previously.

The RG2009 Argo product has temperature values at 2.5, 10 and 20 dbar levels and deeper, the H2008 product has temperatures at 10, 20, 30 dbar and deeper levels, and the APDRC product has temperature values at 0, 5, and 10 meter and deeper levels. We use the 5 meter level for the APDRC product, the 10 dbar (10 meter) level for the H2008 product, and the 2.5 dbar level for the RG2009 product (which represents measurements ranging from 2.5 to 7.5 dbar with a mean level of 5 dbar/meters) to provide the most comparable and highest available depths; estimated 0-meter temperatures from APDRC are not used as they result from interpolation (since no Argo floats sample sea skin temperatures).

Throughout the paper we refer to the record derived from Argo floats as “near-SST”, as the highest level of the ocean measured by most Argo floats is approximately 5 meters below the surface.[22] However, with the exception of satellite radiometer-based estimates, all of the instruments used in this analysis record ocean temperatures at depths between 0 and 20 meters. For example, ships tend to measure temperatures through engine room intake valves at depths of 7-11 meters for large ships and 1-3 meters for small ships.[1] Moored buoys typically measure SSTs at 3 meters of depth, while drifting buoys measure SSTs at around 0.5 meters. Recent work[35] found no long term difference in warming rates between 0-4 meter and 4-9 meter depths in a CMIP5 model; similarly we have established that our results are robust when using the next deeper level of each Argo dataset (Figures S19). The different depths sampled by the different observational systems provide a basis for assessing whether depth plays a role in the rate of recent warming.

Argo data have been used to create SST analogues in the past; for example, Roemmich and Gilson[30] compared ARGO "near"-SST to NOAA's OISST v1, while Roemmich et al[31] compared a 5-meter Argo-based SST record to OISST v2. Here we perform a similar analysis using the Argo-based fields provided by RG2009, APDRC, and H2008.

Satellites

The Along-Track Scanning Radiometer (ATSR) instruments provide infrared images of the earth from which skin temperatures may be derived. ATSR data are incorporated into two gridded datasets, the ATSR Reprocessing for Climate (ARC)[23] spanning the period from 1996 through 2012, and the experimental NCEO/ESA SST CCI Analysis L3S version EXP-1.2 (ESA-CCI or CCI)[29] which also incorporates data from the Advanced Very High Resolution Radiometers (AVHRR) and spans the period from 1996 through present (end of 2015). Coverage between 60S and 60N is largely complete (except for a few cells each month in the ATSR record which are affected by cloud, typically in the Southern Ocean or North Atlantic). Both the ATSR-only (through mid-2012) and ATSR + AVHRR (through present) CCI data were analyzed, and the CCI data are used in the paper as it extends through present (and differences between the two are minor during the period of overlap, as shown in Figure 6).

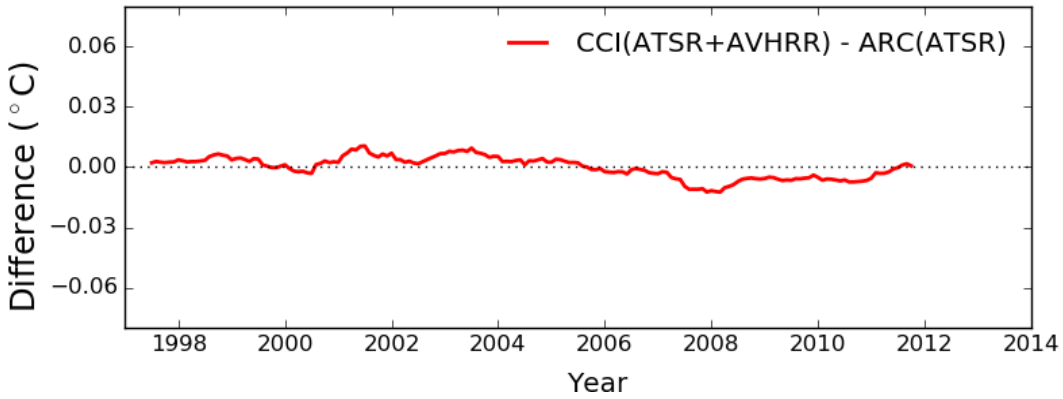


Figure 6: *12-month centered moving average of differences between CCI ATSR + AVHRR and ATSR-only ARC SST records during the period of overlap. The earlier instrumentally-homogeneous ARC SST shows small differences to the newer combined version, however the difference are minor compared to the differences relative to the composite SST records.*

Spatial Coverage

The main figures in the paper were generated by limiting difference series to common spatial coverage between the four composite SST series and the IHSST in question. For example, a difference series between ERSSTv4 and the buoy-only record would show the difference for all grid cells for each month where all four composite SST series and the buoy-only record had data available. The requirement that all four composite series share the same coverage is intended to remove the effects of interpolation on the results, as all rely on largely the same ICOADS data.

Two additional tests described in the discussion were undertaken to ensure that the results were robust to choices of how coverage was handled. In the first test the analysis was done for the two periods of interest (1997-2015 and 2005-2015) restricting the analysis to only grid cells where all series available for that period had coverage. During the 1997-2015 period, this means that only 1x1 lat/lon grid cells where the four composite series, the buoy-only record, and the CCI record all had coverage for any given month were used. During the 2005-2015, grid cells required coverage by the four composites, buoys, CCI, and all three Argo records to be used.

In the second test we created fully spatially and temporally-complete fields to control for both difference in coverage for any given time period as well as changes in coverage over time. Infilling is performed on the gridded data using the original grid sampling for that record: for the buoy record this is on the 550km equal area grid, and for the HadSST3 on the 5x5 degree grid. The resulting infilled field is then copied onto a 1x1 degree grid as before. Infilling is performed using the method of kriging,[38] by which the values at unobserved locations are inferred from the observed values. Each observation is weighted on the basis of distance from the target location using a variogram relating the expected variance between two grid cells to the distance between them, which is determined from grid cells for which observations are available, fitted with an exponential model controlled by a single range parameter which is the e-folding distance of the variance. The kriging calculation also uses the covariance between locations where observations are present to estimate the amount of independent information in each observation. The buoy record shows longer range autocorrelation that the HadSST3 data, with respective e-folding distances of 1400 km and 900 km, suggesting that the buoy record shows more spatial autocorrelation.

Infilled temperature observations will therefore be a weighted combination of the nearest observations if there are observations within a small multiple of the e-folding distance. Locations which are very distant from any observation will tend towards an optimal estimate of the global mean of the temperature field.

Uncertainty estimation

SST reconstructions include uncertainties due to limitations of both the data and the methods. Differences between reconstructions may arise due to random errors in the data or introduced during processing, or due to uncorrected biases in the observational data. Identification of a bias requires that the difference between reconstructions must be shown to be larger than can be accounted for by random errors alone. To that end we now examine different methods for the determination of the uncertainty in a reconstruction. Two approaches are used. Firstly, co-located temperature difference series are used to estimate the significance of the differences. Secondly a method is outlined for the use of independent temperature series to directly estimate the structural uncertainty in each series.

Significance of the temperature difference series trends

In order to assess the significance of differences in trends between temperature series, we first calculate the difference temperature series from the difference map series in order to eliminate differences in coverage. The trend in the difference series is then compared to the uncertainty in that trend estimated using an appropriate autoregression model, and used to determine whether the trend difference is significantly different from zero.

The trend in the difference series is identical to the difference in the trends between the two series, assuming both map series are reduced to common coverage. However calculation of the trend in the difference series offers a benefit when determining the uncertainty in that trend.[24] If the trend difference is calculated from the trends of the individual series, the uncertainty in the trend difference requires the determination of the covariance between the model residuals. The respective residuals contain common internal variability and so are strongly correlated, therefore the covariance term is positive. Omission of the covariance term leads to the uncertainty in the trend difference being dramatically overestimated. With the covariance term included, estimates

of the uncertainty in the trend difference from either the difference series, or from the two individual series, give identical results.

The difference series linear trends are estimated with Ordinary Least Squares (OLS) with correction of the standard error to account for serial correlation of the residuals[25-27]. The general approach is to estimate the effective sample length (and thus the effective degrees of freedom) from an estimate of the positive autocorrelation of the residuals:

$$n_e = n_t / (1 + 2 \sum_{j=1}^{n-1} \varrho_j) \quad (1)$$

where n_t is the original series length, n_e is the effective sample length and ϱ_j is the autocorrelation at lag j of an autoregressive (AR) or autoregressive moving-average (ARMA) noise model estimated from the OLS residuals. An ARMA(1, 1) model was used for global SST all gridded and global difference series (e.g. ERSSTv4 - buoys). The ARMA model coefficients were estimated with maximum likelihood for global series and Yule-Walker (moments) for gridded series trends. An ARMA(1, 1) series $X_t$with white noise series ϵ_t satisfies:

$$X_t = \phi X_{t-1} + \epsilon_t + \theta \epsilon_{t-1} \quad (2)$$

Then the autocorrelation function (ACF) of an ARMA(1, 1) series is given by:

$$\varrho_0 = 1$$
$$\varrho_1 = (\phi + \theta)(1 + \phi\theta) / (1 + 2\phi\theta + \theta^2)$$
$$\varrho_j = \varrho_1 \phi^{j-1}, \quad j >= 2 \quad (3)$$

where ϕ and $\theta$are the respective AR and MA coefficients.

Since the assessed trends cover only 11-19 years (132-228 months), a bias correction was also applied to the global difference series trends in order to account for the underestimate of autocorrelation in such short series[26,28]. The original Tjostheim and Paulsen correction for the AR(1) estimated coefficient ϕ is given by:

$$\phi_{bc} = \phi + (1 + 4\phi) / n_t \quad (4)$$

The bias correction of ARMA(1, 1) estimated ACF coefficients ϱ_1, ϕ generalizes (4) by also accounting for the positive difference between ϕ and ϱ_1. Note the AR(1) bias correction in (4) then becomes the special case where $\theta = 0$ and $\varrho_1 = \phi$.

$$\phi_{bc} = \phi + (1 + 4(2\phi - \varrho_1)) / n_t$$
$$\varrho_{1bc} = \varrho_1 + (1 + 4(2\varphi - \varrho_1)) / n_t \quad (5)$$

The ARMA coefficient estimates ϕ_{bc} and ϱ_{1bc} then can be substituted into the appropriate specific form of equation (1). The ARMA(1, 1) formulation in equation (3) can then be simplified[27]:

$$n_e = n / \left(1 + 2 \sum_{j=1}^{n-1} \varrho_{1bc} {\phi_{bc}}^{j-1}\right) \approx n/\left(1 + 2\varrho_{1bc}/(1 - \phi_{bc})\right) \quad (6)$$

IHSST Uncertainty estimation

The methods presented so far allow us to estimate the significance of the differences between temperature series. However it would also be useful to be able to estimate the uncertainty in each individual IHSST series. Two methods will be used, the first based on internal consistency of the buoy data, and the second on intercomparison of the IHSST temperature datasets.

The uncertainty in the buoy data may be estimated by dividing the buoys into two random subsets, and calculating gridded temperature data from each subset of the data. Global temperature series are then calculated from the collocated values from each map series. A 120-month moving root-mean-squared difference between the two temperature series provides an estimate of the uncertainty in the global temperature for the region of common coverage (after scaling by 1/√2). This uncertainty estimate includes the effects of random measurement errors, as well as a sampling error which increases with decreasing coverage, however it does not include coverage uncertainty or systematic biases affecting all of the buoys.

In the second approach, an estimate of the uncertainties in each of the IHSST series is obtained from the difference temperature series for the overlap period 2005-2011. The uncertainty in the

difference series between the buoy and Argo data arises from the sum of the variances of the two series, assuming that the series are independent:

$$\sigma^2_{\text{buoy-Argo}} = \sigma^2_{\text{buoy}} + \sigma^2_{\text{Argo}} \quad (7)$$

and similar expressions for the remaining two series, where σ^2 is the squared uncertainty in the given temperature series. The squared uncertainty in the difference temperature may be estimated from the variance of the difference series, adjusting the number of degrees of freedom to account for the removal of the annual cycle from the difference series.

The uncertainty in a given series may then be estimated using equations of the following form:

$$\sigma^2_{\text{buoy}} = ½(\sigma^2_{\text{buoy-Argo}} + \sigma^2_{\text{buoy-CCI}} - \sigma^2_{\text{Argo-CCI}}) \quad (8)$$

The resulting uncertainty estimates include the effects of random measurement errors and any biases in the independent data sources which are not correlated across the data sources, however as before they do not include coverage bias. This is similar to the approach outlined in O'Carroll et al.[36]

The uncertainty in the trend in an IHSST series may be estimated from the uncertainty in the monthly temperatures obtained from equation (8) using the equation:

$$\sigma_\beta^2 = \frac{\nu\sigma^2}{\sum_i (t_i - t)^2} \quad (9)$$

where σ_β^2 is the variance of the trend, σ is the standard deviation of the time series values, t_i is the date of the i'th value in fractional years, and ν is the number of months of data per effective degree of freedom[27]. Note that this differs from the ordinary equation for the uncertainty in a trend in the use of the standard deviation of the time series in place of the standard deviation of the residuals - this is because the difference in trends between a pair of series also contributes to the uncertainty. For the trend of a set of contiguous monthly values this simplifies to:

$$\sigma_\beta^2 = \frac{\nu\sigma^2}{\Delta t^3} \quad (10)$$

Where Δt is the length of the period in years. ν is about 2 for the buoy series, or about 8 for the smoother Argo or CCI series.

ACKNOWLEDGEMENTS

We thank four anonymous reviewers for feedback on the article, Owen Embury for assistance with ESA-CCI SST data, Masayoshi Ishii for information regarding COBE-SST data, Willis Eschenbach for help in locating APDRC Argo data, and Boyin Huang for providing gridded ERSSTv4 ensemble data. Argo data used in the paper was collected and made freely available by the International Argo Program and the national programs that contribute to it. (http://www.argo.ucsd.edu, http://argo.jcommops.org). The Argo Program is part of the Global Ocean Observing System.

FUNDING

Z.H. and R.R. were funded by Berkeley Earth. M.R.'s research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. P.J. was funded by George Mason University. No specific grants were allocated to support this project.

AUTHOR CONTRIBUTIONS

K.C. and Z.H. conceived of the project; K.C. and R.R. produced the Buoy record. Z.H. and K.C. analyzed the Argo data. M.R. analyzed the CCI data. K.C., D.C. and Z.H. analyzed time series and trends. Z.H., K.C., P.J., D.C. and M.R. wrote the paper, with all authors providing input.

All authors declare that they have no competing interests.

DATA AND CODE

Buoy and ship data is available from ICOADS at: http://icoads.noaa.gov/products.html
ESA-CCI data is available from the Centre for Environmental Data Analysis at: http://gws-access.ceda.ac.uk/public2/nceo_uor/sst/L3S/EXP1.2/

ARC-SST data is available at:
http://catalogue.ceda.ac.uk/uuid/ff8a7f27b827c108dd9756adffaaa942
Argo data from the Asia-Pacific Data Research Center (APDRC) is available at:
http://apdrc.soest.hawaii.edu/projects/Argo/data/gridded/On_standard_levels/index-1.html
Argo data from the Japan Agency for Marine-Earth Science and Technology (H2008) is available at: http://www.jamstec.go.jp/ARGO/argo_web/argo/?page_id=83&lang=en
Argo data from Roemmich and Gilson (RG2009) is available at: http://sio-argo.ucsd.edu/RG_Climatology.html
ERSSTv4 data is available at: http://www1.ncdc.noaa.gov/pub/data/cmb/ersst/v4/netcdf/
ERSSTv3 data is available at: http://www1.ncdc.noaa.gov/pub/data/cmb/ersst/v3b/netcdf/
HadSST3 data is available at: http://www.metoffice.gov.uk/hadobs/hadsst3/data/download.html
COBE-SST data is available at: http://ds.data.jma.go.jp/tcc/tcc/products/elnino/cobesst/cobe-sst.html
Python code used in this analysis is available at:
http://www-users.york.ac.uk/~kdc3/papers/ihsst2016/
All datasets used were most recently accessed in early July, 2016.

REFERENCES

1. E. C. Kent, J. J. Kennedy, D. I. Berry, R. O. Smith, Effects of instrumentation changes on sea surface temperature measured in situ. *WIREs Clim Chg*. **1**, 718–728 (2010).

2. B. Huang *et al.*, Extended Reconstructed Sea Surface Temperature Version 4 (ERSST.v4). Part I: Upgrades and Intercomparisons. *J. Climate*. **28**, 911–930 (2015).

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

4. M. Ishii, A. Shouji, S. Sugimoto, T. Matsumoto, Objective analyses of sea-surface temperature and marine meteorological variables for the 20th century using ICOADS and the Kobe Collection. *Int. J. Climatol*. **25**, 865–879 (2005).

5.

T. R. Karl *et al.*, Possible artifacts of data biases in the recent global surface warming hiatus. *Science*. **348**, 1469–1472 (2015).

6.

J. J. Kennedy, A review of uncertainty in in situ measurements and data sets of sea surface temperature. *Rev. Geophys.* **52**, 1–32 (2014).

7.

S. D. Woodruff *et al.*, ICOADS Release 2.5: extensions and enhancements to the surface marine meteorological archive. *Int. J.Climatol*. **31**, 951–967 (2011).

8.

W. J. Emery, D. J. Baldwin, P. Schlüssel, R. W. Reynolds, Accuracy of in situ sea surface temperatures used to calibrate infrared satellite measurements. *J. Geophys. Res. Oceans*. **106**, 2387–2405 (2001).

9.

J. Hansen, R. Ruedy, M. Sato, K. Lo, Global Surface Temperature Change. *Rev. Geophys.* **48** (2010), doi:10.1029/2010RG000345.

10.

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

11.

T. M. Smith, R. W. Reynolds, T. C. Peterson, J. Lawrimore, Improvements to NOAA's Historical Merged Land–Ocean Surface Temperature Analysis (1880–2006). *J. Climate*. **21**, 2283–2296 (2008).

12.

R. S. Vose *et al.*, NOAA's Merged Land–Ocean Surface Temperature Analysis. *Bull. Am. Meteorol. Soc.* **93**, 1677–1685 (2012).

13.

K. Ishihara, Calculation of Global Surface Temperature Anomalies with COBE-SST. *(Japanese) Weather Service Bulletin* **73,** (2006).

14.

R. Rohde, R. A. Muller, R. Jacobsen, E. Muller, C. Wickham, A New Estimate of the Average Earth Surface Land Temperature Spanning 1753 to 2011. *Geoinfor. Geostat.: An Overview*. **1** (2013), doi:10.4172/2327-4581.1000101.

15.

K. Cowtan, R. G. Way, Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. *Q. J. R. Meteorol. Soc.* **140**, 1935–1944 (2014).

16. B. Huang *et al.*, Further Exploring and Quantifying Uncertainties for Extended Reconstructed Sea Surface Temperature (ERSST) Version 4 (v4). *J. Climate* (2015), doi:10.1175/JCLI-D-15-0430.1.

17. W. Liu *et al.*, Extended Reconstructed Sea Surface Temperature Version 4 (ERSST.v4): Part II. Parametric and Structural Uncertainty Estimations. *J. Climate*. **28**, 931–951 (2015).

18. E. C. Kent *et al.*, Global analysis of night marine air temperature and its uncertainty since 1880: The HadNMAT2 data set. *J. Geophys. Res. Atmos.* **118**, 1281–1298 (2013).

19. G. J. Flato *et al,* Evaluation of Climate Models. In: Climate Change 2013: The Physical Science Basis (eds T. F. Stocker, *et al*.) (Cambridge Univ. Press, 2013).

20. S. Hirahara, M. Ishii, Y. Fukuda, Centennial-Scale Sea Surface Temperature Analysis and Its Uncertainty. *J. Climate*. **27**, 57–75 (2014).

21. R. W. Reynolds *et al.*, Daily High-Resolution-Blended Analyses for Sea Surface Temperature. *J. Climate*. **20**, 5473–5496 (2007).

22. D. Roemmich, J. Gilson, The 2004–2008 mean and annual cycle of temperature, salinity, and steric height in the global ocean from the Argo Program. *Progress in Oceanography*. **82**, 81–100 (2009).

23. C. J. Merchant *et al.*, A 20 year independent record of sea surface temperature for climate from Along-Track Scanning Radiometers. *J. Geophys. Res. Oceans.* **117**, C12013 (2012).

24. P. J. Klotzbach, R. A. Pielke, R. A. Pielke, J. R. Christy, R. T. McNider, An alternative explanation for differential temperature trends at the surface and in the lower troposphere. *J. Geophys. Res. Atmos.* **114**, D21102 (2009).

25.

B. D. Santer *et al.*, Statistical significance of trends and trend differences in layer-average atmospheric temperature time series. *J. Geophys. Res. Atmos.* **105**, 7337–7356 (2000).

26.

J. Lee, R. Lund, Revisiting simple linear regression with autocorrelated errors. *Biometrika*. **91**, 240–245 (2004).

27.

G. Foster, S. Rahmstorf, Global temperature evolution 1979–2010. *Environ. Res. Lett.* **6**, 44022 (2011).

28.

D. Tjøstheim, J. Paulsen, Bias of some commonly-used time series estimates. *Biometrika*. **70**, 389–399 (1983).

29.

C. J. Merchant *et al.*, Sea surface temperature datasets for climate applications from Phase 1 of the European Space Agency Climate Change Initiative (SST CCI). *Geoscience Data Journal*. **1**, 179–191 (2014).

30.

D. Roemmich, J. Gilson, The global ocean imprint of ENSO. *Geophys. Research Letters*. **38**, (2011).

31.

D. Roemmich, J. Church, J. Gilson, D. Monselesan, P. Sutton, S. Wijffels, Unabated planetary warming and its ocean structure since 2006. *Nature Climate Change*. **5**, 240–245 (2015).

32.

W. Tang, S. H. Yueh, A. G. Fore, A. Hayashi, Validation of Aquarius sea surface salinity with in situ measurements from Argo floats and moored buoys. *J. Geophys. Res. Oceans.* **119**, 6171–6189 (2014).

33.

S. Hosoda, T. Ohira, K. Sato, T. Suga, Improved description of global mixed-layer depth using Argo profiling floats. *Journal of Oceanography*. **66**, 773-787 (2010).

34.

S. Hosoda, T. Ohira, T. Nakamura, A monthly mean dataset of global oceanic temperature and salinity derived from Argo float observations. *JAMSTE°C Rep. Res. Dev.* **8**, 47–59 (2008).

35.

M. Richardson, K. Cowtan, E. Hawkins, M.B. Stolpe, Reconciled climate response estimates from climate models and the energy budget of Earth. *Nature Climate Change*. **6**, 931–935 (2016).

36. A.G. O'Carroll, J.R. Eyre, R.W. Saunders, Three-Way Error Analysis between AATSR, AMSR-E, and In Situ Sea Surface Temperature Observations. *Journal of Atmospheric and Oceanic Technology*. **33** (2008).

37. Argo. Argo float data and metadata from Global Data Assembly Centre (Argo GDAC). SEANOE. http://doi.org/10.17882/42182 (2000).

38. N. Cressie, The origins of kriging. *Mathematical Geology.* **22**, 239–252 (1990).

SUPPLEMENTARY MATERIALS

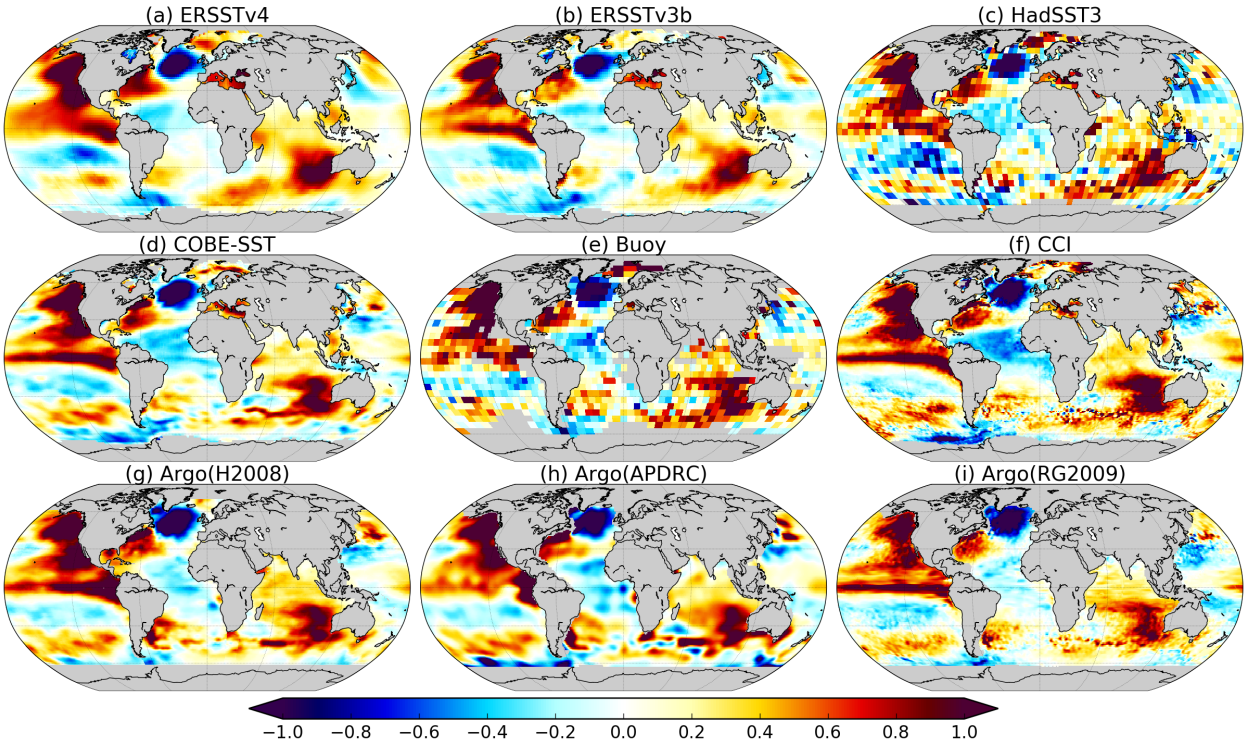


Figure S1. *Trend maps on the period 2005-2015 for all of the composite records, and for the buoy, Argo, and CCI records. The HadSST3 and buoy records are determined with no fitting or smoothing, and so show sharp grid cell boundaries, in contrast to the other records. The general*

features are similar between all records, with the most noticeable difference being in the mid-Atlantic and South Pacific.

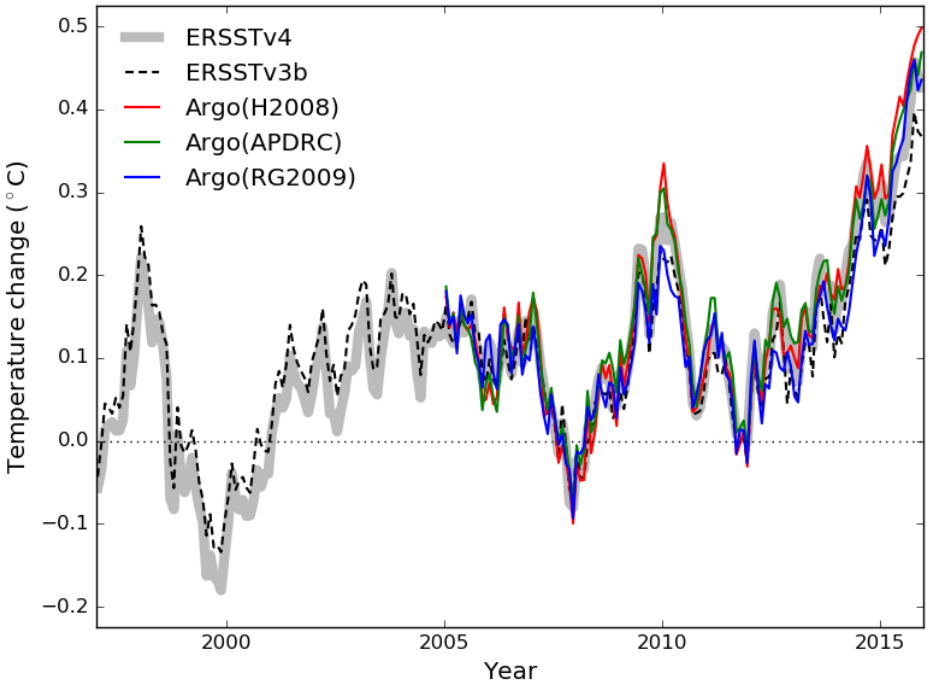


Figure S2. *Comparison of ERSSTv3b and v4 with three different Argo-based near-SST records, using the same spatial restrictions as in Figure 1, but with ERSSTv4 aligned to 1997-2001 (inclusive), with all other series aligned onto v4 using the 2005-2007 period due to the limited timespan with Argo data.*

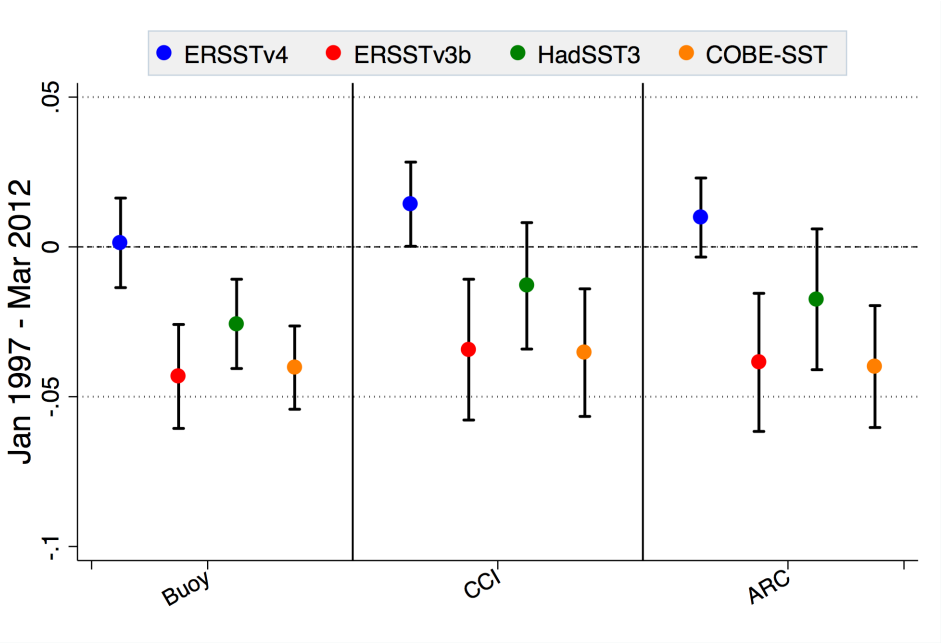


Figure S3. *Trends and 95% confidence intervals (degrees °C per decade) in difference series for each IHSST and composite SST series, masked to common composite SST coverage.*

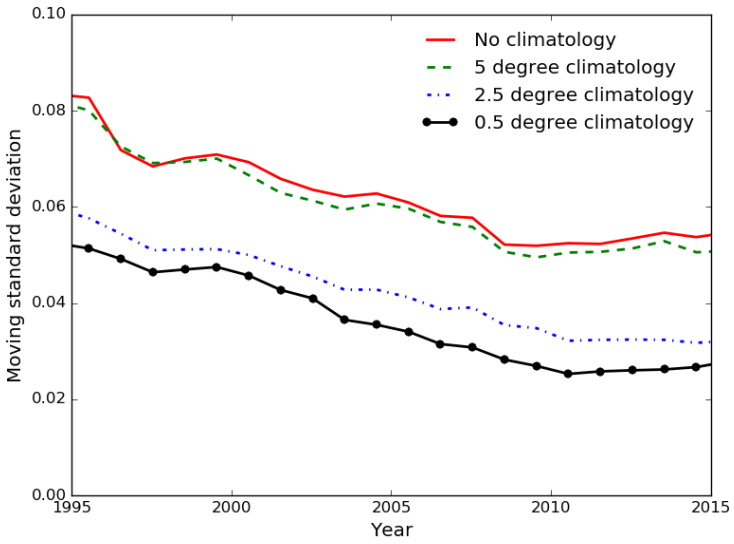


Figure S4. *Cross validated uncertainties for the buoy record, whether with no climatology or with daily climatologies derived from the OISSTv2 daily reanalysis data. The results of using a daily climatology on a coarse 5 degree do not differ significantly from the results using no climatology, so adjusting for the day of the month has little impact in reducing the errors in the global temperature estimates. However adjusting for position within the cell by using a finer grid*

climatology has a substantial impact, reducing the errors in the global temperature estimates by a factor of two.

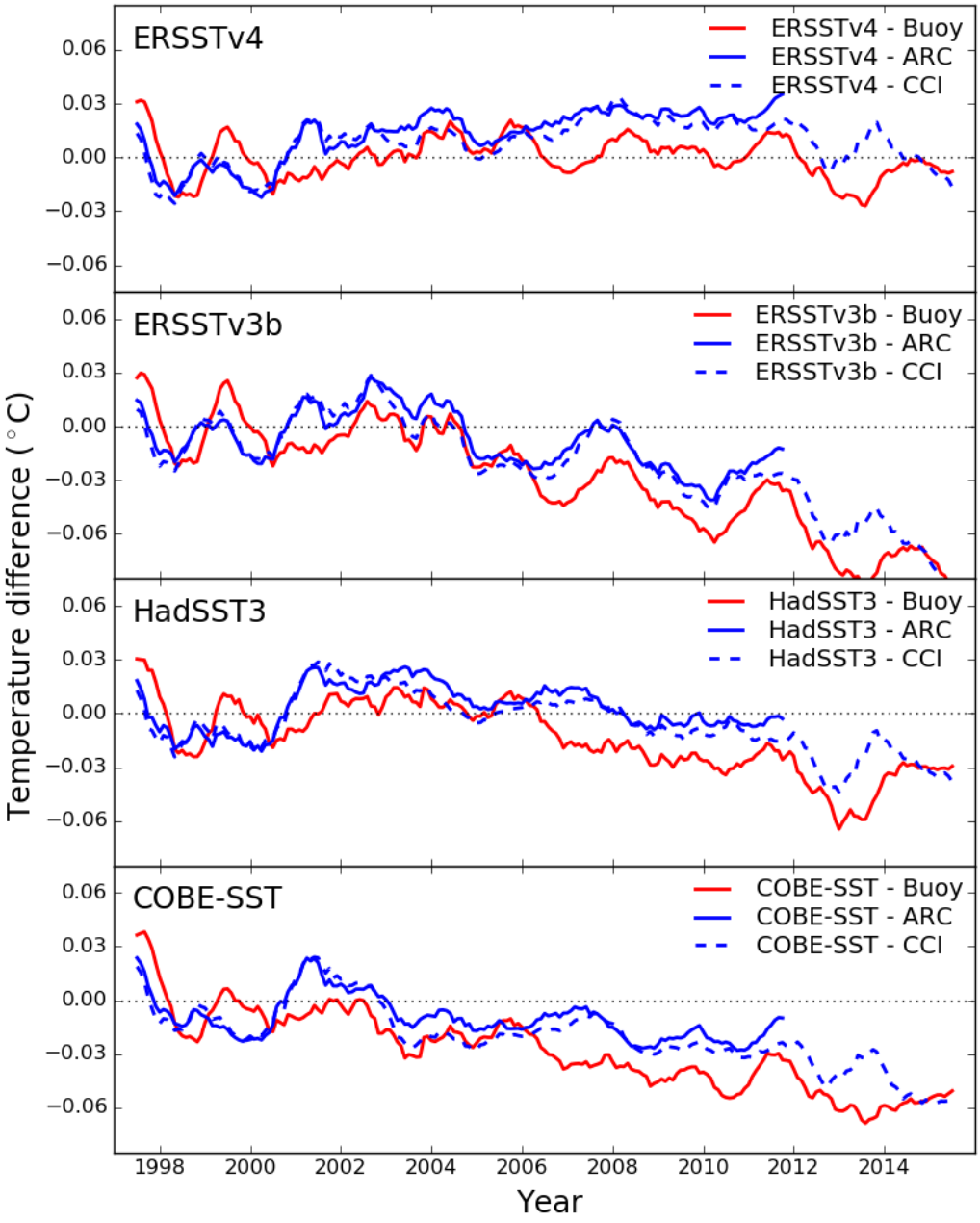


Figure S5. *12-month centered moving average of temperature difference series between composite and buoy-only, CCI, and ARC SST anomalies restricted to common coverage across all series shown (four composites, buoys, and ARC/CCI).*

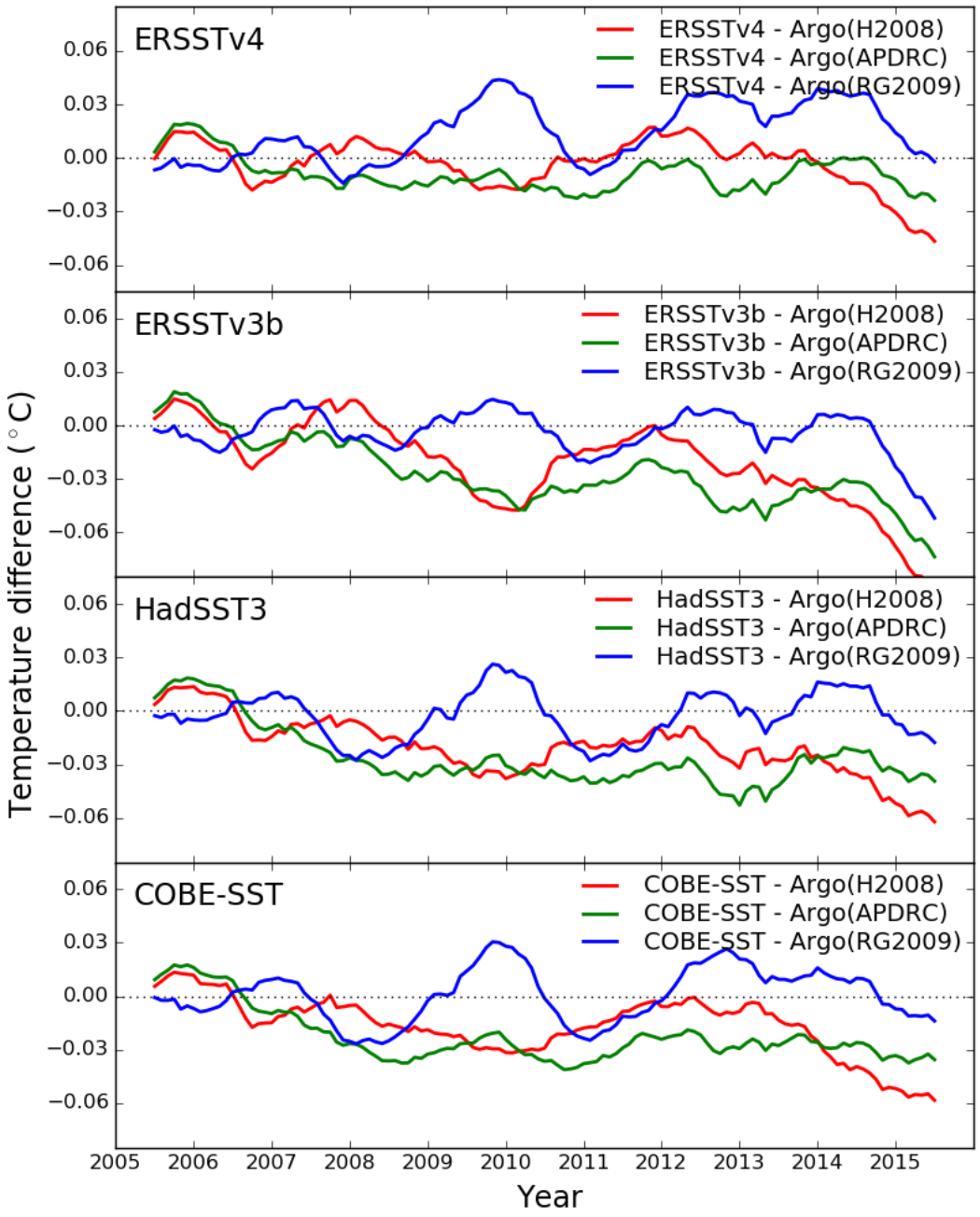


Figure S6. *12-month centered moving average of temperature difference series between composite and Argo near-SST anomalies restricted to common coverage across all series with 2005-2015 records (four composites, three Argos, buoy-only, and CCI).*

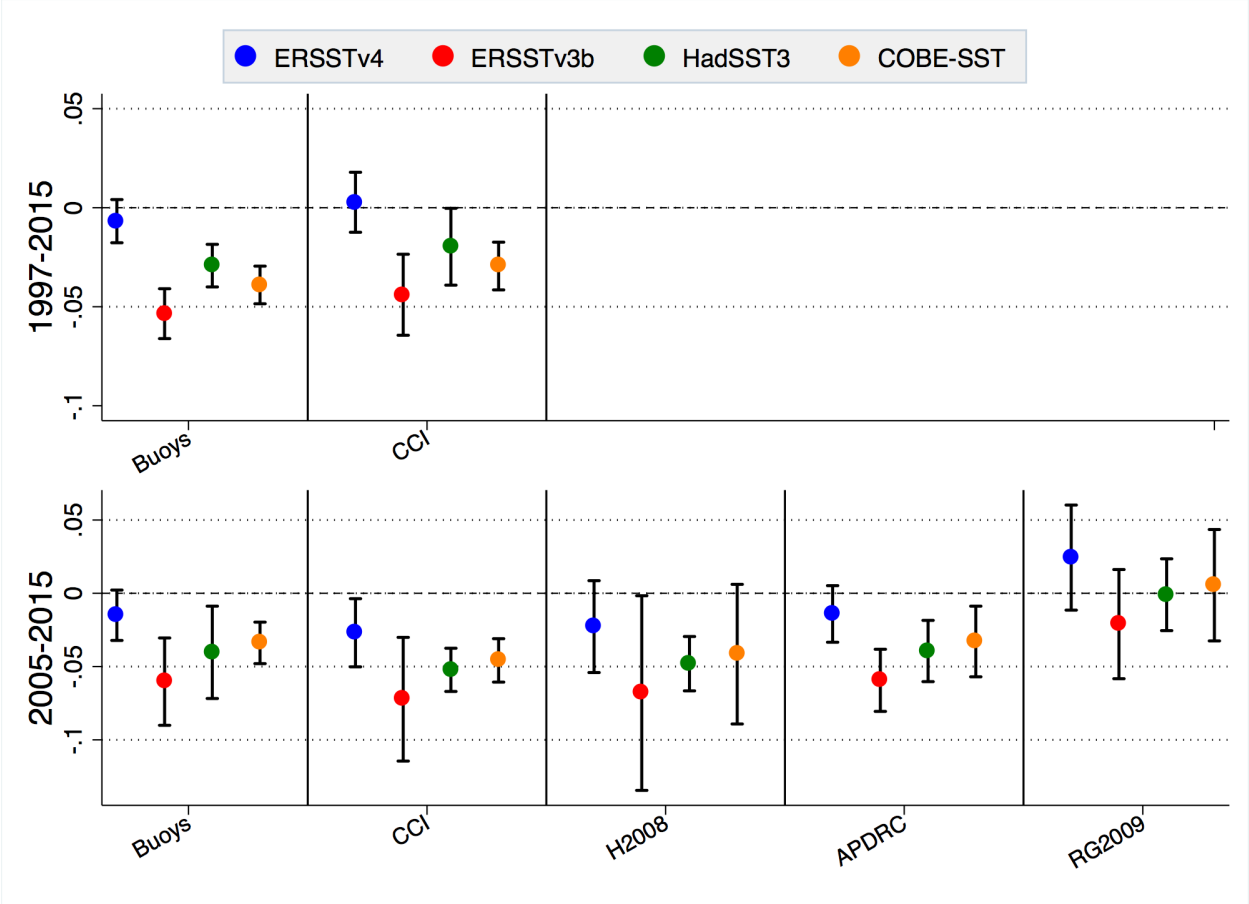


Figure S7. *Trends and 95% confidence intervals (degrees °C per decade) in difference series for each IHSST and composite SST series, masked to common coverage for all series available. 1997-2015 trends are masked to common coverage for the four composite series, buoys, and CCI. 2005-2015 trends are masked to common coverage for the four composites, buoys, CCI, and the three Argo series. Confidence intervals for trends are calculated using an ARMA[1,1] autocorrelation model.*

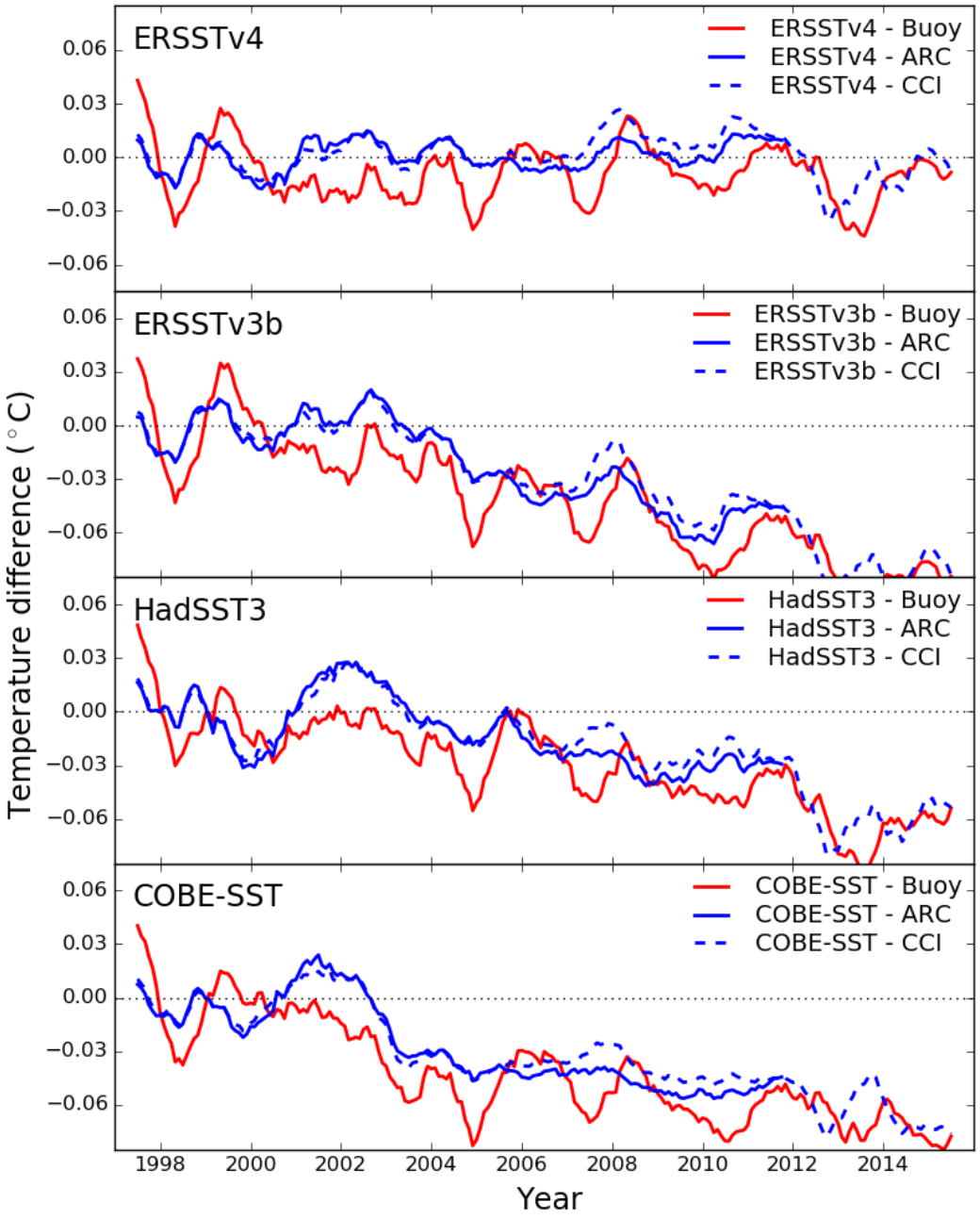


Figure S8. *12-month centered moving average of temperature difference series between composite and buoy-only, CCI, and ARC SST anomalies with the buoy and HadSST3 series kriged and all series reduced to common coverage to ensure consistent complete spatial and temporal coverage.*

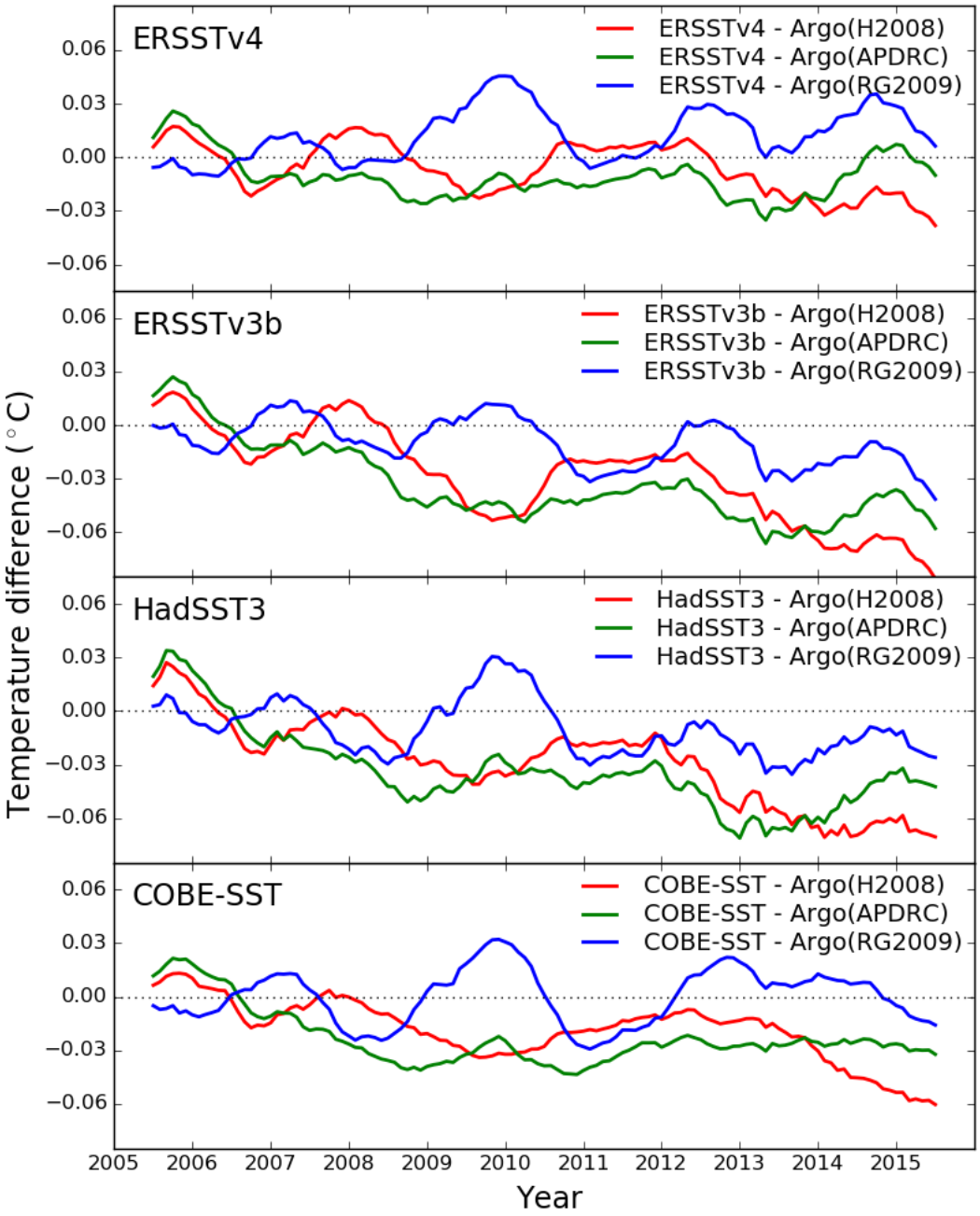


Figure S9. *12-month centered moving average of temperature difference series between composite and Argo near-SST anomalies with the buoy and HadSST3 series kriged and all series reduced to common coverage to ensure consistent complete spatial and temporal coverage.*

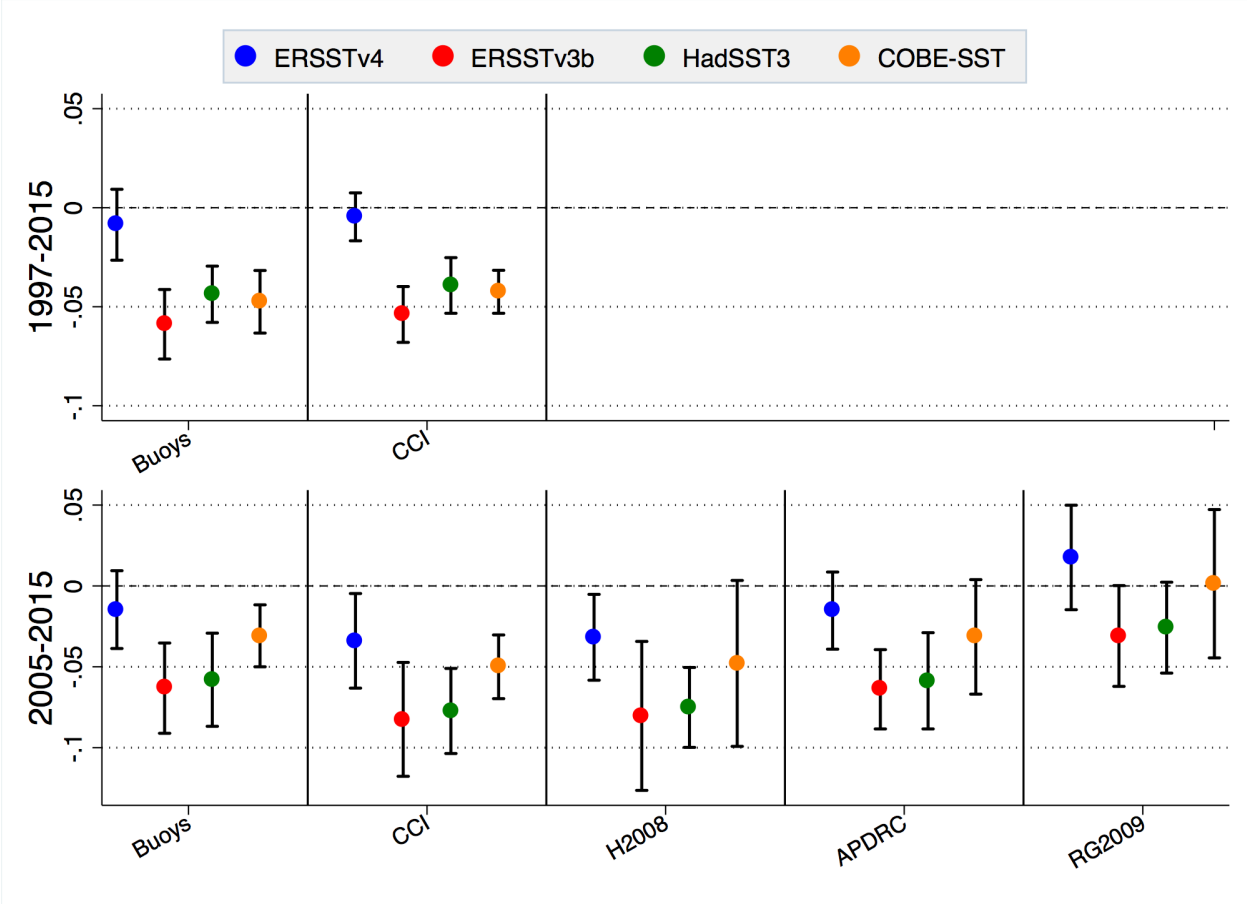


Figure S10. *Trends and 95% confidence intervals (degrees °C per decade) in difference series for each IHSST and composite SST series, with the buoy and HadSST3 series kriged and all series reduced to common coverage to ensure consistent complete spatial and temporal coverage.*

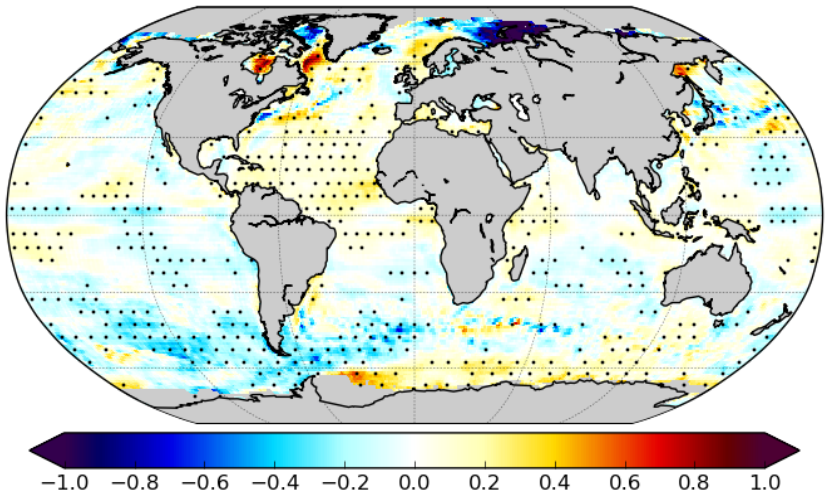


Figure S11. *Trend difference maps over January 1997 through December 2015 for the difference between ERSSTv4 and CCI. Unlike the other IHSSTs, the differences between ERSSTv4 and CCI shows significant long range autocorrelation. CCI shows faster warming over much of the southern extratropics, but slower warming elsewhere.*

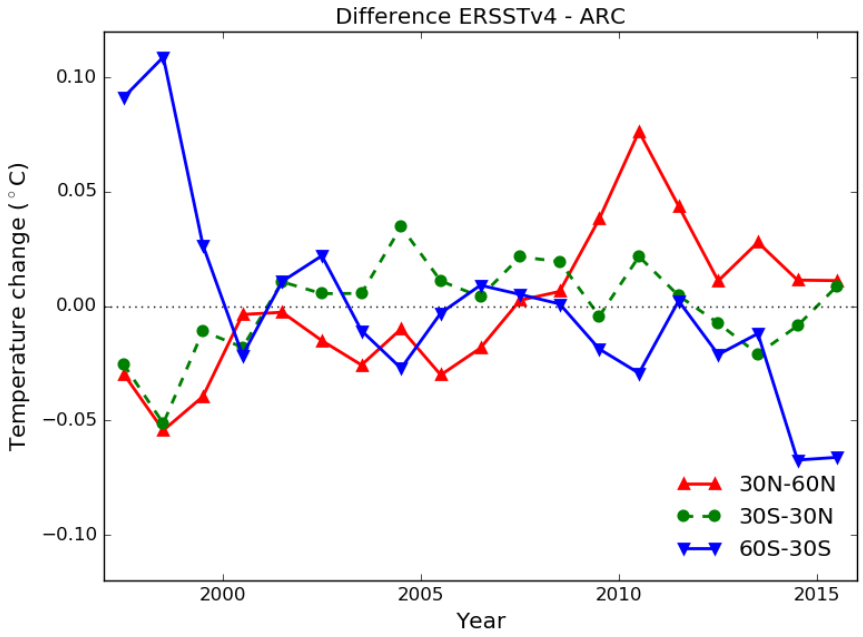


Figure S12. *Differences between ERSSTv4 and CCI by latitude zone. The regional differences between ERSSTv4 and CCI are clear in the zonal temperature series. The biggest differences occur before 2002 and after 2012, however the zonal trends differ across the whole record. CCI shows faster warming over much of the southern extratropics, but slower warming elsewhere.*

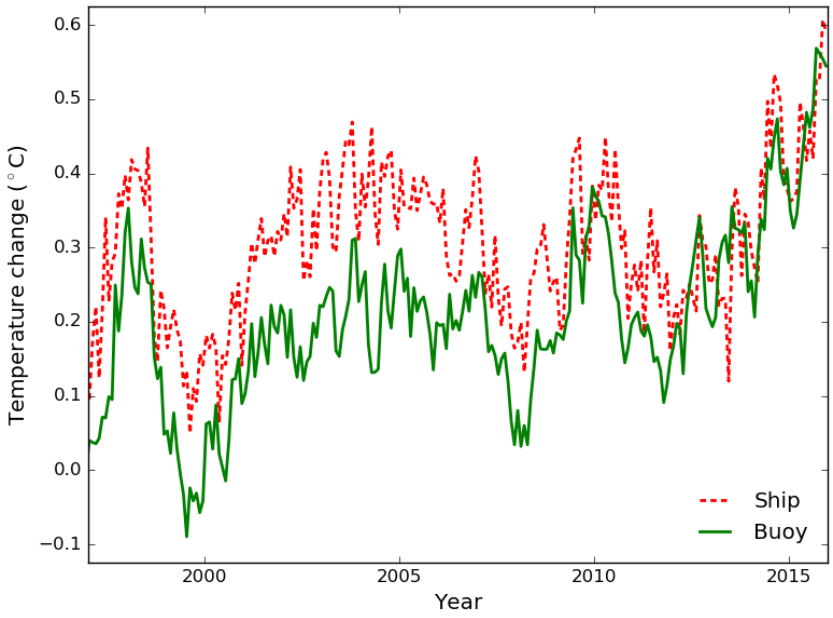


Figure S13. *Buoy-only and ship-only temperature anomalies from January 1997 through December 2015, with no matching of coverage. A crude ship-only record was determined using the same gridding method as for the buoy record, but with no quality control to eliminate erroneous records. Each ship or buoy observation is normalised by subtracting the same OISST daily climatology, producing a consistent estimate of the offset in temperature measurements. The anomalies shown have the annual cycle removed and the mean of the annual cycle added back in. Similar to Huang et al 2015 the results suggest that the bias arises from both (a) ships showing lower trends and (b) ships being warmer. Either of the difference due to the transition, or the difference in trend, would create a bias in the trends in the uncorrected composite record.*

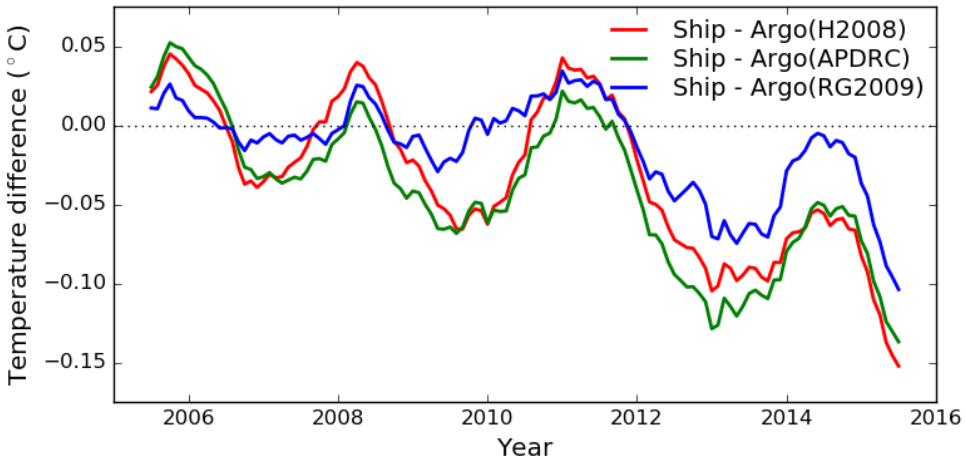


Figure S14. *Difference between ship-only record and the three Argo series using a 12-month centered moving average. All three Argo records show greater warming than the ship-only record.*

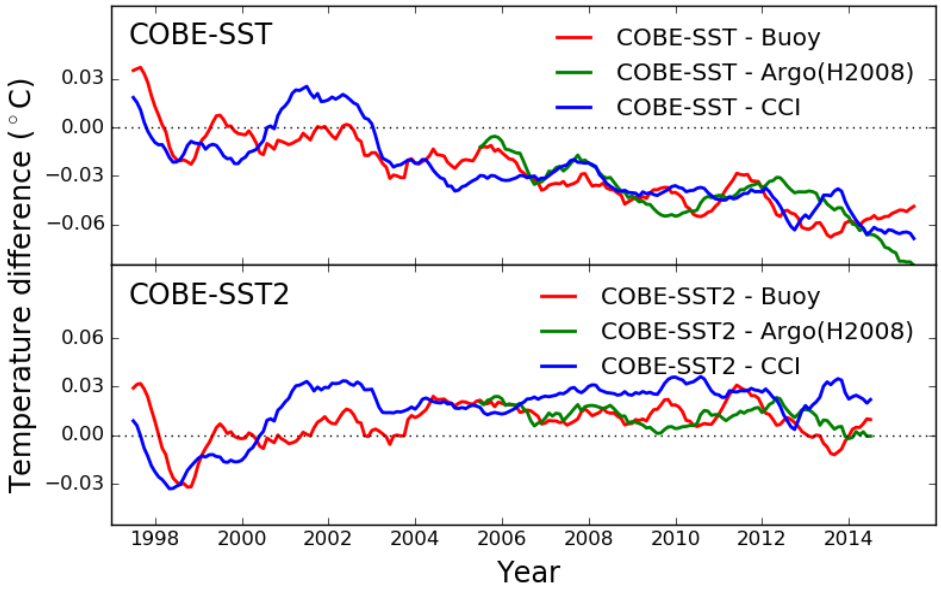


Figure S15. *Comparison of COBE-SST and COBE-SST2 to the IHSSTs using a 12-month centered moving average. The COBE-SST version 2 temperature record is currently only available to 2014, however this newer product is much more similar to IHSST records than the older version. COBE-SST2 shows slightly more warming than the buoy record, with the largest differences in early period when buoy observations are sparse. Agreement with CCI is good after 2001.*

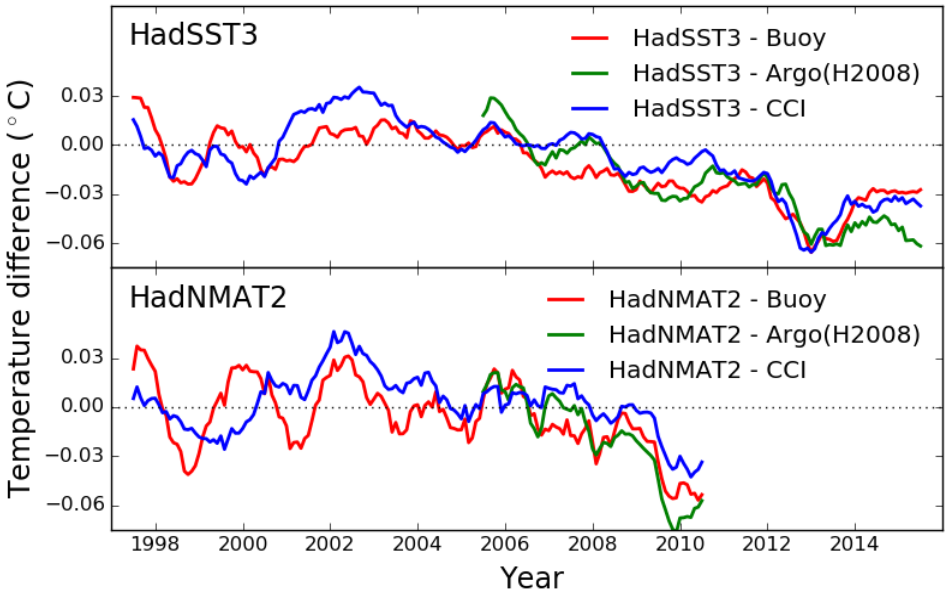


Figure S16. *Comparison of HadSST3 and HadNMAT2 to the IHSSTs using a 12-month centered moving average. The HadNMAT2 temperature record is currently only available through 2010. HadNMAT2 appears to show a cool bias comparable to if not larger than that of HadSST3 relative to the IHSSTs in the period from 2003 through the end of 2010.*

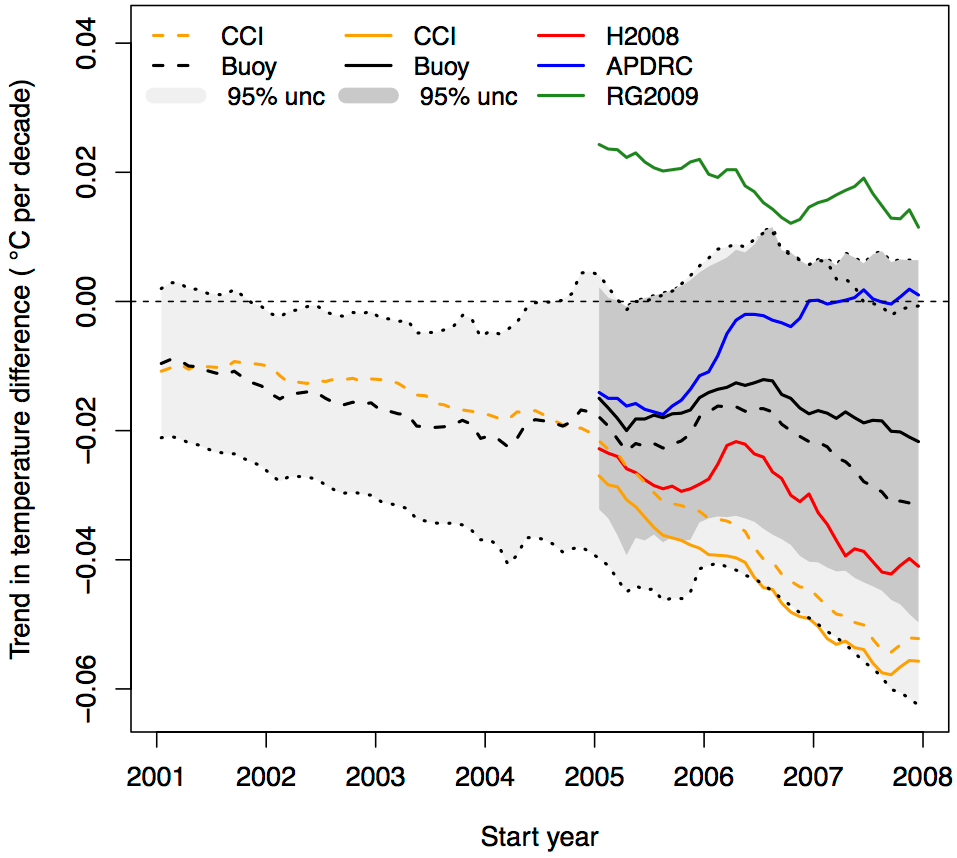


Figure S17. *Trends in differences for ERSSTv4 vs. IHSST records with common coverage from 1997 (buoys and CCI only; dashed lines) and common coverage from 2005 (buoys, CCI and Argos as solid lines). Trends to 2015 are shown as a function of start year, with 95% uncertainties for trend in differences for ERSSTv4 vs. buoys.*

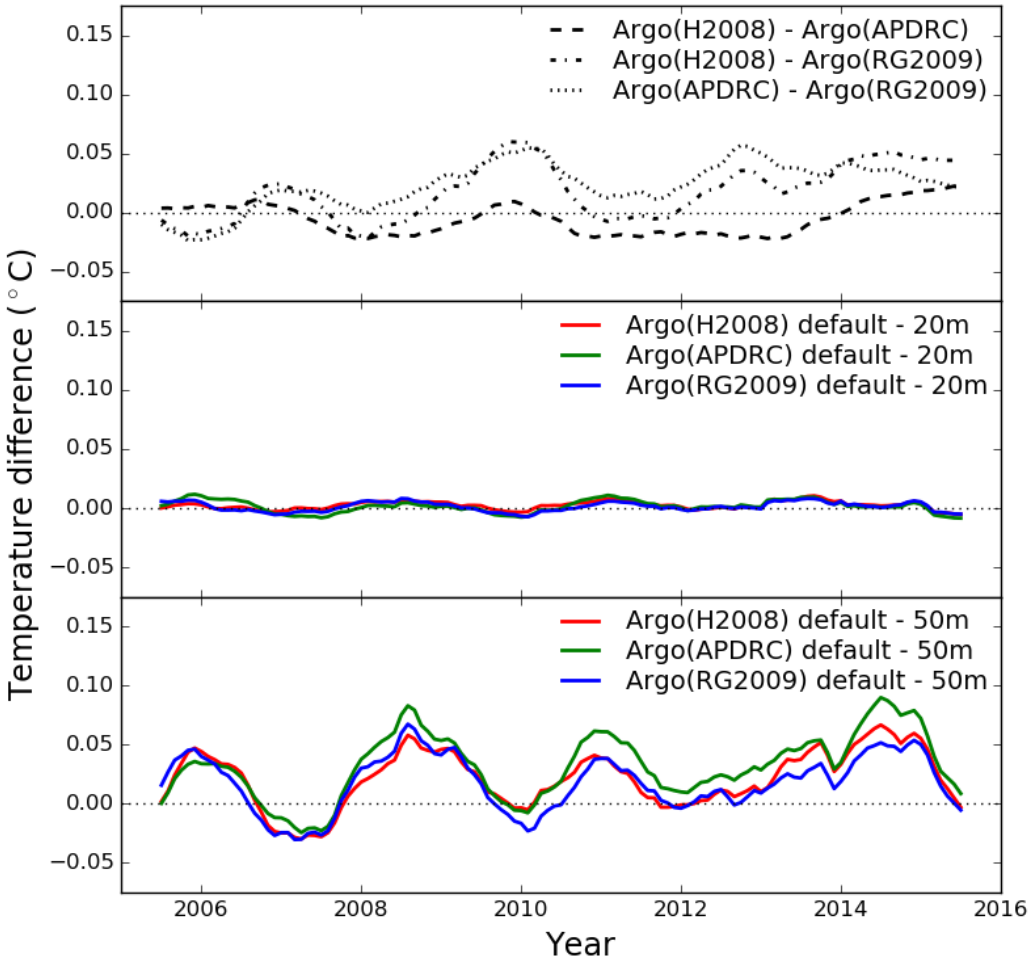


Figure S18. *Top panel shows 12-month smoothed differences between default Argo records used in the paper (e.g. 5 meter depths for RG2009 and APDRC; 10 meter depths for H2008). Middle panel shows differences between temperatures at default depths and a depth of 20 meters. Bottom panel shows differences between default depths and a depth of 50 meters. In both cases the change in temperatures with depth is smaller than the difference in temperatures between Argo records at the highest available (default) depth.*

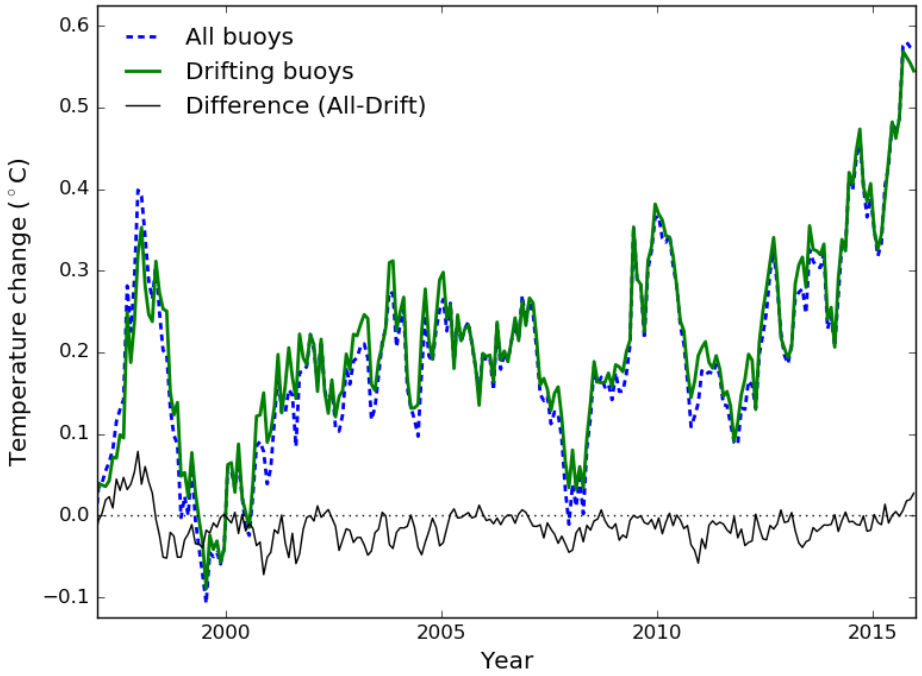


Figure S19. *Comparison of buoy records comprised of all buoys (drifting + moored) and only drifting buoys. Each buoy observation is normalised by subtracting the same OISST daily climatology, producing a consistent estimate of the offset in temperature measurements. The anomalies shown have the annual cycle removed and the mean of the annual cycle added back in. This suggests that moored buoys are cool-biased relative to drifting buoys, and that the conflation of the two into a single record could be biased due to the changing composition of each over time. However, the choice of buoy records to include has little effect on the difference series relative to composite SST records.*

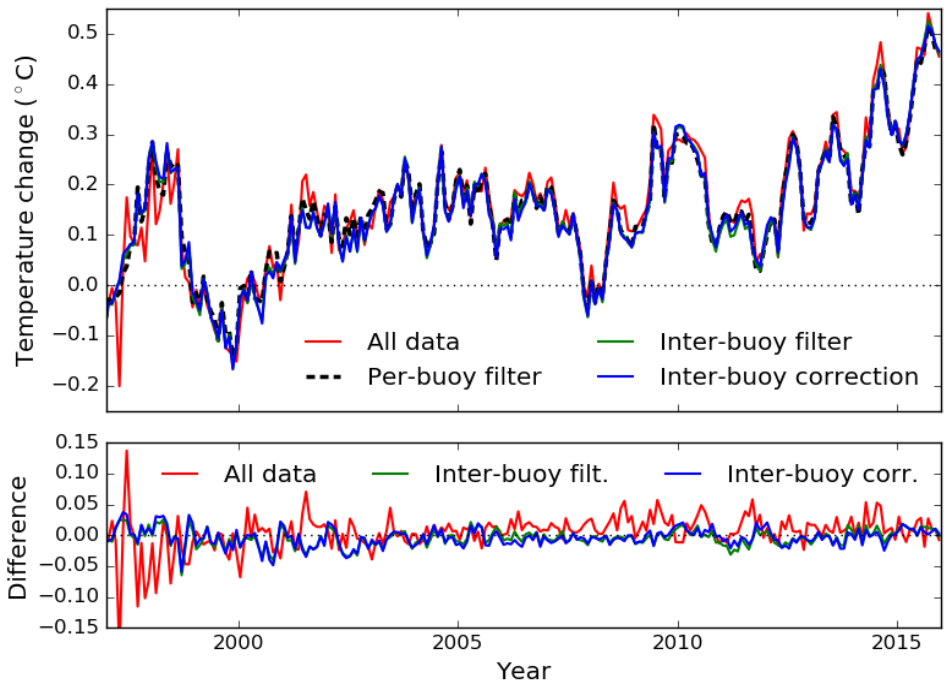


Figure S20. *Comparison of drifting buoy-based IHSST records for different quality control and homogenization choices. The lower panel shows the difference between the alternative schemes and the default per-buoy filtering.*

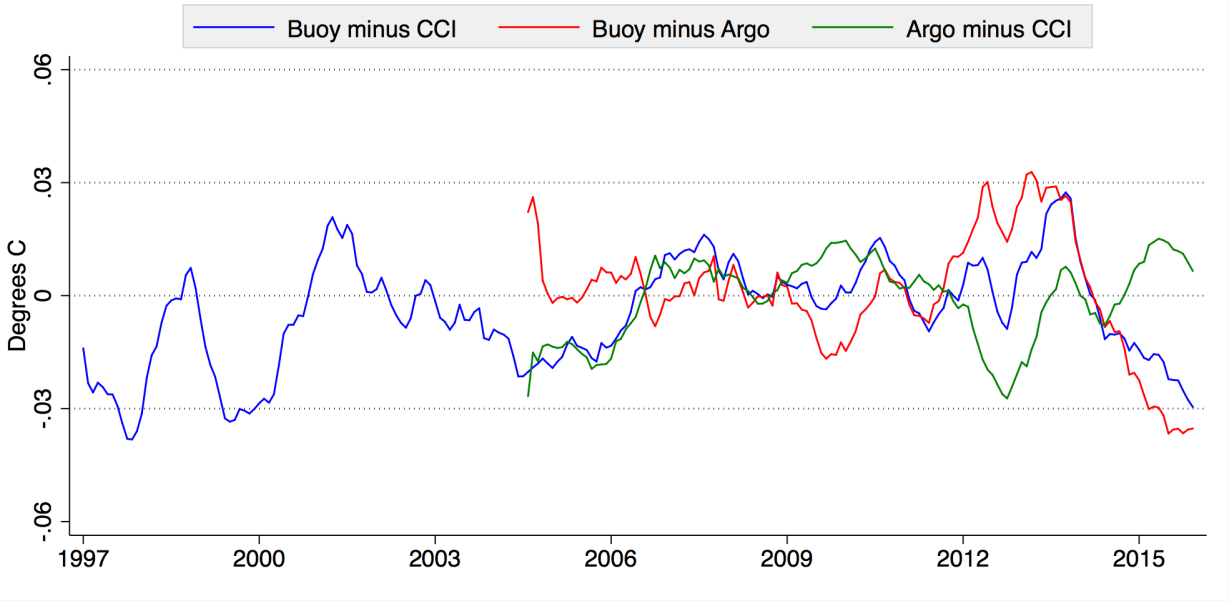


Figure S21. *12-month centered moving average of differences between IHSST series from January 1997 - December 2015 when reduced to common coverage for each separate pairing. The Argo series shown in H2008.*

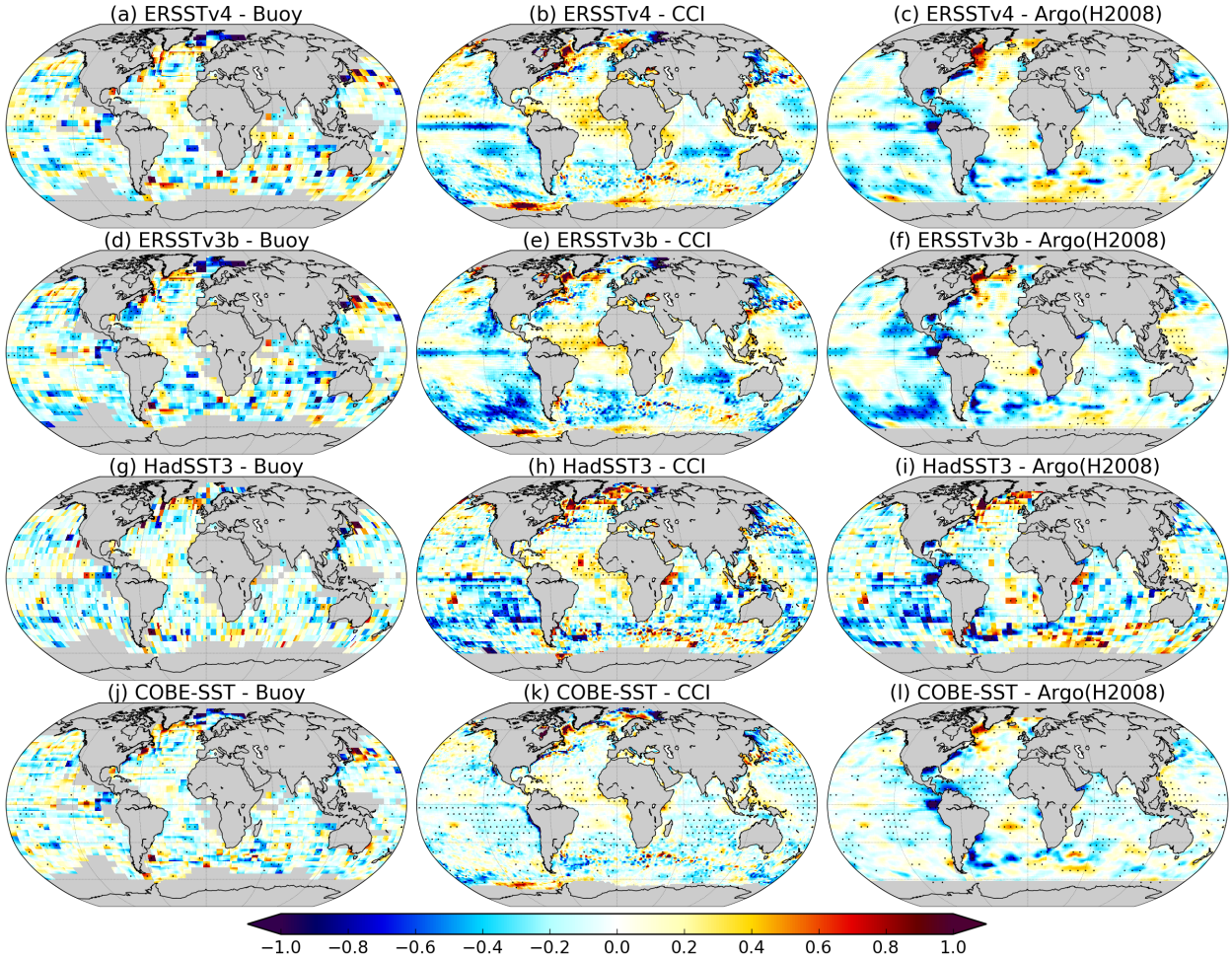


Figure S22. *Trend difference maps over 2005-2015 for the composite records versus Buoy, CCI, and Argo (H2008). Dots represent trends which are significant at the 2σ level using an ARMA(1,1) autoregressive model and no bias correction. Differences with the buoy record are localised and noise-like. Differences with the Argo and CCI records are more significant and show greater spatial correlation due to smoothing, although the differences still tend to be localised. Differences between smoothed map series are more likely to be significant, since the noise in an unsmoothed record increases the trend uncertainty.*

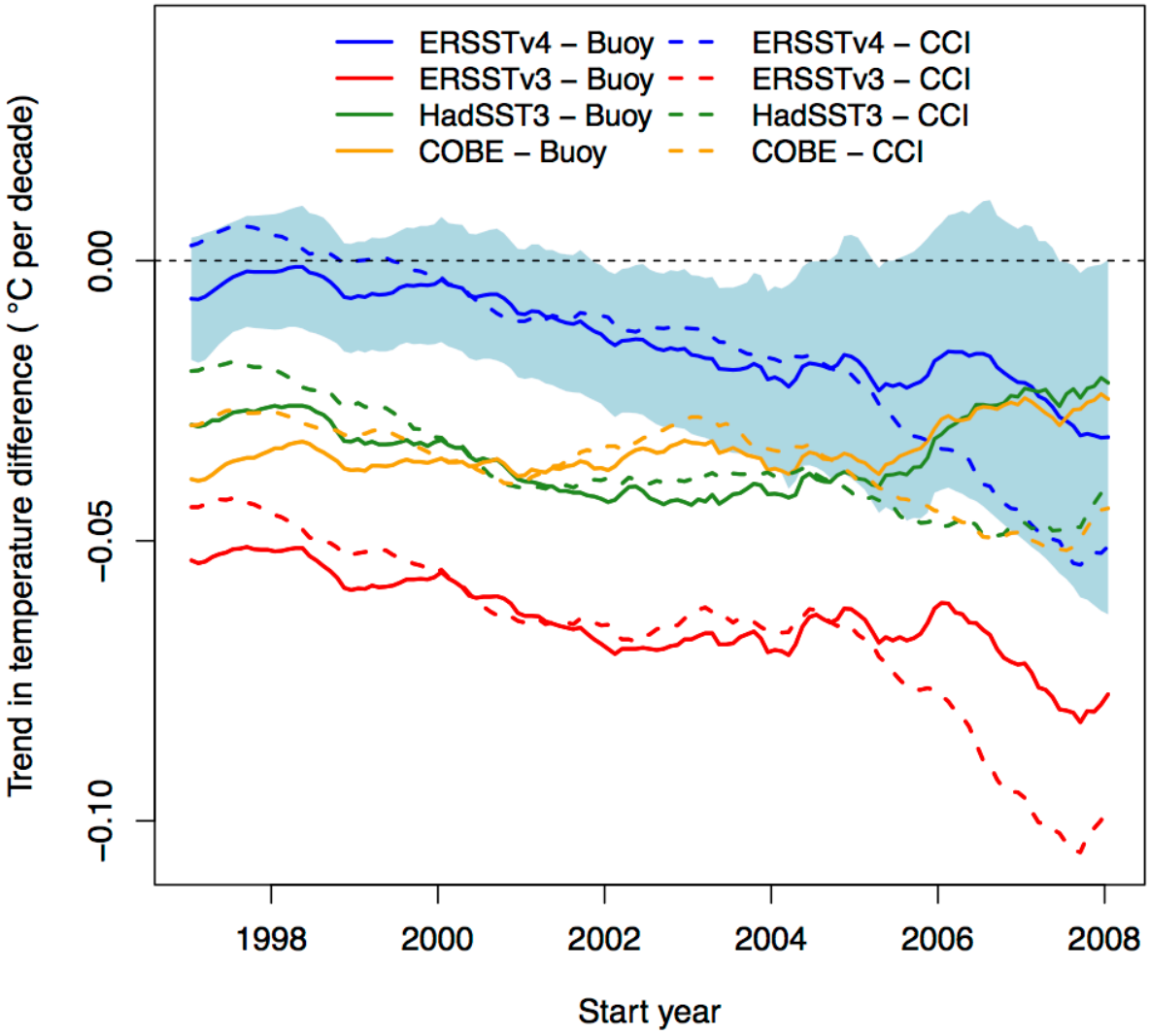


Figure S23. *Trends in differences for composite vs. buoy (solid lines) and CCI (dashed lines) IHSST records with common coverage. Trends to 2015 are shown as a function of start year, with 95% uncertainties for trend in differences for ERSSTv4 vs. buoys.*

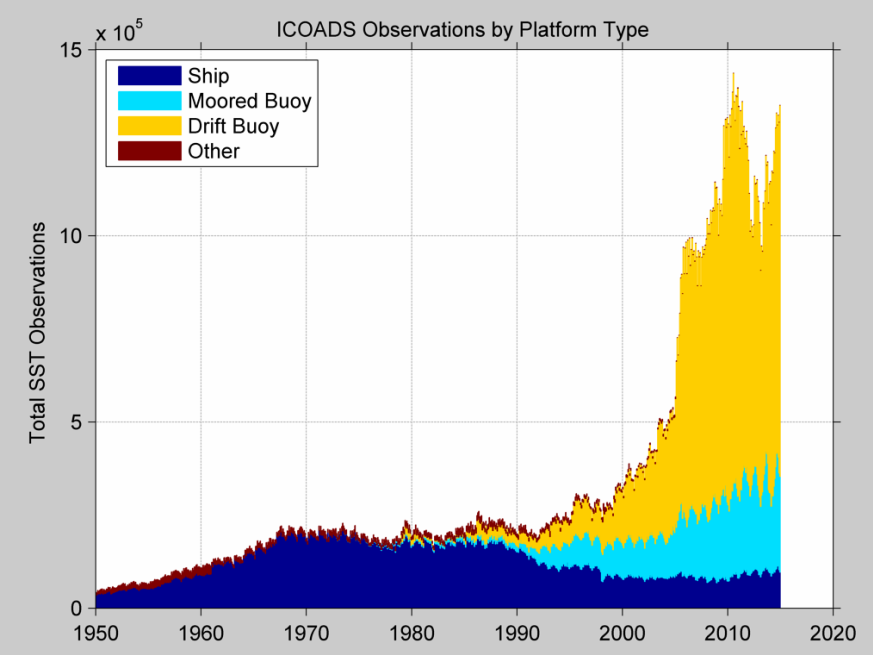


Figure S24. *Number of observations over time by instrument type in the ICOADS (v2.5) database. The number of observations from buoys has increased dramatically in recent years, while the number of measurements from ships has decreased. About 75 percent of current buoy measurements come from drifting rather than moored buoys.*

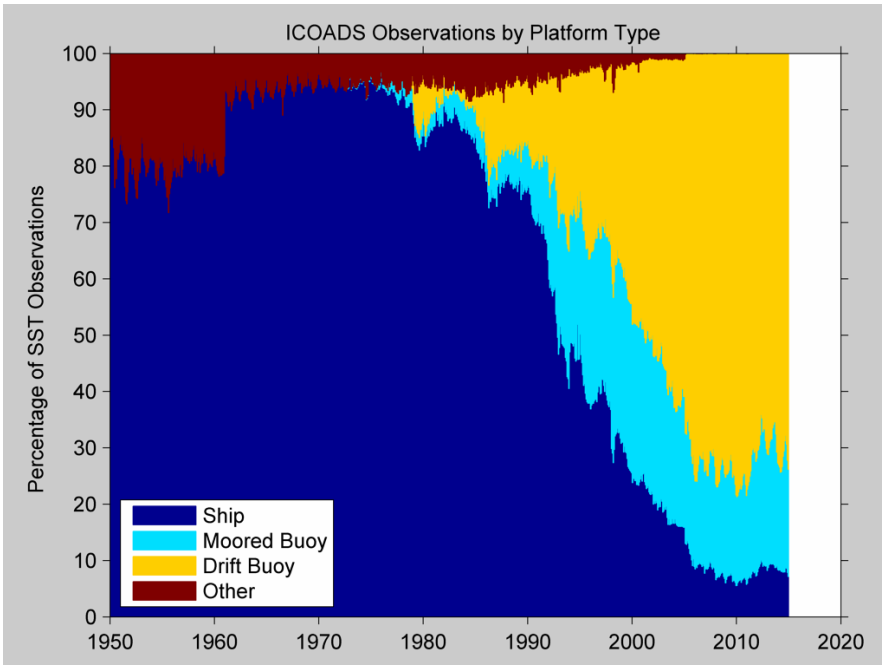


Figure S25. *Similar to Figure S24, but showing the percent of ICOADS observations in each year from each instrument type.*

IV. EVALUATING CLIMATE MODEL PERFORMANCE

Climate models represent one of the most important tools we have for understanding both past and future changes to Earth's climate. Assessing the performance of climate models compared to observations can help identify where current models may be performing poorly, and can inform future model improvements. Understanding the historical performance of the current generation of climate models can also increase (or decrease) confidence in the accuracy of future projections.

Research in this dissertation on evaluating climate model performance has focused on three distinct areas:

1) Accurately comparing climate models and observations through masking, like-to-like blended model fields, and improvements to observational records.
2) Evaluating the performance of post-publication projections of past climate models.
3) Evaluating the role of internal variability in 20th century temperatures using a simple climate model.

1. ACCURATE COMPARISONS OF CLIMATE MODELS AND OBSERVATIONS

Climate models have frequently been criticized for overestimating the rate of warming compared to observations in media reports and Congressional testimony. These critiques are based on peer reviewed papers that evaluate temperature hindcasts and limited forecasts (e.g. post-2005) of CMIP5 models, and either examine the 1998-2014 hiatus period [59] or focus on a narrowly defined test – such as tropical (20S to 20N) tropospheric temperatures form 1979 to present.[60]

When comparing models and observations it is also essential to make like-to-like comparisons of observational and model fields. For example, many observational temperature products have gaps associated with limited spatial coverage, particularly in regions like the Arctic and in the pre-satellite era.[9] Model fields should be masked to ensure the same temporal/spatial coverage as observations, instead of comparing globally complete estimates to more fragmentary ones.

Observations often comprise a combination of different measurement techniques. For example, the observational GMST record is actually a combination of SAT over land and SST over the

ocean, while climate model global SAT fields are frequently used in comparisons, including in the IPCC AR5.[4]

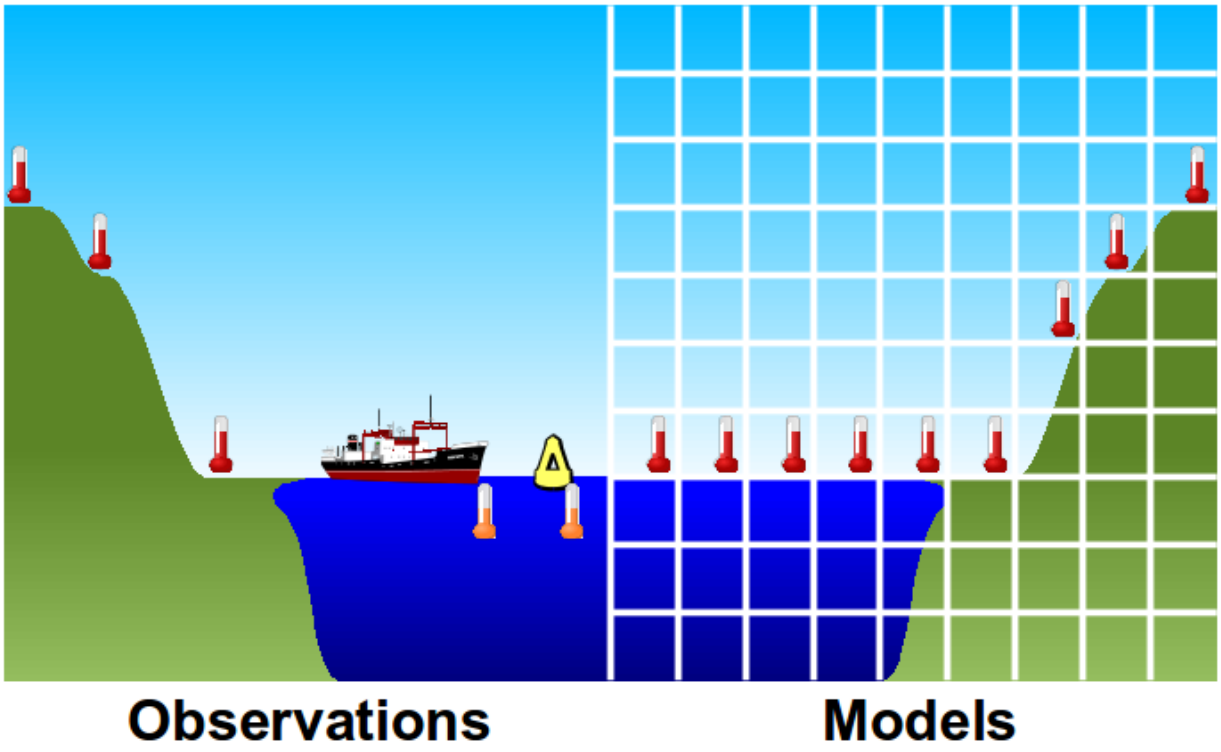


Figure IV.1. *Diagram showing the blended SAT/SST fields measured by global mean surface temperature observations compared to the global SAT fields generally used by models. Figure courtesy of Kevin Cowtan.*

There is also a secondary effect at work here. Global surface temperature records generally incorporate air temperatures over sea ice (since the water under sea ice is insulated from the surface). However with the loss of Arctic sea ice over recent decades, regions of the ocean which were previously given air temperatures switch to sea surface temperatures. This doesn't matter if the two temperatures are measured in the same way. However climate scientists usually work in terms of temperature changes (or anomalies) with respect to some reference period. Air and water temperatures may not be comparable outside of this reference period. In practice, because air temperatures have warmed faster than sea surface temperatures, and so the loss of sea ice introduces a cool bias in the temperature record at the point when the ice melts.

A like-with-like comparison of models to observations reduces the discrepancy in GMST warming trend since 1975 between models and observations by more than a third. Furthermore,

the discrepancy is very recent in origin: until the middle of the last decade there is no discrepancy in trend between the models and observations. The difference between SAT and blended SAT/SST model fields is shown in the figure below.

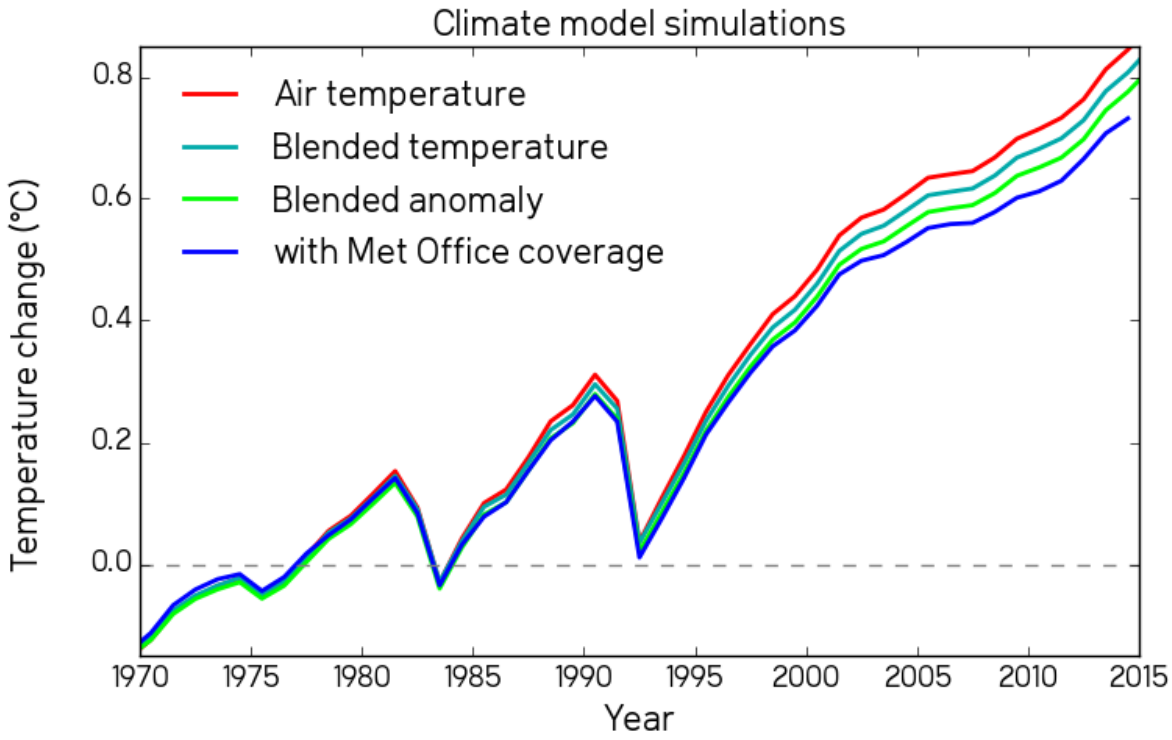


Figure IV.2. *CMIP5 historical + future projected global mean surface temperatures constructed from SAT fields, blended absolute SAT/SST fields, blended anomaly SAT/SST fields, and blended anomaly SAT/SST fields masked to HadCRUT4 coverage. From Cowtan et al 2016.*[4]

When blended fields are used and observations through present are considered, agreement between climate models and observations is substantially improved. The figure below shows the CMIP5 multimodel mean and model spread compared to surface temperature observations from six different groups; of these NASA, Berkeley, Cowtan & Way, and Copernicus (ERA5) all provide globally-complete temperature fields over the 1970-2019 period.

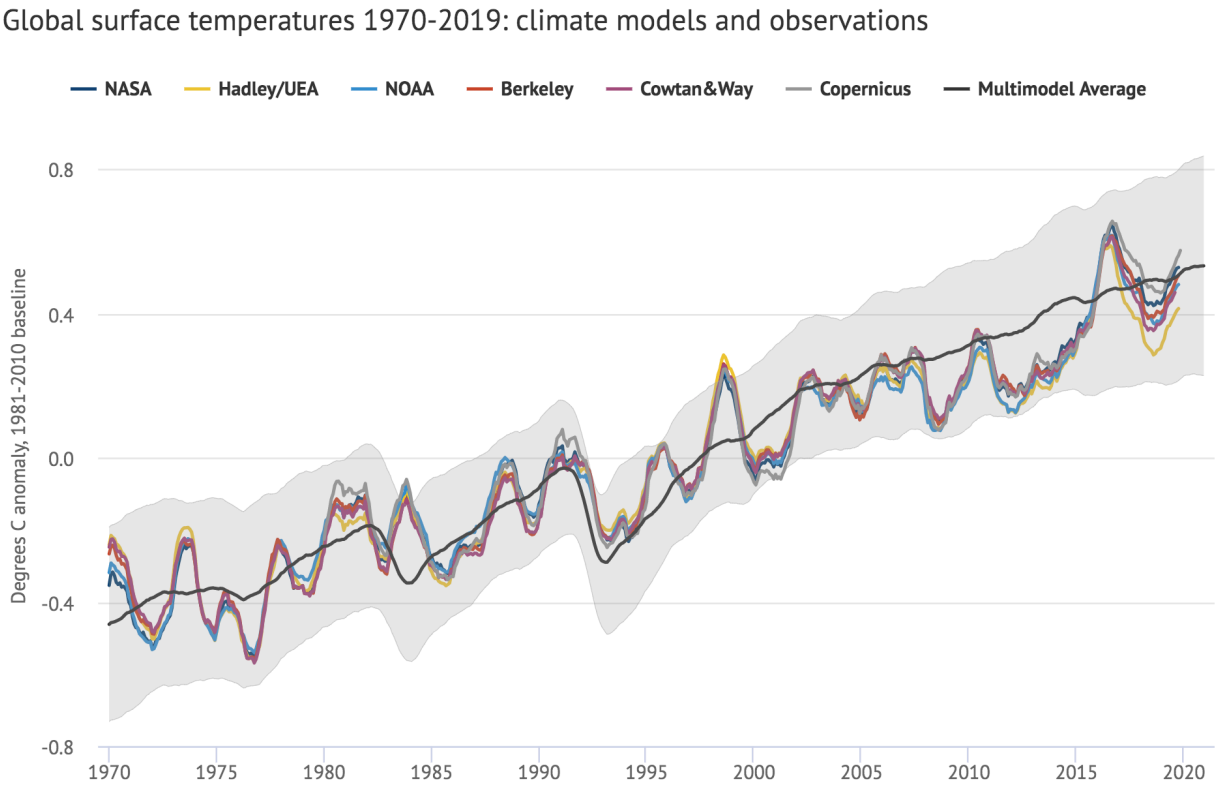


Figure IV.3. *12-month average global average surface temperatures from blended SAT/SST CMIP5 model fields and observations between 1970 and 2020 (observations extend through October 2019). Models use RCP4.5 forcings after 2005. Anomalies plotted with respect to a 1981-2010 baseline.*

Even with like-to-like comparisons between climate models and observations, disagreements over temperature changes may remain. It is always important to consider observational temperature estimates from different providers – as well as their published uncertainties – to better capture structural uncertainty in observations. For example, some of the apparent post-1998 differences between models and observations were due to uncorrected biases in sea surface temperature records. The transition between HadCRUT3 and HadCRUT4 in 2012 led to notably improved agreement between modeled and observed GMST, as did the switch from ERSSTv3b to ERSSTv4/v5 in NOAA and NASA temperature records.[61]

Further improvements in model/observation are expected once HadSST4 becomes part of the HadCRUT4 product next year. The hybrid SST record we discuss in the Chapter III also show improved agreement with CMIP5 models compared to HadSST3, as shown in the figure below.

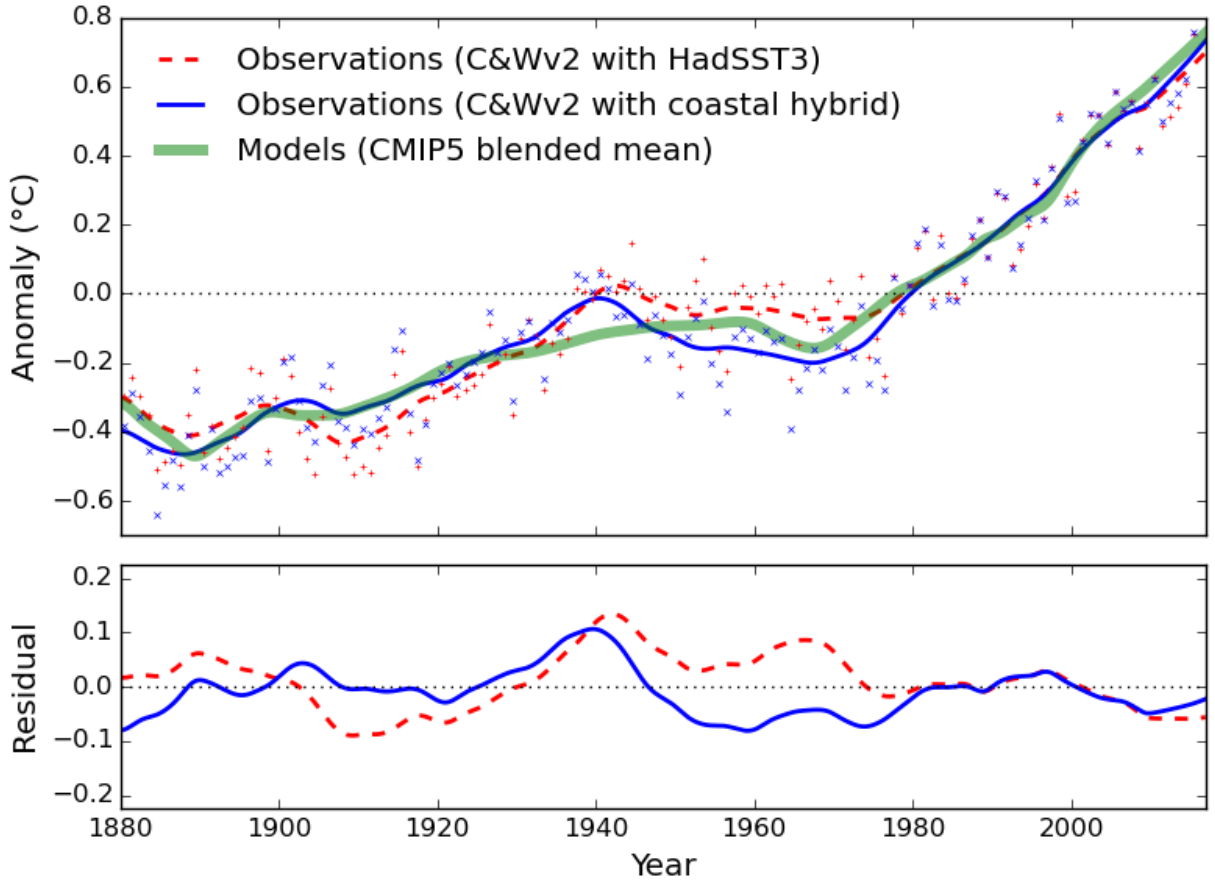
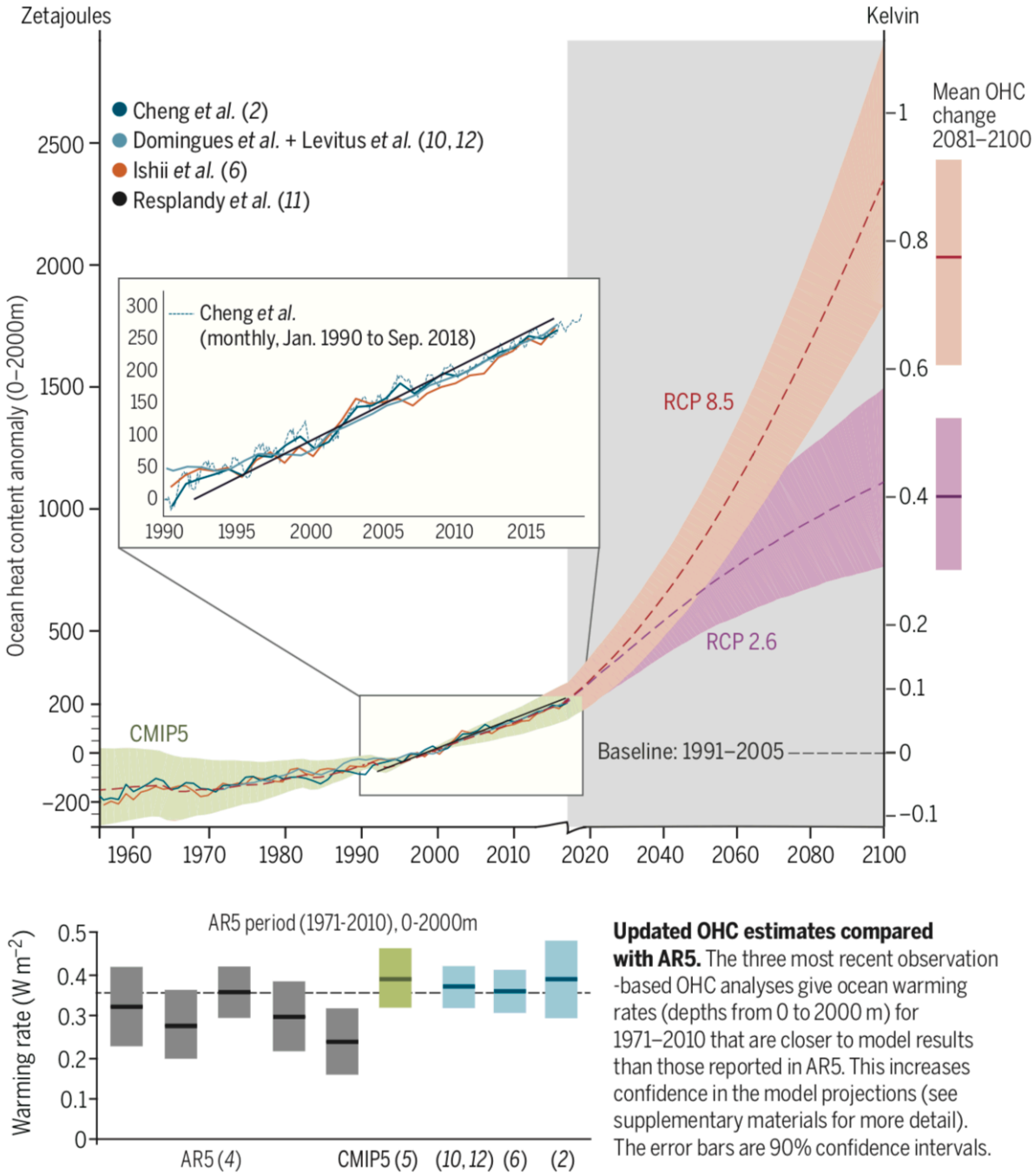


Figure IV.3. *(top) Comparison between observational GMST records using HadSST3 and the coastal hybrid SST for ocean temperatures with the blended CMIP5 multimodel mean. (bottom) Differences of each from the blended CMIP5 multimodel mean.*

A recognition of the uncertainty in observational records is particularly important for temperature series with large structural uncertainties, such as lower tropospheric temperatures [29,30] or ocean heat content.[57,58] Here models may disagree with some observational records and not others, or changes in observational records over time may bring observations into (or out of) agreement with model projections.

In the case of ocean heat content, apparent mismatches between observations and model projections have been due to observational biases later corrected.[57] The figure below, from our 2019 *Science* paper, shows that the three updated or new ocean heat content records published subsequent to the IPCC AR5 show substantially faster OHC warming than most of the five OHC series featured in the AR5 – and agree quite well with the CMIP5 multi-model mean.



Updated OHC estimates compared with AR5. The three most recent observation-based OHC analyses give ocean warming rates (depths from 0 to 2000 m) for 1971–2010 that are closer to model results than those reported in AR5. This increases confidence in the model projections (see supplementary materials for more detail). The error bars are 90% confidence intervals.

Figure IV.4. *Recent 0-2000m observational OHC records compared to CMIP5 model projections (top panel), and OHC records featured in the AR5 compared to newer/updated post-AR5 OHC estimates and CMIP5 models over the 1971-2010 period highlighted in the AR5 (bottom panel). Figure 1 from Cheng et al 2019.*[57]

2. EVALUATING HISTORIC MODEL PERFORMANCE

Current-generation climate models are often compared to observations through "hindcasts" where observationally-based radiative forcing estimates are used to project historical temperatures from the mid-1800s onward. However, these hindcasts are not always an independent test of model skill. Some modeling groups have explicitly selected tunable parameters to improve GMST hindcast performance,[32] while others have implicitly done so, using poor hindcast performance as a reason to reassess parameter choices.[33]

Evaluating the performance of future GMST projections from past climate models provides a more robust test of model skill. However, given the magnitude of internal climate variability, it is difficult to evaluate model projections until at least 15 years after the model was published. Other metrics, such as OHC, may have shorter emerge times and allow faster evaluation of model performance.

One challenge of evaluating future model projections is that they are subject to two independent sources of uncertainty: uncertainty in the model's representation of physically processes governing GMST, and uncertainty in the future forcings projected by the model. Even if a physically-perfect model existed 50 years ago, it still could provide a poor projection of future warming if it substantially over or under-estimated future atmospheric CO_2 concentrations or other forcings. Climate modelers should not necessarily be judged on their ability to forecast future emissions – indeed, these are dependent on human factors that are inherently much less predictable than atmospheric physics. Many early climate models simply assumed that atmospheric CO_2 would increase by 1% per year, for example.

In our 2019 Geophysical Research Letters paper (included later in this chapter), we undertook an assessment of all the climate model projections published since the first climate models were published in the late 1960s.[34] We identified 17 projections from 14 different models published between 1970 and 2001 (the original Manabe and Wetherald 1967 model had no temporally-specific warming projection, while models post-2001 had too short a forecast period for evaluation). These include the energy balance models used for the main-text projections of the first three IPCC assessment reports.

We compared the temperature trends in models and observations in the years after they were published. Using this metric (and accounting for uncertainties both in models and observations), we found that 10 of 17 model projections were consistent with observations – e.g. differences in trends between models and observations over the model future projection period were not significantly different from zero.

To account for mismatches between model and observed forcings post-publication, we conducted an additional test where we compared the implied transient climate response (TCR) for both models and observations. Implied TCR is essentially just the ratio of the change in temperature to the change in radiative forcing, but provides a useful way to control for forcing mismatches without having to rerun old climate models using modern observational forcing estimates. It is an imperfect metric, as it will not work well when model and observational forcings differ dramatically and push the system further from or closer to equilibrium conditions (as discussed in more detail in the paper), but is better than not accounting for forcing mismatches. Under this metric 14 of 17 models were consistent with observations, though the uncertainty in observational implied TCR is relatively large.

While the old climate models are functionally obsolete and do not include the many improvements made by modern Earth System Models, the fact that both classes of climate model did so well in projecting future warming should increase our confidence that current climate models are getting things right for mostly the right reasons. While there are still real uncertainties in future warming associated with climate sensitivity, we can confidently state that the rate of surface warming we are experiencing today is pretty much what past climate models projected it would be.

3. ROLE OF INTERNAL VARIABILITY IN 20TH CENTURY TEMPERATURES

The role of variability due to natural ocean cycles in global warming is a long-standing debate in climate science. The scientific community overwhelmingly agrees that human activities are responsible for the observed increase in temperatures for the last half-century. However, the relative influences of natural drivers of climate change – such as volcanic eruptions, ocean cycles, and the sun – on warmer and cooler phases superimposed on the long-term warming trend is still an area of active research.

The “early warming period” between 1915 and 1945 has long been a challenge for scientists to explain. Prior studies have suggested that about half the observed warming during this period is attributable to factors that are “external” to the climate – such as human-caused greenhouse gas emissions, volcanic eruptions and variability in the sun’s output.[63] The remaining half are attributed to “internal” factors – natural fluctuations within the climate system itself. This has led to suggestions that there may be long-term ocean cycles operating over 60- to 70-year periods which influence global temperatures.[64] They are commonly associated with the Atlantic Multidecadal Variability index (AMV).

In a paper published in the Journal of Climate we challenged this prevailing view.[65] We found that virtually all of the observed changes in global average temperatures over the past 170 years are caused by external drivers, leaving little room for an “unforced” internal ocean contribution. This means that ocean cycles on timescales of 60-70 years are unlikely to be a factor in the observed evolution of global temperatures since 1850. Instead, external factors, such as periods of strong volcanic activity and the release of aerosol particles (air pollution), have caused temperatures to fluctuate.

To determine the effects of external drivers on global temperatures, we used a two-box impulse response model, which transfers the forcing estimate into an associated temperature response. This allows us to include both fast and slow climate responses to the different drivers, and reflects the role that the ocean plays in buffering the rate of warming observed. Technical details of the model used can be found in Haustein et al 2019 in Appendix B.

The figure below shows observed global temperatures (in black) compared to the model using climate forcings (yellow) and climate forcings that include ENSO conditions (light blue).

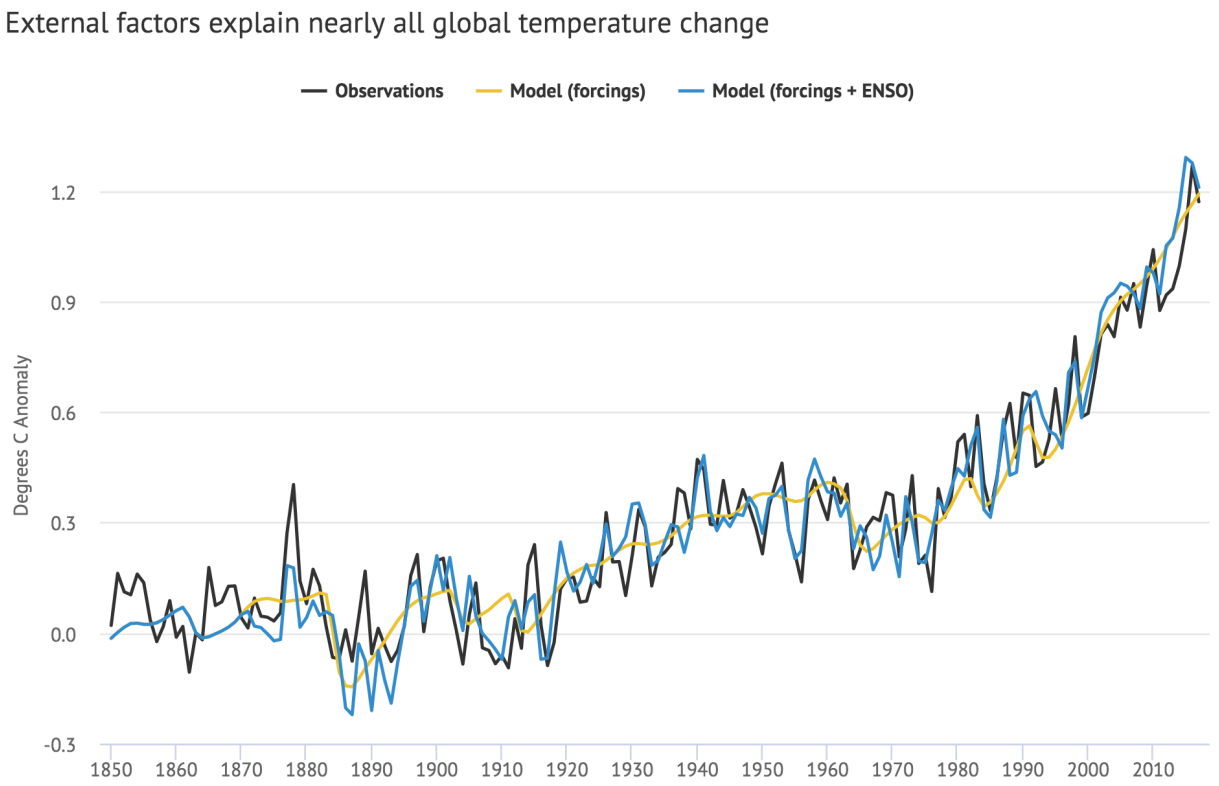


Figure IV.5. *Global surface temperatures from observations (Cowtan/Way land temperature data combined with HadISST2 sea surface temperature data over ocean, in black) and model results for forcings-only (yellow) and forcings plus ENSO (light blue). The temperature anomalies are expressed relative to 1850-1879. Based on Figure 5 in Haustein et al 2019.*[65]

In our model, virtually all (97-98%) of the long-term changes in temperature can be explained by external forcing. This approach uses a more precise description of the anthropogenic aerosol feedback processes (warming effect of black-carbon pollution and cooling effect of sulphate particles from industrial combustion) and removes biases in sea surface temperature (SST) records caused by a change in the way measurements were taken around the second world war. However, even without these updated forcings and observational estimates, this approach captures a substantial portion of the variability in global temperature.

The model effectively matches temperatures over both land and ocean. The figure below shows model results for land (orange) compared to land temperature observations (red), as well as similar values for sea surface temperatures (dark blue for observations, light blue for model results).

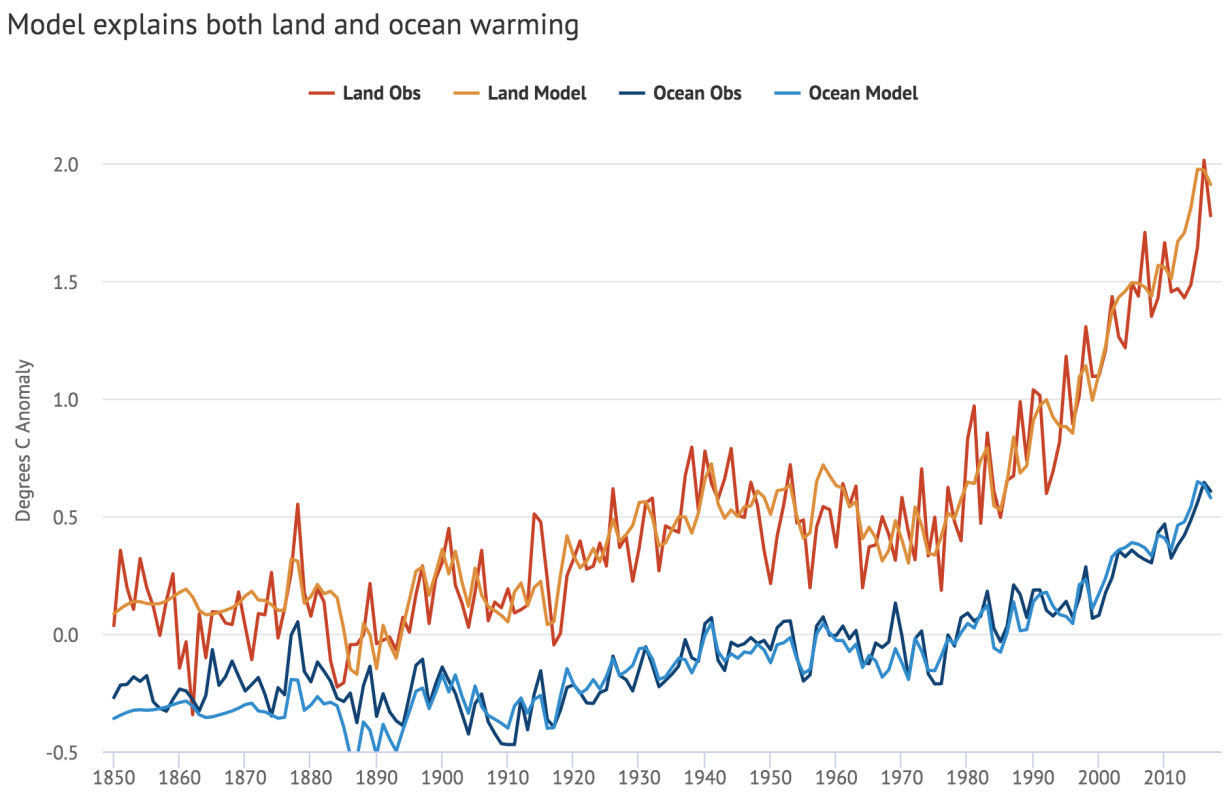


Figure IV.5. *Global land and ocean surface temperatures from observations and model results for forcings plus ENSO (light blue). The temperature anomalies are expressed relative to 1850-1879; ocean temperatures have been offset by -0.3°C to avoid overlap.*

The model also allows us to attribute temperature changes to different forcings. The figure below shows a breakdown of the different factors contributing to global surface temperatures, including human forcings (greenhouse gases and aerosols), natural forcings (volcanoes and solar) and short-term variations due to ENSO. The black dots show the observed temperature record and the grey line shows the model simulation that incorporates all the different drivers.

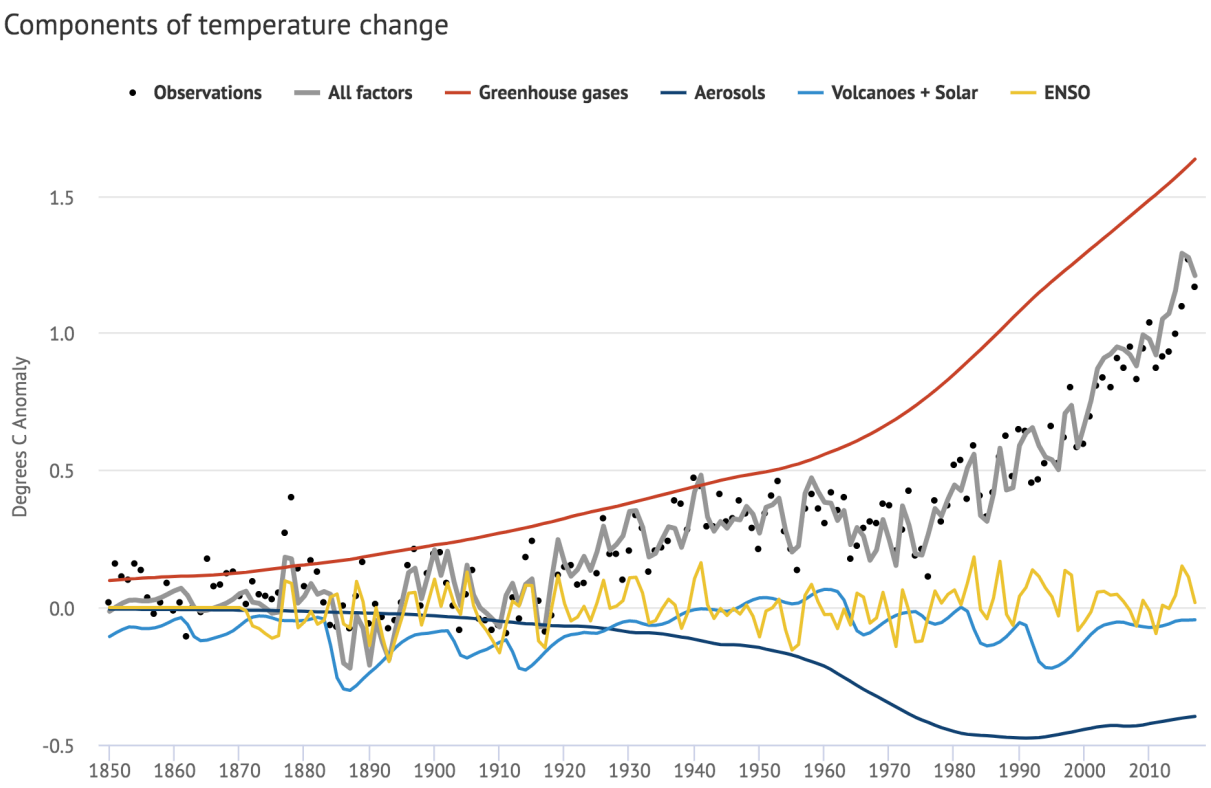


Figure IV.6. *Global surface temperatures from observations (black) and model results for all factors (grey), greenhouse gases (red), aerosols (dark blue), natural forcings (light blue) and the short-term variability due to ENSO (yellow). The temperature anomalies are expressed relative to 1850-1879.*

In addition to comparisons with observed temperatures between 1850 and present, we can extend the model into the more distant past – back to the year 1500 – using estimates of past climate forcings. The results can be compared to temperature reconstructions based on climate proxies.

The figure below shows the results when we extend the model back to 1500 (in red), compared both to a palaeoclimate reanalysis dataset (NTrend2015 – in the bold orange line) and individual proxy records (orange).[66] Temperatures for the Northern Hemisphere are shown, as that is where a large number of palaeoclimate temperature reconstructions are available.

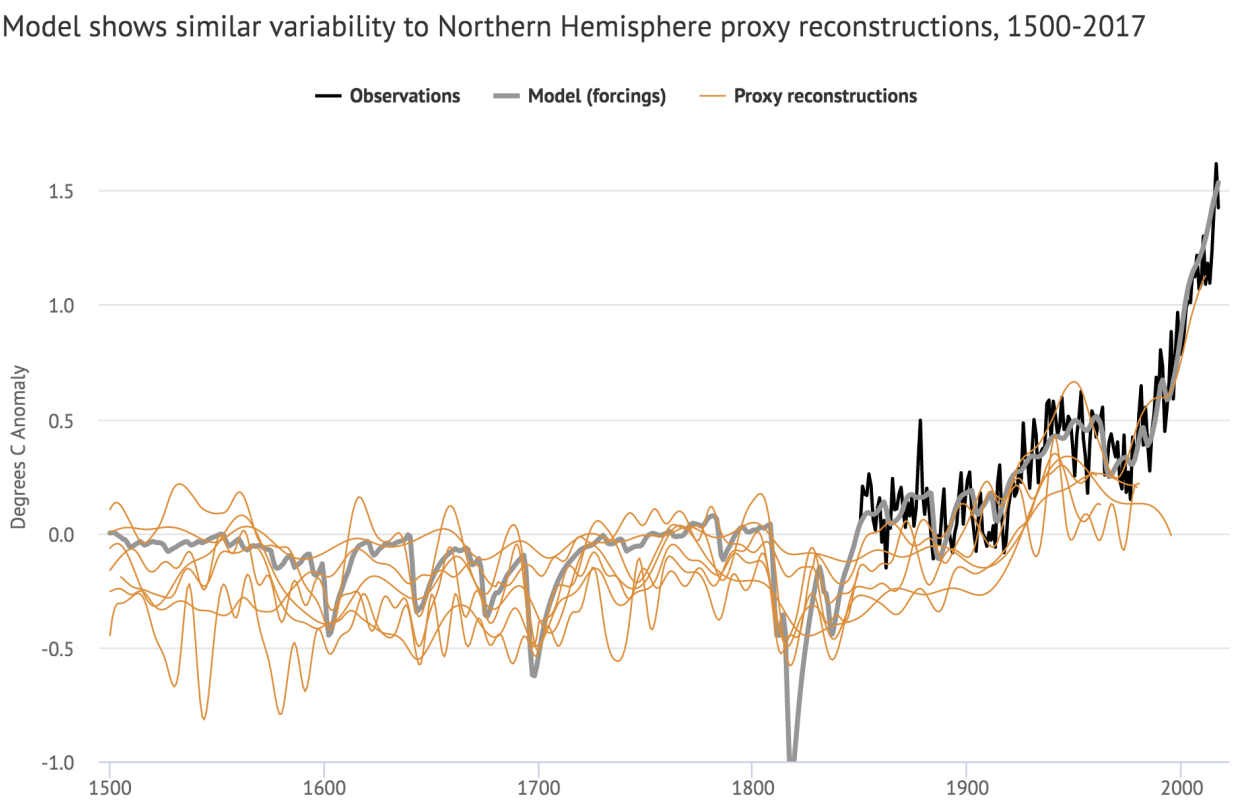


Figure IV.7. *Northern Hemisphere surface temperatures from observations (black) and model results for all factors (grey), along with seven different proxy-based paleoclimate estimates.*

During that period from 1500 to present, the model captures most of the multidecadal variability present in the proxy data. This improves our confidence that there are not large sources of internal variability missing from the model – at least over the past 500 years or so.

While the climate system continues to be influenced by short-term natural variability from El Niño and La Niña events, the hypothesis that oceans have been driving the climate into colder or warmer periods for multiple decades in the past – and that they may do so in the future – is unlikely to be correct. Most of the complex global climate models strongly support the hypothesis that oceans have only limited ability to alter global temperatures on multidecadal timescales. This study provides a support for those model results.

This means that we can expect future warming to be primarily driven by external forcing factors – such as human-caused greenhouse gas emissions – along with the variability associated with ENSO.

There are still some differences between past complex climate model simulations and observations. However, in our Journal of Climate paper we suggest that these models should

use an earlier model start date that includes strong volcanic eruptions in the early 1800s – which are still impacting global temperatures in the mid-to-late 1800s and likely even longer – which in turn would help improve the agreement between the two. Updated climate forcings – which are being included in the upcoming CMIP6 modelling project – will also help resolve some of the historical disagreements.

4. PAPER 4: EVALUATING THE PERFORMANCE OF PAST CLIMATE MODEL PROJECTIONS

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

[1] Energy and Resources Group, University of California, Berkeley. 310 Barrows Hall, Berkeley, CA 94720, 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, 77 Massachusetts Ave., Cambridge, MA, USA.
[4] NASA Goddard Institute for Space Studies, 2880 Broadway, New York, USA.

Geophysical Research Letters, 2019

KEY POINTS

- Evaluation of uninitialized multi-decadal 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.

ABSTRACT

Retrospectively comparing future model projections to observations provides a robust and independent test of model skill. Here we analyse 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 projects 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.

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 (Schmidt et al. 2017; Mauritsen et al. 2019; Gettelman et al. 2019). 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 and Wetherald 1967) or general circulation models (e.g. Manabe and Wetherald 1975). Simple energy balance models have since been gradually side-lined 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, Stouffer et al. 1989, IPCC 2013).

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 and Sutton, 2012).

Researchers have previously evaluated prior model projections from the Hansen et al. (1988) NASA Goddard Institute for Space Studies model (Rahmstorf et al. 2007, Hargreaves et al. 2010), the Stouffer and Manabe (1989) Geophysical Fluid Dynamics Laboratory model (Stouffer and Manabe 2017), the IPCC First Assessment Report (IPCC 1990; Frame and 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 analyses 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 modellers 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 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 and 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 (Reichler and Kim 2008, Knutti et al. 2013). 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 post-publication projection periods.

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 fourteen 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 First Assessment Report (FAR – IPCC 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), Rascool and Schneider 1971 (RS71), Sawyer 1972 (S72), Broecker 1975 (B75), Nordhaus 1977 (N77), Schneider and Thompson 1981 (ST81), Hansen et al. 1981 (H81), Hansen et al. 1988 (H88), and Manabe and Stouffer 1993 (MS93). The energy balance model (EBM) 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 supplementary materials.

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 supplementary materials.

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 – NASA GISTEMP (Lenssen et al. 2019), NOAA GlobalTemp (Vose et al. 2012), Hadley/UEA HadCRUT4 (Morice et al. 2012), Berkeley Earth (Rohde et al. 2013), and Cowtan and Way (Cowtan and 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 vs time, and change in temperature vs 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 introducing sharp inter-annual changes in forcing that would complicate the interpretation of TCR over shorter time periods. For the observational record, ΔF_{anthro} is based on a 1000-member ensemble of observationally-informed forcing estimates (Dessler and Forster 2018). Model forcings are recomputed from published formulas and tables when possible and otherwise digitized from published figures (see supplementary section S2 for details). Details on the approach used to calculate implied TCR can be found in supplementary materials 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 20th century anthropogenic forcing falls within this regime (Gregory and Mitchell 1997, Gregory and Forster 2008, 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 1.3x 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 between the modelled and observed metrics includes 0, the two are deemed consistent; otherwise, they are inconsistent. Additionally, we follow the approach of Hargreaves (2010) in calculating a skill score for each model for both temperature vs time and implied TCR metrics. This skill score is based on the root-mean-squared errors of the model projection trend vs observations compared to a zero-change null-hypothesis projection. See supplementary materials section S1.3 for details on calculating consistency and skill scores.

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 8 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.

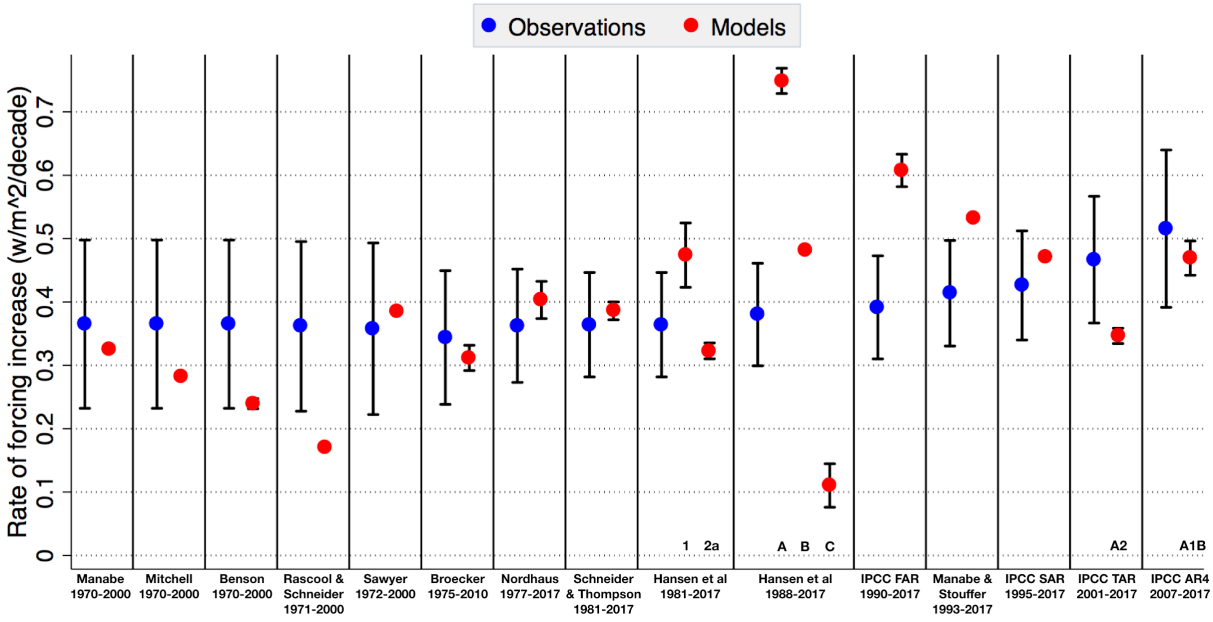


Figure 1. *Rate of external forcing increase (in watts per meter squared per decade) in models and observations over model projection periods.*

Comparisons between climate models and observations over model projection periods are shown in Figure 2 for both temperature vs. 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 overvations under both metrics. Using the temperature vs time metric, 10 of the 17 model projections show results consistent with observations. Of the remaining 7 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.

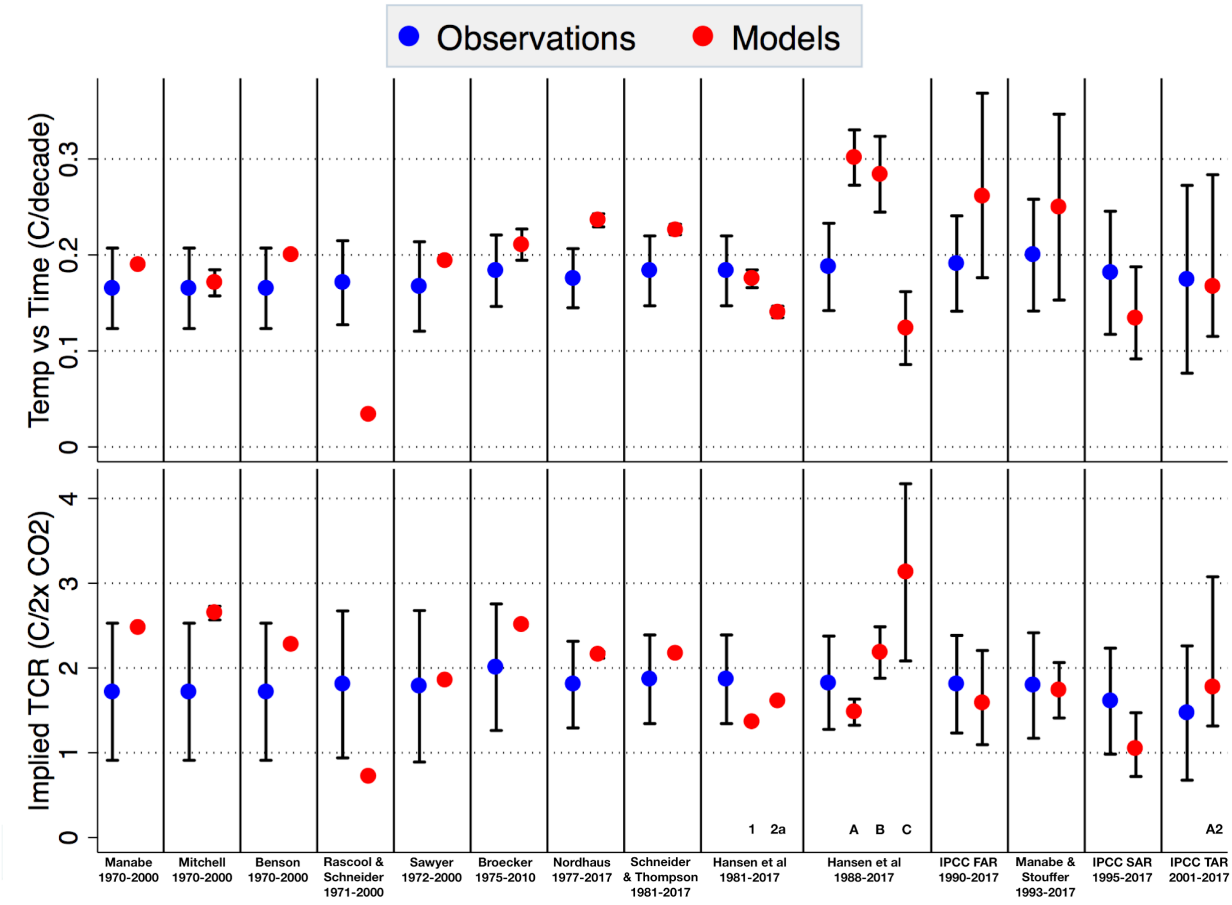


Figure 2: *Comparison of trends in temperature vs time (top panel) and implied TCR (bottom panel) between observations and models over the model projection periods displayed at the bottom of the figure. Figure S1 shows a variant of this figure with the AR4 projections included.*

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 (see supplementary text S2 for a discussion of the anomalously low-ECS model used in RS71).

A number of model projections were inconsistent with observations on a temperature vs 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, 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, 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 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 30th anniversary of the event in 2018 focused considerable attention on the accuracy of the projection (United States. Cong. Senate 1988; Borenstein and Foster, 2018). 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).

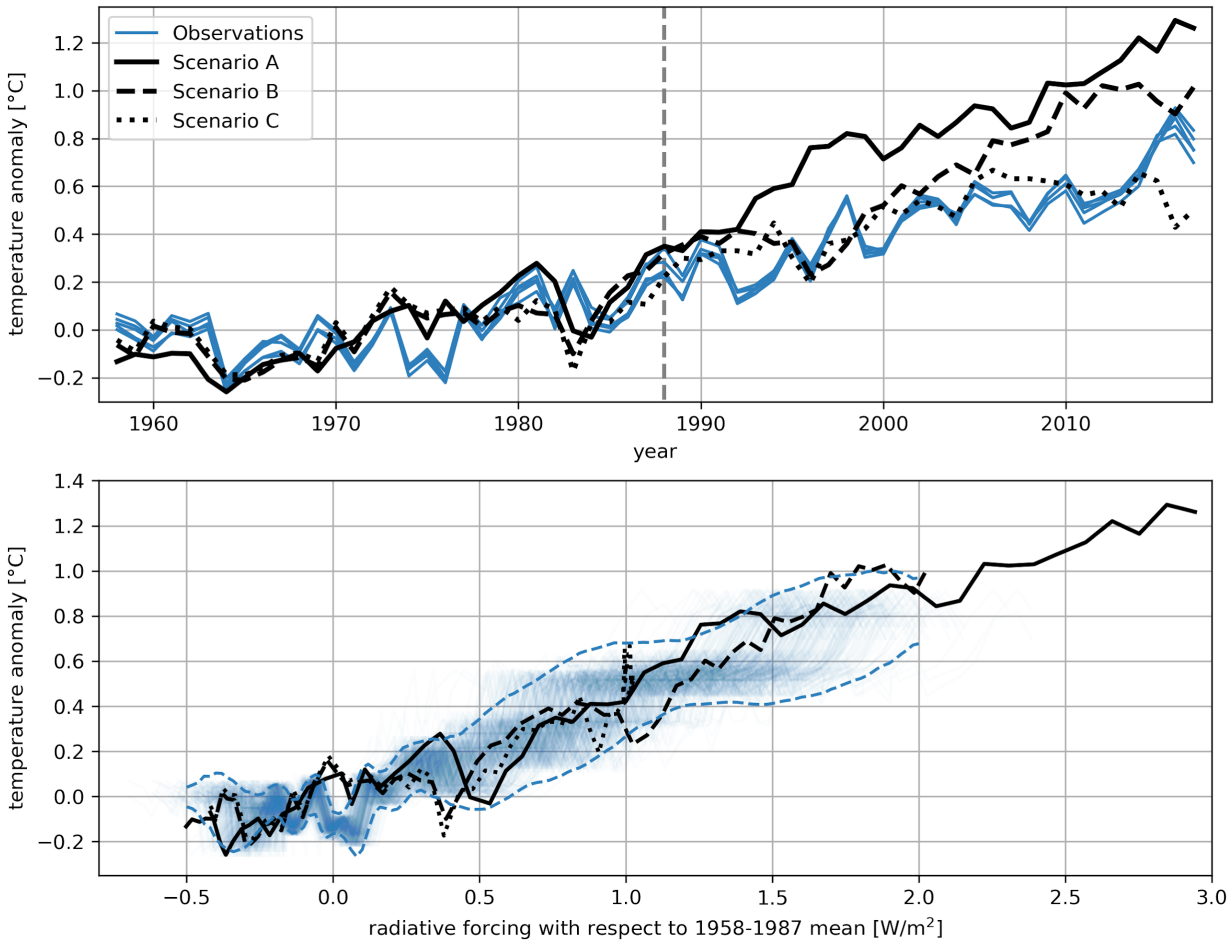


Figure 3: *Hansen et al. 1988 projections compared with observations on a temperature vs. time basis (top) and temperature vs external forcing (bottom). The dashed grey line in the top panel represent the start of the projection period. The transparent blue lines in the lower panel represent 500 random samples of the 5000 combinations of the 5 temperature observation products and the 1000 ensemble members of estimated forcings (the full ensemble is subsampled for visual clarity). The dashed blue lines show the 95% confidence intervals for the 5000 member ensemble (see supplementary text S1.4 for details). Anomalies for both temperature and forcing are shown relative to a 1958-1987 pre-projection baseline.*

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

Model	Timeframe	ΔT / Δt skill	ΔT / ΔF 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]

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. Both temperature vs time (ΔT / year) and implied TCR (ΔT / ΔF) median scores and uncertainties are shown.*

The average of the median skill scores across all the model projections evaluated is 0.69 for the temperature vs 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 parts per million over current levels, with projected CO_2 concentrations increasing up to twice as fast as actually observed. 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 GCM results, as relatively few individual model runs were available at the time. From the AR4 onward IPCC projections were based on the multi-model mean and model spread. We examine individual models from the first three IPCC reports on both a temperature vs time and implied TCR basis in Figure S5.

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).

A number of high-profile model projections – H88 scenarios A and B and the IPCC FAR in particular – have been criticised for projecting higher warming rates than observed (e.g. Michaels and 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 multi-decadal 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.

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 and N. Foster (2018) Warned 30 years ago, global warming ‘is in our living room’, Associated Press, https://www.apnews.com/dbd81ca2a7244ea088a8208bab1c87e2 , June 18, 2018. (last accessed Aug 22, 2019)

Broecker, W. S. (1975). Climatic Change: Are We on the Brink of a Pronounced Global Warming? Science, 189(4201), 460 LP-463. https://doi.org/10.1126/science.189.4201.460

Cowtan, K. and Way, R. G. (2014), Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. Q.J.R. Meteorol. Soc., 140: 1935-1944. doi:10.1002/qj.2297

Cowtan, K., Hausfather, Z., Hawkins, E., Jacobs, P., Mann, M. E., Miller, S. K., Steinman, B. A., Stolpe, M. B., and Way, R. G. (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.

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., and Stone, D. A. (2012). Assessment of the first consensus prediction on climate change. Nature Climate Change, 3, 357. Retrieved from 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., and Forster, P. M. (2008), Transient climate response estimated from radiative forcing and observed temperature change, J. Geophys. Res., 113, D23105, doi:10.1029/2008JD010405.

Gregory, J. M., and Mitchell, J. F. B. (1997). The climate response to CO2 of the Hadley Centre coupled AOGCM with and without flux adjustment. Geophysical Research Letters, 24(15), 1943–1946. https://doi.org/10.1029/97GL01930

Gregory, J. M., Ingram, W. J., Palmer, M. A., Jones, G. S., Stott, P. A., Thorpe, R. B., Lowe, J. A., Johns, T. C., and Williams, K. D. (2004), A new method for diagnosing radiative forcing and climate sensitivity, Geophys. Res. Lett., 31, L03205, doi:10.1029/2003GL018747.

Hansen, J., Fung, I., Lacis, A., Rind, D., Lebedeff, S., Ruedy, R., Russell, G., and Stone, P., (1988). Global climate changes as forecast by Goddard Institute for Space Studies three-dimensional model. J. Geophys. Res., 93, 9341-9364, doi:10.1029/JD093iD08p09341.

Hansen, J., Johnson, D., Lacis, A., Lebedeff, S., Lee, P., Rind, D., and Russell, G. (1981). Climate Impact of Increasing Atmospheric Carbon Dioxide. Science, 213(4511), 957 LP-966. https://doi.org/10.1126/science.213.4511.957

Hansen, J., Russell, G., Lacis, A., Fung, I., Rind, D., and Stone, P. (1985). Climate Response Times: Dependence on Climate Sensitivity and Ocean Mixing. Science, 229(4716), 857 LP-859. https://doi.org/10.1126/science.229.4716.857

Hargreaves, J.C. (2010). Skill and uncertainty in climate models. Wiley Interdisciplinary Reviews: Climate Change, vol. 1, pp. 556-564, 2010. http://dx.doi.org/10.1002/wcc.58

Hawkins, E., and Sutton, R. (2012), Time of emergence of climate signals, Geophys. Res. Lett., 39, L01702, doi:10.1029/2011GL050087.

Held, I.M., M. Winton, K. Takahashi, T. Delworth, F. Zeng, and G.K. Vallis (2010): Probing the Fast and Slow Components of Global Warming by Returning Abruptly to Preindustrial Forcing. J. Climate, 23, 2418–2427, https://doi.org/10.1175/2009JCLI3466.1

IPCC (AR4) (2007). Climate Change 2007: The Physical Science Basis, Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Solomon, S.; Qin, D.; Manning, M.; Chen, Z.; Marquis, M.; Averyt, K.B.; Tignor, M.; and Miller, H.L. (eds.), Cambridge University Press, ISBN 978-0-521-88009-1 (pb: 978-0-521-70596-7).

IPCC (AR5) (2013). Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1535 pp.

IPCC (FAR) (1990). Climate Change: The IPCC Scientific Assessment. Report prepared by Working Group I. Houghton, J.T., Jenkins, G.J., and Ephraums, J.J. (eds). Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 365 pp.

IPCC (SAR) (1996), 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. Houghton, J.T.; Meira Filho, L.G.; Callander, B.A.; Harris, N.; Kattenberg, A., and Maskell, K. (eds.), Cambridge University Press, ISBN 0-521-56433-6 (pb: 0-521-56436-0).

IPCC (TAR) (2001). Climate Change 2001: The Scientific Basis, Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change. Houghton, J.T.; Ding, Y.; Griggs, D.J.; Noguer, M.; van der Linden, P.J.; Dai, X.; Maskell, K.; and Johnson, C.A. (eds.), Cambridge University Press, ISBN 0-521-80767-0 (pb: 0-521-01495-6).

Knutti, R., Masson, D., and Gettelman, A. (2013), Climate model genealogy: Generation CMIP5 and how we got there, Geophys. Res. Lett., 40, 1194– 1199, doi:10.1002/grl.50256.

Lenssen, N. J. L., Schmidt, G. A., Hansen, J. E., Menne, M. J., Persin, A., Ruedy, R., & Zyss, D. (2019). ImpImprovements 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: Singer S.F. (eds) Global Effects of Environmental Pollution. Springer, Dordrecht.

Manabe, S. and Strickler, R.F. (1964). Thermal Equilibrium of the Atmosphere with a Convective Adjustment. J. Atmos. Sci., 21, 361–385, https://doi.org/10.1175/1520-0469(1964)021<0361:TEOTAW>2.0.CO;2

Manabe, S. and Wetherald, R.T. (1975). The Effects of Doubling the CO_2 Concentration on the climate of a General Circulation Model. J. Atmos. Sci., 32, 3–15, https://doi.org/10.1175/1520-0469(1975)032<0003:TEODTC>2.0.CO;2

Manabe, S., and Stouffer, R. J. (1993). Century-scale effects of increased atmospheric C02 on the ocean–atmosphere system. Nature, 364(6434), 215–218. https://doi.org/10.1038/364215a0

Manabe, S., and 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., Smagorinsky, J., and Strickler, R.F. (1965). Simulated Climatology Of A General Circulation Model With A Hydrologic Cycle. Mon. Wea. Rev., 93, 769–798, https://doi.org/10.1175/1520-0493(1965)093<0769:SCOAGC>2.3.CO;2

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-ESM1.2) and its response to increasing CO_2. Journal of Advances in Modeling Earth Systems, 11, 998– 1038. https://doi.org/10.1029/2018MS001400

Michaels, P., and 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: Singer S.F. (eds) Global Effects of Environmental Pollution. Springer, Dordrecht, p. 139.

Morice, C.P., Kennedy, J.J., Rayner, N.A. and 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, doi: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., … Allen, M. R. (2013). Energy budget constraints on climate response. Nature Geoscience, 6, 415. Retrieved from https://doi.org/10.1038/ngeo1836

Rahmstorf, S., Cazenave, A., Church, J. A., Hansen, J. E., Keeling, R. F., Parker, D. E., and Somerville, R. C. J. (2007). Recent Climate Observations Compared to Projections. Science, 316(5825), 709 LP-709. https://doi.org/10.1126/science.1136843

Rahmstorf, S., Foster, G., and 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. and Schneider, S.H. (1971) Atmospheric Carbon Dioxide and Aerosols: Effects of Large Increases on Global Climate. Science, 173, 138-141.

Reichler, T. and Kim, J. (2008). How Well Do Coupled Models Simulate Today's Climate?. Bull. Amer. Meteor. Soc., 89, 303–312, https://doi.org/10.1175/BAMS-89-3-303

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

Robert Rohde, Richard A. Muller, et al. (2013) A New Estimate of the Average Earth Surface Land Temperature Spanning 1753 to 2011. Geoinfor Geostat: An Overview 1:1.. doi:10.4172/gigs.1000101

Rohrschneider, T., Stevens, B., & Mauritsen, T. (2019). On simple representations of the climate response to external radiative forcing. Climate Dynamics. 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., Molod, A., Neale, R., and Saha, S. (2017). Practice and philosophy of climate model tuning across six U.S. modeling centers. Geosci. Model Dev., 10, 3207-3223, doi:10.5194/gmd-10-3207-2017.

Schneider, S. H., and Thompson, S. L. (1981), Atmospheric CO_2 and climate: Importance of the transient response, J. Geophys. Res., 86(C4), 3135– 3147, doi:10.1029/JC086iC04p03135.

Stouffer, R. J., and Manabe, S. (2017). Assessing temperature pattern projections made in 1989. Nature Climate Change, 7, 163. Retrieved from https://doi.org/10.1038/nclimate3224

Stouffer, R. J., and Manabe, S. (2017). Assessing temperature pattern projections made in 1989. Nature Climate Change, 7, 163. Retrieved from https://doi.org/10.1038/nclimate3224

Stouffer, R. J., Manabe, S., and 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., … Rose, S. K. (2011). The representative concentration pathways: an overview. Climatic Change, 109(1), 5. https://doi.org/10.1007/s10584-011-0148-z

Vose, R. S., Arndt, D., Banzon, V. F., Easterling, D. R., Gleason, B., Huang, B., … Wuertz, D. B. (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

ACKNOWLEDGEMENTS

ZH conceived of the project, ZH and HFD created the figures, and ZH, HFD, TA, and GS helped gather data and wrote the article text. A public GitHub repository with code used to analyze the data, generate figures, and csv files containing the data shown in the figures is available here: https://github.com/hausfath/OldModels. Additional information on the code and data used in the analysis can be found in the supplementary materials. We would like to thank Piers Forster for providing the ensemble of observationally-informed radiative forcing estimates.

FUNDING

No dedicated funding from any of the authors supported this project

SUPPLEMENTARY MATERIALS

Contents of this file

Links to data and code

Text S1. Detailed methods.

Text S2. Detailed description of how each climate model projection was assessed.

Supplementary Figures S1-S6.

DATA AND CODE

A spreadsheet with tabs containing data from all of the models evaluated in this study is available here:
https://github.com/hausfath/OldModels/blob/master/Model%20data%20spreadsheet.xlsx

A public GitHub repository with code used to analyze the data, generate figures, and csv files containing the data shown in the figures is available here:
https://github.com/hausfath/OldModels

The 1000-member ensemble of observationally-informed radiative forcing estimates can be found here: https://github.com/hausfath/OldModels/tree/master/forcing_data

Observational temperature datasets can be found at the following links:
NASA GISTEMP – https://data.giss.nasa.gov/gistemp/
NOAA GlobalTemp – https://www.ncdc.noaa.gov/monitoring-references/faq/anomalies.php
Hadley HadCRUT4 – http://www.metoffice.gov.uk/hadobs/hadcrut4/
Berkeley Earth – http://berkeleyearth.org/data/
Cowtan and Way – http://www-users.york.ac.uk/~kdc3/papers/coverage2013/series.html

TEXT S1: DETAILED METHODS

S1.0: Additional methods notes

The choice to start the future projection period at the date of publication was made as a conservative choice to avoid any possibility of observed temperatures informing model development or parameterization. While in some cases the specific date on which the model was run prior to the paper publication is known, in most cases (particularly for earlier studies) this is not readily available. In other cases (e.g. for the IPCC AR4) models were run with projected future forcings that start well before the model was developed, which does not completely preclude the knowledge of observed temperatures in the intervening period from informing the development and tuning of parameterizations but is unlikely given the multi-year timescale of model development.

While a more complex two-layer model with ocean heat uptake would be able to better capture the relationship between forcing and temperature response than our simple implied TCR metric (Rohrschneider et al. 2019), we have purposefully chosen to avoid a situation where we are using a more complex model with its own somewhat uncertain parameterizations to evaluate the performance of historical climate models. A more complex model may also not provide an effective comparison with early climate model projections, many of which (prior to ST81) did not include ocean heat uptake dynamics. is not an optimal metric in all cases, but will provide a more accurate evaluation of model projection performance than the conventional approach of analyzing changes in temperature over time without accounting for differences in the time evolution of modeled and observationally-estimated radiative forcings.

While our analysis uses instantaneous radiative forcings either calculated from modeled CO2 concentrations or based on published data from past climate models, it does not account for differing forcing efficacies (Hansen et al 2005; Marvel et al 2016). When taken into account – based on efficacies from the GISS model – efficacy-adjusted forcings are around 3% higher at present, and representing an additional source of uncertainty in our analysis.

S1.1: Temperature vs time

To evaluate the performance of model projections against observed temperatures, the linear trend in both observations and model projections was calculated over the future projection period. An ordinary least squares approach was used to calculate the trend coefficient of all five observational temperature records over the future projection period. A first-order autoregressive model (AR1) was further used to estimate trend uncertainties, similar to the approach used in Hausfather et al. (2017).

Specifically, trend coefficients of temperature with respect to time,β, were estimated with the ordinary least squares model:

$$y_i = \beta x_i + \varepsilon_i \quad (1)$$

The uncertainty introduced by the choice of observational estimate was calculated from the variance of the five coefficients ($\beta_1 \ldots \beta_5$):

$$var(obs) = \sum(\beta_i - \mu)^2/N \tag{2}$$

whereμ is the average of the five coefficients and $N = 5$.

An AR1 model was used to estimate the regression confidence intervals:

$$X_t = c + \rho X_t + \varepsilon_t \tag{3}$$

where c is a constant, ε_t is white noise, and$\rho$is the model parameter. The variance for the regression of a given observational temperature record i can be calculated by:

$$var(X_t)_i = {\sigma_e}^2/(1 - \rho^2). \tag{4}$$

The ordinary least squares model provides a more physically meaningful coefficient than the AR(1) model, while the AR(1) model provides a better estimate of the variance (accounting for autocorrelation). These were calculated separately for each of the five observational temperature datasets. In cases where the confidence intervals of the regression coefficient from the AR(1) approach were smaller than those from an OLS model, the OLS coefficients were used to provide a more conservative estimate.

A mean and combined uncertainty were estimated by averaging the five coefficients and by adding the (two-sigma) coefficient uncertainty and the mean AR(1) (two-sigma) trend uncertainty in quadrature, assuming that the two are independent:

$$\bar{\beta} \pm \sqrt{4 \cdot var(obs) + 4 \cdot \overline{var(X_t)}}. \tag{5}$$

As the standard deviation is the square root of the variance, $\left(2 \cdot \sqrt{var}\right)^2 = 4 \cdot var$. For models, where only a single realization of projected temperatures is available, the same approach was used except with a single β and $var(X_t)$.

S1.2: Implied TCR

Implied TCR is defined as the ratio between the change in temperature and the change in external forcing over the model projection period, for both models and observations. It is referred to as 'implied' as it differs from the traditional definition of TCR, which is typically based on idealized experiments where CO_2 is increased by 1% per year (Cubasch et al. 2001).

When explicit external forcing values were not available, they were estimated from model greenhouse gas concentrations using the simplified radiative forcing functions from the IPCC AR5 (Myhre et al. 2013). Forcing from a change in atmospheric concentration of CO_2 is given by:

$$\Delta F_{co_2} = 5.35 \cdot ln \frac{(P_{CO_2} + \alpha_{CO_2})}{P_{CO_2}}. \tag{7}$$

Here P_{CO_2} represents the initial concentration of CO_2 in the atmosphere when the model projection period began, while α_{CO_2} represents the additional parts per million CO_2 added through the end of 2017 (or when the model run ended if prior to 2017).

The direct radiative forcing of a given increase of CH_4 and/or N_2O in the atmosphere can be approximated by:

$$\Delta F_{CH_4} = 0.036(\sqrt{P_{CH_4} + \beta_{CH_4}} - \sqrt{P_{CH_4}}) - f(P_{CH_4} + \beta_{CH_4},\ P_{N_2O}) + f(P_{CH_4}, P_{N_2O})$$

$$\Delta F_{N_2O} = 0.12(\sqrt{P_{N_2O} + \beta_{N_2O}} - \sqrt{P_{N_2O}}) - f(P_{CH_4}, P_{N_2O} + \beta_{N_2O}) + f(P_{CH_4}, P_{N_2O})$$

where:

$$f(M, N) = 0.47\ ln(1 + 2.01 \cdot 10^{-5}(MN)^{0.75} + 5.31 \cdot 10^{-15} M(MN)^{1.52}). \tag{8}$$

In this equation P_{CH_4} is the initial concentration of atmospheric CH_4, while β_{CH_4} is the addition being evaluated. P_{N_2O} is the initial concentration of N_2O, and β_{N_2O} is the addition being evaluated. The radiative forcing of both CH_4 and N_2O is a function of the combination of both, reflecting their interacting atmospheric chemistry.

We use a 1000-member ensemble of observationally-informed radiative forcing estimates from

Forster and Dessler (2018) to account for uncertainties in forcing associated with aerosol, land albedo, and other factors that are relatively poorly observationally constrained. The ensemble members are combined with each of the five observational temperature records to regress the change in temperature against the change in radiative forcing, following the approach used in Eqs. 1-5 but substituting radiative forcing for time and using an OLS rather than AR(1) approach for trend uncertainties given the absence of a time variable needed for an autoregressive model.

Specifically, a set of 5000 $\Delta T/\Delta F_{anthro}$estimates are calculated for each model projection period across the five observational temperature estimate and 1000 radiative forcing ensemble members. The mean of these estimates is calculated, and uncertainties are estimated based on both the variation across these 5000 estimates and on the mean confidence intervals of the regression coefficients. These uncertainties and added in quadrature as the two are independent:

$$\overline{TCR_{implied}} \pm \sqrt{4 \cdot var(TCR_{implied}) + 4 \cdot \overline{var(\beta)}} \quad (9)$$

where $var(\beta)$ is the variance of the OLS regression in Eq. 1, but regressing temperature against anthropogenic forcing rather than time.

Models, in turn, have a single realization of $\Delta T$and ΔF_{anthro} over their projection period, and the uncertainties are only estimated from $var(\beta)$.

S1.3: Calculating consistency and skill scores

We refer to model projections as consistent or inconsistent with observations based on a comparison of the differences between the two. Specifically, when comparing models on a temperature vs time basis, we difference the model and observation global mean surface temperature time series for each of the five observational time series. These difference series will remove any common variability between model projections and observations (Hausfather et al. 2017). Trends and trend confidence intervals for these difference series are then calculated following the approach in Eq. 5. Model projections and observations are considered consistent if the trend 95% confidence interval of the difference series is inclusive of zero, indicating that zero difference in trends cannot be ruled out.

When comparing model projections and observations on an implied TCR basis (e.g. change in temperature compared to change in forcing), using linear regressions on difference series is more problematic given the lack of shared time axis between the two. Instead, we assume that the trend uncertainties for each are independent, and add the uncertainties in quadrature to the difference in trends. Specifically, we calculate:

$$TCR_{diff} \pm \sqrt{4 \cdot var(TCR_{models}) + 4 \cdot var(TCR_{obs})}$$

where:

$$TCR_{diff} = TCR_{implied,model} - \overline{TCR_{implied,obs}}$$
$$var(TCR_{model}) = \sqrt{var(TCR_{implied,model})^2 + var(\beta_{i,model})^2}$$
$$var(TCR_{obs}) = \sqrt{var(TCR_{implied,obs})^2 + var(\beta_{i,obs})^2}.$$

(10)

Here we similarly consider model projections and observations to be consistent if the 95% confidence interval of the difference between the two is inclusive of zero. This approach produces results quite similar to those from the difference series approach used in the temperature vs time case, suggesting that the phase of internal variability in model projections and observations are largely independent.

Skill scores are calculated following the approach of Hargreaves (2010). The root-mean-squared errors of the projectioned trend, E_f, is compared to a reference technique E_{refr}, where E_{refr} is simply the assumption of temperature persistence (e.g. zero trend over time). As Hargreaves points out, the assumption of persistence generally outperforms the extrapolation of recent trends over any given interval in the historical global mean surface temperature record, at least prior to the last few decades. This serves as a reasonable counterfactual, particularly for early 1970s and 1980s models where the modern warming trend was less apparent to researchers at the time (Broecker 1975).

Skill scores, SS, are defined as:

$$SS = 1 - \sqrt{\frac{E_f}{E_{refr}}}$$

where:

$$E_f = \left(\beta_{i,obs} - \beta_{i,model}\right)^2$$
$$E_{refr} = \left(\beta_{i,obs} - 0\right)^2 .$$

(11)

Skill score uncertainties are estimated based on calculating skill scores separately for each model projection using the five different observational temperature records (for the temperature vs time metric) and the 5000 permutations of observational temperature record and observational forcing ensemble (for the implied TCR metric). The median skill score is calculated across all available runs for each metric. This is shown rather than the mean as the absolute value nature of the skill score means that a few ensemble members with very low skill can drag the mean skill score disproportionately down.

The uncertainties shown span the 5th to 95th percentile of resulting skill scores, accounting for both uncertainties from the choice of observational record and forcing series and the trend uncertainty due to temporal variability in the underlying time series. These are calculated via a Monte Carlo approach that takes the trend coefficient uncertainties into account. For the temperature vs time metric, 100 permutations of each of the five observational temperature records are estimated, where each randomly samples a value from the Gaussian distribution of the resulting regression trend coefficients. For the implied TCR metric, 100 values were randomly sampled from the Gaussian distribution of the resulting regression trend coefficients for each of the 5000 permutations of temperature record and observationally-based forcing series.

S1.4: Temperature uncertainties at a given forcing

The combination of 5 observed temperature time series and 1000 observationally-informed forcing time series gives an ensemble of 5000 estimates of how temperature varies as a function of radiative forcing. We can not immediately estimate uncertainty in temperature as a function of forcing because the forcing data points are not co-located. Thus, we define a regular grid of forcings with fine spacing of 0.02 W/m^2 and linearly interpolate the 5000 temperature values from the annual forcings values to the fine grid. We then calculate the sample standard

deviation across the 5000 member ensemble and estimate a 95% confidence interval at each forcing value. These confidence intervals for Hansen et al. 1988 and IPCC FAR are shown by the dashed blue lines in the lower panels of Figure 3 and Figure S6, respectively.

CLIMATE MODEL PROJECTION ASSESSMENT

This section provides detailed descriptions of how each historical climate model projection was digitized and analyzed, including what data points were available, if and what interpolation of data was applied, and what scenarios were used. When model projections were not available in a digital form, they were digitized from published figures using the free OS X application plotDigitizer: http://plotdigitizer.sourceforge.net/

Manabe 1970

Citing results from their previously published Manabe and Wetherald 1967 model, Manabe calculates the equilibrium surface air temperature in a one-dimensional radiative convective equilibrium model for a given distribution of relative humidity. Citing an increase in surface air temperature of 2.3°C as CO_2 concentrations are doubled from 300 ppm to 600 ppm, he uses an independent prediction of external forcing to predict transient warming in 2000, relative to 1900: "suppose the concentration of CO_2 increases by about 25% from AD1900 to AD2000 as the U.N. Department of Social and Economic Affairs predicts, the resulting increase of surface temperature would be about 0.8°C". It is unclear whether he carried out additional runs of the Manabe and Wetherald model to arrive at this number or simply scaled their previously calculated ECS of 2.36°C (table 5 of Manabe and Wetherald 1967) using the logarithmic dependence of CO_2 radiative forcing on CO_2 concentrations (equation 7), which gives

$$2.36 \times log(1.25) \,/\, log(2.0) \;= 0.759 \;\simeq 0.8°C.$$

To express this prediction as a change in radiative forcing and GMST between 1970 and 2000, we assume a CO_2 concentration of 300 ppm in 1900 and 320 ppm in 1970. The 320 ppm value is consistent with other papers published at the time (Mitchell et al. 1970; Benson et al. 1970; Rascool and Schneider 1971; Sawyer 1972), though lower than our current estimate of 1970 global CO_2 concentrations (325 ppm). The referenced prediction of a 25% increase from 1900 to 2000 thus predicts 375 ppm of CO_2 in 2000. Using equation 7 to convert CO_2 concentrations into a radiative forcing, we determine a predicted increase of radiative forcing of $\Delta F = 0.85\ W/m^2$between 1970 and 2000. To determine the predicted increase in GMST between 1970 and 2000 from the increase in GMST between 1900 and 2000, we assume, as when calculating implied TCR, that a linear relationship between temperature and forcing holds. Then, the

increase in GMST between 1970 and 2000 can be calculated by linearly interpolating between $T = 0ºC$at $F = 0\ W/m^2$and $T = 0.8ºC$at $F = 1.20\ W/m^2$to $T = 0.23ºC$at$F = 0.35º\ W/m^2$. The resulting changes in radiative forcing and GMST are $\Delta F_{2000-1970} = 0.85\ W/m^2\ \Delta T_{2000-1970} = 0.57ºC$. We linearly interpolate the forcing and temperature to arrive at annual values.

Link: https://link.springer.com/chapter/10.1007%2F978-94-010-3290-2_4

Note: Manabe and Wetherald 1967 itself is not included here because it did not provide a prediction for when CO_2 would reach a given level, only for the amount of warming that would result once that level was reached. It simply provided an equilibrium response to doubled CO_2 rather than a timeseries of transient response.

Mitchell 1970

Similar to Manabe 1970, Mitchell 1970 uses the estimate of ECS from Manabe and Wetherald 1967 and projections of CO_2 levels, implicitly assuming the system instantaneously reaches equilibrium, to determine future changes in GMST.

Mitchell states that the increase in CO_2 concentrations, “relative to a 19th century base level of 290 ppm, [...] is projected to accumulate to 11% by 1970, 15% by 1980, 20% by 1990, and 27% by 2000 A.D”. We convert CO_2 concentrations into radiative forcings using equation 7, with a reference of 320 ppm in 1969. Temperatures are taken from Mitchell’s statement that “temperature contribution of CO_2 changes anticipated in the future, neglecting all other mechanisms of climatic change, will consist of a further warming (above 1969 temperature levels) of about 0.1°C (0.2ºF) by 1980, 0.3ºC (0.5ºF) by 1990, and 0.5ºC (0.8ºF) by 2000 A.D”. We linearly interpolate the forcing and temperature to arrive at annual values.

Link: https://link.springer.com/chapter/10.1007/978-94-010-3290-2_15

Benson 1970

Benson predicts that CO_2 concentrations will increase linearly at the contemporaneous rate of 0.7 ppm per year. Linearly extrapolating from a value of 320 ppm in 1970, he arrives at a

concentration of 384 ppm in 2000. Using equation 7, we translate this into an increase in CO_2 radiative forcing of 0.98 W/m^2 from 1970 to 2000.

Using Manabe and Wetherald 1967's estimate of climate sensitivity, expressed as a warming of 0.3ºC per 10% increase in CO_2 concentrations, he finds that temperatures should increase by "about 0.6ºC" from 1970 to 2000. Presumably, he used some form of equation 7, which expresses the approximately logarithmic dependence of radiative forcing on CO_2, to get

$$\Delta T \ = \ ECS \times \frac{\Delta F2000 - 1970}{\Delta F2\times} = 2.36 \times log(1.2) \ / \ log(2.0) \ = 0.62 \ \simeq 0.6°C.$$

We linearly interpolate the forcing and temperature to arrive at annual values.

Link https://doi.org/10.1073/pnas.67.2.898

Rasool and Schneider 1971

Rasool and Schneider (1971)'s method of projecting GMST change based on an ECS and radiative forcing is similar to the above studies but both their projected increases in CO_2 concentrations of 10% from 1971 to 2001 and their estimate $ECS = 0.8ºC$ are less than half those of all other contemporaneous projections discussed above and below (see note below on why this disagrees so much with the Manabe and Wetherald 1967 estimate of ECS). They state: "if CO_2 is augmented by another 10 percent in the next 30 years, the increase in the global temperature may be as small as 0.1ºK". We can reproduce this calculation, following equation 7, if we assume the system is always at equilibrium and is described by a constant feedback parameter, such that

$$\Delta T = 0.8 \times log(1.1) \ / \ log(2.0) \ = 0.11ºC \ \simeq 0.1ºC.$$

We linearly interpolate the forcing and temperature to arrive at annual values.

Note: Schneider (1975) discusses the difference between the Rasool and Schneider (1971) and Manabe and Wetherald (1967)'s estimate of ECS at length, based on simulations by Manabe and Wetherald who generously replicated their simulations with Rasool and Schneider (1971)'s

assumptions. The differences between their estimates can be summarized by the following: 1) Rasool and Schneider assume an isothermal stratosphere, which allows too much radiation to space in optically-thick bands as CO_2 is increased, and thus limits the amount of heating at the surface; 2) Rasool and Schneider do not include near-infrared solar absorption by water vapor and CO_2, resulting in less heating at the surface; 3) Manabe and Wetherald 1967's infrared radiation transfer scheme was less elaborate than that of Rasool and Schneider 1971, resulting in a 0.4ºC warm bias in their ECS relative to the radiation scheme used in Rasool and Schneider 1971.

Link: http://dx.doi.org/10.1126/science.173.3992.138

Sawyer 1972

Citing Manabe 1970, he assumes an ECS of 2.4ºC. He speculates that a 25% increase in CO_2 concentrations from 319 in 1969 to 399 ppm in 2000 would lead to a warming of 0.6ºC. We are unclear how Sawyer arrived at this value, since the typical scaling would provide a temperature change of:

$$2.4 \times log(1.25) \,/\, log(2.0) \;=\; 0.77 \;\simeq\; 0.8\,^{\circ}C.$$

Given the fact that

$$2.4 \times 1.25 \,/\, 2.0 \;=\; 0.6^{\circ}C,$$

it seems possible that Sawyer mistakenly approximated CO_2 forcing as a linear function of CO_2 concentrations, resulting in a spurious underestimate of the temperature change. It is not clear to us how else Sawyer could come up with a temperature change of 0.6ºC from the cited values.

Link: https://www.nature.com/articles/239023a0

Broecker 1975

Citing the calculation of $ECS = 2.4ºC$ from the general circulation model in Manabe and Wetherald 1975 as the most reliable estimate of ECS (which coincidentally differs only slightly from their previous column-model calculations of $ECS = 2.36ºC$ in Manabe and Wetherald 1967), Broecker follows the same approach as Manabe 1970 and projects GMST changes forward to 1980, 1990, 2000, and 2010 (see his Table 1 and Figure 1), using a variant of equation 7 that gives nearly identical results. We linearly interpolate the forcing and temperature between values reported in Table 1 to arrive at annual values between 1975 and 2010.

Link: https://science.sciencemag.org/content/189/4201/460

Note: we only consider the anthropogenically forced response, ignoring projected contributions from the assumed sinusoidal cycles of natural variability, which Broecker himself later admitted were flawed (Broecker 2017).

Nordhaus 1977

The temperature response for a given trajectory of CO_2 concentrations is given by equation 7, same as all of the above, using a value of $ECS = 2ºC$, the choice of which seems to be mostly informed by the Manabe and Wetherald 1967 model but reflects the large spread of estimates in the 1970s literature (see above).

Nordhaus 1977 differs from the above models in that he calculates CO_2 trajectories based on decoupled linear economic and carbon cycle models. While Nordhaus 1977 explores scenarios with constraints on the level of allowable CO_2 concentrations in the atmosphere, we only consider the temperature time-series for the uncontrolled case, which eventually reaches CO_2 concentrations four to five times pre-industrial levels. We note that by 2020, the uncontrolled scenario results in CO_2 concentrations that, by 2020, are only slightly higher than that of a scenario where CO_2 concentrations are constrained to never go beyond double the CO_2 concentration from the year 1974.

The temperature time-series is digitized from Figure 1 while the radiative forcing is calculated according to equation 7 by digitizing the time series of carbon dioxide content in the atmosphere in Figure 9 and converting to parts per million.

Note: There appears to be a typo in Nordhaus 1977 monograph in the first footnote of page 5, which cites Manabe and Wetherald 1969 when referencing the model for the temperature response, which does not appear in the bibliography and is elsewhere cited as Manabe and Wetherald 1967.

Links: http://cowles.yale.edu/sites/default/files/files/pub/d04/d0443.pdf (long version) https://www.jstor.org/stable/pdf/1815926 (short version)

Schneider and Thompson 1981

In contrast to all of the above, which consider the case of instantaneous thermal equilibrium, Schneider and Thompson 1981 consider the transient evolution of surface temperatures in a two-box energy balance model. When diffusive heat uptake by the deep ocean (represented by the lower box) is included, the transient warming is reduced relative to the instantaneous equilibrium case (equivalent, the small thermal inertia or short radiative relaxation timescale case). Here, we only consider the case of a diffusive timescale of 550 years for the global deep ocean, which is the scenario in Schneider and Thompson 1981 that most corresponds to modern understanding of the diffusive and advective timescales for the deep ocean circulation and is the middle of the range of diffusive timescales considered in the paper. The ratio r_c of instantaneous CO_2 concentration over the 1925 value relative is assumed to increase quadratically according to:

$$1.443\ ln(r_c)\ =\ 7.03 \times 10^{-5}\ t^2.$$

External forcing is estimated from the CO_2 concentrations using equation 7. The temperature time series is digitized from Figure 3 for a diffusive timescale $\tau_d = 550$ years.

Link: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JC086iC04p03135

Hansen et al. 1981

We consider two forcing scenarios for Hansen et al. 1981: scenario 1 (“fast growth”) and scenario 2a (“slow growth” without coal phaseout). In both cases, natural gas, oil, and coal consumption increases according to a prescribed growth rate (4% and 2% for scenarios 1 and

2, respectively). While the prescribed growth rates do technically taper in time, the tapering does not come into effect until 2020 so it does not affect the results shown here. When the relatively limited gas and oil reserves are depleted by ever-increasing energy consumption, they are in principle replaced by coal though in practice none of the reserves are depleted until after 2020 in either scenario. Energy consumption in Joules is converted to ppm of CO_2 according to the conversion factors in Table 2 of Hansen et al. 1981. Hansen et al. 1981 do not discuss the potential for future carbon sinks; following them, we thus unrealistically assume all emitted CO_2 remains in the atmosphere perpetually. Some of the excess forcing due to the permanence of anthropogenic CO_2 in the atmosphere is likely offset by observed increases in other greenhouse gases, which are not included in the Hansen et al. 1981 projections. CO_2 concentrations are converted to a radiative forcing using only the terms involving a change in CO_2 concentrations from equation 9 of Hansen et al. 1981, which agrees with our equation 7 to within 2% for historical changes in CO_2 concentrations.

Temperature time series corresponding to the forcing scenarios 1 and 2a are digitized from Figure 6 of Hansen et al. 1981.

Link: https://pubs.giss.nasa.gov/abs/ha04600x.html

Hansen et al. 1988

We consider the three forcing scenarios for H88: a rapid growth scenario A, a slow growth scenario B, and is a scenario C wherein emissions are so dramatically curtailed by the year 2000 that net emissions vanish. Both annual temperature and forcing values calculated from the model were obtained from NASA Goddard Institute for Space Studies (GISS).

Link: https://pubs.giss.nasa.gov/abs/ha02700w.html

Manatabe and Stouffer 1993

Forcings are calculated by digitizing the $4xCO_2$ time series (1% increase per year) in their Figure 1a and converting to a radiative forcing using equation 7. Temperature changes are calculated by digitizing panel the time series corresponding to the $4xCO_2$ experiment in Figure 1b.

Link: https://www.nature.com/articles/364215a0

IPCC First Assessment Report (FAR)

The main text of the IPCC FAR featured projections from a simple box-diffusion upwelling energy balance model (EBM) tuned to the individual climate models featured in the supplement to the report. We digitized EBM temperature valuesfrom Figure 8 in the Policymakers Summary. As the original values are unavailable in a digital form, the IPCC AR5 took a similar approach in digitizing old figures. The values we obtained through digitization were comparable to those in the AR5 WG1 Chapter 1 appendix. We chose not to directly use the digitized values reported in the AR5 as they only provided a high and low range of projections and did not include a best-estimate, and digitizing the best-estimate while relying on the digitized high and low values in the AR5 would introduce potential inconsistencies in the digitization approach.

The AR5 chose an unusual set of bounds for its reported FAR values, relying on a stringent mitigation scenario (Scenario D in Figure 9) as its lower bound and the best-estimate business-as-usual scenario as its upper bound. We instead use the values reported in Figure 8, which show a low estimate, best estimate, and high estimate of temperature change in the FAR business-as-usual scenario. The low and high estimates are used as the uncertainty bounds on the best estimate. External forcing values for the EMB were digitized from Figure 6 (also Figure A.6) using the business-as-usual scenario, with all three scenarios (low, best, and high) relying on the same underlying set of forcings.

Individual climate model projections featured in Figure S5 were obtained from the FAR supplementary materials. Climate models in the IPCC FAR from UKMET and GFDL use only CO_2 changes for future forcings. They did not have model years specified, so were aligned such that their 1990 value was the model year in which CO_2 concentrations were closest to 1990 observations. GCMs included in the IPCC FAR employ scenarios where CO_2 or GHG forcing increases by 1% per year, while the simple energy balance models featured in the report used the IS92a scenario.

IPCC Second Assessment Report (SAR)

We digitized EBM valuesfrom Figure 19, using the 2.5ºC ECS run (including aerosols) as the best estimate and 1.5ºC/4.5ºC ECS runs as the low and high-end estimates. Similar to the FAR, the original values are unavailable in digital form and the values we obtained through digitization were comparable to those in the AR5 WG1 Chapter 1 appendix. Projected FAR EBM CO_2 concentrations were digitized from Figure 5 (IS92a scenario), while total external forcing was digitized from Figure 6.

GHG-only model runs (excluding aerosols) were used from the IPCC SAR for the individual models shown in Figure S5, as the specific aerosol forcings used differ by models and are poorly documented. SAR climate models mostly employ scenarios where CO_2 increases by 1% per year, while the simple energy balance models featured in the report used the IS92a scenario.

IPCC Third Assessment Report (TAR)

Decadal values for both temperatures, total forcings, and CO_2 used in the EBM featured in the TAR main text were obtained from Appendix I:
https://www.ipcc.ch/site/assets/uploads/2018/03/TAR-APPENDICES.pdf

These decadal values were transformed into annual estimates via linear interpolation.

Individual climate models featured in the TAR using the A2 SRES scenario were selected and shown in Figure S5, as that is the scenario with the most unique model runs available.

IPCC Fourth Assessment Report (AR4)

Coupled Model Intercomparison Project 3 (CMIP3) temperature projections featured in the AR4 were obtained from KNMI climate explorer. A1B runs were used as they were readily available, though over the 2007-2017 period differences between A1B and A2 in CMIP3 are minor. External forcing values used in the CMIP3 A1B scenario were based on those used in GISS model E, as precise forcing values used by each model are not readily available (and the differences within a given SRES scenario in forcings used between models should be small):
https://data.giss.nasa.gov/modelE/transient/dangerous.html

The AR4 best estimate shown in the paper is based on the A1B multimodel mean, while the low and high scenarios reflect the 5th and 95th percentile of the ensemble of A1B model runs for any given year.

SUPPLEMENTARY FIGURES S1-S5.

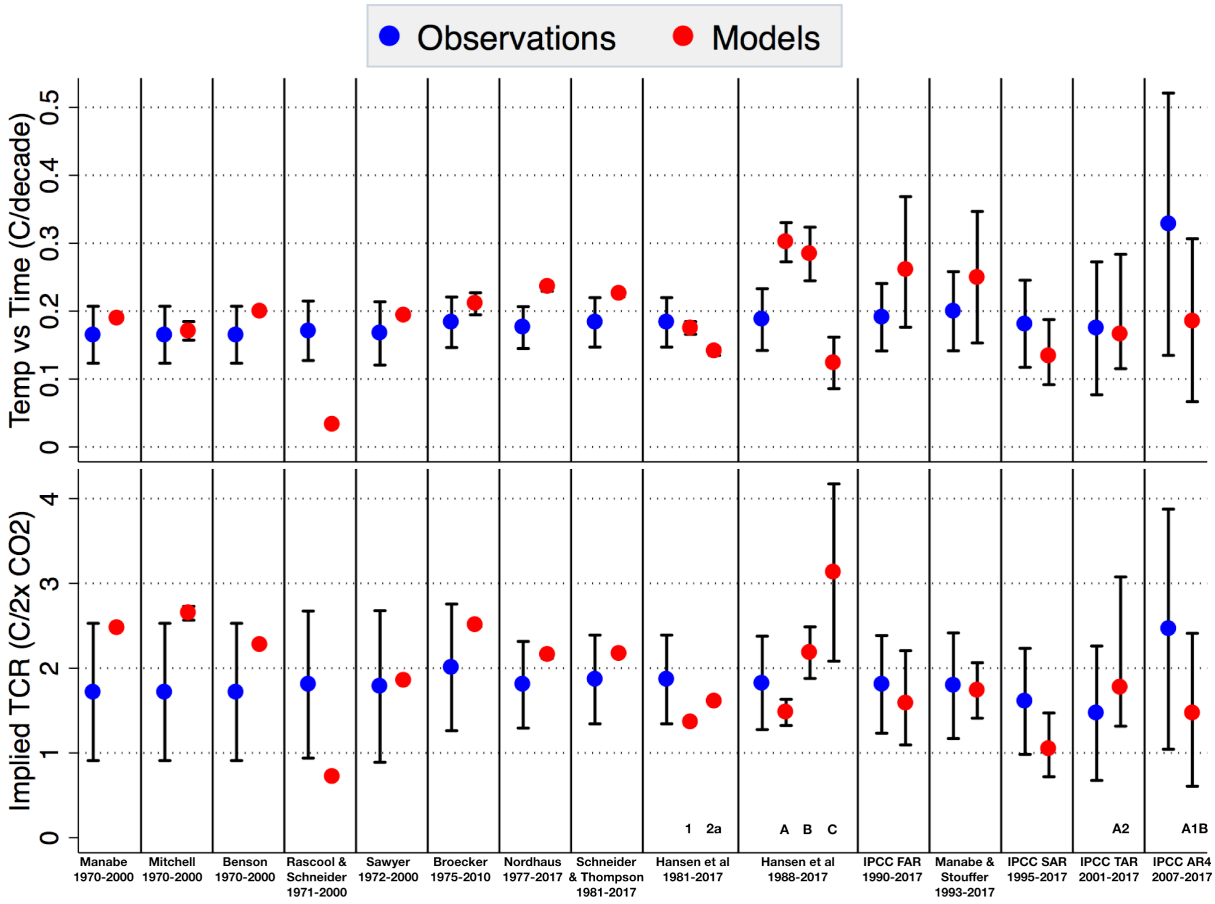


Figure S1. Comparison of trends in temperature vs time (top panel) and implied TCR (bottom panel) between observations and models over the model projection periods displayed at the bottom of the figure. As in Figure 1, but with the 2007-2017 IPCC AR4 projection included.

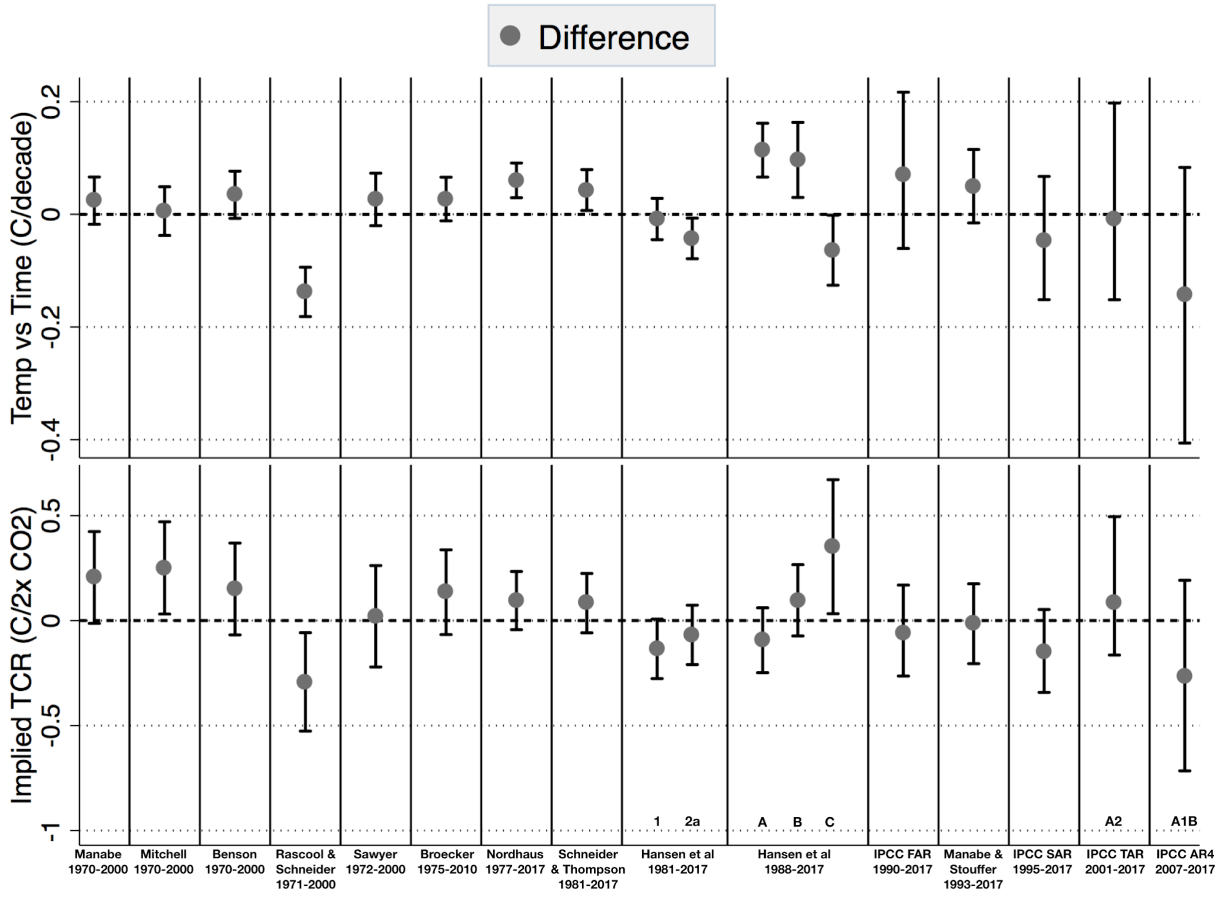


Figure S2. Difference between climate models and observations on a temperature vs time (top panel) and implied TCR (bottom panel) basis over the model projection periods displayed at the bottom of the figure. Values higher than zero indicate that the model projected more warming (or higher TCR) than observed.

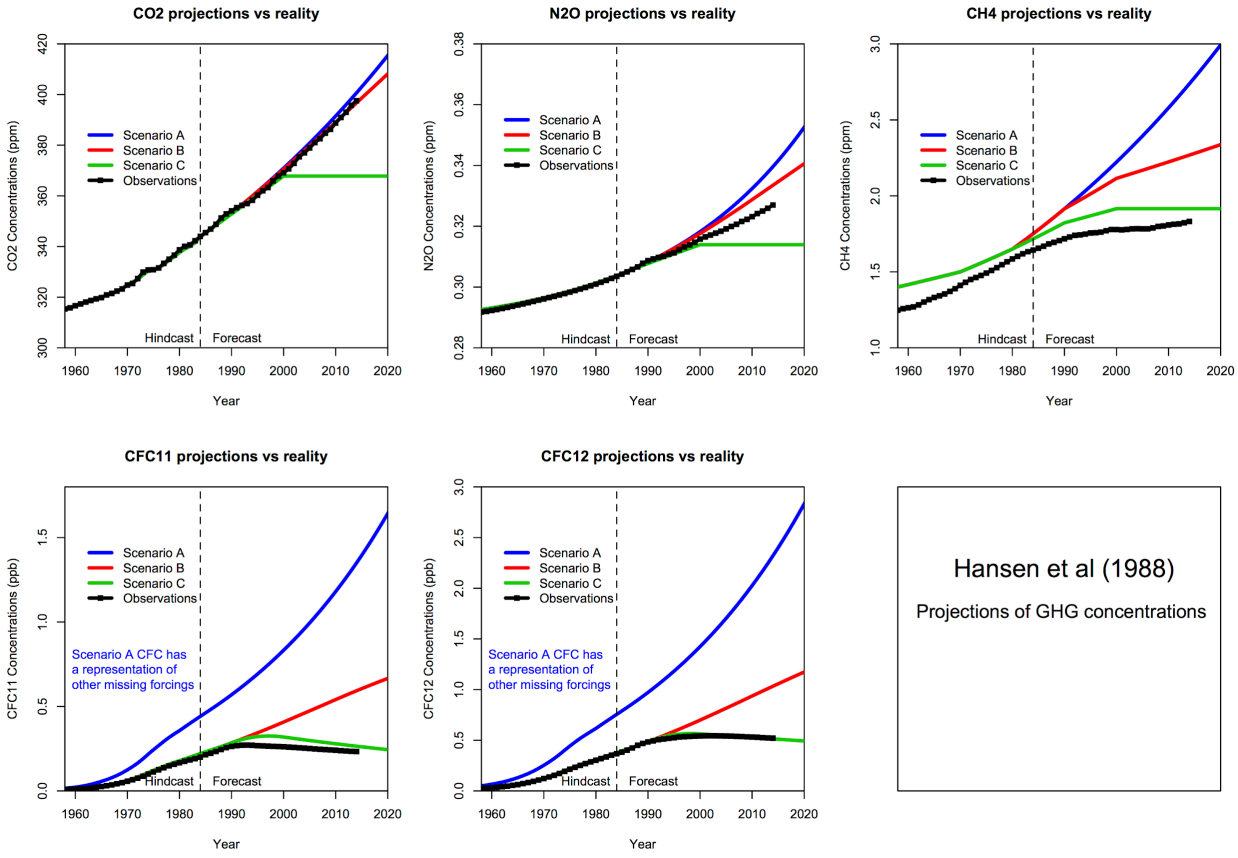


Figure S3. Greenhouse gas concentrations in Hansen et al. (1988) scenarios compared to observations.

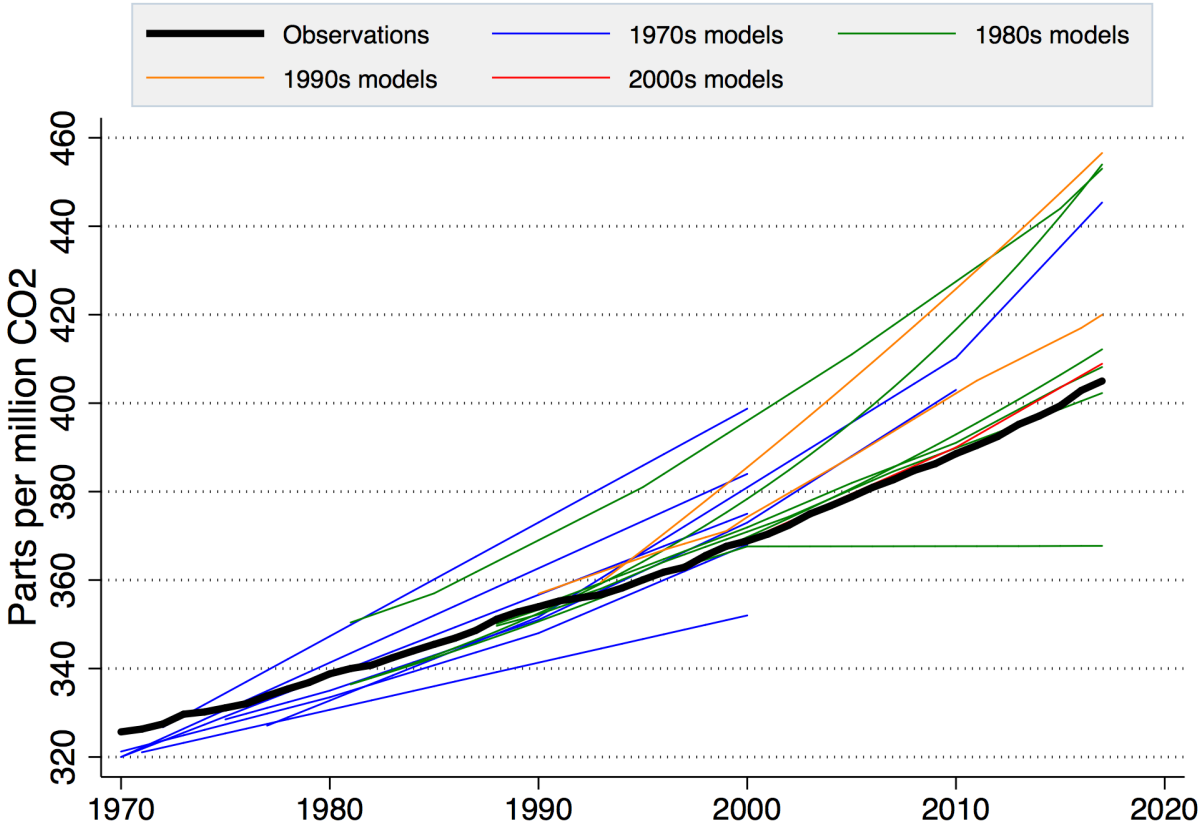


Figure S4*:* Model projected CO_2 concentrations colored by decade in which the model was published compared to observations (black). Observed CO_2 concentrations were taken from Meinshausen et al. (2017).

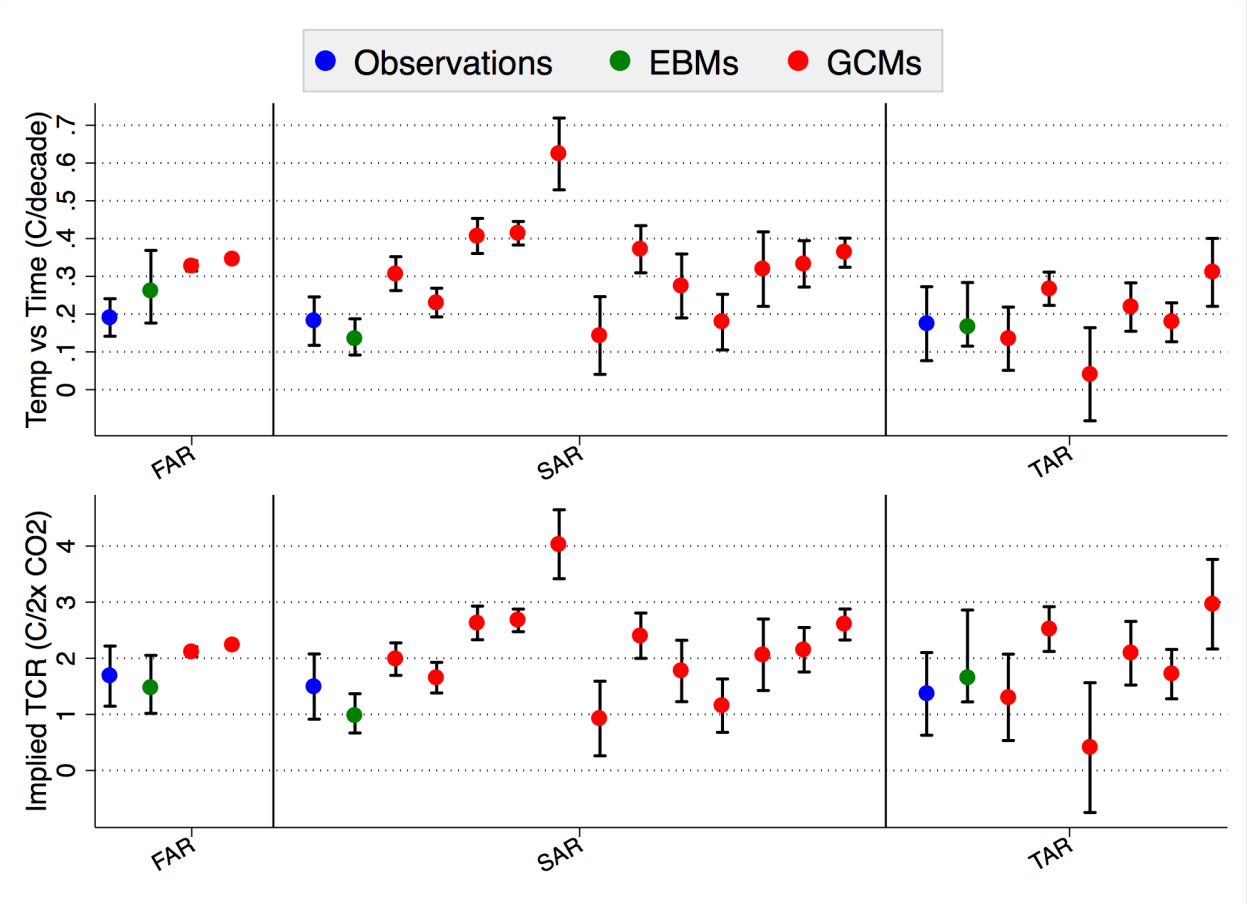


Figure S5. Comparison of trends in temperature vs time (top panel) and implied TCR (bottom panel) between observations and models included in the first three IPCC assessment reports over model future projection periods. Main-text projections based on simple energy balance models are shown in green (those are also included in Figure 1).

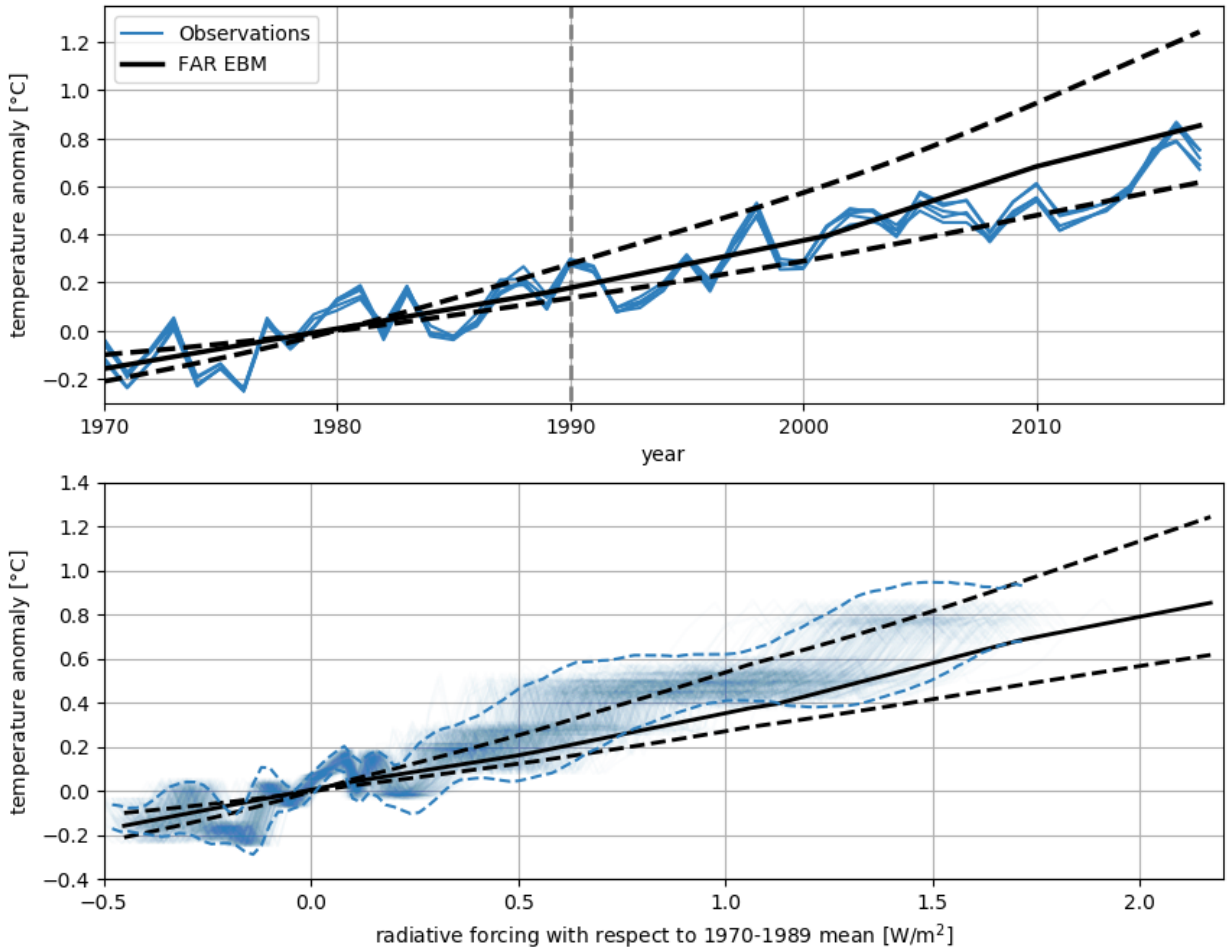


Figure S6: IPCC FAR projections compared with observations on a temperature vs. time basis (top) and temperature vs forcing (bottom). The dashed grey line in the top panel represent the start of the future projection period. The probability distribution in the lower panel represents the 5000 combinations of the 5 temperature observation products and the 1000 ensemble members of estimated forcings. Anomalies for both temperature and forcing are shown relative to a 1970-1989 pre-future-projection baseline.

REFERENCES (SUPPLEMENTARY MATERIALS)

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

Cubasch U., Meehl G.A., Boer G.J., Stouffer R.J., Dix M., Noda A., Senior C.A., Raper S.C.B., Yap K.S. (2001). Projections of future climate change. In Climate change 2001: the scientific

basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change (eds JT Houghton, Y Ding, DJ Griggs, M Noguer, P van der Linden, X Dai, K Maskell, CI Johnson), pp. 525–582. Cambridge, UK: Cambridge University Press.

G. Myhre, D. Shindell, F.-M. Bréon, W. Collins, J. Fuglestvedt, J. Huang, D. Koch, J.-F. Lamarque, D. Lee, B. Mendoza, T. Nakajima, A. Robock, G. Stephens, T. Takemura, H. Zhang, Anthropogenic and natural radiative forcing, 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, T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Doschung, A. Nauels, Y. Xia, V. Bex, P.M. Midgley, Eds. (Cambridge Univ. Press, 2013), pp. 659–740

Hansen, J., Sato, M., Ruedy, R., Nazarenko, L., Lacis, A., Schmidt, G. A., … Zhang, S. (2005). Efficacy of climate forcings. Journal of Geophysical Research: Atmospheres, 110(D18). https://doi.org/10.1029/2005JD005776

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

Marvel, K., Schmidt, G. A., Miller, R. L., & Nazarenko, L. S. (2015). Implications for climate sensitivity from the response to individual forcings. Nature Climate Change, 6, 386. Retrieved from http://dx.doi.org/10.1038/nclimate2888

Meinshausen et al. (2017). M., et al. Historical greenhouse gas concentrations for climate modelling (CMIP6), Geosci. Model Dev., 10, 2057-2116, doi.org/10.5194/gmd-10-2057-2017, 2017.)

Schneider, S.H. (1975). On the Carbon Dioxide–Climate Confusion. J. Atmos. Sci., 32, 2060–2066, https://doi.org/10.1175/1520-0469(1975)032<2060:OTCDC>2.0.CO;2

V. CONCLUSION

Uncertainties in estimates of global surface temperatures, both during the 20th century and 21st century, are of critical importance to a number of pressing questions in climate science such as how well are models reproducing observed warming, has there been any evidence of a "hiatus" in warming temperatures during the past two decades, and what is the remaining carbon budget to avoid surpassing 1.5°C and 2°C warming targets. Reducing these uncertainties will improve our understanding of prior changes to the climate and our confidence in estimates of future projected changes.

Work done in this dissertation has had a notable impact on our understanding and estimates of temperatures. This includes ensuring that urbanization is not biasing our record of land temperatures,[20] testing the performance of land temperature homogenization,[41] resolving differences between ocean temperature records in recent decades,[15] developing a novel sea surface temperature record to help better understand WW2-era uncertainties,[54] and evaluating recent changes in ocean heat content.[57] In an encouraging sign of the impact of our work, the new HadSST4 temperature product from the UK Met Office prominently features comparisons with the instrumentally homogenous sea surface temperature records we developed.[53]

Similarly, the work that I and coauthors have undertaken has changed the approach used in evaluating the performance of GMST climate model projections, demonstrating the need to use common coverage and blended SAT/SST fields to ensure like-to-like comparisons with observations.[4] Evaluating the future projections of old climate models improves our confidence that the current generation of models is accurately capturing the physical processes driving GMST change.[34] This work on evaluating old climate models will be featured prominently in Chapter 1 of the upcoming IPCC 6th Assessment Report, where I am serving as a contributing author.

There are a number of next steps that can be taken to improve our understanding of historical temperatures, improve temperature measurements going forward, and better compare climate models and observations. One essential step is to convince World Meteorological Organization member countries to fund the creation of the global land Climate Reference Network, using the one developed over the past 15 years in the US as an example. As we pointed out in our recent paper on the subject,[47] a network of as few as 160 well-separated monitoring stations would be sufficient to provide an accurate estimate of global mean surface land temperature changes

going forward. The correlation lengths are large enough for temperature that a relatively small number of stations are needed; though for other climate variables – such as precipitation – a denser station network is desirable.

A global climate reference network would serve a number of purposes. It would anchor our estimates of global land temperatures, helping reduce uncertainties going forward, particularly in areas currently under-sampled such as parts of Africa, the Arctic, and Antarctica. It would also provide a known homogenous test case to use in the evaluation of homogenization of larger local weather station networks. Using modern technology, stations can be largely self-powering and automatically provide data to a central international repository via satellite uplink, similar to the current USCRN.

A second area of global temperature record improvement would be to explore new methods to resolve disagreements between historical sea surface temperature record reconstructions, particularly during the first half of the 20th century.[67] Our work using coastal and island stations is a first step in this direction,[54] but more work needs to be undertaken using novel methods to detect and correct for often poorly-documented changes in measurement technique (e.g. wooden vs canvas buckets), deck height, ship speed, and other factors that affect historical measurements. One promising method recently published was to compare spatially and temporally-proximate sets of ship-based measurements from different shipping fleets over time to empirically determine offsets, assuming similar practices across members of national fleets.[48,68] This is conceptually similar to the pairwise homogenization approach used for land-based measurements, and could potentially be expanded in the future to develop ship-specific (rather than fleet-specific) corrections.

Extending satellite radiometer-based sea surface temperature estimates further back in time could also help resolve some disagreements between records in recent decades. A new satellite sea surface temperature was published in summer 2019 that extends back to the start of the satellite era – in 1981 – and could be compared to other composite sea surface temperature records.[69]

Third, evaluating the performance of climate models through observational comparisons requires having model fields that mimic the spatial coverage and field types used by observations.[4] These fields have been calculated for CMIP5 climate models, but do not yet exist for CMIP6. These blended SAT/SST fields also have implications for estimates of climate sensitivity, both for ECS and TCR. For example, observationally-based estimates of TCR agree

quite well with models when adjusted for differences between SAT and SST warming rates over the oceans.[24] I've developed similar estimates of global SATs for use in ECS calculations in a upcoming review paper on climate sensitivity that attempts to reconcile model, emergent constraint, observational, and paleo estimates to narrow the likely range of climate sensitivity.

Fourth, there is work to do to resolve differences between tropospheric and surface temperature records. Climate models agree well with surface temperatures, but show less warming than most tropospheric temperature records, particularly in the tropics. The disagreement between surface and troposphere records is unusual, as even in the absence of anthropogenic forcing (e.g. in a hypothetical solar-forced warming plant) tropospheric temperatures over the ocean should be amplified relative to the surface. The absence of this behavior in most satellite-based records suggest that either large biases remain in the observational record, or our basic understanding of tropospheric amplification is inaccurate. Given the large structural uncertainties in MSU-based satellite records, and the much smaller uncertainties in surface records, if the observations are problematic the fault is likely in the tropospheric record.

Finally, notable difference among surface temperature records have emerged in the past few years, and appear to be driven primarily by differences in Arctic coverage. With the advent of modern relatively-homogenous reanalysis products like ERA5, there is the possibility of using other types of atmospheric measurements to more effectively interpolate temperatures across the sparsely-measured Arctic. This is particularly important for temperature records like NOAA and Hadley that have limited coverage in the region and use spatial interpolation approaches – such as lat/lon grid cell averages – that are poorly suited to the polar regions.

Global temperatures are our most iconic indicator the changing climate, and improvements to our historical estimates also improve our understanding of likely future changes. The temperature record also provides an important means to visualize and communicate the changing climate to policymakers and the broader public – as the temperature spiral, climate stripes, and my own visualizations have shown over the past few years. The work in this dissertation – as well as ongoing projects – have played both scientific and science communication roles, leading to hundreds of articles in the popular press, including a front page story in the New York Times.[70] Substantial uncertainties in temperature records have been resolved in the papers included in this dissertation, and we have a good roadmap to try and resolve remaining differences moving forward.

VI. ADDITIONAL REFERENCES

1. Callendar, G. S. (1938). The artificial production of carbon dioxide and its influence on temperature. Q.J.R. Meteorol. Soc., 64: 223–240. doi:10.1002/qj.49706427503
2. Jones, P. (2016). The reliability of global and hemispheric surface temperature records. Adv. Atmos. Sci. 33: 269. doi.org/10.1007/s00376-015-5194-4
3. E. C. Kent, J. J. Kennedy, D. I. Berry, R. O. Smith (2010). Effects of instrumentation changes on sea surface temperature measured in situ. Wiley Interdiscip. Rev. Clim. Change 1, 718–728.
4. Cowtan, K., Hausfather, Z., Hawkins, E., Jacobs, P., Mann, M. E., Miller, S. K., … Way, R. G. (2015). Robust comparison of climate models with observations using blended land air and ocean sea surface temperatures. Geophysical Research Letters, 42(15). doi.org/10.1002/2015GL064888
5. United States. Cong. House. Committee on Science, Space & Technology. Hearings, Mar. 29th, 2017. 115th Cong. Washington, DC.
6. Hansen, J., Ruedy, R., Sato, M., & Lo, K. (2010). Current GISS Global Surface Temperature Analysis. NASA Goddard Institute for Space Studies, 1–34.
7. Morice, C.P., Kennedy, J.J., Rayner, N.A. and 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, doi:10.1029/2011JD017187
8. R. S. Vose, D. Arndt, V. F. Banzon, D. R. Easterling, B. Gleason, B. Huang, E. Kearns, J. H. Lawrimore, R. W. Reynolds, T. M. Smith, C. N. Williams, D. B. Wuertz, M. J. Menne, T. C. Peterson (2012). NOAA's merged land–ocean surface temperature analysis. Bull. Am. Meteorol. Soc. 93, 1677–1685
9. K. Cowtan, R. G. Way (2014). Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. Q. J. R. Meteorol. Soc. 140, 1935–1944 (2014).
10. R. Rohde, R. A. Muller, R. Jacobsen, E. Muller, C. Wickham (2013). A New Estimate of the Average Earth Surface Land Temperature Spanning 1753 to 2011. Geoinfor. Geostat.: An Overview. 1, doi:10.4172/2327-4581.1000101.
11. Copernicus Climate Change Service (C3S) (2017). ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate . Copernicus Climate Change Service Climate Data Store (CDS).

12. Risbey, J. S., Lewandowsky, S., Cowtan, K., Oreskes, N., Rahmstorf, S., Jokimäki, A., & Foster, G. (2018). A fluctuation in surface temperature in historical context: reassessment and retrospective on the evidence. Environmental Research Letters, 13(12), 123008. https://doi.org/10.1088/1748-9326/aaf342
13. Kennedy J.J., Rayner, N.A., Smith, R.O., Saunby, M. and Parker, D.E. (2011). Reassessing biases and other uncertainties in sea-surface temperature observations since 1850 part 1: measurement and sampling errors. J. Geophys. Res., 116, D14103, doi:10.1029/2010JD015218
14. 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, 2014: 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
15. 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), e1601207. doi.org/10.1126/sciadv.1601207
16. Huang, J., Zhang, X., Zhang, Q., Lin, Y., Hao, M., Luo, Y., … Zhang, J. (2017). Recently amplified arctic warming has contributed to a continual global warming trend. Nature Climate Change, 7(December), 1–5. doi.org/10.1038/s41558-017-0009-5
17. Menne, M. J., Williams, C. N., & Vose, R. S. (2009). The U.S. Historical Climatology Network Monthly Temperature Data, Version 2. Bulletin of the American Meteorological Society, 90(7), 993–1007. http://doi.org/10.1175/2008BAMS2613.1
18. Quayle, R. G., D. R. Easterling, T. R. Karl, and P. Y. Hughes (1991), Effects of recent thermometer changes in the Cooperative Station Network, Bull. Am. Meteorol. Soc., 72, 1718–1723, doi:10.1175/15200477(1991)072<1718:EORTCI>2.0.CO;2.
19. Vose, R. S., et al., (2003). An evaluation of the time of observation bias adjustment in the U.S. Historical Climatology Network. Geophysical Research Letters, 30(20), 1–4. doi.org/10.1029/2003GL018111
20. Hausfather, Z., M. J. Menne, C. N. Williams, T. Masters, R. Broberg, and D. Jones (2013). Quantifying the effect of urbanization on U.S. Historical Climatology Network temperature records, J. Geophys. Res. Atmos., 118, 481–494, doi:10.1029/2012JD018509.
21. Pielke, R. A., G. Marland, R. A. Betts, T. N. Chase, J. L. Eastman, J. O. Niles, D. Niyogi, and S. Running, (2002). The influence of land-use change and landscape dynamics on the climate system—Relevance to climate change policy beyond the radiative effect of greenhouse gases. Philos. Trans., 360A, 1705–1719.

22. Medhaug, I., Stolpe, M. B., Fischer, E. M., & Knutti, R. (2017). Reconciling controversies about the “global warming hiatus.” Nature, 545(7652), 41–47. doi.org/10.1038/nature22315
23. T. R. Karl, A. Arguez, B. Huang, J. H. Lawrimore, J. R. McMahon, M. J. Menne, T. C. Peterson, R. S. Vose, H.-M. Zhang. (2015). Possible artifacts of data biases in the recent global surface warming hiatus. Science 348, 1469–1472.
24. M. Richardson, K. Cowtan, E. Hawkins, M.B. Stolpe (2016). Reconciled climate response estimates from climate models and the energy budget of Earth. Nature Climate Change. 6, 931–935.
25. Otto, A. et al. (2013). Energy budget constraints on climate response. Nat. Geosci. 6, 415–416.
26. Schurer, A. P., Mann, M. E., Hawkins, E., Tett, S. F. B., & Hegerl, G. C. (2017). Importance of the pre-industrial baseline for likelihood of exceeding Paris goals. Nature Climate Change, 7, 563. Retrieved from dx.doi.org/10.1038/nclimate3345
27. Millar, R. J., Fuglestvedt, J. S., Friedlingstein, P., Rogelj, J., Grubb, M. J., Matthews, H. D., … Allen, M. R. (2017). Emission budgets and pathways consistent with limiting warming to 1.5 °c. Nature Geoscience, 10(10), 741–747. doi.org/10.1038/NGEO3031
28. Knutti, R., Masson, D., and Gettelman, A. (2013). Climate model genealogy: Generation CMIP5 and how we got there, Geophys. Res. Lett., 40, 1194– 1199, doi:10.1002/grl.50256.
29. Spencer, R.W., Christy, J.R. & Braswell, W.D. (2017). UAH Version 6 global satellite temperature products: Methodology and results. Asia-Pacific J Atmos Sci 53: 121. https://doi.org/10.1007/s13143-017-0010-y
30. Mears, C. A., & Wentz, F. J. (2017). A satellite-derived lower-tropospheric atmospheric temperature dataset using an optimized adjustment for diurnal effects. Journal of Climate, 30(19), 7695–7718. doi.org/10.1175/JCLI-D-16-0768.1
31. Cheng, L., Abraham, J., Hausfather, Z., & Trenberth, K.E. (2019). How fast are the oceans warming? Science 363 (6423), 128-129.
32. 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-ESM1.2) and its response to increasing CO2. Journal of Advances in Modeling Earth Systems, 11, 998–1038. https://doi.org/10.1029/2018MS001400
33. 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

34. Hausfather, Z., Drake, H. F., Abbott, T., & Schmidt, G. A. (2019). Evaluating the performance of past climate model projections. Geophysical Research Letters, n/a(n/a). https://doi.org/10.1029/2019GL085378
35. Thorne, P. W., et al, (2011). Guiding the Creation of A Comprehensive Surface Temperature Resource for Twenty-First-Century Climate Science. Bull. Amer. Meteor. Soc., 92, ES40–ES47. doi.org/10.1175/2011BAMS3124.1
36. Menne, M. J., & Williams, C. N. (2009). Homogenization of Temperature Series via Pairwise Comparisons. Journal of Climate, 22(7), 1700–1717. doi.org/10.1175/2008JCLI2263.1
37. Karl, T.R., H.F. Diaz, and G. Kukla (1988), Urbanization: its detection and effect in the United States climate record, J. Climate, 1, 1099-1123.
38. Hausfather, Z., M. J. Menne, C. N. Williams, T. Masters, R. Broberg, and D. Jones (2013), Quantifying the effect of urbanization on U.S. Historical Climatology Network temperature records, J. Geophys. Res. Atmos., 118, 481–494, doi:10.1029/2012JD018509.
39. Vose, R.S., S. Applequist, M. Squires, I. Durre, M.J. Menne, C.N. Williams, Jr., C. Fenimore, K. Gleason, and D. Arndt, 2014: Improved historical temperature and precipitation time series for U.S. climate divisions. Journal of Applied Meteorology and Climatology, 53,1232-1251, doi:10.1175/JAMC-D-13-0248.1.
40. Diamond, H. J., et al. (2013). U.S. Climate Reference Network after one decade of operations: Status and assessment, Bull. Am. Meteorol. Soc., 94, 485–498.
41. Hausfather, Z., K. Cowtan, M. J. Menne, and C. N. Williams Jr. (2016). Evaluating the impact of U.S. Historical Climatology Network homogenization using the U.S. Climate Reference Network, Geophys. Res. Lett., 43, doi:10.1002/2015GL067640.
42. Williams, C. N., M. J. Menne, and P. W. Thorne (2012), Benchmarking the performance of pairwise homogenization of surface temperatures in the United States, J. Geophys. Res., 117, D05116, doi:10.1029/2011JD016761
43. Hausfather, Z., Rohde, R.A., Menne, M.J., Williams, C.N., Zhang J., (2012). A comparative analysis of monthly temperature homogenization in the conterminous United States. AGU Fall Meeting Abstracts.
44. Willett, K., Williams, C., Jolliffe, I. T., Lund, R., Alexander, L. V., Brönnimann, S., Vincent, L. A., Easterbrook, S., Venema, V. K. C., Berry, D., Warren, R. E., Lopardo, G., Auchmann, R., Aguilar, E., Menne, M. J., Gallagher, C., Hausfather, Z., Thorarinsdottir, T., Thorne, P. W. (2014). A framework for benchmarking of homogenisation algorithm performance on the global scale. Geoscientific Instrumentation, Methods and Data Systems, Volume 3, Issue 2, 2014, pp.187-200 doi:10.5194/gi-3-187-2014

45. Venema, V., Willett, K., Auchmann, R., Aguilar, E., Thorne, P., Williams, C., Menne, M., Vincent, L., Killick, R., Brönnimann, S., Hausfather, Z, Jolliffe, I., Thorarinsdotir, T., Easterbrook, S., Lund, R., Gallagher, C., Lopardo, G., Berry, D., Alexander, L., (2018). The error worlds of the global benchmarks for the International Surface Temperature Initiative (ISTI). European Geophysical Union conference abstract.
46. Venema, V. K. C., Mestre, O., Aguilar, E., Auer, I., Guijarro, J. A., Domonkos, P., … Brandsma, T. (2012). Benchmarking homogenization algorithms for monthly data. Clim. Past, 8(1), 89–115. https://doi.org/10.5194/cp-8-89-2012
47. Thorne, P.W., Diamond, H.J., Goodison, B., Harrigan, S., Hausfather, Z., Ingleby, N.B., Jones, P.D., Lawrimore, J.H., Lister, D.H., Merlone, A., Oakley, T., Palecki, M., Peterson, T.C., de Podesta, M., Tassone, C., Venema, V., Willett, K.M., (2018). Towards a global land surface climate fiducial reference measurements network. International Journal of Climatology.
48. Hausfather, Z. Corrections to ocean-temperature record resolve puzzling regional differences. Nature 571 (7765), 328-329.
49. M. Ishii, A. Shouji, S. Sugimoto, T. Matsumoto, (2005). Objective analyses of sea-surface temperature and marine meteorological variables for the 20th century using ICOADS and the Kobe Collection. Int. J. Climatol. 25, 865–879.
50. 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), e1601207. doi.org/10.1126/sciadv.1601207
51. C. J. Merchant, O. Embury, N. A. Rayner, D. I. Berry, G. K. Corlett, K. Lean, K. L. Veal, E. C. Kent, D. T. Llewellyn-Jones, J. J. Remedios, (2012). A 20 year independent record of sea surface temperature for climate from Along-Track Scanning Radiometers. J. Geophys. Res. Oceans 117, C12013.
52. C. J. Merchant, O. Embury, J. Roberts-Jones, E. Fiedler, C. E. Bulgin, G. K. Corlett, S. Good, A. McLaren, N. Rayner, S. Morak-Bozzo, C. Donlon, (2014). Sea surface temperature datasets for climate applications from Phase 1 of the European Space Agency Climate Change Initiative (SST CCI). Geosci. Data J. 1, 179–191.
53. Kennedy, J. J., Rayner, N. A., Atkinson, C. P., & Killick, R. E. (2019). An Ensemble Data Set of Sea Surface Temperature Change From 1850: The Met Office Hadley Centre HadSST.4.0.0.0 Data Set. Journal of Geophysical Research: Atmospheres, 124(14), 7719–7763. https://doi.org/10.1029/2018JD029867

54. Cowtan, K., Rohde, R., & Hausfather, Z. (2017). Evaluating biases in Sea Surface Temperature records using coastal weather stations. Quarterly Journal of the Royal Meteorological Society. doi.org/10.1002/qj.3235
55. Cheng, L., K. E. Trenberth, J. Fasullo, J. Abraham, T. P. Boyer, K. von Schuckmann, and J. Zhu (2017), Taking the pulse of the planet, Eos, 98, https://doi.org/10.1029/2017EO081839.
56. M. Rhein et al. (2013). Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, T. F. Stocker et al., Eds. (Cambridge Univ. Press, 2013), pp. 215–315.
57. Cheng, L., Abraham, J., Hausfather, Z., & Trenberth, K.E. (2019). How fast are the oceans warming? Science 363 (6423), 128-129.
58. Cheng, L., Trenberth, K. E., Fasullo, J., Boyer, T., Abraham, J., & Zhu, J. (2017). Improved estimates of ocean heat content from 1960 to 2015. Science Advances, 3(3), e1601545. https://doi.org/10.1126/sciadv.1601545
59. Fyfe, J. C., Meehl, G. A., England, M. H., Mann, M. E., Santer, B. D., Flato, G. M., … Swart, N. C. (2016). Making sense of the early-2000 global warming slowdown. Nature Climate Change, 6, 224–228. https://doi.org/10.1038/nclimate2938
60. McKitrick, R., & Christy, J. (2018). A Test of the Tropical 200- to 300-hPa Warming Rate in Climate Models. Earth and Space Science, 5(9), 529–536. https://doi.org/10.1029/2018EA000401
61. Risbey, J. S., Lewandowsky, S., Cowtan, K., Oreskes, N., Rahmstorf, S., Jokimäki, A., & Foster, G. (2018). A fluctuation in surface temperature in historical context: reassessment and retrospective on the evidence. Environmental Research Letters, 13(12), 123008. https://doi.org/10.1088/1748-9326/aaf342
62. Manabe, S., and 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
63. Hegerl, G. C., Brönnimann, S., Schurer, A., & Cowan, T. (2018). The early 20th century warming: Anomalies, causes, and consequences. WIREs Climate Change, 9(4), e522. https://doi.org/10.1002/wcc.522
64. Gan, Z., Guan, X., Kong, X., Guo, R., Huang, H., Huang, W., & Xu, Y. (2019). The Key Role of Atlantic Multidecadal Oscillation in Minimum Temperature Over North America During Global Warming Slowdown. Earth and Space Science, 6(3), 387–397. https://doi.org/10.1029/2018EA000443

65. Haustein, FEL Otto, V Venema, P Jacobs, K Cowtan, Z Hausfather, ... (2019). A Limited Role for Unforced Internal Variability in Twentieth-Century Warming. Journal of Climate 32 (16), 4893-4917
66. Wilson, R., Anchukaitis, K., Briffa, K. R., Büntgen, U., Cook, E., D'Arrigo, R., … Zorita, E. (2016). Last millennium northern hemisphere summer temperatures from tree rings: Part I: The long term context. Quaternary Science Reviews, 134, 1–18. https://doi.org/https://doi.org/10.1016/j.quascirev.2015.12.005
67. Kent, E. C., Kennedy, J. J., Smith, T. M., Hirahara, S., Huang, B., Kaplan, A., … Zhang, H.-M. (2016). A Call for New Approaches to Quantifying Biases in Observations of Sea Surface Temperature. Bulletin of the American Meteorological Society, 98(8), 1601–1616. https://doi.org/10.1175/BAMS-D-15-00251.1
68. Chan, D., Kent, E. C., Berry, D. I., & Huybers, P. (2019). Correcting datasets leads to more homogeneous early-twentieth-century sea surface warming. Nature, 571(7765), 393–397. https://doi.org/10.1038/s41586-019-1349-2
69. Merchant, C. J., Embury, O., Bulgin, C. E., Block, T., Corlett, G. K., Fiedler, E., … Donlon, C. (2019). Satellite-based time-series of sea-surface temperature since 1981 for climate applications. Scientific Data, 6(1), 223. https://doi.org/10.1038/s41597-019-0236-x
70. Pierre-Louis, K. 2019. Ocean Warming Is Accelerating Faster Than Thought, New Research Finds. New York Times, Section A, Page 1, Jan. 11th.

VII. ACKNOWLEDGEMENTS

This dissertation would not be possible without the help of my coauthors over the years. Kevin Cowtan, Matt Menne, Ken Caldeira, and Gavin Schmidt all provided extensive mentorship and collaborative assistance in the papers I have published over the years.

My work at Carbon Brief has also helped me become a better writer and communicator. I would like to thank Richard Muller, Robert Rohde, Steven Mosher, Elizabeth Muller, and the rest of the Berkeley Earth team for assistance and many useful discussions of my research projects over the past decade. My dissertation committee members – Margaret Torn, William Collins, and Lara Kueppers – provided helpful feedback throughout the process. Finally, I would like to thank my family – Avary, David, and Tavi – for their love and support.

Some passages included in this Chapters 2 and 3 of the dissertation have been quoted verbatim from previous writing I have done for Carbon Brief that provide a more accessible description of the more technical papers I coauthored that are included in Appendix B. These are creative commons licensed and are used with permission from Carbon Brief.

APPENDIX A: PEER-REVIEWED PUBLICATIONS IN THE DISSERTATION

Citations for the four published lead-authored papers included in the dissertation can be found below:

1. Hausfather, Z., M. J. Menne, C. N. Williams, T. Masters, R. Broberg, and D. Jones (2013). Quantifying the effect of urbanization on U.S. Historical Climatology Network temperature records, J. Geophys. Res. Atmos., 118, 481–494, doi:10.1029/2012JD018509.
2. Hausfather, Z., K. Cowtan, M. J. Menne, and C. N. Williams Jr. (2016). Evaluating the impact of U.S. Historical Climatology Network homogenization using the U.S. Climate Reference Network, Geophys. Res. Lett., 43, doi:10.1002/2015GL067640.
3. 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), e1601207. doi.org/10.1126/sciadv.1601207
4. Hausfather, Z., Drake, H. F., Abbott, T., & Schmidt, G. A. (2019). Evaluating the performance of past climate model projections. Geophysical Research Letters. https://doi.org/10.1029/2019GL085378

APPENDIX B: ADDITIONAL PAPERS

This section contains additional published papers that touch on issues raised in the dissertation that were coauthored rather than lead-authored. Supplementary materials are not included here for brevity; they can be found on the respective journal websites. The additional papers included are listed below:

LAND TEMPERATURE RECORDS/HOMOGENIZATION

Thorne, P.W., Diamond, H.J., Goodison, B., Harrigan, S., Hausfather, Z., Ingleby, N.B., Jones, P.D., Lawrimore, J.H., Lister, D.H., Merlone, A., Oakley, T., Palecki, M., Peterson, T.C., de Podesta, M., Tassone, C., Venema, V., Willett, K.M., (2018). Towards a global land surface climate fiducial reference measurements network. International Journal of Climatology.

Willett, K., Williams, C., Jolliffe, I. T., Lund, R., Alexander, L. V., Brönnimann, S., Vincent, L. A., Easterbrook, S., Venema, V. K. C., Berry, D., Warren, R. E., Lopardo, G., Auchmann, R., Aguilar, E., Menne, M. J., Gallagher, C., Hausfather, Z., Thorarinsdottir, T., Thorne, P. W. (2014). A framework for benchmarking of homogenisation algorithm performance on the global scale. Geoscientific Instrumentation, Methods and Data Systems, Volume 3, Issue 2, pp.187-200 doi:10.5194/gi-3-187-2014

Thorne, P.W., Donat, M.G., Dunn, R.J.H., Williams, C.N., Alexander, L.V., Caesar, J., Hausfather, Z.,... (2016) Reassessing changes in diurnal temperature range: Intercomparison and evaluation of existing global data set estimates. Journal of Geophysical Research: Atmospheres 121 (10), 5138-5158

OCEAN TEMPERATURES

Cowtan, K., Rohde, R., & Hausfather, Z. (2017). Evaluating biases in Sea Surface Temperature records using coastal weather stations. Quarterly Journal of the Royal Meteorological Society. doi.org/10.1002/qj.3235

Cheng, L., Abraham, J., Hausfather, Z., & Trenberth, K.E. (2019). How fast are the oceans warming? Science 363 (6423), 128-129.

Hausfather, Z. Corrections to ocean-temperature record resolve puzzling regional differences. Nature 571 (7765), 328-329.

Cheng, L., Foster, G., Hausfather, Z., Trenberth, K.E., Abraham, J. (2020). Acceleration of Ocean warming. Nature Geoscience (submitted).

MODEL/OBSERVATION COMPARISONS

Cowtan, K., Hausfather, Z., Hawkins, E., Jacobs, P., Mann, M. E., Miller, S. K., … Way, R. G. (2015). Robust comparison of climate models with observations using blended land air and ocean sea surface temperatures. Geophysical Research Letters, 42(15). doi.org/10.1002/2015GL064888

Haustein, FEL Otto, V Venema, P Jacobs, K Cowtan, Z Hausfather, ... (2019). A Limited Role for Unforced Internal Variability in Twentieth-Century Warming. Journal of Climate 32 (16), 4893-4917

OTHER SUBJECTS

Hausfather, Z. & Peters, G. (2020). Rethinking “business as usual” emission scenarios. Nature (under review).

Mayer, A., Hausfather, Z., Jones, A.D., & Silver W.L. (2019). The potential of agricultural land management to contribute to lower global surface temperatures. Science advances 4 (8).

Roe, S., Streck, C., Obersteiner, M., Frank, S., Griscom, B., Drouet, L., Fricko, O., Hausfather, Z., ... (2019). Contribution of the land sector to a 1.5° °C world. Nature Climate Change, 1-12.

Brown, P.T., Moreno-Cruz, J., Saunders, H., Hausfather, Z., Davis, S.J., Tong, F., Caldeira, K., (2020). Net economic impact of UN global warming mitigation targets under heightened damage estimates. Earth's Future (submitted).

Hausfather, Z. (2015). Bounding the climate viability of natural gas as a bridge fuel to displace coal. Energy Policy, 86, 286–294. https://doi.org/10.1016/j.enpol.2015.07.012

Zhang, X., Myhrvold, N. P., Hausfather, Z., & Caldeira, K. (2016). Climate benefits of natural gas as a bridge fuel and potential delay of near-zero energy systems. Applied Energy, 167, 317–322. https://doi.org/https://doi.org/10.1016/j.apenergy.2015.10.016