http://advances.sciencemag.org/content/3/1/e1601207

CLIMATE CHANGE

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] Robert Rohde[2]

Sea surface temperature (SST) records are subject to potential biases due to changing instrumentation and measurement practices. Significant differences exist between commonly used composite SST reconstructions from the National Oceanic and Atmospheric Administration'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) from 2003 to the present. The update from ERSST version 3b to version 4 resulted in an increase in the operational SST trend estimate during the last 19 years from 0.07° to 0.12°C per decade, indicating a higher rate of warming in recent years. 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 ERSST version 3b and smaller but significant cooling biases in HadSST3 and COBE-SST from 2003 to the 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 data sets.

INTRODUCTION

Accurate sea surface temperature (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 have been hampered by large inhomogeneities due to a marked increase in the use of buoy-based measurements and changing characteristics of ships taking measurements (*1*, *2*). Up until the last two decades, most SST measurements were taken by ships, first with buckets thrown over the side and increasingly through engine room intakes (ERIs) after 1940. Since 1990, the number of buoy-based SST measurements has increased around 25-fold, whereas the number of observations from ships has fallen by around 25% (*3*, *4*). In the last 25 years, SST assay methods have changed from 80% ship-based in 1990 to 80% buoy-based in 2015. Modern ship-based measurements (primarily ERI, although hull contact sensors and other devices are also used) tend to generate temperature readings around 0.12°C higher than those of buoys, whose sensors are directly in contact with the ocean's surface (*1*, *5*, *6*). As the number of ships actively taking measurements available in the International Comprehensive Ocean-Atmosphere Data Set (ICOADS) database (*4*) has fallen, a growing portion of ships are also using non-ERI systems that may introduce further changes in the combined record (*1*). Although 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 (*3*).

The National Oceanic and Atmospheric Administration's (NOAA) Extended Reconstruction Sea Surface Temperature (ERSST) (*5*), the Hadley Centre SST data set (HadSST3) (*1*), and the Japanese Meteorological Agency's Centennial Observation-Based Estimates of SSTs (COBE-SST) (*7*) 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 (Goddard Institute for Space Studies Surface Temperature Analysis) (*8*), the Met Office Hadley Centre's and the University of East Anglia's Climatic Research Unit's HadCRUT (*9*), NOAA's GlobalTemp (*10*, *11*), the Japan Meteorological Agency (*12*), Berkeley Earth (*13*), and Cowtan and Way (*14*). Because the oceans cover 71% of Earth's surface, changes to SST series have large impacts on the resulting global temperature records.

ERSST was recently updated from version 3b (ERSSTv3b) 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 the National Oceanography Centre's HadNMAT2 (*5*, *15*–*17*). ERSSTv3b did not include any SST bias adjustments after 1941, whereas ERSSTv4 continues these adjustments through the present. Although the largest changes to the ERSST record occurred during World War II, ERSSTv4 also indicated a higher rate of warming after 2003. This led Karl *et al.* (*18*) to conclude that the central estimate of the rate of global mean surface temperature change during the 1998–2012 period was comparable to that during the 1951–2012 period, in contrast to the Intergovernmental Panel on Climate Change characterization of the recent period as a "hiatus" (*19*). These updates also created a notable divergence between ERSSTv4, HadSST3, and COBE-SST from 2003 to the 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 SST (IHSST) measurements are available from drifting buoys, Argo floats (*20*), and satellites (see Materials and Methods for details on each IHSST series). To assess how well the composite SST

[1]Energy and Resources Group, University of California, Berkeley, Berkeley, CA 94720, USA. [2]Berkeley Earth, Berkeley, CA 94705, USA. [3]Department of Chemistry, University of York, York, U.K. [4]Independent Researcher, Montreal, Quebec, Canada. [5]Department of Environmental Science and Policy, George Mason University, Fairfax, VA 22030, USA. [6]NASA Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA.
*Corresponding author. Email: hausfath@berkeley.edu

records correct for biases due to the changing instrumentation, we compare each of them in turn to IHSST series that were 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, although 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) spans a different period of time. Buoy data have reasonably complete spatial coverage of the oceans from the late 1990s to the present. Argo floats achieve sufficient coverage for analysis from January 2005, whereas reliable satellite data span from 1996 to the present. Two sources of infrared radiometer–based satellite sea skin temperature are considered: the ARC [ATSR (Along Track Scanning Radiometer) Reprocessing for Climate] SST product (*21*) from ATSR data, which provided data only through the end of 2011, and the European Space Agency Climate Change Initiative experimental record (hereafter CCI) (*22*), 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 homogeneous and is not fully independent from in situ buoy 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 data sets—from the Asia-Pacific Data Research Center (APDRC) (*23*), the Japan Agency for Marine-Earth Science and Technology (hereafter H2008) (*24*, *25*), and Roemmich and Gilson (hereafter RG2009) (*26*)—are examined, with a number of different data sets chosen to reflect the uncertainty introduced by attempting to reconstruct near-SSTs using Argo data.

RESULTS

From January 1997 through December 2015, 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 at 0.08°C per decade, whereas 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) (autoregressive moving average) model to correct for autocorrelation], as shown in Fig. 1. ERSSTv3b is comparable to ERSSTv4 and the buoy and satellite records before 2003, but notable divergences are apparent thereafter.

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.

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 to the present relative to the buoy-only record, although 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 Fig. 2. Each difference series is constructed by restricting all four composite SST series to common grid cells for each month and by 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 Materials and Methods.

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 (Fig. 3). The RG2009 series is more ambiguous, with trends that are not significantly different ($P > 0.05$) from either ERSSTv3b or ERSSTv4. Similarly, both APDRC and H2008 suggest cool biases in HadSST3 and COBE-SST, whereas RG2009 does not show a significant trend in the difference series with any of the composite temperature records (see Fig. 4). Differences between the Argo series emerge through different interpolation techniques and additional data incorporation: APDRC uses Aviso satellite altimetry for sea surface height estimates, H2008 uses a small amount of data from the Triangle Trans-Ocean Buoy Network and conductivity-temperature-depth profilers (mostly before 2005) (*25*), whereas RG2009 relies solely on Argo data.

To assess the significance of differences between composite series and IHSSTs, we examined whether trends in differences between the data sets were statistically different from 0 (that is, $P < 0.05$), as shown in Fig. 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

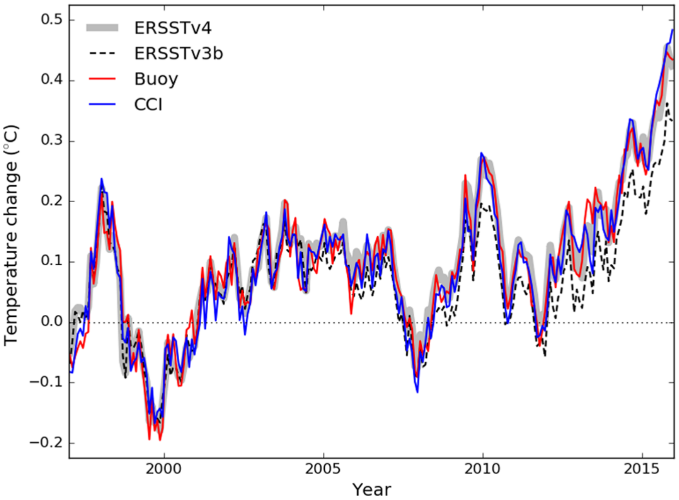

Fig. 1. Comparison of the different ERSSTv3b, ERSSTv4, buoy-only, and CCI SST monthly anomalies from January 1997 to December 2015, restricting all series to common coverage. ERSSTv4 is shown as a broad band for visualization purposes; this band does not represent an uncertainty range. The series are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

trends in differences between the data sets except in the case of H2008, which showed slightly greater warming over the 2005–2015 period. ERSSTv3b, HadSST3, and COBE-SST 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 to 2012, as shown in fig. S3). During 2005–2015, ERSSTv3b, HadSST3, and COBE-SST have significantly lower warming 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 (*15*) and HadSST3 (*1*) 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 (*16*). HadSST3 provides a 100-member ensemble to assess parametric uncertainty but separately treats sampling and measurement uncertainty. We derived a 1000-member ensemble from the HadSST3 ensemble, with each member expanded to 10 members by adding an AR1 time series with SD 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 (Fig. 5).

The ERSSTv4 ensemble is not symmetric around the operational "best" estimate, which is based on the most empirically justified

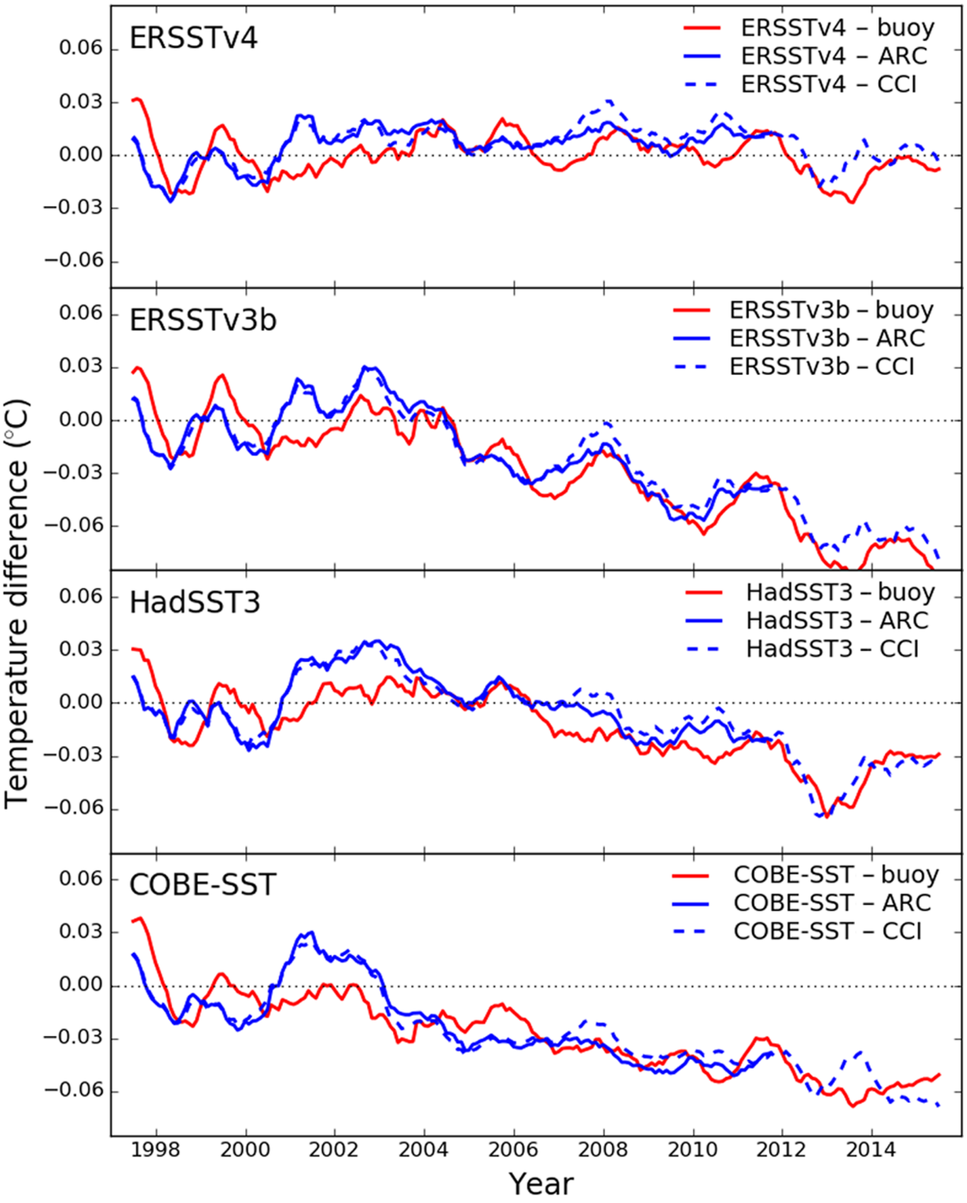

Fig. 2. Twelve-month centered moving average of temperature difference series between composite and buoy-only, CCI, and ARC SST anomalies. Values below 0 indicate that the composite series has a cool bias relative to the IHSST record.

combination of parameter settings (*5*); most of the 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 HadSST3 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 HadSST3 and CCI, coupled with the time-varying coverage of HadSST3.

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

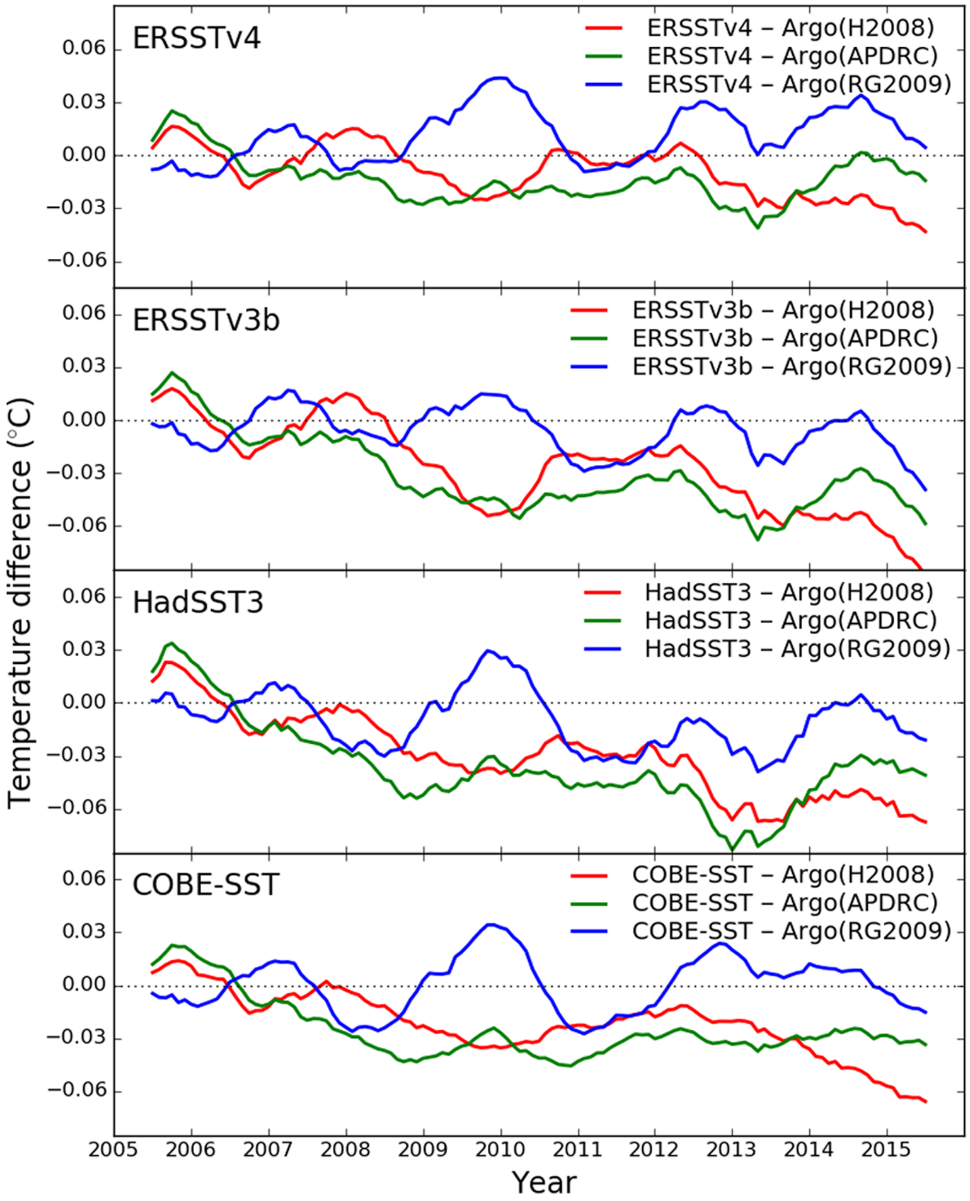

Fig. 3. Twelve-month centered moving average of temperature difference series between composite and Argo near-SST anomalies.

ensemble, because the methods are independent: The ERSST ensemble relies on a bottom-up estimation of uncertainty from the different uncertainties in the methodology, whereas Eq. 8 yields a top-down estimate based on the differences between independent data sources. The trend uncertainties estimated from Eq. 8 are 10 to 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 on the basis of how spatial coverage is handled. For each IHSST difference series, we restrict coverage for each month to the coverage shared in common between the IHSST series in question and the four composite records. This not only serves to maximize the spatial overlap between the data sets 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, because 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 regardless how spatial coverage is handled, we performed two additional tests to account for both spatial and temporal-spatial consistency across the series. In the first test, we restricted all series examined for the two time periods in question (1997–2015 and 2005–2015) to only 1 × 1 latitude/longitude grid cells containing records from all series examined over those time frames. During the 1997–2015 period, we only looked at grid cells with common coverage across the four composite series, buoys, and CCI, whereas 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 figs. S5 to 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 a lower CCI trend during the 2005–2015 period as the only notable differences.

In the second coverage test, we applied a kriging spatial interpolation approach to 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, whereas 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 figs. S8 to S10. Here, the cooling bias in ERSSTv3b, COBE-SST, and HadSST3 is more pronounced with

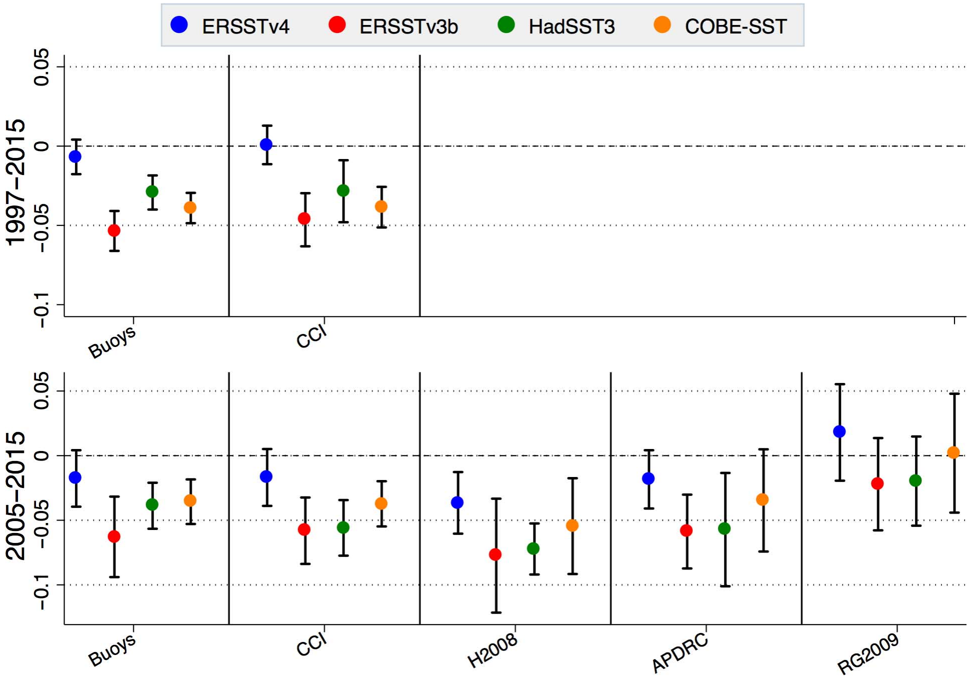

Fig. 4. Trends and 95% confidence intervals (°C per decade) in difference series for each IHSST and composite SST series, masked to common composite SST coverage. Each difference series represents a composite series minus an IHSST series. Confidence intervals for trends are calculated using an ARMA(1, 1) autocorrelation model. Values below 0 indicate that the composite series has a lower trend than the IHSST series over the period examined. The two trend periods examined are January 1997 to December 2015 and January 2005 to December 2015.

respect to the buoy-only and CCI records, although 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 (not apparent in Figs. 2 and 4) that is only apparent when the CCI coverage is reduced to match the buoy coverage (see figs. 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 warming in the northern mid-latitudes. The Southern Ocean is consistently cloud-covered; thus, CCI might be expected to be less accurate in these regions. Winds can also affect skin temperature retrievals

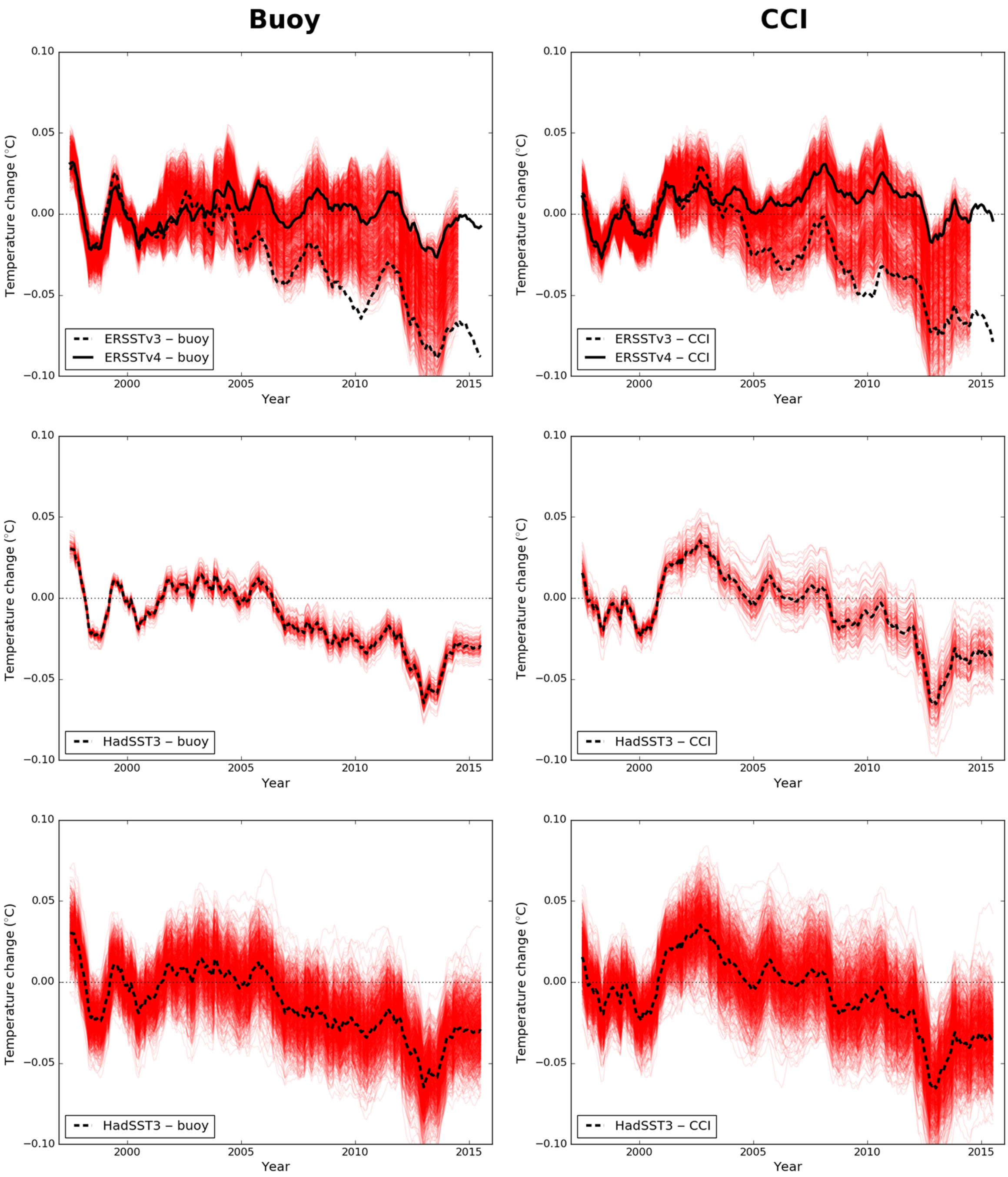

Fig. 5. Twelve-month centered moving average of temperature difference series between collocated ERSSTv4/HadSST3 ensemble realizations and IHSST anomalies. The left column shows the difference series with the buoy-only record. The right column shows the difference series with the CCI record. The top 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.

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 caution.

Coverage biases are also affected by the choice of baseline for geographical map series. The results presented use a 19-year (1997–2015) baseline for both the ERSSTv3b data to which the other series are then matched and 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 1997–2015 period but are significantly higher ($P < 0.01$) than the ERSSTv3b trend, supporting the conclusions of Karl *et al.* (*18*). 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 cool-biased. By contrast, four of five IHSST series suggest cool biases in both ERSSTv3b and HadSST3, whereas three of five IHSST series 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 is particularly noteworthy, because HadSST3 includes explicit buoy-ship offset adjustments comparable to those used by ERSSTv4 and continues ship SST corrections through the present (*1*). The source of the apparent cooling bias in recent years in HadSST3 is unclear, although 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 (*4*). When comparing IHSSTs to a ship-only SST record (restricting to common coverage), we have identified a strong cool bias in the ship record, particularly since 2010. Not only are ship temperatures higher than buoy temperatures at the start of the study period (due to an approximately 0.1°C offset), the ship record substantially underestimates the rate of warming over the later part of the period as well (fig. 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 2.5 to 20 m (fig. S14). ERSSTv4 mostly avoids this potential bias in ship records by assigning an increased weight to buoys in recent years (*5*), although the slightly higher trends in buoys, CCI, and two of the three Argo series vis-a-vis ERSSTv4 during 2005–2015 (Fig. 4, bottom) 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, because 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 equally weighting ships and buoys. NMATs (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 (fig. S15), possibly because of the residual inhomogeneities in NMAT records. Whereas 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 the present and is not yet in operational use in the Japanese Meteorological Agency global land/ocean temperature product (fig. S16) (*27*).

Interpreting the Argo results

The Argo records cover a shorter period (11 years rather than 19), and their results are less clear-cut than the buoy and CCI IHSSTs. The H2008 and APDRC records support ERSSTv4 (and even suggest that it might be a bit too cool), although APDRC results are somewhat sensitive to the choice of start year (fig. S17). RG2009 falls between ERSSTv3b and ERSSTv4 in trend and does not reject either. Similarly, H2008 and APDRC suggest a cool bias in HadSST3 and (to a lesser extent) in COBE-SST over the 2005–2015 period, whereas the results of RG2009 are ambiguous and do not allow any differentiation between composite record trends.

The brevity of the Argo records and their divergence from other records limit the weight that can be placed on them. If the faster warming H2008 and APDRC records are accurate, then all of the IHSSTs (buoys, satellites, and Argo floats) are in basic agreement in rejecting the slower warming ERSSTv3b record. However, if the slower warming RG2009 record is correct, then this would imply either that the buoy and CCI IHSSTs are too warm during 2005–2015 or that there may be a variation in temperature trend with depth: The skin record and the top meter show faster warming, whereas 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, because 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. However, if the RG2009 record is correct, it may suggest that the slower warming ship record arises from a combination of both depth and the bias in the ship record (because the ship record exhibits less warming than even RG2009, as shown in fig. 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 range of 2.5 to 7.5 m. Although the Argo records show no discernable reduction in trends between depths of 5, 10, and 20 m (fig. S18), they 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 SST product from the combination of ship and buoy records. Similarly, most of the Coupled Model Intercomparison Project Phase 5 (CMIP5) climate models have a top layer spanning 0 to 10 m and so may not resolve the top meter of the ocean. This could present a challenge both in testing for the depth effect in models and in comparing the models to observations. However, because two of the three Argo-based records analyzed show no significant difference with buoy and CCI surface records and the Argo series is 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 IHSST 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 present in ship-based records in recent years.

Some uncertainty remains, particularly in Argo-based near-SST reconstructions. Two of the three Argo reconstructions examined agree well globally with the buoy and radiometer-based IHSSTs, whereas the third does not allow for any effective differentiation between composite SST series. Similarly, although CCI and ARC-SST radiometer-based estimates agree quite well globally with the buoy-only record, 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 are only available after 2005. Nonetheless, SST time series from drifting buoys, satellite radiometers, and two of the three Argo series strongly suggest a cool bias present in ERSSTv3b, HadSST3, and COBE-SST. Overall, these results suggest that the new ERSSTv4 record represents the most accurate composite estimate of global SST trends during the past two decades and thus support the finding (*14*) that previously reported rates of surface warming in recent years have been underestimated.

MATERIALS AND METHODS

We compared 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 obtained existing spatially gridded fields for each SST series (and created novel ones in the case of buoy-only and ship-only records) and converted each to standardized 1° latitude by 1° longitude uniform grid (hereafter 1° × 1° 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 0 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 to a more complete record containing the same weather signal. The spatially complete ERSSTv3b record was therefore aligned to 0 during the 1997–2015 period, and then the other data sets are aligned to the normalized 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 toward it.

Data series were carefully aligned to ensure accurate intercomparisons of SST series. The process was as follows: Optimum Interpolation SST (OISST) was used to construct 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 was then calculated using this high-resolution daily climatology, yielding 228 monthly fields (19 years × 12 months). ERSSTv3b was also aligned to the 1997–2015 baseline. All of the composite series and IHSSTs (including the buoy series) were then aligned to the baselined ERSSTv3b on the basis of whatever months are available for each grid cell. These were then masked to common coverage and plotted in Fig. 1. This made use of the spatial completeness of ERSSTv3b to avoid artifacts due to baselining temporally incomplete cells on an incomplete baseline period; we used ERSSTv3b for this purpose to avoid biasing our results toward ERSSTv4. Pairwise difference map series were calculated between the aligned maps. The study was restricted to the 1997–2015 period, with the start date determined by buoy coverage and a data break in the ATSR-based SST data. Details of how each data set was obtained and processed are provided below.

ERSST, HadSST3, and COBE-SST

Both ERSSTv3b (*10*) and ERSSTv4 (*11*) were produced on a 2° × 2° 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 1° × 1° grid, repeating each value from the original grid to the four corresponding cells in the finer grid. HadSST3 (*1*) was produced on a 5° × 5° grid with no values for sea ice cells and was expanded to the 1° × 1° grid by repeating each value from the original grid to the 25 corresponding cells in the finer grid. COBE-SST (*7*) and COBE-SST2 (*27*) were distributed as a 1 × 1 gridded product; cells with sea ice were recorded as −1.8°C, similar to ERSST, and were set to missing. Because both HadSST3 and ERSSTv4 included ensembles of realizations with different parameterizations, for the main analysis in the paper (for example, Figs. 1 to 4), the operational version of each series was used. This is the ensemble median in the case of HadSST3, whereas ERSSTv4 provides a preferred realization.

Different approaches were used in the construction of the gridded SST products. In the HadSST3 record, observations only contributed to the grid cell and month in which they occurred, leading to some cells for which no temperature estimate was available. In the COBE-SST records, optimal interpolation was used in both space and time to create a spatially complete field from the available data. The ERSST and COBE-SST2 data sets combined a low-resolution reconstruction with the fitting of empirical orthogonal teleconnections to the observations to produce a spatially complete field, in which local temperatures could be inferred from distant observations (up to a specified distance) through teleconnections. All the records included data from ICOADS (albeit some from different releases of the database); however, in addition to differences in the processing methods, ERSSTv4 attached an increased weight to buoy observations on the basis of their lower estimated uncertainty.

Because some of the composite SST series included 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 data sets for any given month. Because HadSST3 included no explicit interpolation (apart from that implicit in its use of relatively large 5° × 5° 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 were obtained from the ICOADS Release 2.5 data (*4*). Drifting buoys were selected by the World Meteorological Organization

(WMO) 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 because of an offset in temperatures between drifting and moored buoys (perhaps due to measurement depth; see fig. 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 was determined by averaging buoy observations over the span of a month for each cell in a global grid. The grid consisted of cells of equal area, with equatorial cells spanning a range of 5° in both longitude and latitude. At higher latitudes, the longitudinal width of a cell in degrees was increased by calculating the area of the latitude band, dividing by the area of a 5 × 5 cell at the equator, and using that many cells in the latitude band to maintain a constant area.

The data were processed 1 month at a time. For each buoy, data were divided into days. The (typically hourly) temperature, latitude, and longitude data for that day were averaged. Buoys that showed temperature variations with an SD exceeding 1°C or positional variation with an SD exceeding 0.5° of latitude or longitude during a single day were excluded for the whole month: This can occur if a buoy is beached or picked up by a ship. The temperature was then converted to an anomaly using climatology calculated from OISST version 2 (OISSTv2) (*28*) for that day of the year and for the corresponding latitude and longitude on a finer 0.5° grid. This mitigated the biasing effects of temperature observations at the beginning or end of a month or the northern or southern edges of a 5° latitude band. The daily mean temperature anomaly for the buoy was then added to a list for the corresponding grid cell. After all buoy records were processed, all temperature anomalies for a given cell were 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 artifacts, 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 miscalibration 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 miscalibrated 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 was no sign of such a bias in the comparisons between different IHSSTs, and the cross-validated uncertainties were lowest for the recent period where the composite records show most difference.

Additional interbuoy comparisons were performed to address this possibility. For each grid cell and month where at least three 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 SD 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:

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

(2) Filtering on the basis of daily variability only (the default per-buoy filter, described at the start of this section).

(3) Filtering on the basis of daily variability and interbuoy variability (that is, the additional filter described in the previous paragraph).

(4) Filtering on the basis of monthly and interbuoy variability and application of the bias correction [as in (3) but then recalculating the buoy record after applying a correction to each buoy on the basis of its mean difference with passing buoys].

The resulting temperature series are shown in fig. S20, along with the differences of the other methods from the default method. The largest difference arose from using all of the data without filtering for daily variability. Interbuoy variability and bias correction made a smaller difference. The differences between the methods were small compared to the differences between the composite records. The default method using a per-buoy filter showed the lowest trend during 1997–2015 and was therefore a conservative choice.

The buoy coverage was limited, particularly in the 1990s, and comparisons to other data sets may have been affected by coverage bias. To produce an unbiased comparison to other data sets, all the data sets were expanded onto a 1° × 1° grid. Comparisons were made using only the cells for which the data sets being compared had values. The area-weighted mean temperature was then calculated for each record using the common coverage cells. The percentage 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 was 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 were masked for security reasons (although this has been improved in Release 3 of ICOADS). The test to detect excessive motion or variation within a single day was therefore omitted. The only quality control applied to the ship record therefore arose from the calculation of the global mean of the SST field, which excluded observations that fell in land areas. The ship observations were subject to significant quality issues, and the limited quality control implemented in this record therefore provided 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 APDRC (*23*), the Japan Agency for Marine-Earth Science and Technology (H2008) (*24*, *25*), and Roemmich and Gilson (RG2009) (*26*) were used. These data were produced on a monthly 1° × 1° grid and were 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, whereas 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 had temperature values at 2.5, 10, and 20 dbar and deeper levels; the H2008 product had temperatures at 10, 20, and 30 dbar and deeper levels, and the APDRC product had temperature values at 0, 5, and 10 m and deeper levels. We used the 5-m level for the APDRC product, the 10-dbar (10 m) level for the H2008 product, and the 2.5-dbar level for the RG2009 product (which represented measurements ranging from 2.5 to 7.5 dbar with a mean level of 5 dbar/m) to provide the most comparable and highest available depths; estimated 0-m temperatures from APDRC were not used because they resulted from interpolation (because no Argo floats sampled sea skin temperatures).

Throughout the paper, we refer to the record derived from Argo floats as "near-SST," because the highest level of the ocean measured by most Argo floats is approximately 5 m below the surface (*26*). However, with the exception of satellite radiometer-based estimates, all of the instruments used in this analysis recorded ocean temperatures at depths between 0 and 20 m. For example, ships tend to measure temperatures through ERI valves at depths of 7 to 11 m for large ships and 1 to 3 m for small ships (*3*). Moored buoys typically measure SSTs at a depth of 3 m, whereas drifting buoys measure SSTs at around 0.5 m. Recent work (*29*) found no long-term difference in warming rates between depths of 0 to 4 m and depths of 4 to 9 m in a CMIP5 model; similarly, we have established that our results are robust when using the next deeper level of each Argo data set (fig. S19). The different depths sampled by the different observational systems provide a basis to assess whether depth plays a role in the rate of recent warming.

Argo data have been used to create SST analogs in the past; for example, Roemmich and Gilson (*30*) compared ARGO "near"-SST to NOAA's OISSTv1, whereas Roemmich *et al.*(*31*) compared a 5-m Argo-based SST record to OISSTv2. Here, we performed a similar analysis using the Argo-based fields provided by RG2009, APDRC, and H2008.

Satellites

The ATSR instruments provided infrared images of Earth, from which skin temperatures may be derived. ATSR data were incorporated into two gridded data sets, the ATSR ARC (*21*) spanning the 1996–2012 period and the experimental National Center for Earth Observation/European Space Agency SST CCI Analysis L3S version EXP-1.2 (ESA-CCI or CCI) (*22*), which also incorporates data from the AVHRR and spans the period from 1996 to the present (end of 2015). Coverage between 60°S and 60°N was largely complete (except for a few cells each month in the ATSR record, which were affected by cloud, typically in the Southern Ocean or North Atlantic). Both the ATSR-only (through mid-2012) and ATSR + AVHRR (through present time) CCI data were analyzed, and the CCI data were used in the paper because they extend to the present (and differences between the two were minor during the period of overlap, as shown in Fig. 6).

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, because all largely rely on 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 carried out for the two periods of interest (1997–2015 and 2005–2015), restricting the analysis to only grid cells, where all series available for those periods had coverage. During the 1997–2015 period, this means that only 1 × 1 latitude/longitude 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 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 and changes in coverage over time. Infilling was performed on the gridded data using the original grid sampling for that record: For the buoy record, this was on the 550-km equal area grid, and for the HadSST3, this was on the 5° × 5° grid. The resulting infilled field was then copied onto a 1° × 1° grid as before. Infilling was performed using the method of kriging (*32*), by which the values at unobserved locations were inferred from the observed values. Each observation was 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 was determined from grid cells for which observations were available, fitted with an exponential model controlled by a single range parameter, which was the e-folding distance of the variance. The kriging calculation also used the covariance between locations where observations were present to estimate the amount of independent information in each observation. The buoy record showed longer range autocorrelation than the HadSST3 data, with respective e-folding distances of 1400 and 900 km, suggesting that the buoy record showed more spatial autocorrelation.

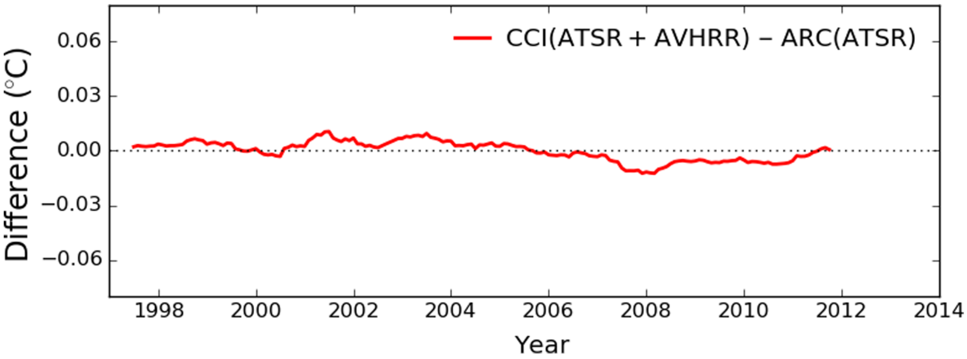

Fig. 6. Twelve-month centered moving average of differences between CCI ATSR + AVHRR and ATSR-only ARC SST records during the period of overlap. The earlier IHSST ARC shows small differences to the newer combined version; however, the differences are minor compared to the differences relative to the composite SST records.

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 that are very distant from any observation will tend toward 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 because of random errors in the data or errors introduced during processing or because of 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 examined different methods for the determination of the uncertainty in a reconstruction. Two approaches were used. First, collocated temperature difference series were used to estimate the significance of the differences. Second, a method was outlined for the use of independent temperature series to directly estimate the structural uncertainty in each series.

Significance of the temperature difference series trends.

To assess the significance of differences in trends between temperature series, we first calculated the difference temperature series from the difference map series to eliminate differences in coverage. The trend in the difference series was then compared to the uncertainty in that trend estimated using an appropriate autoregression model and used to determine whether the trend difference was significantly different from 0.

The trend in the difference series is identical to the difference in the trends between the two series, assuming that 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 (*33*). 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 markedly 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 were estimated with ordinary least squares (OLS), with SE correction to account for serial correlation of the residuals (*34*–*36*). 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

$$n_e = n_t \Big/ \left(1 + 2\sum_{j=1}^{n-1} \varrho_j\right) \tag{1}$$

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 ARMA noise model estimated from the OLS residuals. An ARMA(1, 1) model was used for all gridded and global difference series (for example, 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

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

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

$$\begin{aligned} \varrho_0 &= 1 \\ \varrho_1 &= (\phi+\theta)(1+\phi\theta)/(1+2\phi\theta+\theta^2) \\ \varrho_j &= \varrho_1\phi^{j-1}, \ j \geq 2 \end{aligned} \tag{3}$$

where ϕ and θ are the respective AR and MA coefficients.

Because the assessed trends cover only 11 to 19 years (132 to 228 months), a bias correction was also applied to the global difference series trends to account for the underestimate of autocorrelation in these short series (*35*, *37*). The original Tjøstheim and Paulsen correction for the AR(1) estimated coefficient ϕ is given by

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

The bias correction of ARMA(1, 1) estimated ACF coefficients ϱ_1, and ϕ generalizes (*7*) by also accounting for the positive difference between ϕ and ϱ_1. Note that the AR(1) bias correction in Eq. 4 then becomes a special case where θ = 0 and ϱ_1 = ϕ [AR(1) is employed in the few cases where this results in more conservative uncertainties].

$$\begin{aligned} \phi_{bc} &= \phi + \left(1 + 4\left(2\phi - \varrho_1\right)\right)/n_t \\ \varrho_{1bc} &= \varrho_1 + \left(1 + 4\left(2\varphi - \varrho_1\right)\right)/n_t \end{aligned} \tag{5}$$

The ARMA coefficient estimates ϕ_{bc} and ϱ_{1bc} can then be substituted into the appropriate specific form of Eq. 1. The ARMA(1, 1) formulation in Eq. 3 can then be simplified as (*36*)

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

IHSST uncertainty estimation.

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

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 were then calculated from the collocated values from each map series. A 120-month moving root mean square 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 that 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

$$\sigma^2{}_{\text{buoy-Argo}} = \sigma^2{}_{\text{buoy}} + \sigma^2{}_{\text{Argo}} \tag{7}$$

and assuming 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

$$\sigma^2{}_{\text{buoy}} = {}^1\!/_2(\sigma^2{}_{\text{buoy}-\text{Argo}} + \sigma^2{}_{\text{buoy}-\text{CCI}} - \sigma^2{}_{\text{Argo}-\text{CCI}}) \tag{8}$$

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.* (*38*).

The uncertainty in the trend in an IHSST series may be estimated from the uncertainty in the monthly temperatures obtained from Eq. 8 using the equation

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

where σ^2_β is the variance of the trend, σ is the SD of the time series values, t_i is the date of the ith value in fractional years, and ν is the number of months of data per effective degree of freedom (*36*). Note that this differs from the ordinary equation for the uncertainty in a trend in the use of the SD of the time series in place of the SD 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

$$\sigma^2_\beta = \frac{\nu\sigma^2}{\Delta t^3} \tag{10}$$

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.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/3/1/e1601207/DC1

fig. S1. Trend maps on the 2005–2015 period for all of the composite records, and for the buoy, Argo, and CCI records.

fig. S2. Comparison of ERSSTv3b and ERSSTv4 with three different Argo-based near-SST records, using the same spatial restrictions as in Fig. 1, but with ERSSTv4 aligned to 1997–2001 (inclusive), with all other series aligned onto ERSSTv4 using the 2005–2007 period because of the limited time span with Argo data.

fig. S3. Trends and 95% confidence intervals (°C per decade) for the 1997–2012 period for buoy, ARC, and CCI IHSSTs and each composite SST series, masked to common composite SST coverage.

fig. S4. Cross-validated uncertainties for the buoy record, whether with no climatology or with daily climatologies derived from the OISSTv2 daily reanalysis data.

fig. S5. Twelve-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).

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

fig. S7. Trends and 95% confidence intervals (°C per decade) in difference series for each IHSST and composite SST series, masked to common coverage for all series available.

fig. S8. Twelve-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.

fig. S9. Twelve-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.

fig. S10. Trends and 95% confidence intervals (°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.

fig. S11. Trend difference maps from January 1997 to December 2015 for the difference between ERSSTv4 and CCI.

fig. S12. Differences between ERSSTv4 and CCI by latitude zone.

fig. S13. Buoy-only and ship-only temperature anomalies from January 1997 to December 2015, with no matching of coverage.

fig. S14. Difference between ship-only record and the three Argo series using a 12-month centered moving average.

fig. S15. Comparison of COBE-SST and COBE-SST2 to the IHSSTs using a 12-month centered moving average.

fig. S16. Comparison of HadSST3 and HadNMAT2 to the IHSSTs using a 12-month centered moving average.

fig. S17. Trends in differences for ERSSTv4 records versus IHSST records, with common coverage from 1997 (buoys and CCI only as dashed lines) and common coverage from 2005 (buoys, CCI, and Argos as solid lines).

fig. S18. Differences between Argo series at minimum reported depth, and differences within each Argo series as minimum reported, 20- and 50-m depths.

fig. S19. Comparison of buoy records composed of all buoys (drifting + moored) and only drifting buoys.

fig. S20. Comparison of drifting buoy–based IHSST records for different quality control and homogenization choices.

fig. S21. Twelve-month centered moving average of differences between IHSST series from January 1997 to December 2015 when reduced to common coverage for each separate pairing.

fig. S22. Trend difference maps during 2005–2015 for the composite records versus Buoy, CCI, and Argo (H2008).

fig. S23. Trends in differences for composite versus buoy (solid lines) and CCI (dashed lines) IHSST records with common coverage.

fig. S24. Number of observations over time by instrument type in the ICOADS (version 2.5) database.

fig. S25. Similar to fig. S24, but showing the percentage of ICOADS observations in each year from each instrument type.

REFERENCES AND NOTES

1. 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**, D14104 (2011).
2. J. J. Kennedy, A review of uncertainty in in situ measurements and data sets of sea surface temperature. *Rev. Geophys.* **52**, 1–32 (2014).
3. E. C. Kent, J. J. Kennedy, D. I. Berry, R. O. Smith, Effects of instrumentation changes on sea surface temperature measured in situ. *Wiley Interdiscip. Rev. Clim. Change* **1**, 718–728 (2010).
4. S. D. Woodruff, S. J. Worley, S. J. Lubker, Z. Ji, E. Freeman, D. I. Berry, P. Brohan, E. C. Kent, R. W. Reynolds, S. R. Smith, C. Wilkinson, ICOADS Release 2.5: Extensions and enhancements to the surface marine meteorological archive. *Int. J. Climatol.* **31**, 951–967 (2011).
5. B. Huang, V. F. Banzon, E. Freeman, J. Lawrimore, W. Liu, T. C. Peterson, T. M. Smith, P. W. Thorne, S. D. Woodruff, H.-M. Zhang, Extended reconstructed sea surface temperature version 4 (ERSST.v4). Part I: Upgrades and intercomparisons. *J. Clim.* **28**, 911–930 (2015).
6. 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).

7. 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).
8. J. Hansen, R. Ruedy, M. Sato, K. Lo, Global surface temperature change. *Rev. Geophys.* **48**, RG4004 (2010).
9. 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** 10.1029/2011JD017187 (2012).
10. 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. Clim.* **21**, 2283–2296 (2008).
11. 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, NOAA's merged land–ocean surface temperature analysis. *Bull. Am. Meteorol. Soc.* **93**, 1677–1685 (2012).
12. K. Ishihara, Calculation of global surface temperature anomalies with COBE-SST. *Weather Service Bulletin* **73**, S19–S25 (2006).
13. R. Rohde, R. A. Muller, R. Jacobsen, E. Muller, S. Perlmutter, A. Rosenfeld, J. Wurtele, D. Groom, C. Wickham, A new estimate of the average Earth surface land temperature spanning 1753 to 2011. *Geoinfor. Geostat. An Overview* **1**, 1–7 (2013).
14. 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).
15. B. Huang, P. W. Thorne, T. M. Smith, W. Liu, J. Lawrimore, V. F. Banzon, H.-M. Zhang, T. C. Peterson, M. Menne, Further exploring and quantifying uncertainties for extended reconstructed sea surface temperature (ERSST) version 4 (v4). *J. Clim.* **29**, 3119–3142 (2015).
16. W. Liu, B. Huang, P. W. Thorne, V. F. Banzon, H.-M. Zhang, E. Freeman, J. Lawrimore, T. C. Peterson, T. M. Smith, S. D. Woodruff, Extended reconstructed sea surface temperature version 4 (ERSST.v4): Part II. Parametric and structural uncertainty estimations. *J. Clim.* **28**, 931–951 (2015).
17. E. C. Kent, N. A. Rayner, D. I. Berry, M. Saunby, B. I. Moat, J. J. Kennedy, D. E. Parker, Global analysis of night marine air temperature and its uncertainty since 1880: The HadNMAT2 data set. *J. Geophys. Res. Atmos.* **118**, 1281–1298 (2013).
18. 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, Possible artifacts of data biases in the recent global surface warming hiatus. *Science* **348**, 1469–1472 (2015).
19. G. Flato, J. Marotzke, B. Abiodun, P. Braconnot, S. Chou, W. Collins, Evaluation of climate models, in *Climate Change 2013: The Physical Science Basis*, T. F. Stocker, D. Qin, G. K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, P. M. Medley, Eds. (Cambridge Univ. Press, 2013), pp. 741–866.
20. Argo, Argo float data and metadata from Global Data Assembly Centre (Argo GDAC). SEANOE, 10.17882/42182 (2002).
21. 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, A 20 year independent record of sea surface temperature for climate from Along-Track Scanning Radiometers. *J. Geophys. Res. Oceans* **117**, C12013 (2012).
22. 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, 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 (2014).
23. 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).
24. S. Hosoda, T. Ohira, K. Sato, T. Suga, Improved description of global mixed-layer depth using Argo profiling floats. *J. Oceanogr.* **66**, 773–787 (2010).
25. S. Hosoda, T. Ohira, T. Nakamura, A monthly mean dataset of global oceanic temperature and salinity derived from Argo float observations. *JAMSTEC Rep. Res. Dev.* **8**, 47–59 (2008).
26. 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. *Prog. Oceanogr.* **82**, 81–100 (2009).
27. S. Hirahara, M. Ishii, Y. Fukuda, Centennial-scale sea surface temperature analysis and its uncertainty. *J. Clim.* **27**, 57–75 (2014).
28. R. W. Reynolds, T. M. Smith, C. Liu, D. B. Chelton, K. S. Casey, M. G. Schlax, Daily high-resolution-blended analyses for sea surface temperature. *J. Clim.* **20**, 5473–5496 (2007).
29. M. Richardson, K. Cowtan, E. Hawkins, M. B. Stolpe, Reconciled climate response estimates from climate models and the energy budget of Earth. *Nat. Clim. Change* **6**, 931–935 (2016).
30. D. Roemmich, J. Gilson, The global ocean imprint of ENSO. *Geophys. Res. Lett.* **38**, L13606 (2011).
31. D. Roemmich, J. Church, J. Gilson, D. Monselesan, P. Sutton, S. Wijffels, Unabated planetary warming and its ocean structure since 2006. *Nat. Clim. Change* **5**, 240–245 (2015).
32. N. Cressie, The origins of kriging. *Math. Geol.* **22**, 239–252 (1990).
33. P. J. Klotzbach, R. A. Pielke Sr., R. A. Pielke Jr., 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).
34. B. D. Santer, T. M. L. Wigley, J. S. Boyle, D. J. Gaffen, J. J. Hnilo, D. Nychka, D. E. L. Parker, E. Taylor, Statistical significance of trends and trend differences in layer-average atmospheric temperature time series. *J. Geophys. Res. Atmos.* **105**, 7337–7356 (2000).
35. J. Lee, R. Lund, Revisiting simple linear regression with autocorrelated errors. *Biometrika* **91**, 240–245 (2004).
36. G. Foster, S. Rahmstorf, Global temperature evolution 1979–2010. *Environ. Res. Lett.* **6**, 44022 (2011).
37. D. Tjøstheim, J. Paulsen, Bias of some commonly-used time series estimates. *Biometrika* **70**, 389–399 (1983).
38. 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. *J. Atmos. Ocean Tech.* **33**, 1197–1207 (2008).

Acknowledgments: We thank four anonymous reviewers for feedback on the article, O. Embury for assistance with ESA-CCI SST data, M. Ishii for information regarding COBE-SST data, W. Eschenbach for help in locating APDRC Argo data, and B. Huang for providing gridded ERSSTv4 ensemble data. Argo data used in the study were collected and made freely available by the International Argo Program and the national programs that contribute to it (www.argo.ucsd.edu). 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 NASA. 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 the project; K.C. and R.R. produced the buoy records. Z.H. and K.C. analyzed the Argo data. M.R. analyzed the CCI data. K.C., D.C.C., and Z.H. analyzed time series and trends. Z.H., K.C., P.J., D.C.C., and M.R. wrote the paper, with all authors providing input. **Competing interests:** All authors declare that they have no competing interests. **Data and materials availability:** Buoy and ship data are available from ICOADS at http://icoads.noaa.gov/products.html; ESA-CCI data are 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 are available at http://catalogue.ceda.ac.uk/uuid/ff8a7f27b827c108dd9756adffaaa942; Argo data are available from the APDRC at http://apdrc.soest.hawaii.edu/projects/Argo/data/gridded/On_standard_levels/index-1.html; Argo data are available from the Japan Agency for Marine-Earth Science and Technology (H2008) at http://www.jamstec.go.jp/ARGO/argo_web/argo/?page_id=83&lang=en; Argo data are available from Roemmich and Gilson (RG2009) at http://sio-argo.ucsd.edu/RG_Climatology.html; ERSSTv4 data are available at http://www1.ncdc.noaa.gov/pub/data/cmb/ersst/v4/netcdf/; ERSSTv3b data are available at http://www1.ncdc.noaa.gov/pub/data/cmb/ersst/v3b/netcdf/; HadSST3 data are available at http://www.metoffice.gov.uk/hadobs/hadsst3/data/download.html; COBE-SST data are available at http://ds.data.jma.go.jp/tcc/tcc/products/elnino/cobesst/cobe-sst.html; and the Python code used in this analysis is available at http://www-users.york.ac.uk/~kdc3/papers/ihsst2016/. All data sets used were most recently accessed in early July 2016. All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Submitted 27 May 2016
Accepted 22 November 2016
Published 4 January 2017
10.1126/sciadv.1601207

Citation: Z. Hausfather, K. Cowtan, D. C. Clarke, P. Jacobs, M. Richardson, R. Rohde, Assessing recent warming using instrumentally homogeneous sea surface temperature records. *Sci. Adv.* **3**, e1601207 (2017).

Science Advances

Assessing recent warming using instrumentally homogeneous sea surface temperature records

Zeke Hausfather, Kevin Cowtan, David C. Clarke, Peter Jacobs, Mark Richardson and Robert Rohde

Sci Adv **3** (1), e1601207.
DOI: 10.1126/sciadv.1601207