Evaluating biases in Sea Surface Temperature records using coastal weather stations

Kevin Cowtan[a], Robert Rohde[b], Zeke Hausfather[b,c]

[a]Department of Chemistry, University of York, Heslington, York
YO10 5DD, United Kingdom.
[b]Berkeley Earth, Berkeley CA 94705.
[c]University of California Berkeley, Berkeley CA 94720.

Abstract

Sea surface temperatures form a vital part of global mean surface temperature records. Historical observation methods have changed substantially over time from buckets to engine room intake sensors, hull sensors and drifting buoys, rendering their use for climatological studies problematic. There are substantial uncertainties in the relative biases of different observations which may impact the global temperature record.

Island and coastal weather stations can be compared to coastal sea surface temperature observations to obtain an assessment of changes in bias over time. The process is made more challenging by differences in the rate of warming between air temperatures and sea surface temperatures, and differences across coastal boundaries. A preliminary sea surface temperature reconstruction homogenized using coastal weather station data suggests significant changes to the sea surface temperature record, although there are substantial uncertainties of which only some can be quantified. A large warm excursion in versions 4 and 5 of the NOAA Extended Reconstructed Sea Surface Temperature during World War 2 is rejected, as is a cool excursion around 1910 present in all existing records. The mid-century plateau is cooler than in existing reconstructions.

1 Introduction

Historical estimates of global mean surface temperature are generally constructed from a blend of land surface air temperature from weather stations and sea surface temperature (SST) estimates from ships and buoys. Changes to weather station equipment have had only a modest effect over the past one and a half centuries, which can be largely corrected by use of metadata and interstation comparisons (Menne and Williams Jr., 2009; Hausfather *et al.*, 2016). By contrast sea surface temperatures have been measured using both canvas and insulated buckets, engine room intake sensors, ship hull sensors and free floating buoys, with the different systems measuring temperatures at different depths (Kent *et al.*, 2010). The changing measurement methods require substantial corrections, the largest of which being the 'bucket correction' of about $0.4°C$ around the start of the Second World War.

Different approaches have been used to homogenize sea surface temperature observations. The HadSST3 record from the UK Hadley Centre makes use of metadata to determine the most likely method used for a given observation, along with field replication of measurement methods and reconciliation of different observation types to correct for the heterogenous observation systems (Folland and Parker, 1995; Rayner *et al.*, 2006; Kennedy *et al.*, 2011a,b) and is based on the most complete current analysis of the observational metadata. Temperature fields determined from the unadjusted data are also available. The COBE-SST2 record (Hirahara *et al.*, 2014) also uses metadata but adopts a different approach to dealing with observations where metadata is unavailable, with similar results. By contrast the NOAA Extended Reconstructed Sea Surface Temperature version 4 (ERSSTv4) and 5 (ERSSTv5) products (Huang *et al.*, 2015, 2017) make use of nighttime marine air temperature (NMAT) observations (Kent *et al.*, 2013) as a reference against which to correct the sea surface temperature observations from ships.

Both methods have limitations: the metadata approach depends on inference of the observational method for each observation and the correct determination of the resulting bias. The NMAT approach depends on the assumption that the NMATs themselves are unbiased, or at least less biased than the sea surface temperature observations. Nighttime marine air temperatures are used because they are less influenced by daytime heating of the ship superstructure, however other factors such as the height of the deck above sea level also influence nighttime observations. The metadata and NMAT methods are largely independent, although NMATs have been used indirectly in estimating the prevalence of bucket types (Folland and Parker, 1995). If both methods produced similar results this would increase our confidence in them, however in practice there are substantial differences between the reconstructions prior to 1980. Kent *et al.* (2017) identify this problem and suggest approaches to addressing it, including the comparison of coastal weather stations and sea surface temperature observations.

The substantial differences between the different records can be seen in a common coverage comparison, shown in Figure 1, along with the difference between them. The records show fairly good agreement from the 1970s to the present. However, ERSSTv4 and ERSSTv5 are significantly cooler than HadSST3 and COBE-SST2 over the period 1920-1970, except for the World War 2 period (shown in the shaded area of Figure 1). Both ERSST versions are warmer than HadSST3 or COBE-SST2 prior to 1890 and show further divergence earlier in the 19th century.

The differences around World War 2 are particularly striking, with ERSSTv4 and v5 showing a large spike in temperatures while HadSST3 shows only a modest peak. A drop in the number of observations coupled with changing data sources makes this period particularly problematic (Kennedy *et al.*, 2011b).

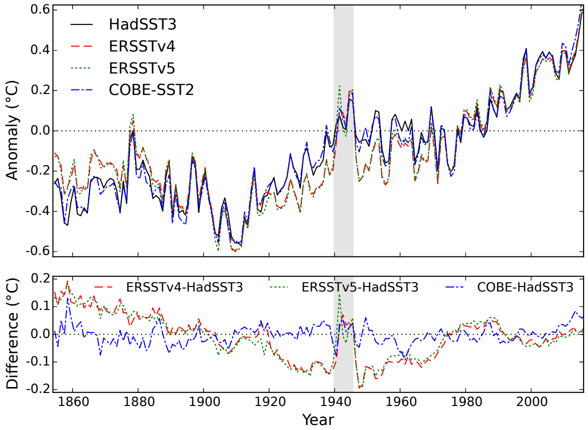

Figure 1: HadSST3 sea surface temperature anomalies with respect to the period 1961-1990, compared to ERSSTv4, ERSSTv5 and COBE-SST2. All records are aligned to HadSST3 on the period 1981-2010 (top panel) when the records show good agreement: this convention will be adopted for the remaining figure. Differences with HadSST3 are show in the bottom panel. All records are masked for common spatial coverage.

While ship-based measurements were greatly impacted by the war, land-based observations were less disrupted. Previous research has taken advantage of the more homogeneous land record during this period; for example Folland (2005) evaluated corrections to bucket observations by using the sea surface temperatures to drive a climate model, and then comparing the modelled land temperatures to observations. Similarly, Jones *et al.* (1991) and Parker *et al.* (1995) used data from coastal or island weather stations to assess the homogeneity of the sea surface temperature observations from ships passing close to those islands. Since ships are mobile platforms that can move between open ocean and coastal waters, a bias in the observations close to shore will generally also correspond to a bias in open ocean observations.

This paper will provide a preliminary evaluation of the use of island and coastal weather stations for the automatic homogenization of global sea surface temperatures across the whole period of the sea surface temperature record. The quality controlled but unadjusted sea surface temperature fields from the HadSST3 dataset will be compared to quality controlled coastal and island weather station data from version 4(beta) of the Global Historical Climatology Network-Monthly (GHCN-M v4) (Lawrimore *et al.*, 2011), and the differences used to correct the sea surface temperature record. The process is complicated by the presence of a climate signal in the difference in temperature between the sea surface and marine air temperatures (Cowtan *et al.*, 2015), and differences in temperature on crossing the coastal boundary, which must be taken into account.

A distinction is generally made between sea surface temperature (SST) of the surface ocean waters, marine air temperature (MAT) of the air at the ocean surface, and land surface air temperature (LSAT) as observed by weather stations. These will be assumed to refer to non-coastal regions, and the new terms coastal SST (CSST), coastal marine air temperature (CMAT) and coastal land surface air temperature (CLSAT) will be used for coastal regions. The differences between MAT and SST will be referred to as air-sea difference. The difference between SST and CSST will be referred to as inshore difference. The differences between CMAT and CLSAT will be referred to as coastal difference. The difference between CLSAT and LSAT will be referred to as inland difference. Not all of these are resolvable in either models or observations due to the limited resolution of climate models and limited spatial coverage of the observations.

HadSST3 temperatures are expressed in terms of anomalies with respect to a 1961-1990 baseline; other records will be aligned to HadSST3 on the period 1981-2010 when the records show good agreement. Absolute temperature differences are ignored and only differences in temperature change between different types of observations will be discussed.

2 Change in coastal land surface air temperature as an indicator of sea surface temperature

The use of weather stations to assess inhomogeneities in SST assumes that change in land surface air temperature measured by coastal weather stations is a good indication of change in sea surface temperature, and this assumption must be evaluated. Globally, land warms faster than oceans, and so it is possible that coastal air temperatures might overestimate sea surface temperature change. Coastal air temperatures are less variable than temperatures in continental interiors, so land based weather stations will be most useful if they are sufficiently close to the coast. Island weather stations may be particularly useful in this regard.

To evaluate the utility of coastal land-based weather stations to estimate coastal sea surface temperatures, surface air temperatures were examined for the high resolution GFDL-HiRAM C360 model runs, which are reported on a fine ∼30 km grid (Harris *et al.*, 2016). Atmospheric Model Intercomparison Project-style historical experiments are available for the period 1979-2008, which is characterized by rapid greenhouse warming. Sea surface temperatures ('tos' in CMIP nomenclature), surface air temperatures ('tas'), and the land mask ('sftlf') are all available on the same grid (Taylor *et al.*, 2012). Two runs of this model are available.

In order to determine whether land-based weather stations can give an indication of marine air temperature, the trend in the difference between surface air temperature and sea surface temperature (i.e. tas-tos) was examined while crossing coastal boundaries. No sea surface temperatures are available for pure land cells, however the variation in temperature difference can be examined as a function of increasing land fraction in cells with up to 99% land.

A map of the trend in the difference between tas and tos was calculated over the period 1979-2008 for every cell for which both values were present. Every pair of adjacent cells between 60°S and 60°N in the trend map were compared. For every pair of adjacent cells where both trend values were present and the land fraction in the two cells was different, the difference in trend and the difference in land fraction were calculated. Ordinary least squares regression was used to determine the contribution of increasing land fraction difference to increasing trend difference.

The data show an increase in tas-tos trend when moving from the cell with 0% land to a cell with 100% land (Figure S1). The coefficient of determination in the regression is small ($R^2 \sim 0.03$), suggesting that geographical variability is large compared to the coastal effect. The t-value of the prediction is large ($t \sim 35$); however it is likely to be overestimated due to spatial autocorrelation. The best indication of uncertainty in the regression coefficient therefore comes from repeating the experiment with different runs of the same climate model. The values of the difference in trend for a change from 0% to 100% land estimated by regression are $0.028°C$/decade and $0.029°C$/decade for the two runs of the HiRAM model.

These values are about 20% of the sea surface temperature trend for the study period. However, the 30 km cells used in the HiRAM model are large compared to typical distances between a coastal weather station and the sea. In practice a coastal weather station is likely to be characterized by a grid cell that is part ocean, so the actual land effect on the air temperature trend may be less than this. If the ratio of land air to sea surface temperature change is roughly constant over time the land surface air temperatures can simply be scaled to address the impact of the coastal effect.

The same calculation was repeated for a selection of CMIP5 historical simulations (described in Table S1) for which the appropriate fields were available. CMIP5 model runs typically use different grids for the land and ocean data, and so the sea surface temperatures were first transferred onto the surface air temperature grid using inverse distance weighting. Historical runs typically end in 2005, so the period 1976-2005 was used. The CMIP5 model grids are generally much coarser than the HiRAM grid (typically 100-200km), and so the air temperatures of high land fraction coastal cells will sample regions further inland than for the HiRAM model.

The trend and regression calculations were repeated for each model, with the results shown in Figure 2. There is significant variation between models, with the MIROC5 model showing a rather higher coastal effect than the HiRAM runs. Given that the coastal difference in air temperature trend moving from sea to land is non-negligible, the coastal weather stations will require adjustment before they are used to homogenize the sea surface temperature data. The coastal trend difference appears to increase roughly linearly with cell land fraction, and so a scaling should be applied to the weather station data that is

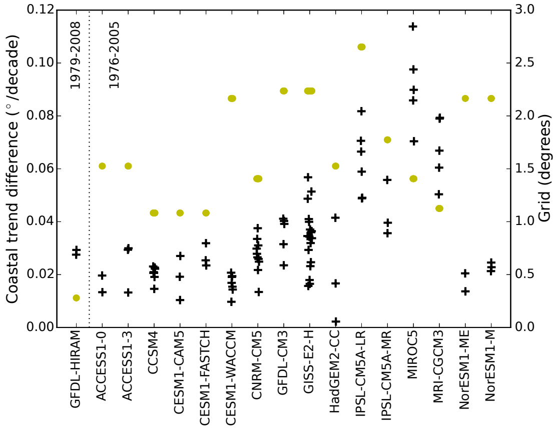

Figure 2: Coastal 30 year temperature trend differences for different climate models. Black crosses indicate the regression coefficient between the trend difference and the sea fraction between neighbouring cells with different land fractions for individual runs of a given model, calculated over the latitude range 60°S to 60°N. Spots indicate the average of the latitude and longitude dimensions of a grid cell for that model at the equator for that model, with the scale on the right hand axis.

linearly dependent on the land fraction around the weather station.

3 Coastal weather station record

A coastal weather station record was constructed using the GHCN-M v4 temperature data (Lawrimore *et al.*, 2011), which uses data from the International Surface Temperature Initiative (Rennie *et al.*, 2014) and includes data from 26,182 weather stations. The raw data were used in preference to the homogenized data, because (a) homogenization is expected to be of limited use for isolated island stations, and (b) homogenization may potentially increase coastal trends and reduce inland trends in order to bring them into agreement.

Information on station environment is not currently included in the GHCN-M version 4 data, and so coastal and island stations were identified using using a quarter degree global land mask from Jet Propulsion Laboratory (2013). Stations north of 60°N or south of 60°S were omitted to avoid the effects of sea ice, and stations in the Baltic and Mediterranean region were omitted since these may not reflect the global oceans. Stations were also omitted that lie more than 10km from the nearest coast according to metadata from Mosher (2017). Two station selections were used:

1. An island station list, consisting of 428 stations for which the average land fraction for the 8 cells surrounding the cell containing the weather station

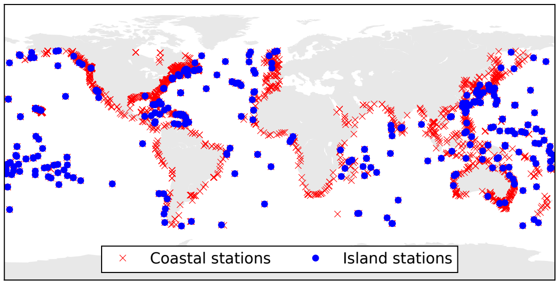

Figure 3: Map of coastal and island weather stations from GHCNv4 used in the construction of the coastal weather station record. Crosses show coastal stations, while dots show the subset of stations that are included on the island list on the basis of the land fraction in the surrounding cells.

was less than 10%. By chance all 428 stations fall in cells for which the land fraction is recorded as zero. The station list includes a few stations that are not on islands but have large exposure to the ocean, however it provides no coverage prior to the 1920s.

2. A coastal station list, consisting of 2386 stations for which the land fraction in the station cell was less than 50% or the land fraction in one of the four orthogonally adjacent cells was 0%. Some stations are available back to the start of the HadSST3 data in 1850. The coastal station list is a superset of the island station list.

The two station selections are shown in Figure 3, and the selection criteria are illustrated in Figure S2.

To address the different warming rates of coastal air and sea surface temperatures, the CLSAT temperature observations for each land weather station were scaled according to equation 1, in accordance with the climate model results.

$$T_{CLSAT,scaled} = T_{CLSAT,anom}(a - bl(\phi, \lambda)) \tag{1}$$

$T_{CLSAT,anom}$ is the original temperature anomaly, $T_{CLSAT,scaled}$ is the scaled anomaly, $l(\phi, \lambda)$ is the land fraction in the given grid cell and a and b are coefficients whose determination will be described later.

The station records for the selected stations are first aligned using the Climatic Anomaly Method (Jones, 1994), using a baseline period of 1961-1990 for consistency with HadSST3. For stations with at least 25 months of data present in the 30 year baseline period for a given month of the year, temperature anomalies were determined by subtracting a constant from all data for that month of the year to bring the mean on the baseline period to zero. If insufficient months of data were available, data were not used for that station for that month of the year. Data for 851 of the 2386 coastal stations were aligned in this way. A gridded temperature field was then calculated from the initial set of temperature anomalies, using a 5 × 5 degree grid.

A limitation of the climatic anomaly method is that stations or months cannot be used if insufficient data are available during the baseline period, reducing the number of available station records. Additional stations were therefore added iteratively by determining the offset required for each month of the year to fit the new station to the initial stations by the following method: The scaled station anomalies in each grid cell were averaged for each month of the record. The resulting sparse temperature field was extended to global coverage using kriging (Cressie, 1990) following the method of Cowtan and Way (2014). Anomalies were calculated for additional stations for which at least 15 months of data were available during the baseline period by fitting them to the temperature record for the appropriate grid cell, yielding 1328 aligned stations. A second global temperature field was determined from the expanded station list. In a third step, anomalies were calculated for further additional stations for which 15 months of data were available at any time between 1850 and the present by fitting them to the temperature record for the appropriate grid cell, yielding 2196 aligned stations. A spatially incomplete coastal temperature field was calculated from the resulting anomalies.

4 Coastal station homogenization of the sea surface temperature record

In addition to the corrected HadSST3 record, Kennedy *et al.* also distribute raw sea surface temperature fields with no adjustments for instrument type. The coastal weather station record was used to determine a time dependent (and optionally spatially dependent) correction to the raw sea surface temperature observations to bring them into agreement with the scaled coastal weather station record. The correction field T_{corr} is based on the difference field between the (sparse) coastal weather station field and the raw sea surface temperature field, given by equation 2.

$$T_{corr} = T_{CLSAT,scaled} - T_{CSST} \tag{2}$$

In order to ensure maximum coverage, the more complete sea surface temperature field was first infilled by kriging using the method of Cowtan and Way.

Both air-sea and coastal temperature differences can be influenced by weather (for example due to the greater heat capacity of the ocean), and so the differences between the coastal weather station and sea surface temperature anomaly fields show significant spatial and month-on-month variability. The correction to the raw sea surface temperatures must therefore be averaged both spatially, and over a moderate time window.

The HadSST3 corrections are spatially relatively uniform over most of the record, except for the periods where the sea surface temperatures come primarily from buckets, when there is a significant zonal variation in the bias arising from the varying air-sea temperature differential with latitude (Kent *et al.*, 2017). The primary component of the zonal variation is a contrast between the tropics and higher latitudes, however during some periods (such as the late 1940s) hemispheric differences are also apparent due to differences in the shipping fleets in different regions. This suggests that the correction field T_{corr} might be modelled by some combination of the zonally invariant spherical harmonics, Y_{00}, Y_{01} and Y_{02}, illustrated in Figure S3:

1. Y_{00} is a constant field. Fitting Y_{00} is equivalent to fitting the global mean of the correction field.

2. Y_{01} changes sign between the hemispheres, and so captures hemispheric differences.

3. Y_{02} changes sign between the equator and the poles, and so captures differences between the tropics and the higher latitudes.

In the early record, the available weather stations are clustered in developed regions with varying concentrations, and so a naive fitting method would overweight the regions with more observations. To address this issue, the spherical harmonics were fitted to the coastal difference map using generalised least squares (GLS), which includes information about the expected covariances of the observations in order to weight each observation according to the amount of independent information it provides. The covariance matrix of observations was constructed as an exponentially-declining function of distance in the same way as the variogram in Cowtan and Way, with an e-folding range of 800km determined empirically from the data over the period 1981-2010 when the coastal stations have good geographical coverage.

Three different models are fitted, the first using just Y_{00}; the second using Y_{00} and Y_{02}, and the third using Y_{00}, Y_{01} and Y_{02}. The coefficients for each spherical harmonic in each model are shown as a function of time in Figure S4. The Y_{00} (global mean) coefficient suggests a cool bias in the raw sea surface temperatures relative to the coastal air temperatures in the decades prior to World War 2, and to a lesser extent in the decade following the end of World War 2, consistent with previous analyses. This bias is apparent even without temporal smoothing of the coefficient. The Y_{01} and Y_{02} coefficients show rather greater monthly variability which is of a similar or greater amplitude to any persistent signal, and show very large excursions in the earliest decade of the record.

The coefficients were therefore smoothed using a 36 month linear lowess smooth with a cubic window (Cleveland, 1979), chosen to provide the most smoothing possible without distorting the World War 2 feature in the Y_{00} coefficient (Figure 4). The smoothed Y_{01} (hemispheric) coefficient still does not display a persistent signal, however the Y_{02} (equator-pole) coefficient tends to be negative in the periods dominated by canvas bucket observations (1880-1940 and 1945-1950) (Folland and Parker, 1995), and positive in the 21st century when buoy observations become dominant. Prior to 1880 the Y_{02} coefficient shows large excursions, arising from most of the available coastal temperature data being confined to the mid latitudes. However the weakness of the signal in

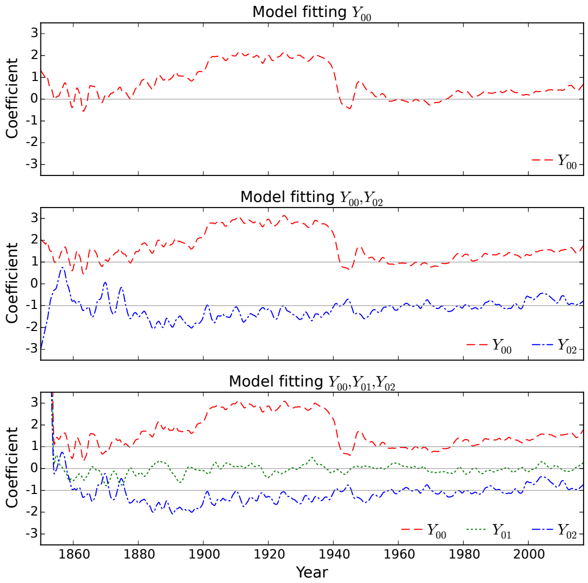

Figure 4: Smoothed coefficients of the spherical harmonics Y_{00}, Y_{01} and Y_{02} used in fitting the coastal temperature difference map for each month of the record, after application of a 36 month lowess smooth. Three different models are fitted, the first using just Y_{00}; the second using Y_{00} and Y_{02}, and the third using Y_{00}, Y_{01} and Y_{02}. Each panel shows a single model, with lines showing the values of each coefficient fitted in that model. The coefficients are offset and distinguished by line style and identified in the key.

the Y_{01} and Y_{02} terms even with smoothing suggests that the coastal temperature differences are only marginally informative with respect to the geographical components of the sea surface temperature bias.

Once the fit to the difference field has been determined, the spherical harmonics are then scaled by the fitted coefficients to determine a global correction field, which is then added to the raw HadSST3 field to produce a corrected sea surface temperature record. The corrected record is dependent on the values of a and b which scale the coastal temperature anomalies to account for the differential warming rates across the air-sea and coastal boundaries. Values for a and b are determined by assuming that the trend in the coastal temperature difference over the period 1976-2015 is dominated by the warming signal, justified by the rapid warming over this period and the comparatively limited metadata based corrections identified by Kennedy *et al.* (2011b). This also represents a long period of good spatial coverage where there is little difference between the bias corrections of the different observational records. The HadSST3 trend is therefore assumed to be correct over this period, and the coefficients a and b determined such that the global mean of the temperatures in the co-located corrected field yields the same trend. The island stations have $l(\phi, \lambda) = 0$ for all stations, and so can be used to determine a value for a, giving $a = 0.86$. A value for b is then determined such that the trend in the corrected record using the coastal station list also matches the HadSST3 trend, giving $b = 0.25$. The coefficients a and b do not vary significantly with the introduction of additional spherical harmonics to the regression. The value of b, however, is contingent on the land fraction assigned to each station, which for this study is the value of the quarter degree cell containing the station in the Jet Propulsion Laboratory (2013) land mask.

The temperature field resulting from adding the correction field (which has global coverage) to the raw HadSST3 temperature field for each month will be referred to as a *coastal hybrid sea surface temperature*, and has the same coverage as HadSST3. The mean sea surface temperature for each month was then calculated from the mean of the cells for which HadSST3 observations were available, weighting each grid cell according to the area of the cell occupied by ocean. The annual means using one, two or three spherical harmonics were then plotted for the whole period of the record (Figure 5).

The number of spherical harmonics makes essentially no difference to the resulting geographical means after 1900, and little difference between 1880 and 1900. However in the earliest decades, the inclusion of additional spherical harmonics increases the annual variability in the record. The remainder of this study will therefore focus primarily on the most parsimonious model where only the global mean of the coastal difference map (Y_{00}) is fitted; this will also allow the sensitivity of the results to different subsets of the coastal temperature record to be evaluated.

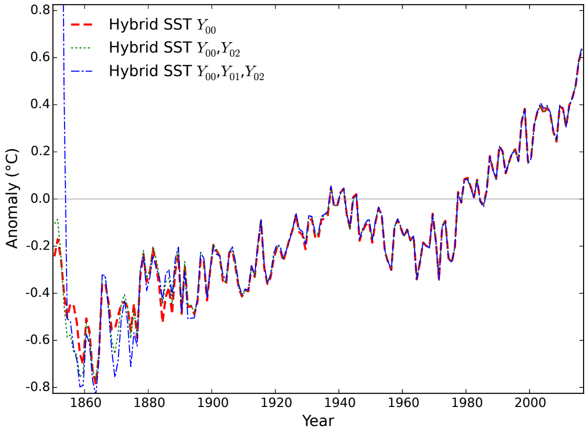

Figure 5: Coastal hybrid temperature reconstructions determined by fitting the coastal temperature difference map for each month of the record and using the resulting model to correct the sea surface temperature field. Three different models are fitted, the first using just Y_{00}; the second using Y_{00} and Y_{02}, and the third using Y_{00}, Y_{01} and Y_{02}. For this and all subsequent figures, temperature anomalies are calculated based on the coverage of HadSST3, with coastal cells weighted according to the fraction of cell occupied by ocean.

5 Results

Global marine temperature reconstructions were determined using the coastal hybrid method fitting a single global term to the coastal temperature difference field, and applying the 36 month lowess smooth to the resulting coefficients. Two temperature reconstructions were calculated as follows:

1. A reconstruction from HadSST3 using just the island stations.

2. A reconstruction from HadSST3 using the full list of coastal stations.

The resulting fields were masked to common coverage with the HadSST3 dataset before calculation of an area weighted monthly mean temperature series for each reconstruction. The island temperature series begins in 1920 due to limited island station coverage. Annual means were calculated from the monthly series, and compared to HadSST3 in Figure 6. Both of the coastal hybrid reconstructions show a cooler mid 20th century plateau than HadSST3. The coastal reconstruction rejects the coolness of the first two decades of the 20th century found in existing SST datasets and also suggests a cooler 19th century. The coastal reconstruction also weakly supports the presence of a cool bias in the HadSST3 dataset during the 2010s (Table S2) previously reported by Hausfather *et al.* (2016).

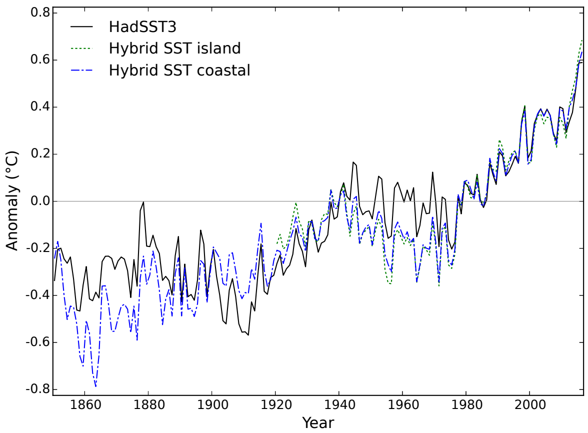

Figure 6: Comparison of two versions of the coastal hybrid temperature record to HadSST3. The two hybrid records use only island stations to correct HadSST3 over the period 1920-2016, or all coastal stations to correct HadSST3 over the period 1850-2016.

5.1 Sensitivity of the hybrid SSTs to the coastal temperature record

If the corrections to the sea surface temperature arise from global biases in the observational platforms and procedures, they should be detectable across the globe rather than arising from just one region. To test this the calculation was repeated omitting a hemisphere of data from the coastal difference field. The generalized least squares calculation reconstructs the missing hemisphere with the optimal average of the remaining hemisphere. The calculation was performed ten times, omitting the northern hemisphere, the southern hemisphere and eight hemispheres centered on points on the equator separated by 45 degrees of longitude. The resulting ensemble of 10 reconstructions is compared to HadSST3 in Figure 7. The ensemble members show cooler temperatures for most of the mid 20th century plateau, but are spread around HadSST3 in the 1930s. The ensemble members show warmer temperatures around 1910, and cooler temperatures in the mid 19th century. The ensemble is somewhat bimodal in the late 19th century, with some members much cooler than and others similar to HadSST3.

Global sea surface temperature reconstructions based on just the equatorial or mid latitude data show a somewhat greater contrast, with the mid-latitude data showing a cooler mid-century plateau than the equatorial data (which is still cooler than HadSST3). The bucket bias is greatest at the equator, and so correction using mid latitude data leads to a smaller correction and therefore cooler temperatures than HadSST3 in the 19th century, while the tropical data lead to a reconstruction that is similar to or slightly warmer than HadSST3 for

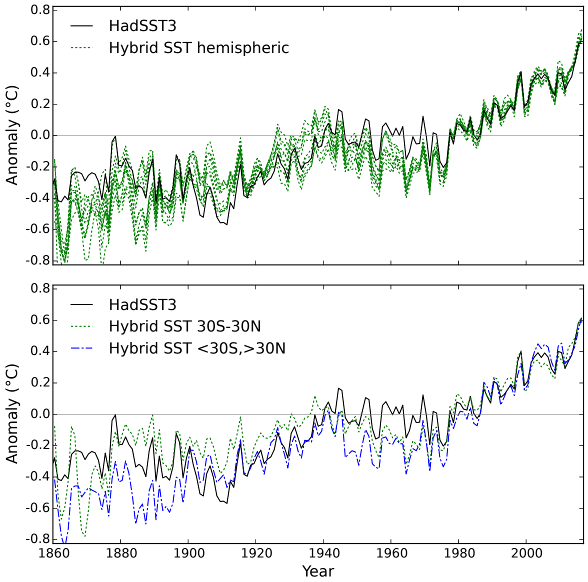

Figure 7: Coastal hybrid temperature reconstructions using different subsets of the coastal weather station record. The correction field is determined by fitting the Y_{00} coefficient to each of ten hemispheric subsets of the coastal difference field (top panel), or to just the equatorial or mid latitude cells of the coastal difference field (lower panel).

most of the early period. Prior to 1880, the tropical data are very sparse so the coastal hybrid record is likely to be cool biased due to the lower corrections from the mid-latitude stations.

The coastal hybrid temperature reconstruction is strongly determined by the coastal weather station record, which is in turn dependent on both the station selection (which has already been explored through the island-only record and the hemispheric and zonal subsets), and the scale terms a and b which account for the difference in warming rate between sea surface temperatures and weather stations with different degrees of exposure to the sea. Since only an ad-hoc estimate of the values of these parameters is available, the sensitivity of the resulting record to those values must be explored.

Reducing the parameter a (which controls the scaling of all weather stations relative to coastal sea surface temperatures) while holding b at zero (or more generally, scaling a and b together), reduces the amount of warming fairly uniformly across the whole record (Figure 8). Thus a misestimation of a could lead to a misestimation of the total amount of warming since the 19th century, but the resulting record would maintain its shape, still showing a cooler mid century plateau and no dip around 1910. Increasing the b parameter leads to reduced warming prior to World War 2 but has a rather smaller effect on late

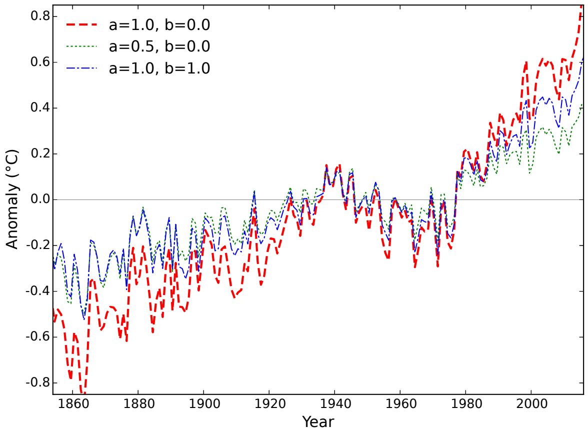

Figure 8: Comparison of hybrid temperature reconstructions using different values of the weather station scaling parameters a and b.

20th century warming. This behaviour arises from the sparsity of island stations in the early record, hence the b term which controls for the inland effect of less exposed stations plays a greater role.

The dependence of the coastal hybrid record on a novel temperature reconstruction using raw rather than homogenized temperature data must also be considered. Hybrid coastal temperature reconstructions were therefore determined using the existing CRUTEM version 4 and GHCN version 3 gridded temperature fields (Jones *et al.*, 2012; Lawrimore *et al.*, 2011), using a single scale factor in each case to preserve the trend in the resulting record on the period 1976-2015 (Figure S5). Using the CRUTEM data produces a coastal hybrid record that is broadly similar to that obtained using the custom coastal weather station record.

If the GHCN gridded data are used the resulting record shows significantly more warming prior to 1970. Part of this difference can be explained by the automated homogenization used in the GHCN record, because a hybrid reconstruction using the GHCN version 4 homogenized data also shows more early warming (Figure S6). The GHCN version 3 based record would imply an implausible sign change in the bucket bias in the early period; this is more likely to arise from the GHCN homogenization algorithm not accounting for the different rates of warming of coastal and inland stations. The remaining differences probably arise from the smaller weather station inventory for GHCN version 3 compared to GHCN version 4, and changes in the mix of coastal and non-coastal stations in the large 5 × 5 degree cell used by the GHCN gridded data.

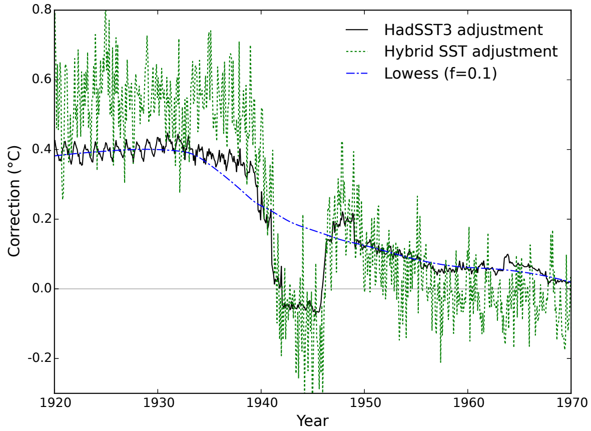

Figure 9: Comparison of the corrections applied to the raw sea surface temperature reconstruction by either the hybrid coastal method, or by the metadata-based HadSST3 method. The dashed line is a lowess smooth through the HadSST3 corrections, smoothed to emulate the smoothing used in the ERSSTv4 algorithm.

5.2 World War 2

The ERSST and HadSST3 records show a large discrepancy during World War 2, with ERSSTv4 and ERSSTv5 showing substantial warmth for most of the war period, while HadSST3 shows only a modest warm period spanning two to three years. To assess this period a coastal hybrid record was constructed without temporal smoothing. The resulting adjustments to the raw record are compared to the corresponding metadata-based HadSST3 adjustments in Figure 9.

Without the temporal smoothing term the adjustments from the coastal hybrid method show greater inter-monthly variability, however the shape of the adjustment matches the metadata-based HadSST3 adjustments well. The size of the adjustment suggested by the coastal hybrid method is larger than that for HadSST3, and falls outside the range of the 100 member HadSST3 ensemble (Kennedy *et al.*, 2011b). The similarity in shape provides a validation of both the metadata assignments of observation type in HadSST3, and the utility of the coastal hybrid method in detecting that bias.

The discrepancy in the size of the World War 2 bias between HadSST3 and the hybrid record could arise from non-uniformity in the zonal distribution of coastal observations, given the latitude dependence of the bucket bias. To test this possibility the World War 2 period was also examined in reconstructions based on hemispheric subsets of the coastal temperature data, or on the tropical or mid-latitude data alone (Figure 10). The use of a hemispheric or zonal subset of the coastal stations can lead to an estimate of the post-war bias that is larger

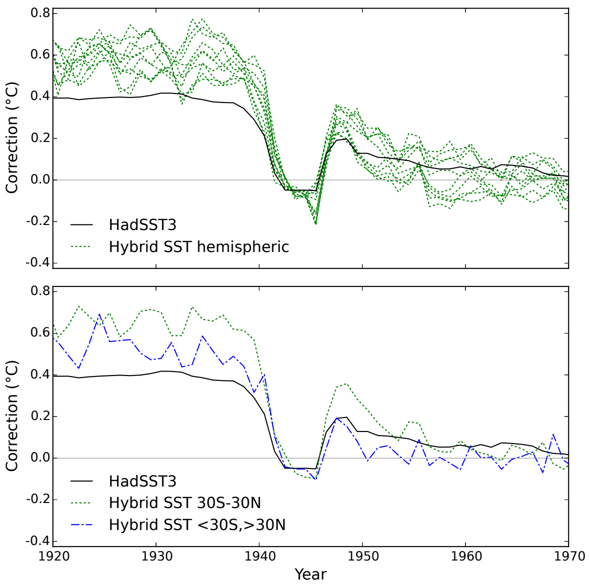

Figure 10: Coastal hybrid temperature corrections for the World War 2 period, using different subsets of the coastal weather station record. The correction field is determined by fitting the Y_{00} coefficient to each of ten hemispheric subsets of the coastal difference field (top panel), or to just the equatorial or mid latitude cells of the coastal difference field (lower panel).

or smaller than the HadSST3 estimate. The equatorial data lead to a larger estimate of the pre-war bias than the mid-latitude data, however in both cases the estimated bias is larger than in HadSST3.

The wartime warmth in the ERSSTv4 reconstruction arises from a failure to correct for the sharp changes in bias during this period. ERSSTv4 applies a lowess smooth to the difference between the SST and NMAT data to determine the bias correction using a window of 10% (i.e. about 200 months) of the data. The same smoothing operation applied to the HadSST3 adjustments is shown in Figure 9: the smoothed correction does not capture the World War 2 bias. Both the metadata adjustments of HadSST3 and the coastal hybrid method reject the World War 2 warmth in ERSSTv4, and the smoothing term provides a sufficient explanation for the bias.

ERSSTv5 attempts to address this issue by allowing for a discontinuity in the bias correction at the start of World War 2, however like ERSSTv4 it also shows a large warm excursion spanning the years of the war. The ERSST datasets are dependent on the HadNMAT2 data to provide an estimate of the SST bias. Application of the coastal hybrid calculation to HadNMAT2 suggests that the NMATs are also subject to a significant warm bias during the wartime period, which could arise from the same changes in the observing fleet that cause the

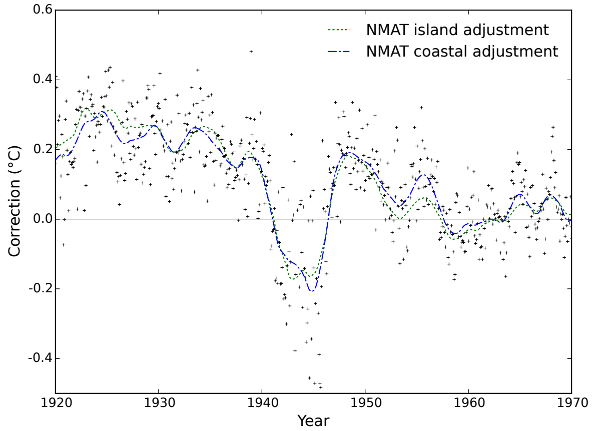

Figure 11: Comparison of the corrections applied to the HadNMAT2 night-time marine air temperature reconstruction by the hybrid coastal method, using either coastal or island stations. Lines show the 36 month lowess-smoothed correction, while crosses indicate the monthly values of the correction using the coastal station set.

SST bias (Figure 11). If this is the case then the NMAT data are not capable of addressing the wartime bias, even allowing for discontinuous changes in bias at both the start and end of the war.

6 Discussion

The homogenization of the sea surface temperature record is challenging, owing to a constantly changing fleet of mobile observation platforms and variability in the observation protocols. Both metadata and external temperature data sources have been used to homogenize the data by HadSST3 and ERSSTv4 respectively, with differing results. Coastal weather stations provide an alternative and independent check on those homogenizations, but are subject to uncertainties and biases due to the temperature differences across coastal boundaries as well as any uncorrected biases in the weather station observations.

This study presents a preliminary attempt at the use of coastal weather station records to correct inhomogeneities in the sea surface temperature record. The challenges of removing the climate signal from the coastal temperature differences are substantial, and so the results should be considered an indication of possible problems in existing series rather than a definitive temperature history. The new record suggests, in decreasing order of confidence, that:

1. The World War 2 warm spike in ERSSTv4 and ERSSTv5 is spurious. The coastal temperature data support the shape of the meta-data based correction of HadSST3, providing evidence for the wartime corrections. The

coastal temperature data suggest more tentatively that the size of the correction (due to a transition between bucket and engine room observations) is slightly underestimated in HadSST3.

2. The mid-century plateau spanning the 1940s to the 1970s is cooler in the coastal hybrid record than in HadSST3. This supports the cooler temperatures of the ERSST records over this period, although the details differ. The same result is obtained when using all of the coastal weather station data or spatially distinct subsets of the data.

3. The larger estimate of the size of the bucket correction in the coastal hybrid record leads to a greater upward correction of pre-World War 2 temperatures, leading to warmer temperatures since 1900 and an earlier start to the mid-century plateau. The large dip in temperatures around 1910 in existing records is largely eliminated in the coastal hybrid record.

4. The rate of warming in HadSST3 since 1998 is likely to be underestimated, consistent with previous work showing less warming in ship observations over that period than in more reliable buoy measurements (Kennedy *et al.*, 2011b; Karl *et al.*, 2015; Hausfather *et al.*, 2017).

5. The coastal hybrid record is also cooler than existing records between 1880 and 1900, however this result is contingent on the station selection, with some subsets of the data yielding temperatures similar to HadSST3.

6. The coastal hybrid record shows cooler temperatures between 1850 and 1880 than the existing SST records. However coastal weather station coverage in the tropics is poor and island station coverage non-existent during this period.

The sparsity of data in the tropics in the earliest part of the record presents a problem in estimating the bias in the sea surface temperature observations due to the zonal dependence of the air-sea temperature difference. When the Y_{02} coefficient is included in the model, the resulting temperature record only shows significantly different behaviour prior to about 1880 (Figure 5). While the coastal hybrid method is likely to have a cool bias at the start of the record, the agreement of the different spherical harmonic models after 1880 point is consistent with the cool bias being confined to the period prior to that date.

The coastal hybrid record is compared to co-located data from both HadSST3 and ERSSTv5 in Figure 12, and shows significant differences with both. The existing records disagree over the warmth of the mid 20th century plateau with ERSSTv5 being cooler than HadSST3, however the hybrid record is cooler than both. The hybrid record rejects the warm spike in ERSSTv5 during World War 2. The hybrid record is broadly consistent with HadSST3 between 1915 and 1935, however it rejects the unexplained cool period between 1900 and 1915 in the existing records. Prior to 1900 HadSST3 is generally cooler than ERSSTv5, however the hybrid record is cooler than both.

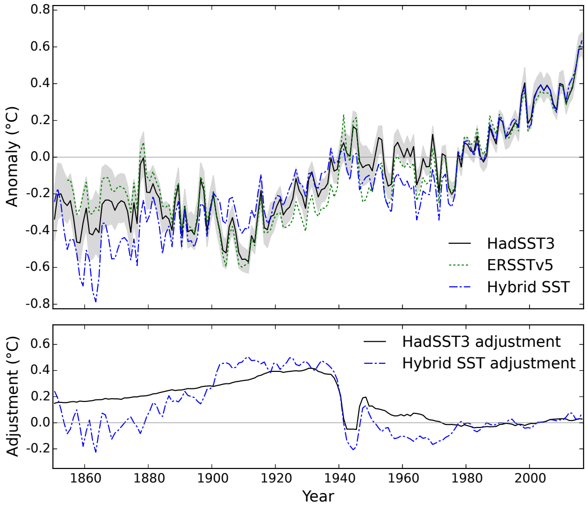

Figure 12: Comparison of the coastal hybrid temperature reconstruction (using all coastal stations and fitting the global mean of the coastal temperature differences only) to co-located data from HadSST3 and ERSSTv5 for the period 1850-2016. Spatial coverage is that of HadSST3 for all of the records, with coastal cells weighted by ocean fraction.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. A comparison with the ERSSTv4 ensemble is shown in Figure S7

The late 19th century and early 20th century periods are of particular interest, with the coastal hybrid record showing a gradual warming that is more consistent with climate model simulations than the existing records. The bucket bias is estimated by Folland and Parker (1995) to increase linearly from 1850 to 1920, however the coastal hybrid suggests a bias that remains small until around 1890 and then increases rapidly until 1910 (Figure 12, lower panel), similar to the results of Jones *et al.* (1991) after 1880.

The differences between the coastal hybrid and existing sea surface temperature reconstructions are not necessarily indicative of problems in the existing records, although divergence between the existing records means that they cannot all be correct. The coastal record may be more realistic if the coastal weather station record is reliable and if the relationship between coastal air temperature and offshore sea surface temperature is correctly modelled.

Possible problems with the coastal temperature record include changing weather station coverage and the use of raw rather than homogenized temperature data. For the period after 1920, the similarity of the hybrid record when using the more strict island station selection provides addition support for the results, but does not address the earlier period. Use of homogenized data in the preparation of the coastal hybrid record leads to much greater warming in the 19th century, however this is unlikely to be correct because it would require a change in the sign of the bucket bias. It is more likely that homogenization exaggerates the trend for coastal stations.

The differences between the coastal hybrid record and HadSST3 could arise from changes in the air-sea temperature difference, inshore temperature difference or coastal temperature difference which are not accounted for by the simple scaling scheme of equation 1. The inshore temperature difference may be partially captured in the HadSST3 record due to the presence of vessels traversing coastal waters, however the large 5 × 5 degree grid cells may offset this. Uncertainties in the scaling of the coastal weather station data relative to sea surface temperatures and changes in coverage will affect the evaluation of long term changes in sea surface temperature bias, but are less likely to explain rapid changes like those around World War 2, in the 1970s, or between 1890 and 1910.

It is notable that there are large changes in difference between HadSST3 and the coastal hybrid reconstruction in the 1940s and the late 1970s, corresponding roughly to changes in the sign of the Pacific Decadal Oscillation (PDO). While the corresponding wind changes may affect inshore or coastal temperature differences, the coastal corrections are largely conserved between hemispheres so cannot be driven by Pacific variability alone. Furthermore the ERSST records also show a somewhat cooler mid-century plateau, suggesting that the PDO on its own cannot explain all of the differences between the coastal hybrid and HadSST3 records.

Given the inherent uncertainties it would be premature to adopt the coastal hybrid record as a historical record of sea surface temperature. The limited spatial resolution of the correction limits the utility of the record for estimating temperatures at a sub-global scale, and the changing station coverage in the 19th century certainly biases the record prior to 1880. Metadata-based analyses like that of HadSST3 still provide the best tools for evaluating regional sea surface temperature variation, however it is possible that the approach presented here may provide a useful tool in improving the parameterisation of the metadata-based corrections.

If the coastal hybrid record were correct, there would be implications both for the estimation of climate sensitivity and for the assessment of multidecadal internal variability from the historical temperature record. Estimates of climate sensitivity that rely on a 19th century temperature baseline (Otto *et al.*, 2013; Richardson *et al.*, 2016) would be too low due to the warm bias in the early sea surface temperature record. Differences between temperature observations and the mean of an ensemble of climate model simulations are often attributed to internal variability in the real climate system, because internal variability is expected to cancel out when averaging multiple simulations. Observation-model differences typically show a peak in the late 19th century, and a dip in the early 20th century (Mann *et al.*, 2016). Both of these are reduced if the coastal hybrid record is used in place of existing records, which might suggest a reduced role for multidecadal internal variability in the observed temperature record.

The consequences for the climate sensitivity and internal variability highlight the importance of possible inhomogeneities in the sea surface temperature record. The differences between existing sea surface temperature reconstructions demonstrate that there is a problem to be addressed. The coastal hybrid sea surface temperature reconstruction cannot solve this problem outright because the results are contingent on correctly combining inhomogeneous observations across coastal boundaries; however the method does bring an additional source of observational data to help assess the biases in the sea surface temperature record.

Data and methods for this paper are available at doi://10.15124/6ba8c951-d7d6-40e9-8175-4e40e7c320 with updates at http://www-users.york.ac.uk/~kdc3/papers/evaluating2017.

Acknowledgements

We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups (listed in Table S1) for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals.

We are grateful to S. Mosher for providing metadata for the GHCNv4 station inventory, and K. Haustein and E. Kent for helpful discussion.

References

Cleveland W. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**(368): 829–836, doi:10.1080/01621459.1979.10481038.

Cowtan K, Hausfather Z, Hawkins E, Jacobs P, Mann M, Miller S, Steinman B, Stolpe M, Way R. 2015. Robust comparison of climate models with observations using blended land air and ocean sea surface temperatures. *Geophysical Research Letters* **42**(15): 6526–6534, doi:10.1002/2015GL064888.

Cowtan K, Way R. 2014. Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. *Quarterly Journal of the Royal Meteorological Society* **140**(683): 1935–1944, doi:10.1002/qj.2297.

Cressie N. 1990. The origins of kriging. *Mathematical geology* **22**(3): 239–252, doi:10.1007/BF00889887.

Folland C. 2005. Assessing bias corrections in historical sea surface temperature using a climate model. *International Journal of Climatology* **25**(7): 895–911, doi:10.1002/joc.1171.

Folland C, Parker D. 1995. Correction of instrumental biases in historical sea surface temperature data. *Quarterly Journal of the Royal Meteorological Society* **121**(522): 319–367, doi:10.1002/qj.49712152206.

Harris L, Lin SJ, Tu C. 2016. High-resolution climate simulations using GFDL HiRAM with a stretched global grid. *Journal of Climate* **29**(11): 4293–4314, doi:10.1175/JCLI-D-15-0389.1.

Hausfather Z, Cowtan K, Clarke DC, Jacobs P, Richardson M, Rohde R. 2017. Assessing recent warming using instrumentally homogeneous sea surface temperature records. *Science Advances* **3**(1): e1601 207, doi:10.1126/sciadv.1601207.

Hausfather Z, Cowtan K, Menne M, Williams CN J. 2016. Evaluating the impact of U.S. historical climatology network homogenization using the U.S. climate reference network. *Geophysical Research Letters* **43**(4): 1695–1701, doi:10.1002/2015GL067640.

Hirahara S, Ishii M, Fukuda Y. 2014. Centennial-scale sea surface temperature analysis and its uncertainty. *Journal of Climate* **27**(1): 57–75, doi:10.1175/JCLI-D-12-00837.1.

Huang B, Banzon V, Freeman E, Lawrimore J, Liu W, Peterson T, Smith T, Thorne P, Woodruff S, Zhang HM. 2015. Extended reconstructed sea surface temperature version 4 (ERSST.v4). part i: Upgrades and intercomparisons. *Journal of Climate* **28**(3): 911–930, doi:10.1175/JCLI-D-14-00006.1.

Huang B, Thorne P, Banzon V, Boyer T, Chepurin G, Lawrimore J, Menne M, Smith T, Vose R, Zhang HM. 2017. Extended reconstructed sea surface temperature, version 5 (ersstv5): Upgrades, validations, and intercomparisons. *Journal of Climate* **30**(20): 8179–8205, doi:10.1175/JCLI-D-16-0836.1.

Jet Propulsion Laboratory. 2013. ISLSCP II land and water masks with ancillary data. http://daac.ornl.gov/, doi:10.3334/ORNLDAAC/1200. Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA.

Jones P. 1994. Hemispheric surface air temperature variations: a reanalysis and an update to 1993. *Journal of Climate* **7**(11): 1794–1802, doi:10.1175/1520-0442(1994)007⟨1794:HSATVA⟩2.0.CO;2.

Jones P, Lister D, Osborn T, Harpham C, Salmon M, Morice C. 2012. Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010. *Journal of Geophysical Research Atmospheres* **117**(5), doi:10.1029/2011JD017139.

Jones P, Wigley T, Farmer G. 1991. Marine and land temperature data sets: a comparison and a look at recent trends. In: *Greenhouse-gas-induced climatic change*, Schlesinger M (ed). Elsevier, pp. 153–172.

Karl T, Arguez A, Huang B, Lawrimore J, McMahon J, Menne M, Peterson T, Vose R, Zhang HM. 2015. Possible artifacts of data biases in the recent global surface warming hiatus. *Science* **348**(6242): 1469–1472, doi:10.1126/science.aaa5632.

Kennedy J, Rayner N, Smith R, Parker D, Saunby M. 2011a. Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 1. measurement and sampling uncertainties. *Journal of Geophysical Research Atmospheres* **116**(14), doi:10.1029/2010JD015218.

Kennedy J, Rayner N, Smith R, Parker D, Saunby M. 2011b. Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. measurement and sampling uncertainties. *Journal of Geophysical Research Atmospheres* **116**(14), doi:10.1029/2010JD015220.

Kent E, Kennedy J, Berry D, Smith R. 2010. Effects of instrumentation changes on sea surface temperature measured in situ. *Wiley Interdisciplinary Reviews: Climate Change* **1**(5): 718–728, doi:10.1002/wcc.55.

Kent E, Rayner N, Berry D, Saunby M, Moat B, Kennedy J, Parker D. 2013. Global analysis of night marine air temperature and its uncertainty since 1880: The HadNMAT2 data set. *Journal of Geophysical Research Atmospheres* **118**(3): 1281–1298, doi:10.1002/jgrd.50152.

Kent EC, Berry DI, Carella G, Kennedy JJ, Parker DE, Atkinson CP, Rayner NA, Smith TM, Hirahara S, Huang B, *et al.* 2017. A call for new approaches to quantifying biases in observations of sea-surface temperature. *Bulletin of the American Meteorological Society* **98**(8): 1601–1616, doi:10.1175/BAMS-D-15-00251.1, URL http://dx.doi.org/10.1175/BAMS-D-15-00251.1.

Lawrimore J, Menne M, Gleason B, Williams C, Wuertz D, Vose R, Rennie J. 2011. An overview of the global historical climatology network monthly mean temperature data set, version 3. *Journal of Geophysical Research Atmospheres* **116**(19), doi:10.1029/2011JD016187.

Mann M, Rahmstorf S, Steinman B, Tingley M, Miller S. 2016. The likelihood of recent record warmth. *Scientific Reports* **6**, doi:10.1038/srep19831.

Menne M, Williams Jr C. 2009. Homogenization of temperature series via pairwise comparisons. *Journal of Climate* **22**(7): 1700–1717, doi:10.1175/2008JCLI2263.1.

Mosher S. 2017. Station metadata for GHCN version 4. Personal communication.

Otto A, Otto F, Boucher O, Church J, Hegerl G, Forster P, Gillett N, Gregory J, Johnson G, Knutti R, Lewis N, Lohmann U, Marotzke J, Myhre G, Shindell D, Stevens B, Allen M. 2013. Energy budget constraints on climate response. *Nature Geoscience* **6**(6): 415–416, doi:10.1038/ngeo1836.

Parker D, Folland C, Jackson M. 1995. Marine surface temperature: Observed variations and data requirements. *Climatic Change* **31**(2-4): 559–600, doi:10.1007/BF01095162.

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

Rennie J, Lawrimore J, Gleason B, Thorne P, Morice C, Menne M, Williams C, Almeida WG, Christy J, Flannery M, *et al.* 2014. The international surface temperature initiative global land surface databank: Monthly temperature data release description and methods. *Geoscience Data Journal* **1**(2): 75–102, doi:10.1002/gdj3.8.

Richardson M, Cowtan K, Hawkins E, Stolpe M. 2016. Reconciled climate response estimates from climate models and the energy budget of earth. *Nature Climate Change* **6**(10): 931–935, doi:10.1038/nclimate3066.

Taylor K, Stouffer R, Meehl G. 2012. An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society* **93**(4): 485–498, doi:10.1175/BAMS-D-11-00094.1.