A review of uncertainty in in situ measurements and data sets of sea-surface temperature

John J. Kennedy, Met Office Hadley Centre, FitzRoy Road, Exeter, EX1 3PB, UK
(john.kennedy@metoffice.gov.uk)

Abstract

Archives of in situ sea-surface temperature (SST) measurements extend back more than 160 years. Quality of the measurements is variable and the area of the oceans they sample is limited, especially early in the record and during the two World Wars. Measurements of SST and the gridded data sets that are based on them are used in many applications so understanding and estimating the uncertainties are vital. The aim of this review is to give an overview of the various components that contribute to the overall uncertainty of SST measurements made in situ and of the data sets that are derived from them. In doing so, it also aims to identify current gaps in understanding. Uncertainties arise at the level of individual measurements with both systematic and random effects and, although these have been extensively studied, refinement of the error models continues. Recent improvements have been made in the understanding of the pervasive systematic errors that affect the assessment of long-term trends and variability. However, the adjustments applied to minimize these systematic errors are uncertain and these uncertainties are higher before the 1970s and particularly large in the period surrounding the Second World War owing to a lack of reliable metadata. The uncertainties associated with the choice of statistical methods used to create globally complete SST data sets have been explored using different analysis techniques but they do not incorporate the latest understanding of measurement errors and they want for a fair benchmark against which their skill can be objectively assessed. These problems can be addressed by the creation of new end-to-end SST analyses and by the recovery and digitization of data and metadata from ship log books and other contemporary literature.

1. Introduction

Measurements of the temperature of the sea surface have been made for more than 200 years for a wide variety of purposes. The earliest measurements of sea-surface temperature (SST) in the eighteenth century were taken out of pure scientific interest. Later, after the connection between SST and ocean currents was made, large numbers of measurements were made for the construction of navigational charts. In the twentieth century, the needs of weather forecasting and, to an extent, the need to produce marine climate summaries determined the quantity and quality of observations. Most historical SST measurements were not made by dedicated scientific vessels, but by voluntary observing ships (VOS) on the basis that they would contribute to the safety of life at sea. This is reflected in the geographical distribution of observations, which are largely confined to major shipping lanes.

Nowadays, in situ measurements of SST – those made at the surface as opposed to those made remotely by satellites or aircraft – are used in diverse applications. They are used directly in calibration and validation of satellite retrievals and they are assimilated into ocean analyses [*Roberts-Jones et al.*, 2012]. They are also used to construct data sets of summaries of SST on regular grids and globally-complete SST fields are created using statistical techniques to impute SSTs in regions where there are no observations. The SST data sets and statistical SST ‘reconstructions’ or ‘analyses’ are widely used, for example as an index of global climate change [*Morice et al*., 2012], as a boundary condition for climate simulations [*Folland*, 2005] and reanalyses [*Simmons et al.*, 2010], as initial conditions for decadal forecasts [*Smith et al.*, 2007], in studies of hurricane formation [*Saunders and Harris*, 1997] and in studies of the impact of climate change on marine ecosystems [*Sheppard and Rayner*, 2002].

As the demands for SST measurements have changed, so have the instruments used to make them, and so have the ships and other vessels from which the measurements were made. The first systematic observations were made using buckets to collect a water sample. Buckets made of wood, canvas, tin, leather, brass, rubber and plastic – of designs as various as the materials employed in their construction – have all been used to measure the temperature of the surface layers of the ocean. There are two problems with this approach. The first is that during the collection and hauling, the temperature of the water sample can be modified by the combined actions of latent and sensible heat transfer and the warmth of the Sun. Even in the best conditions, an accurate measurement requires diligence on the part of the sailor; that is the second problem. Improvements to minimize the physical effects were made to bucket designs during the 1950s, but as ships became larger and faster, the making of the measurements became not just thankless, but dangerous.

After the advent of steam ships in the late nineteenth century, it was routine to measure the temperature of the sea water that was circulated through the steam condenser. Condenser inlet measurements and later, engine room inlet (ERI) measurements, were often recorded in ship logbooks, but they were not entered into meteorological logs until the 1930s. The convenience of using measurements that were made as a matter of routine, and the attendant reduction in the risk of losing a bucket or sailor overboard, meant that ERI measurements became the preferred method for measuring SST on board ships during the latter half of the twentieth century. That is not to say that the method was without its difficulties. Modification of the temperature of the water between inlet and thermometer was still a problem and it was now compounded by the varying depth of the measurements.

Since the 1970s, a growing number of ships have been fitted with dedicated sensors either outside or inside the hull. These have been joined by a growing array of moored and drifting buoys which make automated measurements that are relayed by satellite. At present, around 90% of all SST observations come from buoys. In calm conditions drifting buoys measure at a nominal depth of between 10 and 20 cm depending on their size. However, wave motion means that in some conditions the buoy will be submerged for part of the time and report temperatures that are representative of something like the upper 2 m.

Moored buoys are fixed platforms, akin, in some ways, to meteorological stations on land. They come in a variety of shapes and sizes. Most are a few meters in height and width, but the largest in regular use are the 12 m Discus buoys designed to weather the wilder climates of the northern oceans. There are two loose groupings of moored buoys: the Global Tropical Moored Buoy Array (GTMBA) and a more diverse group of coastal moorings mostly around the US. The GTMBA has regular arrays of moorings in the tropical Pacific, Atlantic and Indian Oceans. The majority of moored buoys measure SST at a nominal depth of 1 m. Some measure slightly deeper and some moorings make measurements at a range of depths.

SST measurements from ships and buoys together with near-surface measurements made by oceanographic cruises have been gathered in digital archives. The largest and most comprehensive of these is the International Comprehensive Ocean-Atmosphere Data Set (ICOADS, *Woodruff et al.* [2011]). The latest release of ICOADS, release 2.5, contains individual marine reports from 1662 to 2007, but air and sea temperature measurements only start to appear in the 19th Century. Metadata giving information about some of the measurements and the ships that make them is also provided and is now complemented by information from regular bulletins such as WMO publication 47 (http://www.wmo.int/pages/prog/www/ois/pub47/pub47-home.htm).

Other digital archives exist. Research vessel (RV http://coaps.fsu.edu/RVSMDC/index.shtml) data are gathered at the Research Vessel Surface Meteorology Data Center at Florida State University. Woods Hole Oceanographic Institute (http://www.whoi.edu/) maintains an archive of research mooring data and the OceanSites website (http://www.oceansites.org/data/index.html) provides links to other mooring data. The Pacific Marine Environmental Laboratory maintains an archive of water temperature measurements from the GTMBA at a range of depths and time resolutions that are not available in ICOADS (http://www.pmel.noaa.gov.tao/global/global.html). Near-surface measurements from other sub-surface sources such as the Argo array of autonomous profiling floats also exist.

Despite being comprehensive, ICOADS is incomplete. Large archives of paper records exist around the world and many of these have yet to be digitized. It is not possible yet to know exactly how many undigitized records remain because there is no definitive catalogue of global archives. What is known is that many archives that have been identified are far from being exhausted. The potential for reducing the uncertainty in SST analyses as well as in reconstructions of other marine variables is clear, but funding, particularly sustained funding for the efforts to identify, image and key the data has proved difficult to find. Nonetheless, there have been some successes such as a project to crowd source the keying of Royal Navy logbooks from the First World War. Volunteers on the OldWeather.org project keyed pages from the logbooks online. In the three years since the project started more than 1.6 million weather observations have been digitized, by around 16,400 volunteers.

The observing network was not created with a single purpose in mind. It was certainly not intended to meet the stringent criteria demanded for monitoring long-term environmental change. Nonetheless, historical SST measurements have been widely used in such studies. In a 2010 paper, *Jones and Wigley* [2010] identified uncertainties associated with pervasive systematic errors in SST data sets as an important uncertainty in the estimation of global temperature trends. The obvious gulf between the ideal and the reality leads naturally to questions about the reliability of the SST record. Often this question is couched as a yes/no dichotomy: “are SST records reliable?” A more useful question is "How reliable are they?" Although historical measurements were not made for climate research, or any single purpose, it does not mean that it is impossible to derive from them a record that is useful to a particular end. However, it does mean that special care must be taken in identifying and, as best as possible, quantifying uncertainties.

In using SST observations and the analyses that are based on them, it is important to understand the uncertainties inherent in them and the assumptions and statistical methods that have gone into their creation. In this review I aim to give an overview of the various components that contribute to the overall uncertainty of SST measurements made in situ and of the data sets that are derived from them. In doing so, I also aim to identify current gaps in understanding.

Section 2 provides a classification of uncertainties. The classifications are not definitive, nor are they completely distinct. They do, however, reflect the way in which uncertainties have been approached in the literature and provide a useful framework for thinking about the uncertainties in SST data sets. The uncertainties have been tackled in ascending order of abstraction from the random errors associated with individual observations to the generic problem of unknown unknowns. In this review quoted uncertainties represent one standard deviation of the relevant distribution unless otherwise stated. Section 3 applies this framework to analyze progress and understanding under each of the headings. Some shortcomings of the presentation of uncertainties are discussed in section 4 along with possible solutions. Section 5 reviews how some analyses have used knowledge of likely errors in SST data sets to minimize their exposure to uncertainty. Section 6 briefly discusses SST retrievals from satellites and how these have been used to understand the in situ record. The review concludes with a summary of possible future directions.

2. General Classification of Uncertainties

Throughout this review the distinction will be made between an *error* and an *uncertainty*. The distinction between the two loosely follows the usage in the Guide to the Expression of Uncertainty in Measurement (GUM) [*BIPM*, 2008]. The *error* in a measurement is the difference between some idealized "true value" and the measured value and is unknowable. The GUM defines the uncertainty of a measurement as the "parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand". This is the sense in which uncertainty is generally meant in the following discussion. This is not necessarily the same usage as is found in the cited papers. It is common to see the word error used as a synonym for uncertainty such as in the commonly used phrases standard error and analysis error.

Broadly speaking, errors in individual SST observations have been split into two groupings: random observational errors and systematic observational errors. Although this is a convenient way to deal with the uncertainties, errors in SST measurements will generally share a little of the characteristics of each.

Random observational errors occur for many reasons: misreading of the thermometer, rounding errors, the difficulty of reading the thermometer to a precision higher than the smallest marked gradation, incorrectly recorded values, errors in transcription from written to digital sources and sensor noise among others. Although they might confound a single measurement, the independence of the individual errors means they tend to cancel out when large numbers are averaged together. Therefore, the contribution of random independent errors to the uncertainty on the global average SST is much smaller than the contribution of random error to the uncertainty on a single observation even in the most sparsely observed years. Nonetheless, where observations are few, random observational errors can be an important component of the total uncertainty.

Systematic observational errors are much more problematic because their effects become relatively more pronounced as greater numbers of observations are aggregated. Systematic errors might occur because a particular thermometer is mis-calibrated, or poorly sited. No amount of averaging of observations from a thermometer that is mis-calibrated such that it reads 1 K too high will reduce the error in the aggregate below this level save by chance. However, in many cases the systematic error will depend on the particular environment of the thermometer and will therefore be independent from ship to ship. In this case, averaging together observations from many different ships or buoys will tend to reduce the contribution of systematic observational errors to the uncertainty of the average.

In the 19th and early 20th century, the majority of observations were made using buckets to haul a sample of water up to the deck for measurement. Although buckets were not always of a standard shape or size, they had a general tendency under typical environmental conditions to lose heat *via* evaporation or directly to the air when the air-sea temperature difference was large. *Folland and Parker* [1995] provide a more comprehensive survey of the problem which was already well known in the early 20th Century (see, for example, the introduction to *Brooks* [1926]). *Pervasive systematic observational errors* like the cold bucket bias are particularly pertinent for climate studies because the errors affect the whole observational system and change over time as observing technologies and practices change. The change can be gradual as old methods are slowly phased out, but they can also be abrupt, reflecting significant geopolitical events such as the Second World War [*Thompson et al.*, 2008]. Rapid changes also arise because the digital archives of marine meteorological reports (ICOADS *Woodruff et al.* [2011]) are themselves discontinuous.

Generally, systematic errors are dealt with by making adjustments based on knowledge of the systematic effects. The adjustments are uncertain because the variables that determine the size of the systematic error are imperfectly known. The atmospheric conditions at the point where the measurement was made, the method used to make the measurement – ERI or bucket – the material used in the construction of the bucket if one was used, as well as the general diligence of the sailors making the observations have not in many cases been reliably recorded. Part of the uncertainty can be estimated by allowing uncertain parameters and inputs to the adjustment algorithms to be varied within their plausible ranges thus generating a range of adjustments (e.g., *Kennedy et al.* [2011c]). This *parametric uncertainty* gives an idea of the uncertainties associated with poorly determined parameters within a particular approach, but it does not address the more general uncertainty arising from the underlying assumptions. This uncertainty will be dealt with later as *structural uncertainty*.

First, however, there are a number of other uncertainties associated with the creation of the gridded data sets and SST analyses that are commonly used as a convenient alternative to dealing with individual marine observations. The uncertainties are closely related because they arise in the estimation of area-averages from a finite number of noisy and often sparsely-distributed observations.

In *Kennedy et al.*, [2011b] two forms of this uncertainty were considered: *grid-box sampling uncertainty* and *large-scale sampling uncertainty* (which they referred to as coverage uncertainty). Grid-box sampling uncertainty refers to the uncertainty accruing from the estimation of an area-average SST anomaly within a grid box from a finite, and often small, number of observations. Large-scale sampling uncertainty refers to the uncertainty arising from estimating an area-average for a larger area that encompasses many grid boxes that do not contain observations. Although these two uncertainties are closely related, it is often easier to estimate the grid-box sampling uncertainty, where one is dealing with variability within a grid box, than the large-scale sampling uncertainty, where one must take into consideration the rich spectrum of variability at a global scale.

Although some gridded SST data sets contain many grid boxes which are not assigned an SST value because they contain no measurements, other SST data sets – oftentimes referred to as SST analyses – use a variety of techniques to fill the gaps. They use information gleaned from data-rich periods to estimate the parameters of statistical models that are then used to estimate SSTs in the data voids, often by interpolation or pattern fitting. There are many ways to tackle this problem and all are necessarily approximations to the truth. The correctness of the *analysis uncertainty* estimates derived from these statistical methods are conditional upon the correctness of the methods, inputs and assumptions used to derive them. No method is correct therefore analytic uncertainties based on a particular method will not give a definitive estimate of the true uncertainty. To gain an appreciation of the full uncertainty it is necessary to factor in the lack of knowledge about the correct methods to use, which brings the discussion back to structural uncertainty.

There are many scientifically defensible ways to produce a data set. For example, one might choose to fill gaps in the data by projecting a set of Empirical Orthogonal Functions (EOFs) onto the available data. Alternatively, one might opt to fill the data using simple optimal interpolation. Both are defensible approaches to the problem, but each will give different results. In the process of creating any data set, many such choices are made. *Structural uncertainty* [*Thorne et al.*, 2005] is the term used to understand the spread that arises from the many choices and foundational assumptions that can be (and have to be) made during data set creation. The character of structural uncertainty is somewhat different to the other uncertainties considered so far. The uncertainty associated with a measurement error, for example, assumes that there is some underlying distribution that characterizes the dispersion of the measured values. In contrast, there is generally no underlying "distribution of methods" that can be used to quantify the structural uncertainty. Furthermore, the diverse approaches taken by different teams might reflect genuine scientific differences about the nature of the problems to be tackled. Consequently, structural uncertainty is one of the more difficult uncertainties to quantify or explore efficiently. It requires multiple, independent attempts to resolve the same difficulties, it is an ongoing commitment, and it does not guarantee that the true value will be encompassed by those independent estimates. Nevertheless, the role that the creation of multiple independent estimates and their comparison has played in uncovering, resolving, and quantifying some of the more mystifying uncertainties in climate analyses is unquestionable. The most obvious – one might say, notorious – examples are those of tropospheric temperature records made using satellites and radiosondes [*Thorne et al.*, 2011] and sub-surface ocean temperature analyses [*Lyman et al.*, 2010; *Abraham et al.,* 2013].

Which leads finally to *unknown unknowns*. On February 12th 2002, at a news briefing at the US Department of Defense, Donald Rumsfeld memorably divided the world of knowledge into three quarters:

"There are known knowns. These are things we know we know. We also know there are known unknowns. That is to say, we know there are some things we do not know. But there are also unknown unknowns, the ones we don't know we don't know."

In the context of SST uncertainty, unknown unknowns are those things that have been overlooked. By their nature, unknown unknowns are unquantifiable; they represent the deeper uncertainties that beset all scientific endeavors. By deep, I do not mean to imply that they are necessarily large. In this review I hope to show that the scope for revolutions in our understanding is limited. Nevertheless, refinement through the continual evolution of our understanding can only come if we accept that our understanding is incomplete. Unknown unknowns will only come to light with continued, diligent and sometimes imaginative investigation of the data and metadata.

3. The Current State of Uncertainty in in situ SST Analyses

The classification of uncertainties outlined in section 2 will now be used as a framework to assess uncertainties in the global data sets based on in situ measurements. Preliminary to this it will be helpful to define what exactly is meant by sea-surface temperature.

3.1 Defining Sea-surface Temperature

Traditionally, in situ SST analyses have been considered representative of the upper ten or so meters of the ocean. However, the near-surface temperature structure of the ocean can be rather complex. Under conditions of low wind speed and high insolation, a stable stratified layer of warm water can form near the surface. For a recent review see *Kawai and Wada* [2007]. The diurnal temperature range of the sea-surface can, under certain conditions, exceed 5 K and, somewhat attenuated, penetrate to many tens of meters [*Prytherch et al.,* 2013]. This can lead to strong temperature gradients in the upper few meters of the ocean and consequently measurements made at the same time and location but at different depths can record quite different temperatures. Temperatures measured at the same depth but at different times of day can also differ markedly.

Donlon et al. [2007] proposed that the depth of the measurement be recorded along with the temperature as a first step to reconciling measurements made at different depths and different times of day. *Donlon et al.* [2007] also introduced the concept of an SST foundation (SST_{fnd}) temperature. The current definition (https://www.ghrsst.org/ghrsst-science/sst-definitions/) of "SST_{fnd}, is the temperature free of diurnal temperature variability, i.e., SST_{fnd} is defined as the temperature at the first time of the day when the heat gain from the solar radiation absorption exceeds the heat loss at the sea surface." It is generally assumed that the upper few meters of the ocean are of approximately constant temperature at this point. SST_{fnd} has proved a practical reference point for comparing and combining satellite observations [*Roberts-Jones et al.*, 2012] and was intended to provide "a more precise, well-defined quantity than the previous loosely-defined bulk SST" *Donlon et al.* [2007].

Unfortunately, such niceties of definition are not readily applicable to historical SST measurements and the effect of the interaction between measurement depth and water temperature on SST measurements in in situ archives is not clear. For many ships that measure the temperature of water drawn in below the surface, the depth of the measurements is not known and is likely to have changed depending on how heavily the ship was loaded. Nor is it clear to what extent any warm surface layer is mixed with cooler subsurface water by the passage of the ship or by the interaction of wind, water, Sun and hull [*Amot*, 1954; *Stevenson*, 1964]. Similar interactions have been noted closer to the surface with moored buoys [*Kawai and Kawamura*, 2000]. *James and Fox* [1972] found that ERI measurements from ships became progressively warmer relative to simultaneous bucket observations as the depth of the ERI measurement increased, a similar pattern to that seen by *Kent et al.* [1993]. *Reynolds et al.* [2010] found that measurements made by ships, which were largely ERI measurements in their study period, were on average warmer than nearby drifting buoy observations made nearer to the surface.

Nonetheless, the concept of the foundation SST can be used to get an idea of how changing measurement depth might have affected SST trends in the absence of other considerations. Figure 1 shows an upper estimate of the potential size of the effect of changing measurement depth on global average SST over time (for calculation details see Appendix A). The assumption is that buckets and buoys measure in the upper 30 cm and engine room measurements are measuring SST_{fnd}. The estimated global average bias (relative to the 1961-1990 average) is less than 0.1 K at all times and from 1945 onwards is less than 0.05 K. The bias is largest in the early record when all measurements were made using buckets which sample in the upper meter of the water column. In the more recent period, the blend of buckets, ERI measurements and buoys leads to a smaller, time-varying bias. Although the size of the effect is modest at a global level, locally the average diurnal warming can exceed 0.5K, which would imply a larger effect.

A related problem is that changing times of observation could potentially interact with the diurnal cycle of temperature leading to spurious trends in the data. *Kent et al.* [2010] note "*The implicit assumption is that the sampling of conditions is regular enough that no regional or time-varying bias is introduced into the datasets by neglecting such effects.*" Ships currently make SST observations at regular intervals throughout the day, typically every four or six hours, which is sufficient to minimize the aliasing of diurnal cycles, particularly if the measurements are made at depth. During earlier periods when buckets were widely used, there were systematic changes in the time of observation that might have a more pronounced effect on average SSTs but this has not been quantified.

Even when the measurement depth is known, there are potential problems. Metadata in WMO Publication 47 show that ships measure water temperatures through a wide range of depths from the near surface down to around 25 m [*Kent et al.*, 2007]. Although the average depth was typically less than 10 m, the deepest measurements could be sampling water that is colder than the SST_{fnd}. How large this effect might be is not yet well understood.

Chiodi and Harrison [2006] identified large-scale warm surface features using SST retrievals from microwave satellite instruments that persisted for several days. The warm layer was observed at night suggesting that the effect was independent from diurnal warming and they hypothesized that the multi-day warming might have been confined to a relatively shallow layer between 1 and 5 m thick. The implication is that the depth of the SST foundation temperature can vary rapidly and that it can be much shallower than the deepest in situ SST measurements. During a two week cruise, *Matthews and Matthews* [2013] found persistent temperature difference between the surface and 3 m depth in the tropical Pacific. Similar warm layers can be seen in data from moored buoys. Figure 2 shows time series from several moorings showing multi-day near-surface warm layers that do not penetrate down to 10 m and in some cases do not reach 5 m. Climatologies of mixed layer depth (MLD, see for example *de Boyer Montégut* [2004]) indicate large areas – in regions of upwelling and in the summer hemisphere – where the average MLD is shallower than 30 m, implying measurable temperature gradients within the depth range of ship SST measurements. *Grodsky et al.* [2008] also found differences between SST and temperatures in the mixed layer, which were largest in areas of persistent upwelling – most notably the eastern Pacific – but they did not consider the possible confounding effects of systematic errors in SST or other measurements.

To isolate the specific effect of multi-day or persistent temperature stratification of the near-surface waters would require regular measurements of near-surface waters at a range of depths. Such an analysis is now possible thanks to the network of Argo floats [*Castro et al.,* 2013]. In what follows, it should be noted that variations in depth will contribute to the variance of measurements and will therefore be partly, or wholly, counted in estimates of random and systematic measurement errors.

3.2 Individual Observational Errors

The general quality of raw SST measurements recorded in digital archives is mixed. Consequently, all SST analyses perform a stage of pre-screening, or quality control (QC) in order to remove observations of low quality and minimize the number of egregious errors. The size of the uncertainties of individual measurements will depend to a certain extent on the QC that is applied but the effects of differences in QC have not been assessed systematically.

3.2.1 Random Measurement Errors

Many estimates of random observational error uncertainty have been made. Although thermometers issued to ships by many port meteorological officers are calibrated, such calibration information is not routinely published, nor is there any guarantee that the temperature of a water sample measured by a well calibrated thermometer is equal to the actual SST when the sample has spent time in a bucket, or passed through the pipe work of a ship. Consequently, estimates of measurement uncertainty from the literature are empirical estimates derived from considerations of the variance of the data: for example, spatial [*Lindau,* 2003; *Kent and Challenor*, 2006; *Emery et al.*, 2001] and temporal [*Stubbs*, 1965] semivariograms, by comparing collocated observations [*O'Carroll et al.*, 2008], by resampling [*Shen et al.*, 2007], by using the variation of the variance with the number of observations [*Rayner et al.*, 2006], or by comparison with a background field [*Kent and Berry*, 2008; *Xu and Ignatov*, 2010; *Ingleby*, 2010; *Kennedy et al.*, 2011a;

Atkinson et al., 2013]. Some of the analyses did not distinguish between random observational errors and systematic observational errors, tending to combine them into one estimate. In addition it is not always easy to separate the effects of spatial sampling from measurement errors particularly in regions of high SST variability [*Castro et al.*, 2012].

A single SST measurement from a ship has a typical combined random and systematic error uncertainty of around 1 K to 1.5 K. Results from individual analyses are summarized in Table 1. The studies are mostly based on data from 1970 onwards.

Measurements are not all of identical quality. *Kent and Challenor* [2006] showed that in the period 1970-1997 the uncertainties of measurements from ships varied with location, time, measurement method and the country that recruited the ship. Uncertainties were estimated to be larger in the mid-1970s probably due to data being incorrectly transmitted in real time in the early days of the Global Telecommunication System. Their estimated uncertainty for engine room measurements was larger than for bucket measurements. *Tabata* [1978a] noted that bucket measurements *could* be accurate to 0.15 K, but that ERI measurements were nearly an order of magnitude worse (1.16 K). *Ingleby* [2010] estimated uncertainties for different subsets of the data and noted that manual VOSclim (a high-quality subset of the VOS fleet) measurements and automated measurements were of slightly higher quality than manual ship measurements in general. *Beggs et al.* [2012] showed that Australia Integrated Marine Observing System ships had uncertainties comparable to those from data buoys. Analyses that have looked at statistics for individual ships and buoys have found that some ships and buoys take much higher quality measurements than others [*Kent and Berry*, 2008; *Brasnett*, 2008; *Kennedy et al.,* 2011a; *Atkinson et al.*, 2013]. The subset of ships (around 40-50% of ship observations) that passed the more stringent quality control procedures of *Atkinson et al.* [2013] had significantly lower measurement uncertainties assessed using the method of *Kennedy et al.* [2011a] than did the full fleet of ships. Early results on hull sensors reported by *Emery et al.* [1997] indicated the potential for these sensors to make accurate measurements. Indeed, *Kent et al.* [1993] found that hull sensors installed on ships in the Voluntary Observing Ships Special Observing Project for the North Atlantic (VSOP-NA) gave consistent measurements during the two year observing period.

Drifting buoy measurements are generally more accurate and consistent than ship measurements, but there is a greater relative spread between the estimates which are summarized in Table 2. In part these differences are likely to arise from the level of pre-screening that is applied to the observations. Where quality control is more stringent, estimated uncertainties are likely to be lower and, where the error variance of the observations is low already, the effects of quality control and processing choices are likely to be more pronounced [*Xu and Ignatov*, 2012]. *Castro et al.* [2012] considered differences between drifting buoys and two different satellite products and found that there was little difference between buoys produced by different manufacturers. There is some evidence that the quality of drifting buoy observations has improved slightly over time [*Merchant et al.*, 2012], but this has not been conclusively demonstrated. As a comparison, temperature measurements from Argo have been reckoned to have an uncertainty of around 0.002K [*Abraham et al.,* 2013].

Moored buoys have received less attention. Estimates of the measurement uncertainties are summarized in Table 3. The two studies [*Kennedy et al.*, 2011a; *Xu and Ignatov*, 2010] that examined moorings from the GTMBA separately from other moorings found that they had lower measurement error uncertainties. *Castro et al.* [2012] found that the standard deviations of differences between moorings and satellite data were lower for tropical moorings than for coastal moorings. They noted that in coastal waters there can be large local variations in temperature, which satellites cannot resolve. Some moorings along coastlines are located in estuaries and river mouths and are therefore less likely to be representative of open ocean areas. This is perhaps one reason why *Wilkerson and Earl* [1990], who studied US coastal buoys, found such large standard deviations between ships and moorings (Table 1). *Merchant et al.* [2012] found that few coastal moorings met their required stability criteria.

As noted in section 2, random observational errors are of relatively minor importance in large-scale averages (see Figure 8 and section 3.5), particularly in the modern period when observations are numerous. For an uncertainty of 1.0 K for a single observation due to random observational error, the resulting uncertainty of a global annual average based on 10000 observations would be of order 0.01 K.

3.2.2 Random and Systematic Measurement Errors

Kent and Berry [2008] and *Kennedy et al.* [2011a, 2011b] decomposed the observational errors into random and systematic components. *Brasnett* [2008] and *Xu and Ignatov* [2010] implicitly used the same error model – their analyses output the same statistics produced by *Kent and Berry* [2008] – and the results are indeed very similar (Figure 3). Estimates are summarized in Table 4. The possibility of correlated measurement errors is also implicitly allowed for by *Ishii et al.* [2003] and *Hirahara et al.* [2013] who merge observations from a single ship into a super observation before calculating uncertainties. Adding the uncertainties in quadrature gives a combined observational uncertainty of between 1 and 1.5 K, consistent with earlier estimates (Table 1) that did not differentiate between the two.

In the studies listed in Table 4, the systematic component of the error was assumed to be different for each ship, but this does not on its own capture the effects of pervasive systematic errors. The data from *Kent and Berry* [2008], *Brasnett* [2008] and *Xu and Ignatov* [2010] also show that the systematic observational error component for some ships varies from month to month suggesting that the partitioning of systematic and random effects is also a function of the time period considered.

The addition of a systematic component has a pronounced effect on the uncertainty of large-scale averages comprising many observations. *Kennedy et al.* [2011b] estimated that the effect of the correlations between errors was to increase the uncertainty of the global annual average SST anomaly due to measurement error from 0.01 K (uncorrelated case) to more than 0.05 K in the 19th Century and to more than 0.01 K even in the well-observed modern period when millions of observations contribute to the annual global average (see Figure 8). Systematic errors could also have a pronounced effect on reconstructions when they project onto large-scale modes of variability, or on the estimation of EOFs. However, because of the assumed independence of the errors between ships, the correlated component of the uncertainty remains relatively unimportant for the analysis of long-term trends of large-scale averages. Pervasive systematic errors, which are correlated across a large proportion of the global fleet, (section 3.2) are far more important from that point of view.

One of the difficulties with estimating the uncertainties associated with systematic errors from individual ships is that not all observations in ICOADS can be associated with an individual ship. Some of the reports have no more information than a location, time and SST measurement. *Kennedy et al.* [2011b] had to make estimates of how the uncertainty arising from systematic errors behaved as the number of observations increased by considering the behavior at times when the majority of reports contained a ship name or call-sign. They assumed that observations without call signs behaved in the same way. *Kent and Berry* [2008] suggested that only ship reports with extant metadata be used in climate analyses of the modern period to minimize such ambiguities. For earlier periods, the gains in improved quantification of uncertainty would need to be balanced against the increased uncertainty arising from reduced coverage.

Many gridded SST data sets and analyses, as well as the studies that depend on them, assume that the observational errors are normally distributed, but this is not necessarily the case for individual observations. *Kennedy et al.* [2011a] investigated the properties of observations that had been quality controlled using the procedures described in *Rayner et al.* [2006]. They found that in comparisons with satellite observations the distributions of errors were 'fat-tailed' with the distribution of errors having a positive kurtosis. In the creation of gridded data sets from SST observations, the effects of outliers can be minimized somewhat by the use of resistant or robust statistics such as Winsorised, or trimmed means (see e.g., *Rayner et al.* [2006]). The effect of outliers is further reduced in large scale averages and the distribution of errors in large scale averages tends towards a normal distribution as the number of observations increases [*Kennedy et al.*, 2011a].

3.2.3 Summary of Individual Observational Errors

Many estimates of uncertainties of ship and buoy SST measurements have been made. A typical SST measurement made by a ship has an uncertainty of around 1-1.5K and a drifting buoy observation a typical uncertainty of around 0.1-0.7K. More recent studies split these uncertainties into random and systematic components, which better describe the error characteristics of these platforms. However, a lack of metadata, most particularly ship call signs, hampers the application of such an error model and it does not capture behavior seen in SST measurements such as non-Normal distributions or systematic errors that vary on time scales from months to years.

3.3 Pervasive Systematic Errors and Biases

Kent et al. [2010] conducted a review of literature on pervasive systematic errors (often termed 'biases') in in situ SST measurements. Many studies have looked at the differences in pervasive systematic errors between measurement methods, but fewer have attempted to adjust SST records to minimize the effects of changes in instrumentation.

3.3.1 Bias Adjustments 1850 to 1941

The need for adjustments to minimize the cold bias associated with bucket measurements in the period from 1850 to 1941 is well established. *Folland and Parker* [1995] calculated adjustments using a simplified physical model of the buckets used to make SST measurements combined with fields of climatological air-temperature, SST, humidity, wind and solar radiation. Some parameters in their model were taken from literature and others were estimated from the data. The length of time between the water sample leaving the sea surface and the measurement was estimated by integrating their model until a seasonal cycle in the SST was minimized. The fractional contributions of canvas and wooden buckets were estimated by assuming a linear change over time from a mix of wooden and canvas buckets to predominantly canvas buckets by 1920. The rate of this change was estimated by minimizing the air-sea temperature difference in the tropics. The same method was also used in *Rayner et al.* [2006] and *Kennedy et al.* [2011c].

Smith and Reynolds [2002] took an alternative approach. They adjusted SSTs based on statistical relationships between Night Marine Air Temperature (NMAT) and SST. The resulting adjustments were different to those produced by *Folland and Parker* [1995] although the magnitude of the global average adjustment was similar. Both *Folland and Parker* [1995] and *Smith and Reynolds* [2002] found a long term increase in the magnitude of the adjustments – that is, an increasing cold bias – from the 1850s to 1941.

The methods employed by *Folland and Parker* [1995] and *Smith and Reynolds* [2002] are not independent as they both rely on NMAT, which have their own particular pervasive systematic errors [*Bottomley et al.,* 1990; *Rayner et al.*, 2003; *Kent et al.*, 2013]. The use of NMAT to adjust SST data is, to an extent, unavoidable as the heat loss from a bucket does depend on the air-sea temperature difference.

In data sets based on a ICOADS release 2.0 and later, the earlier bucket adjustments were found to over-adjust SST in the period 1939-1941. *Rayner et al.* [2006] and *Smith et al.* [2008] ramped the adjustments down to zero over this period. *Kennedy et al.* [2011c] showed that the ramp-down corresponded to new data in that release of ICOADS that included a large fraction of ERI measurements.

3.3.2 Bias Adjustments 1941 to Present

In the post-1941 period, *Folland and Parker* [1995], *Smith and Reynolds* [2003], *Smith and Reynolds* [2005] and *Rayner et al.* [2006] opted not to adjust the data because they found no clear evidence of the need for adjustments. However, *Rayner et al.* [2006] did identify biases in Japanese and Dutch data after the Second World War. *Thompson et al.* [2008] identified a discontinuity in global-average SST associated with a change in the composition of ICOADS release 2.1 in late 1945. *Reynolds et al.* [2010] quantified a relative bias between ship and drifting buoy measurements that they thought could lead to an artificial cooling of the global average SST. *Kent et al.* [1999] applied adjustments to ERI measurements, but removed the adjustment from later versions of their data set.

Kennedy et al. [2011c] and *Hirahara et al.* [2013] developed bias adjustments for the period 1941 onwards. *Kennedy et al.* [2011c] used metadata from ICOADS, WMO Publication 47, observer instructions, technical reports and scientific papers to estimate biases for individual measurement types and to assign a measurement method to as many observations as possible. *Hirahara et al.* [2013] used a narrower range of metadata. By comparing subsamples of the data for which the metadata were known, they could estimate appropriate metadata assignments for the remainder.

To estimate the bias adjustments for long-term analyses, an understanding is needed of how biases varied for individual components of the observing system. Several studies have examined ERI and bucket biases in ship data [*Brooks*, 1926; *Brooks*, 1928; *Lumby*, 1927; *Collins et al.*, 1975; *Wahl*, 1948; *Roll*, 1951; *Kirk and Gordon*, 1952; *Amot*, 1954; *Perlroth*, 1962; *Saur*, 1963; *Walden*, 1966; *Knudsen*, 1966; *Tauber*, 1969; *James and Fox*, 1972; *Tabata,* 1978a, 1978b; *Folland et al.*, 1993; *Kent et al.*, 1993] but only *Kent and Kaplan* [2006] provide information that is time-resolved and traceable back to

ICOADS. There is a single study of pervasive systematic errors in hull sensor measurements [*Kent et al.*, 1993], which analyzed data from a small number of ships over a two year period and found that hull sensors were relatively unbiased and showed no systematic change of bias with depth.

Few studies have looked at the long-term stability and calibration drifts of drifting buoys. *Reverdin et al.* [2010] installed 16 drifters with high quality temperature sensors in addition to their usual temperature sensors and found that the temperatures measured by the drifters showed inaccuracies that were larger than the 0.1 °C target accuracy and that they exhibited significant calibration drifts. This is consistent with the behavior seen by *Atkinson et al.* [2013].

3.3.3 Estimating Uncertainty in Bias Adjustments

Folland and Parker [1995] did not explicitly estimate the uncertainties in their adjustments. *Rayner et al.* [2006] explored the parametric uncertainty in the *Folland and Parker* [1995] adjustments using a Monte-Carlo method. In *Smith and Reynolds* [2004] the uncertainty in the bias adjustments was estimated by taking the mean-squared difference between the *Smith and Reynolds* [2002] adjustments and the *Folland and Parker* [1995] adjustments, a first-order estimate of the structural uncertainty.

Kennedy et al. [2011c] used a Monte-Carlo method to explore the parametric uncertainty within their particular approach to bias adjustment. *Hirahara et al.* [2013] also provide uncertainties on their adjustments that are a combination of analysis uncertainties and regression uncertainty.

An important component of the uncertainty of adjustments for the effects of persistent systematic errors arises from a lack of knowledge concerning how the measurements were made. Metadata are often missing, incomplete or ambiguous and sometimes different sources give conflicting information. *Kent et al.* [2007] assessed metadata from ICOADS and WMO Publication 47. They found disagreement in around 20-40% of cases where metadata were available from both sources. *Kennedy et al.* [2011c] allowed for up to 50% uncertainty in metadata assignments based on the discrepancy between observer instructions and measurement methods recorded in WMO Publication 47. *Hirahara et al.* [2013] used differences between subsets of data to infer the fraction of observations made using different methods.

Figure 6 compares estimated biases and metadata assignments from *Kennedy et al.* [2011c] and *Hirahara et al.* [2013]. It shows that from 1945, the estimated biases agree within their parametric uncertainty ranges (Figure 6a) and that the fractions of measurement methods estimated by *Kennedy et al.* [2011c] from literature and other metadata are consistent with the fractions inferred from the data by *Hirahara et al.* [2013] (Figure 6b). However, there are two key differences that highlight the importance of structural uncertainty for understanding the bias adjustments. The first difference is that the phasing out of uninsulated buckets in *Hirahara et al.* [2013] happens earlier and faster than allowed for in the parametric uncertainty analysis of *Kennedy et al.* [2011c]

(Figure 6c). In *Hirahara et al.* [2013] the changeover starts in the 1940s and is especially rapidly in the early 1960s, being nearly complete by around 1962. The second difference is that the estimated bias during the Second World War is higher in the analysis of *Hirahara et al.* [2013] than in *Kennedy et al.* [2011c]. Further work is needed to understand these differences and more complete, more reliable metadata would help reduce uncertainty in SST records.

In the post-1941 period, *Smith and Reynolds* [2003] and *Smith and Reynolds* [2005] estimated the uncertainty due to pervasive systematic errors by considering the difference in estimated bias between measurements made in the engine rooms of the ships and measurements from all ships between 1994 and 1997. They estimated a minimum 1-sigma standard error in the global average of around 0.015 K. The range is similar to, albeit slightly narrower than, that estimated by *Kennedy et al.* [2011c]. The difficulty with the approach taken by *Smith and Reynolds* [2003], *Smith and Reynolds* [2005] and *Smith et al.* [2008] is that the quoted uncertainty range is considered to be symmetric whereas *Kennedy et al.* [2011c] and *Hirahara et al.* [2013] suggest that the true global mean is consistently higher than *Smith et al.* [2008] in the period 1945-1960 (Figure 9). It also suggests that the estimate of *Smith et al.* [2008] in the post World War 2 period (1945-1950s) was slightly too conservative because it compared ERI measurements with a mixture of ERI and insulated bucket measurements, whereas large numbers of observations were made using buckets [*Kennedy et al.*, 2011c; *Hirahara et al.,* 2013].

3.3.4 Refinements to Estimates of Pervasive Systematic Errors

There are some factors that have not been explicitly considered in estimates of biases. Refinements to the models of pervasive systematic errors will address with factors that are implicitly included in random and systematic measurement uncertainties. If it is possible to estimate the bias on a ship-by-ship, or observation-by-observation basis, taking account of the conditions peculiar to that observation, then it might be expected that uncertainties associated with random and systematic observational error will decrease.

Both *Kennedy et al.* [2011c] and *Hirahara et al.* [2013] make simplifying assumptions about the systematic errors associated with modern insulated buckets. Various bucket designs have been used since the end of the Second World War, which are likely to have different bias characteristics. Physical models could be developed for each type of bucket similar to those used by *Folland and Parker* [1995], or statistical methods could be used to estimate the biases as was done in *Kent and Kaplan* [2006].

Other simplifying assumptions used in all analyses include such things as assuming that changes in the observing system happened linearly. Evidence suggests that changes in measurement method were not always monotonic and sometimes happened abruptly (see Figure 6). Improved metadata or more sophisticated statistical techniques could help assess these uncertainties.

An uncertainty associated with pervasive systematic biases, which is not explicitly resolved by current analyses, arises when the conditions at the time of the measurement deviate from the climatological values assumed by the bias correction scheme. If, for instance, the air sea temperature difference is larger than that assumed by the *Folland and Parker* [1995] scheme, then there will be an additional systematic uncertainty that is correlated strongly across synoptic spatial and temporal scales with a potential long-term component where differences persist for months or years. Likewise conditions vary during the day. Such discrepancies could be assessed by evaluating the systematic error using local conditions. Such information could be taken from reanalyses, or an appropriate bucket model could be explicitly included when SST observations are assimilated into ocean-only and coupled reanalyses.

3.3.5 Assessing the Efficacy of Bias Adjustments

The efficacy of the bias adjustments and their uncertainties are difficult to assess. *Folland and Parker* [1995] presented wind tunnel and ship board tests and also used their adjustments to estimate the differences between bucket and ERI measurements in broad latitude bands. These limited comparisons showed that their model could predict experimental results to better than 0.2 K. *Folland and Salinger* [1995] presented direct comparisons between air temperatures measured in New Zealand and SST measurements made nearby. *Smith and Renyolds* [2002] used oceanographic observations to assess their adjustments and those of *Folland and Parker* [1995]. In regions with sufficient observations they found that the magnitude of the *Smith and Reynolds* [2002] adjustments better explained the differences between SSTs and oceanographic observations, but the phase of the annual cycle was better captured by *Folland and Parker* [1995]. *Hanawa et* *al.* [2000] showed that the *Folland and Parker* [1995] adjustments improved the agreement between Japanese ship data and independent SST data from Japanese coastal stations in two periods: before and after the Second World War. However, the collection of ship data (COADS and Kobe collections) used in *Hanawa et al.* [2000] might not have had the same bias characteristics as assumed by *Folland and Parker* [1995] (based on the Met Office Marine Data Bank) in developing their adjustments. Other long term coastal records of water temperature exist. Some of these [*Hanna et al.,* 2006; *MacKenzie and Schiedek,* 2007; *Cannaby and Hüsrevoğlu*, 2009] have been compared to open ocean SST analyses (though not with the express intention of assessing bias adjustments), others have not [*Maul et al.,* 2001; *Nixon et al.,* 2004; *Breaker et al.* 2005].

More recently, *Matthews* [2013] and *Matthews and Matthews* [2013] reported field measurements of SST made using different buckets and simultaneous thermo-salinograph measurements. They found negligible biases between different buckets, but their experimental design involved larger buckets and shorter measurement times than were used in *Folland and Parker* [1995]. Nevertheless, this highlights the potential for well-designed field experiments to improve understanding of historical biases.

An analysis by *Gouretski et al.* [2012] compared SST observations with near-surface measurements (0-20 m depth) taken from oceanographic profiles. It shows that the overall shape of the global average is consistent between the two independent analyses, but that there are differences of around 0.1 K between 1950 and 1970. These are most likely attributable to residual biases, although, as noted above, actual physical differences between the sea surface and the 0-20 m layer cannot be ruled out. Similar differences are seen when comparing SST with the average over the 0-20 m layer of the analysis of *Palmer et al.* [2007] (not shown).

Since the late 1940s, global and hemispheric average SST anomalies calculated separately from adjusted bucket measurements and adjusted ERI measurements showed consistent long-term and short-term changes [*Kennedy et al.*, 2011c]. From the 1990s, there are also plentiful observations from drifting and moored buoys.

In contrast to the modern period, the period before 1950 is characterized by a much less diverse observing fleet. During the Second World War, the majority of measurements were ERI measurements. Before the war, buckets were the primary means by which SST observations were made. This makes it very difficult to compare simultaneous independent subsets of the data. In periods with fewer independent measurement types, it might be possible to use changes in environmental conditions such as day-night differences or air-sea temperature differences to diagnose systematic errors in the data.

Qualitative agreement between the long-term behavior of different global temperature measures – including NMAT, SST and land temperatures – gives a generally consistent picture of historical global temperature change (Figure 5), but a direct comparison is less informative about uncertainty in the magnitude of the trends. *Kent et al.* [2013] showed similar temporal evolution of NMAT and SST in broad latitude bands in the northern hemisphere and tropics. However there are differences of up to 0.4 K in the band from

55°S to 15°S between 1940 and 1960. Studies such as that by *Folland* [2005] can be used to make more quantitative comparisons. *Folland* [2005] compared measured land air temperatures with land air temperatures from an atmosphere-only climate model that had observed SSTs (with and without bucket adjustments) as a boundary forcing. He found much better agreement when the SSTs were adjusted. Atmospheric reanalyses also use observed SSTs along with other observed meteorological variables to infer a physically consistent estimate of land surface air temperatures. *Simmons et al.* [2010] showed that land air temperatures from a reanalysis driven by observed SSTs were very close to those of CRUTEM3 [*Brohan et al.*, 2006] over the period 1973 to 2008. *Compo et al.* [2013] showed similar results for the whole of the twentieth century although the agreement was not quite so close. Although their intention was to show that land temperatures were reliable, their results indicate that there is broad consistency between observed SSTs and land temperatures.

3.3.6 Summary of Pervasive Systematic Errors and Biases

The need to adjust SST data prior to 1941 to account for a cold bias associated with the use of canvas and wooden buckets is well established. There is also good evidence for the need to adjust data after 1941. Adjustments for these pervasive systematic errors have been developed. There are, at all times, two different estimates of the bias adjustments, which are in general agreement and give a first indication of the structural uncertainty. Evidence for the efficacy of the adjustments comes from wind tunnel tests, comparisons with coastal sites and consistency with subsurface ocean temperatures, marine air temperatures and land air temperatures. Contrary evidence comes from a recent field experiment in the Pacific. Uncertainty could be better understood by: improvements in metadata; carefully designed fields tests of buckets and other measurements methods; the creation of new independent evaluations of the biases; and continued comparison between SST and related variables.

3.4 Sampling Uncertainty

The magnitude of the grid-box sampling uncertainty depends on the correlation and variability of SSTs within the grid box, on the number of observations contributing to the grid-box average and where in the grid box they are located. High average correlations within a grid box, low variability and large numbers of observations lead to lower uncertainty estimates. Conversely areas of high variability or low average correlation, such as frontal regions or western boundary currents, tend to have higher grid-box sampling uncertainties as do grid-box averages based on smaller numbers of observations. The estimation of uncertainties arising from the sparseness of observations at scales from grid box level to global has been approached in a number of ways.

3.4.1 Grid-box Sampling Uncertainty

Weare and Strub [1981] counted the number of observations needed to minimize sampling uncertainty in a 5°x5° grid box by ensuring that the observations were evenly split between all areas of the grid box, month and diurnal cycle. From this, they concluded that even sampling could not be achieved with fewer than eleven observations, but that in practice more than eleven, sometimes many more, would be needed.

Rayner et al. [2006] estimated a combined measurement and grid-box sampling uncertainty by considering how the variance of the grid-box average changed as a function of the number of observations. The technique picked up spatial variations in grid-box sampling uncertainty associated with regions of high variability. *Rayner et al.* [2009] showed results from an unpublished analysis by Kaplan, in which spatially complete satellite data were used to estimate the variability within 1°x1° grid boxes. The same features were seen as in the *Rayner et al.* [2006] analysis, allowing for differences in resolution, although the uncertainties estimated by Kaplan tended to be higher. *She et al.* [2007] also used sub-sampling of satellite data to estimate grid-box sampling uncertainty for the Baltic Sea and North Sea. *Kent and Berry* [2005] showed that separately assessing measurement and sampling uncertainties can help to decide whether more, or better, observations are needed to reduce the average uncertainty in an individual grid box.

Morrissey and Greene [2009] developed a theoretical model for estimating grid-box sampling uncertainty that accounted for non-random sampling within a grid box. This was an extension of the method used to estimate sampling uncertainties in land temperature data and global temperatures by *Jones et al.* [1997]. Land temperatures are measured by stations at fixed locations that take measurements every day. Marine temperature measurements are taken at fixed times, but the ships and drifting buoys move during a particular month. *Morrissey and Greene* [2009] do not provide a practical implementation of their approach, only a theoretical framework. *Kennedy et al.* [2011b] extended the concept of the average correlation within a grid box developed in *Jones et al.* [1997] to incorporate a time dimension. *Kent and Berry* [2008] used a temporal autocorrelation model that took account of the days within the period that were sampled, and the days which were not, to estimate the temporal sampling uncertainty. An alternative to the *Jones et al.* [1997] method for land data was provided by *Shen et al.* [2007], but it has not yet been applied in SST analyses.

It is possible that the locations visited by ships and drifting buoys are related and, to an extent, dictated by meteorological and oceanographic conditions. Ships have long used the prevailing currents in the Atlantic to speed their progress and it is in the interest of almost all shipping to steer clear of hurricanes and other foul weather. Bad weather is also likely to have influenced how and when observations were made. Conversely, the conditions in which a sail ship might become becalmed could lead to over sampling of higher SSTs. Drifting buoys drift, and a drifter trapped in an eddy might persistently measure temperatures that are representative of only a very limited area. Drifters also tend to drift out of areas of upwelling and congregate in other areas.

The effect of uneven sampling can be reduced by the creation of ‘super observations’ during the gridding process [*Rayner et al.,* 2006], or data preparation stage [*Ishii et al.,* 2003], but such processes cannot readily account for the situations where no observations are made at all.

As noted by *Rayner et al.* [2006], the grid-box sampling uncertainties are likely to be uncorrelated or only weakly correlated between grid boxes so the effect of averaging together many grid boxes will be to reduce the combined grid-box sampling uncertainty by a factor proportional to the square root of the number of grid boxes. Consequently the sampling component of the uncertainty will be of minor importance in the global annual average (Figure 8).

3.4.2 Large-scale Sampling Uncertainty

Because *Rayner et al.* [2006] and *Kennedy et al.* [2011b] make no attempt to estimate temperatures in grid boxes which contain no observations, an additional uncertainty had to be computed when estimating area-averages. *Rayner et al.* [2006] used Optimal Averaging (OA) as described in *Folland et al.* [2001] which estimates the area average in a statistically optimal way and provides an estimate of the large-scale sampling uncertainty. *Kennedy et al.* [2011b] subsampled globally complete fields taken from three SST analyses and obtained similar uncertainties from each. The uncertainties of the global averages computed by *Kennedy et al.* [2011b] were generally larger than those estimated by *Rayner et al.* [2006]. *Palmer and Brohan* [2011] used an empirical method based on that employed for grid-box averages in *Rayner et al.* [2006] to estimate global and ocean basin averages of subsurface temperatures.

The *Kennedy et al.* [2011b] large-scale sampling uncertainty of the global average SST anomaly is largest (with a 2-sigma uncertainty of around 0.15°C) in the 1860s when coverage was at its worst (Figure 8). This falls to 0.03 °C by 2006. The fact that the large-scale sampling uncertainty should be so small – particularly in the nineteenth century – may be surprising. The relatively small uncertainty might simply be a reflection of the assumptions made in the analyses used by *Kennedy et al.* [2011b] to estimate the large-scale sampling uncertainty. Indeed, *Gouretski et al.* [2012] found that subsampling an ocean reanalysis underestimated the uncertainty when the coverage was very sparse. However, estimates made by *Jones* [1994] suggest that a hemispheric-average land-surface air temperature series might be constructed using as few as a 109 stations. For SST, the variability is typically much lower than for land temperatures though the area is larger. It seems likely that the number of stations needed to make a reliable estimate of the global average SST anomaly would not be vastly greater.

Another way of assessing the large-scale sampling uncertainty is to look at the effect of reducing the coverage of well-sampled periods to that of the less-well-sampled nineteenth century and recomputing the global average (see for example *Parker* [1987]). Figure 4 shows the range of global annual average SST anomalies obtained by reducing each year to the coverage of years in the nineteenth century. So, for example, the range indicated by the blue area in the upper panel for 2006 shows the range of global annual averages obtained by reducing the coverage of 2006 successively to that of 1850, 1851, 1852... and so on to 1899. The red line shows the global average SST anomaly from data that have not been reduced in coverage. For most years, the difference between the sub-sampled and more fully sampled data is smaller than 0.15K and the largest deviations are smaller than 0.2K. For the large-scale sampling uncertainty of the global average to be significantly larger would require the variability in the nineteenth century data gaps to be different from that in the better-observed period.

3.4.3 Summary of Sampling Uncertainty

Uncertainties arising from under-sampling at a grid-box level are easy to assess if the observations can be assumed to be randomly distributed within a grid box. However, sampling is not random. The effect of this is reduced in most analyses by the calculation of super-observations that combine nearby measurements; however, optimal methods to minimize uncertainty are not generally applied. Simple estimates of large-scale sampling uncertainty in the global-average SST from subsampling well-sampled periods suggest a value of at most 0.2K even in poorly observed years. However, there are potential limitations of these simple methods and they should be considered together with the range of statistical reconstructions to get a more complete idea of uncertainty in large-scale averages.

3.5 Reconstruction Techniques and Other Structural Choices

Creating global SST analyses is challenging because of the relative sparseness of observations before the satellite era and the non-stationarity of the changing climate. A large number of different SST data sets based on in situ data have been produced employing a variety of statistical methods. The structural uncertainties associated with estimating SSTs in data voids and at data-sparse times are therefore somewhat better explored than structural uncertainties in the pervasive systematic errors. Data sets used in this paper have been summarized in Table 5 and global averages for these data sets are shown in Figure 5.

3.5.1 Critique of Reconstruction Techniques

The current generation of SST analyses are the survivors of an evolutionary process during which less effective techniques were discarded in favor of better adapted alternatives. It is worthwhile to ask how, as a group, they address the range of criticisms that have arisen during that time.

One concern is that patterns of variability in the modern era which are used to estimate the parameters of the statistical models might not faithfully represent variability at earlier times [*Hurrell and Trenberth*, 1999]. The concern is allayed somewhat by the range of approaches taken. The method of *Kaplan et al.* [1998] which uses the modern period to define Empirical Orthogonal Functions (EOFs, see *Hannachi et al.*, [2007] for a review of the use of EOFs in the atmospheric sciences) tends to underestimate the long-term trend. This is particularly obvious in the nineteenth and early twentieth century. *Rayner et al.* [2003] extended the method by defining a low-frequency, large-scale EOF that better captured the long-term trend in the data. However, it is possible that a single EOF will fail to capture all the low-frequency changes. *Smith et al.* [2008] allow for a nonstationary low-frequency component in their analysis which contributes a large component of uncertainty in the early record, but their reconstruction reproduces less high-frequency variability at data-sparse epochs. *Ilin and Kaplan* [2009] and *Luttinen and Ilin* [2009, 2012] used algorithms that make use of data throughout the record to estimate the covariance structures and other parameters of their statistical models. The three algorithms use either large-scale patterns (VBPCA, GPFA) or local correlations (GP). Differences between the three methods are generally small at the global level, but they diverge during the 1860s when data are few. There is a caveat that despite using all available observations, such methods will still tend to give a greater weight to periods with more plentiful observations. *Ishii et al.* [2005] use a simply-parameterized local covariance function for interpolation. Their optimal interpolation (OI) method was assessed by *Hirahara et al.* [2013] to have larger analysis uncertainties and larger cross-validation errors than the EOF-based COBE-2 analysis. However, the use of a simple optimal interpolation method has the advantage that it makes fewer assumptions regarding the stationarity of large-scale variability.

Another concern is that methods that use EOFs to describe the variability might inadvertently impose spurious long-range teleconnections that do not exist in the real world [*Dommenget*, 2007]. *Smith et al.* [2008] explicitly limit the range across which teleconnections can act. *Ishii et al.* [2005] used a local covariance structure in their analysis. Analyses such as *Kaplan et al.* [1998] and *Rayner et al.* [2003] make the assumption that the EOFs retained in the analysis capture actual variability in the SST fields, but do not explicitly differentiate between variability that can be characterized purely in terms of local co-variability and large-scale teleconnections. *Karspeck et al.* [2012] note that there is not a clear separation of scales and that joint estimation of local and large scale covariances is the logical way to approach the problem.

Most, if not all, statistical methods have a tendency to lose variance either because they do not explicitly resolve small scale processes [*Kaplan et al.,* 1998; *Smith et al.*, 2008], because the method tends towards the climatological average in the absence of data [*Ishii et al.,* 2005; *Berry and Kent*, 2011], or because they tend to smooth the data. *Rayner et al.* [2003] used the method of *Kaplan et al.* [1998] but blended high-quality gridded averages back into the reconstructed fields to improve small scale variability where observations were plentiful. *Karspeck et al.* [2012] analyzed the residual difference between the observations and the analysis of *Kaplan et al.* [1998] analysis using local non-stationary covariances, and then drew a range of samples from the posterior distribution in order to provide consistent variance at all times and locations.

One assumption common to most of the above analysis methods is that SST variability can be decomposed into a small set of distinct patterns that can be combined linearly to describe any SST field. However, it is well known that phenomena such as El Niño and La Niña are not symmetric and that the equations that describe the evolution of SST are non-linear. Consequently, current analyses might not capture the full range of behavior in real SST fields [*Karnauskas,* 2013]. Current generation SST analyses are based on the assumption that individual measurement errors are uncorrelated and that errors are normally distributed. Analysis techniques that incorporate information about the correlation structure of the errors have not yet been developed. Such techniques are likely to be more computationally expensive and lead to larger analysis uncertainties.

3.5.2 Other Structural Choices

Analyses based on SST anomalies will also have an uncertainty associated with the climatological reference fields used to calculate the anomalies. Sub-surface analyses have been shown to be particularly sensitive to choice of base period [*Lyman et al.*, 2010], due in a large part to the relative sparseness of the data sets. Although the problem is likely to be less severe for the better-observed SST record, there are still regions – the Southern Ocean and Arctic Ocean – where observations are few. *Yasunaka and Hanawa* [2011] found that differences between long-term-average SSTs from different analyses were typically less than 0.5 K, but that they exceeded 1 K in places. The largest differences were at high latitudes and in regions with strong SST gradients. There are also likely to be pervasive systematic errors in the climatological averages [*Kennedy et al.*, 2011c].

Other structural differences arise from the way that SSTs are extended to the edge of the sea ice. SSTs can be estimated from measurements of sea-ice concentration [*Rayner et al.*, 2003; *Smith et al.*, 2008; *Hirahara et al.,* 2013]. Although their global impact is likely to be small, the uncertainties in these relationships and estimates need also to be factored into the uncertainty of SSTs in these regions. At the moment, the uncertainty associated with historical sea-ice concentrations is poorly understood.

3.5.3 Comparisons of Reconstructions

Yasunaka and Hanawa [2011] examined a range of climate indices based on seven different SST data sets. They found that the disagreement between data sets was marked before 1880, and that the trends in large scale averages and indices tend to diverge outside of the common climatology period. For the global average, the differences between analyses were around 0.2 K before 1920 and around 0.1-0.2 K in the modern period. Even for relatively well-observed events such as the 1925/26 El Niño, the detailed evolution of the SSTs in the tropical Pacific varied from analysis to analysis. The reasons for the discrepancies are not completely clear because each data set is based on a slightly different set of observations that have been quality controlled and processed in different ways, a problem that could be alleviated by running analyses on identical input data sets.

Combined with information about large-scale sampling uncertainties estimated in other ways, the spread between analyses suggests that the large-scale sampling uncertainty in global average SST anomaly is around 0.2 K in the late 19^{th} century. For the large-scale sampling uncertainty of the global average to be much larger would require variability in the early record to have been different from variability in the modern period, which is a possibility. The resolution of such a question is most likely to be achieved via the digitisation of more observations from paper records.

Progress in assessing the differences between analysis techniques can also be made by studying the relative strengths and weaknesses of interpolation techniques on carefully prepared test data sets using synthetic data, or on 'withheld' data from well observed regions. By running each analysis on the same carefully-defined subsets and tests, it should be possible to isolate reasons for the differences between the analyses and assess the reliability of analysis uncertainty estimates. The International Surface Temperature Initiative (http://www.surfacetemperatures.org/) has been working on such benchmarking exercises for land surface air temperature data, building on work such as the COST ACTION project [*Venema et al.*, 2012].

3.5.4 Summary of Reconstruction Techniques and Structural Uncertainty

A range of reconstruction techniques exist to make globally-complete or near globally-complete SST analyses. The spread in global mean SST between analyses is at worst around 0.2K. The analyses are based on a variety of different statistical models suggesting that estimates of global average SST are not strongly dependent on such choices. However, current reconstruction techniques do not account for systematic errors in the data – they assume errors are random and uncorrelated – and assume that SST fields can be simply parameterized in terms of limited numbers of patterns or simple covariance relationships. Objective comparison of different reconstruction techniques and their associated uncertainty estimates would be aided by the creation of standard benchmark tests which mimic the distribution and character of observational data.

3.6 Comparing Components of Uncertainty

Figure 7 shows individual components of the overall uncertainty estimated for three months. The components include: estimates of structural uncertainty (in lieu of a formal way to estimate this, it is calculated as the standard deviation of seven near-globally-complete analyses: COBE, Kaplan, ERSSTv3, HadISST, GPFA, GP and VBPCA), sampling uncertainty, combined random and systematic measurement error uncertainty, bias uncertainty (estimated from a 200-member ensemble described in section 4) and analysis uncertainties from ERSSTv3b [*Smith et al.* 2008].

At a monthly, grid-box level, the parametric uncertainty in the *Kennedy et al.* [2011c] systematic error estimates is typically the smallest uncertainty and is nearly always less than 0.2 K. The sampling uncertainty and measurement uncertainty both depend on the number of observations, so they are larger in areas with fewer observations. Of the two, measurement uncertainty is typically larger.

In well-observed periods, the spread between the different analyses is roughly what one might expect: closer agreement in well-observed regions, poorer agreement in data-sparse regions, principally the Southern Ocean and Arctic Ocean. At more poorly-observed times, the spread between analyses is narrower than the climatological standard deviation suggesting that the reconstructions are skilful in the sense that they are providing useful information in data voids. However, the narrow spread is in contrast to those areas where there have been changes in the input observations (see, for example, the Indian Ocean in Figure 7(b) and Figure 7(h)). A small number of observations, which are available to one analysis but not another, lead to a larger spread than is seen in data-free regions implying that, while there is diversity in the approaches, there may still be too little for the best estimates alone to effectively bracket the true uncertainty range.

The ERSSTv3 analysis uncertainties are largest in regions where there are consistent data voids. They show a similar pattern to the structural uncertainty estimate in 1944 and 2003, but there is marked difference in 1891, with the analysis uncertainty being larger than the structural uncertainty in the poorly-observed western Pacific.

Figure 8 shows time series of the different components of uncertainty at different spatial scales from global to grid box. The bias uncertainty is relatively constant and is the smallest component of uncertainty at the grid box level for much of the record. The sampling uncertainty for a grid box is larger than the bias uncertainty when observations are few, but in the recent record they are comparable. In this example, the measurement uncertainty is larger than bias and sampling uncertainties at the grid box level, even when observations are numerous. However, in other grid boxes, characterised by strong SST gradients or high variability, such as the western boundary currents, the sampling uncertainty could be larger.

As the size of the area increases and more observations are included in the average, the sampling and measurement uncertainties decrease. Two estimates of the measurement uncertainty are included. In one, correlations between individual errors are taken into account. In the other, measurement errors are assumed to be random and independent. In the latter case, the measurement uncertainties become small relative to other sources of uncertainty at a basin scale early in the 20th century. However, the effect of correlated errors is such that measurement uncertainty remains a major source of uncertainty at all scales until the 1980s when the global VOS fleet reached its peak and the deployment of drifting and moored buoys began.

The largest component at the scales shown here is the structural uncertainty. In the grid box shown, the structural uncertainty is, at times, larger than the combined uncertainty from other components suggesting that some or all of the analyses are losing information. At a global level, where estimated analysis uncertainties are available for COBE, COBE-2, Kaplan and ERSSTv3b data sets, the structural uncertainty is comparable to the estimated analysis uncertainties. For example, in 1900, the ERSSTv3b analysis uncertainty is 0.03K, the COBE analysis uncertainty is 0.06K, COBE-2 gives 0.05K and Kaplan is around 0.05K.

Because of the nature of the uncertainties arising from the adjustments for pervasive systematic errors, the uncertainties become relatively more important as the averaging scale increases. At a global scale, bias uncertainties are comparable to or larger than all other uncertainty components from the 1940s to the present. There is a caveat: because the SSTs are expressed as anomalies, the size of the bias uncertainty depends on the base period used to calculate the anomalies. In Figure 8, the period used is 1961-1990, which is why there is a local minimum in the bias uncertainty centred on that period.

3.7 Estimates of Total Uncertainty

Smith and Reynolds [2005] attempted to combine all the different uncertainties described above to get a total uncertainty estimate. They combined their analysis uncertainty with measurement uncertainty, bias uncertainty and structural uncertainty. Uncertainty associated with pervasive systematic errors and structural uncertainty in the adjustments were estimated by taking the mean squared difference between the *Smith and Reynolds* [2002] and *Folland and Parker* [1995] bias adjustments in the prewar period. After World War 2, the bias uncertainty was estimated by calculating the average difference between engine room measurements and all measurements. Structural uncertainties were estimated by analysing the spread of three SST analyses.

Figure 9 shows the total uncertainty estimate from the latest version of the ERSST analysis, ERSSTv3b, in blue. A similar estimate was made based on the HadSST3 data set in the following way. Measurement uncertainties, grid-box sampling uncertainties and large-scale sampling uncertainties were estimated using the method of *Kennedy et al.* [2011b, 2011c]. To estimate the uncertainty associated with pervasive systematic errors,an ensemble of 200 data sets, comprising the 100 original ensemble members from HadSST3 and a 100-member ensemble generated by replacing the *Rayner et al.* [2006] bucket-correction fields with the fields from *Smith and Reynolds* [2002]. The adjustment uncertainties on individual months were assumed to be correlated within a year, giving a greater uncertainty range than in *Kennedy et al.* [2011c], particularly before 1941. During the war years 0.2 K was added to reflect the additional uncertainty during that period as described by *Kennedy et al.* [2011c]. As above, structural uncertainties were estimated by taking the standard deviation of area-average time series from seven analyses.

The total uncertainty estimates from these two assessments are comparable between 1880 and 1915. Between 1915 and 1941, the ERSSTv3b uncertainty estimate is larger because the estimated bias uncertainty is larger. The difference is most obvious in the northern hemisphere where the differences between the *Smith and Reynolds* [2002] and *Folland and Parker* [1995] bias adjustments are largest. From 1941 to present, the HadSST3-based uncertainty estimate is the larger because the bias uncertainty is larger than in ERSSTv3b.

The obvious question that arises is "do these assessments span the full uncertainty range?" In this case, it probably pays to err on the side of caution. Although the structural uncertainty is based on a range of methods for infilling missing data, there are still commonalities in the approaches taken and there is little diversity in the approaches to bias adjustment. The lack of diversity is troubling because the differences between the median estimates of HadSST3 and ERSSTv3b are greater than the estimated uncertainties of the ERSSTv3b analysis at times during the period 1950-1970 suggesting that the uncertainties may have been underestimated in the earlier assessment.

4 Presentation of Uncertainty

At present, some groups provide explicit uncertainty estimates based on their analysis techniques [*Kaplan et al.,* 1998; *Smith et al.*, 2008; *Kennedy et al.*, 2011b, 2011c, *Ishii et al.,* 2005; *Hirahara et al.,* 2013]. The uncertainty estimates derived from a particular analysis will tend to misestimate the true uncertainty because they rely on the analysis method and the assumptions on which it is based being correct.

Comparing uncertainty estimates provided with analyses can be difficult because not all analyses consider the same sources of uncertainties. Consequently, a narrower uncertainty range does not necessarily imply a better analysis. One way that data set providers could help users is to provide an inventory of sources of uncertainty that have been considered either explicitly or implicitly. This would allow users to assess the relative maturity of the uncertainty analysis.

There is a further difficulty in supplying and using uncertainty estimates: the traditional means of displaying uncertainties – the error bar, or error range – does not preserve the covariance structure of the uncertainties. Unfortunately, storing covariance information for all but the lowest resolution data sets can be prohibitively expensive. EOF-based analyses, like that of *Kaplan et al.* [1998], could in principle efficiently store the spatial-error covariances because only the covariances of the reduced space of principal components need to be kept. For *Kaplan et al.* [1998], based on a reduced space of only 80 EOFs, this is a matrix of order 80^2 elements for each time step as opposed to 1000^2 elements for the full-field covariance matrix. The difficulty with this approach is that not all variability can be resolved by the leading EOFs and excluding higher-order EOFs will underestimate the full uncertainty.

Karspeck et al. [2012] drew samples from the posterior probability produced by their analysis. Each sample provides an SST field that is consistent with the available observations and the estimated covariance structure. Sampling has the added advantage that it can be combined easily with Monte-Carlo samples from the measurement bias distributions. However, production of samples is not always computationally efficient. *Karspeck et al.* [2012] were able to do it for the North Atlantic region, but the computational costs of extending the analysis unchanged to the rest of the world could be prohibitive. *Kennedy et al.* [2011c] provided an ensemble of 100 interchangeable realizations of their bias-adjusted data set, HadSST3. The ensemble spans parametric uncertainties in their adjustment method.

By providing a set of plausible realizations of a data set, or alternatively by providing plausible realizations of typical measurement errors [*Mears et al.*, 2011], it can be relatively easy for users to assess the sensitivity of their analysis to uncertainties in SST data. For example, individual ensemble members of HadSST3 were used in *Tokinaga et al.* [2012], along with other SST analyses, to show that their results were robust to the estimated bias uncertainties in SSTs.

Another approach [*Merchant et al.* 2013] is to separate out components of the uncertainty that correlate at different scales. Random measurement errors, such as sensor noise, are uncorrelated. Some uncertainties, for example those related to water vapor in a satellite view, are correlated at a synoptic scale. Yet others are correlated at all times and places. Grouping uncertainties in this way allows users to propagate uncertainty information more easily.

5 Minimizing Exposure to Uncertainty

Alternative approaches to using the SST data in a way that is less sensitive to biases and other data errors have been made. The following approaches make use of knowledge concerning the types of errors and uncertainties found in SST data and have been adapted to account for them. They highlight the importance of combining understanding of the measurements and their potential errors, as well as understanding of the phenomenon being analyzed. Perhaps the simplest example is *Schell* [1959] who suggested discarding grid-box averages (in that case Marsden squares) based on small numbers of observations.

Thompson et al. [2008] identified an abrupt drop in the observed global average SST anomaly in late 1945, which they attributed to a rapid change in the composition of ICOADS 2.0 [*Worley et al.*, 2005] from mostly US ships immediately before the 1945 drop to mostly UK ships immediately afterwards. This hypothesis was lent further weight by *Kennedy et al.* [2011c]. In a follow-up paper [*Thompson et al.*, 2010], a drop in northern-hemisphere SSTs was identified. In order to show that the drop was not an artifact of the change in measurement method, they divided the ICOADS data into distinct subsets based on the country of the ships making the measurements, considered a range of different SST analyses, and looked at related variables such as NMAT and land surface air temperatures. The probability of a drop being due to a coincident change in the way that all countries measured SST, simultaneous with a sudden change in NMAT and land temperature bias, is small. The fact that the drop was seen in all the different data sets implied that the drop was real. *Tokinaga et al.* [2012] took a similar approach, using bucket measurements from ICOADS as a quasi-homogeneous estimate of SST change over the period 1950 to 2009.

In detection and attribution studies it is common to reduce the coverage of the models to match that of the data. Doing so reduces the exposure of the study to uncertainties associated with interpolation techniques, but it does not avoid the problem of systematic biases. Recent studies [*Jones and Stott*, 2011] have explicitly used a range of data sets to start to map out the effects of structural uncertainties on detection and attribution studies.

SST data sets are routinely compared to the output of climate simulations. Bearing in mind the discussion in section 2 on the definition of SST it might be necessary to ensure that the modeled output and the measured SST correspond to the same quantity. Many climate models employ a surface mixed layer that is several meters thick. However, models have been run with greater resolution in the near-surface ocean [e.g., *Bernie et al.,* 2008] in order to simulate diurnal variability.

Another common use of SST data for which an understanding of the limitations of the data is important is in the calculation and interpretation of EOFs. In many studies EOFs are calculated from globally complete SST analyses because the lack of missing data makes calculating EOFs easy. However, it seems wise to bear in mind that a good deal of statistical processing has already been applied to the SST analyses to make them globally complete. Extracting EOFs from (or applying any other analysis technique to) what are in some cases EOF analyses already, could lead to difficulties of interpretation on top of the more general problems [*Hannachi et al.,* 2007; *Dommenget,* 2007; *Karnauskas,* 2013]. Techniques exist for estimating EOFs from gridded data sets with missing data and these can also incorporate uncertainty information though many assume that the errors are uncorrelated and will tend to underestimate uncertainty in the EOFs and their principal components. See for example, *Roweis* [1998], *Schneider* [2001], *Beckers and Rixen* [2003], *Rutherford et al.* [2004], *Houseago-Stokes and Challenor* [2004], *Kondrashov and Ghil* [2006], *Ilin and Kaplan* [2009] and *Luttinen and Ilin* [*2009*].

6 Satellites

Although the present review is principally concerned with in situ measurements of SST it is necessary to mention the important role that satellite data play in understanding SST variability and uncertainty. The advantages of satellite data are obvious; particularly the ability to measure large areas of the ocean using a single instrument, giving a more nearly global view of SST.

However, the first thing to note is that satellites monitor radiances and do not directly measure SSTs. The measured radiances are affected by the state and constituents of the atmosphere as well as variations in the state and temperature of the sea-surface. The wavelengths that are sampled are set by the design of the instrument. Retrieving SST from the radiances is a difficult inverse process and sensitive to biases and other errors [*Merchant et al.* 2008b]. The second thing to note is that satellite instruments are sensitive to the skin (upper few microns), or sub-skin (upper few millimeters) temperature depending on the wavelengths measured by the satellite. Because satellite instruments are sensitive to the topmost layer of the ocean, the diurnal range of retrieved SSTs is larger than for measurements made at depth. Thirdly, accurate SST retrievals from infra-red instruments are only possible when the view is not obscured by cloud. Microwave retrievals can penetrate cloud, but suffer from problems near to coastlines, and where precipitation rates are high. They also have coarser spatial resolution and higher measurement uncertainties [*O'Carroll et al.*, 2008].

The longest records of SST from satellite are derived from the AVHRR (Advanced Very High Resolution Radiometer) instruments. These instruments make nadir measurements using two infra-red channels. The retrievals are usually calibrated relative to in situ data. More recent re-processings use optimal estimation to obtain a retrieval that is independent of the in situ record [*Merchant et al.*, 2008b] but these have not yet been extended to calculating global averages. Furthermore, the AVHRR instrument is prone to systematic errors caused by aerosols in the atmosphere and the satellite orbits drift slowly altering the sampling of the diurnal cycle through time. Despite the numerous shortcomings of the AVHRR record, *Good et al.* [2007] showed that there was a long-term warming trend in SSTs as measured by AVHRR.

The Along-Track Scanning Radiometers (ATSR) [*Smith et al.*, 2012] were designed to meet the needs of climate monitoring. The satellite is a dual view instrument, taking nadir and forward views using three infra-red channels. The dual view configuration allows for more effective screening of aerosols and the three channels allow for accurate retrievals across a wider range of conditions. Furthermore, the onboard calibration system allows the stability of the radiance measurements from the instrument to be maintained. The ATSR data have been reprocessed in the ATSR Reanalysis for Climate (ARC) project [*Merchant et al.*, 2008a] and the resulting time series have been shown to have biases of less than 0.1 K and stability better than 5 mK/year since 1993 in the tropics where reliable long term moorings can be found [*Embury et al.*, 2012; *Merchant et al.*, 2012]. The ARC reprocessing is almost independent of the in situ network therefore it can be used to corroborate trends seen in the in situ network. In a comparison between global average SST anomalies (at a nominal depth of 0.2 m) calculated using the ARC data and HadSST3, the two time series agree within the estimated HadSST3 uncertainties except for parts of the ATSR1 record in the early 1990s. The ATSR1 period is believed to be of lower quality as a result of the failure of one of the IR channels, failure of the satellite cooling system as well as the high stratospheric aerosol loadings following the eruption of Mount Pinatubo in 1991.

The nearly global, high-resolution view of the world's oceans provided by satellite instruments can be used as a way of improving and testing many aspects of SST analysis. By combining the more detailed fields produced by satellites with the long records of in situ measurements, more detailed reconstructions are possible over a wider area of the Earth [*Rayner et al.*, 2003; *Smith et al.*, 2008; *Hirahara et al.,* 2013]. Satellite data can also be used to assess the verisimilitude of reconstructions based on sparser in situ data.

7 Concluding Remarks and Future Directions

One of the chief difficulties in assessing the uncertainties in SST data sets is the impossibility of tracing individual observations back via an unbroken chain to international measurement standards. The creation of a global array of reference stations each making simultaneous redundant measurements of a variety of marine variables could solve some of the problems of SST analysis that have bedeviled the understanding of historical SST change and would provide a gold standard against which the *future* wider observing system – incorporating observations from ships, buoys, profiling floats and satellites – can be assessed. Even without such traceability a climate record could be more easily maintained by stricter adherence to the Global Climate Observing System [*GCOS* 2003] climate monitoring principles.

In the absence of such a network the estimation of uncertainties has depended heavily on redundancies in measurement systems and in analysis techniques. Full use of the redundancies is now being made in the modern period via comparisons of the many available satellite sources with each other and with in situ sources [*O'Carroll et al.*, 2008; *Merchant et al.*, 2012] and sub-surface data [*Gille*, 2012]. Analyses that ingest a variety of data sources can produce bias statistics for each of the inputs [*Brasnett*, 2008; *Xu and Ignatov*, 2010]. Such information can be exploited to assess their relative quality and, as the analyses are pushed further back in time [*Roberts-Jones et al.*, 2012], they will help assess uncertainties through a larger part of the record.

SSTs are physically related to other measurements including surface pressures and winds, salinity, air temperatures, sub-surface temperatures and ocean biology amongst others. Information from SST can be supplemented by analyses based on physical understanding of the climate system. It has already been shown that by combining information from night marine air temperatures with SST it was possible to greatly reduce uncertainties in early 20th and late 19th century SST. *Yu et al.* [2004] used a joint estimation method to minimize uncertainties in flux estimates based on a range of different variables mostly based on satellite data. Other studies [*Tung and Zhou*, 2010; *Deser et al.*, 2010] have used physical reasoning based on a host of variables to explore uncertainties in the long-term trends of tropical Pacific SSTs first raised by *Vecchi et al.* [2008]. It has even been suggested that proxy records such as isotope ratios from corals and ice cores could be used, with appropriate care, to understand uncertainties in the longest-term changes in SST [*Anderson et al.*, 2013]. The most advanced exemplars of physical and statistical synthesis are ocean and coupled reanalyses which will play an increasingly important role in understanding observational uncertainty and long-term climate change.

A key barrier to understanding SST uncertainty is a lack of appropriate metadata. Better information is needed concerning how measurements were made, which method was used to make a particular observation, calibration information, the depths at which observations were made, and even basic information such as the call sign or name of the ship that made a particular observation.

Some of this information can be inferred from data already contained in marine reports. Where reports in ICOADS cannot be associated with a particular ship, either because they have a missing ID, or a generic ID, there is much to be gained by grouping observations to give plausible ship tracks, or voyages. By using data association techniques to infer such metadata from the location information and other clues such as how frequently observations were made and which variables were observed, it should be possible to assess systematic and random errors on a ship-by-ship basis going back to the start of the record and even infer likely measurement methods based on characteristic variations of the measurements with the meteorological conditions.

A more systematic approach to the assessment of analysis techniques is needed to elucidate the reasons for the differences between analyses and to assess the verisimilitude of analysis uncertainty estimates. Approaches could include theoretical inter-comparisons of statistical methods, comparisons based on well-defined sets of common input observations, and benchmarks built from datasets (such as model output) where the truth is known a piori. Benchmark tests like those planned by the International Surface Temperature Initiative [*Thorne et al.* 2011b] provide an objective measure against which analysis techniques can be evaluated. Both analysis techniques and benchmarks will have to be tailored appropriately for the particular problems affecting SST measurements and the latest understanding of measurement uncertainties.

A key weakness of historical SST data sets is the lack of attention paid to evaluating the effects of data biases particularly in the post-1941 records. Further independent estimates of the biases produced need to be undertaken using as diverse a range of means as possible and the robust critique of existing methods must continue. Ideally, these would be complemented by carefully-designed field tests of buckets and other measurement methods.

Combining new analysis techniques that have been appropriately benchmarked with novel approaches to assessing uncertainty arising from systematic errors, pervasive systematic errors and their adjustments will give new end-to-end analyses that will help to explore the uncertainties in historical SSTs in a more systematic manner.

For long-term historical analyses, there is no substitute for actual observations and relevant metadata. Efforts to identify archives of marine observations and digitize them are ongoing [*Brohan et al.*, 2009; *Wilkinson et al.*, 2011]. Such programs are labor intensive, first in identifying and cataloguing the holdings in archives around the world, then in creating and storing digital images of the paper books and finally in keying the observations. The difficulty of decoding hand written entries in a variety of languages, formats and scripts means that optical character recognition technologies are of limited use. A number of popular crowd-sourcing projects have been started to key information from ships logs that have historical as well meteorological interest. OldWeather.org has keyed data from Royal Navy logs from the First World War [*Brohan et al.*, 2009] and is now working on logs from polar expeditions. Digitization of data also holds the possibility of extending instrumental records further back in time [*Brohan et al.*, 2010]. New observations, with reliable metadata, can be used not only to reduce uncertainty in SST analyses, but also to test the reliability of existing interpolated products and their uncertainties.

The ultimate destination of newly digitized observations is the International Comprehensive Ocean Atmosphere Data Set (ICOADS) [*Woodruff et al.*, 2011]. The ICOADS repository of marine meteorological data has long been the focus of advances in the understanding of marine climatology. It provides a consistent baseline for a wide range of studies, providing a solid basis for traceability and reproducibility. The continued existence, maintenance and improvement of ICOADS are essential to the future understanding of the global climate.

Finally, the work of identifying and quantifying uncertainties will be pointless, if those uncertainties are not used. Uncertainty estimates provided with data sets have sometimes been difficult to use or easy to use inappropriately. As pointed out by *Rayner et al.* [2009], "more reliable and user-friendly representations of uncertainty should be provided" in order to encourage their widespread and effective use.

Acknowledgements

The author was supported by the Joint DECC/Defra Met Office Hadley Centre Climate Programme (GA01101). Thanks to the TAO Project Office of NOAA/PMEL for providing the TAO array data used in Figure 2. Data from the Subduction array were downloaded from http://uop.whoi.edu/archives/dataarchives.html. I am especially grateful to Boyin Huang, Dave Berry, Alexander Ilin, Masayoshi Ishii and Feng Xu for providing additional data. Thanks also to Nick Rayner, David Parker, Chris Atkinson, Elizabeth Kent and John D. Kennedy for comments on the manuscript and Viktor Gouretski, Alexey Kaplan, Sarah Gille, Robin Matthews, Tony Brown and Greg Goodman for useful discussions. Thanks are also due to the four anonymous reviewers for their constructive comments.

Appendix A

Figure 1 was calculated in the following way. Observations were separated into three groups – shallow, deep and unknown – using the metadata assignments of *Kennedy et al.* [2011c]. Bucket and buoy measurements were considered to be shallow. Engine intake and hull contact measurements were considered to be deep. Shallow measurements were assumed to exhibit a diurnal cycle equal to that measured by drifting buoys [*Kennedy et al.*, 2007]. Deep measurements were assumed to have no diurnal cycle. The two groups were assumed to measure the same temperature just before sunrise. The relative bias between the two was calculated by subtracting the minimum of the diurnal cycle from the daily average. This value varies by location and calendar month. The bias in each grid box was estimated by multiplying the relative bias by the fraction of shallow measurements. The bias was then normalized relative to the period 1961-1990, the anomaly period used for HadSST3. Figure 1 shows the global monthly average of the bias.

References

Abraham, J.P. et al. (2013), A review of global ocean temperature observations: Implications for ocean heat content estimates and climate change. Reviews of Geophysics. doi: 10.1002/rog.20022

Amot, A. (1954), Measurements of sea surface temperature for meteorological purposes. Results of observations from ocean weather station M, Meteorologiske Annaler, 4(1), 1-11.

Atkinson, C.P., N.A. Rayner, J. Roberts-Jones, R.O. Smith (2013), Assessing the quality of sea surface temperature observations from drifting buoys and ships on a platform-by-platform basis. Journal of Geophysical Research - Oceans, 118, 3507–3529, doi:10.1002/jgrc.20257.

Anderson, D. M., E. M. Mauk, E. R. Wahl, C. Morrill, A. J. Wagner, D. Easterling, and T. Rutishauser (2013), Global warming in an independent record of the past 130 years. Geophysical Research Letters, 40, 189–193, doi:10.1029/2012GL054271.

Beckers, J. M., M. Rixen (2003), EOF Calculations and Data Filling from Incomplete Oceanographic Datasets. J. Atmos. Oceanic Technol., 20, 1839–1856. doi: 10.1175/1520-0426(2003)020<1839:ECADFF>2.0.CO;2

Beggs H., R. Verein, G. Paltoglou, H. Kippo and M. Underwood (2012), Enhancing ship of opportunity sea surface temperature observations in the Australian region, Journal of Operational Oceanography, (ISSN: 1755-8778), 5, 59-73.

Bernie, D.J., E. Guilyardi, G. Madec, J.M. Slingo, S.J. Woolnough and J. Cole (2008), Impact of resolving the diurnal cycle in an ocean–atmosphere GCM. Part 2: A diurnally coupled CGCM. Climate Dynamics, 31(7-8), 909-925

Bernstein, R., and D. Chelton (1985), Large-Scale Sea Surface Temperature Variability from Satellite and Shipboard Measurements, Journal of Geophysical Research, 90(C6), 11619-11630.

Berry, D.I. and E.C. Kent (2011), Air-Sea fluxes from ICOADS: the construction of a new gridded dataset with uncertainty estimates. International Journal of Climatology, 31(7) 987-1001 doi:10.1002/joc.2059

BIPM, I., IFCC, I., & ISO, I. (2008). IUPAP and OIML, Evaluation of measurement data-Guide to the expression of uncertainty in measurement. International Organization for Standardization (ISO), Online: http://www. bipm. org/en/publications/guides/gum. html.

Bottomley, M., C.K. Folland, J. Hsiung, R.E. Newell, and D.E. Parker (1990), Global Ocean Surface Temperature Atlas. 20 + iv pp. and 313 color plates, Her Majesty's Stn. Off., Norwich, UK, 1990.

Brasnett, B. (2008), The impact of satellite retrievals in a global sea-surface-temperature analysis. Quarterly Journal of the Royal Meteorological Society, 134(636), 1745-1760

Breaker, L.C., W.W. Broenkow, M.W. Denny, and L.V. Beatman (2005), Reconstructing an 83-Year Time Series of Daily Sea Surface Temperature at Pacific Grove, California. Moss Landing, CA, Moss Landing Marine Laboratories, http://aquaticcommons.org/id/eprint/3129

Brohan P., J.J. Kennedy, I. Harris, S.F.B. Tett and P.D. Jones (2006), Uncertainty estimates in regional and global observed temperature changes: a new dataset from 1850. Journal of Geophysical Research, 111, D12106, doi:10.1029/2005JD006548.

Brohan, P., R. Allan, J.E. Freeman, A.M. Waple, D. Wheeler, C. Wilkinson, S. Woodruff (2009), Marine Observations of Old Weather. Bulletin of the American Meteorological Society, 90, 219-230. doi: 10.1175/2008BAMS2522.1

Brohan, P., C. Ward, G. Willetts, C. Wilkinson, R. Allan and D. Wheeler (2010), Arctic marine climate of the early nineteenth century. Climates of the Past, 6, 315-324, doi:10.5194/cp-6-315-2010

Brooks, C. (1926), Observing water-surface temperatures at sea. Monthly Weather Review, 54(6), 241-253, doi:10.1175/1520-0493(1926)54<241:OWTAS>2.0.CO;2.

Brooks, C. (1928), Reliability of different methods of taking sea-surface temperature measurements. Journal of the Washington Academy of Sciences, 18, 525-545.

Cannaby, H., and Hüsrevoğlu, Y. S. (2009), The influence of low-frequency variability and long-term trends in North Atlantic sea surface temperature on Irish waters. – ICES Journal of Marine Science, 66: 1480-1489.

Castro, S. L., G. A. Wick, and W. J. Emery (2012), Evaluation of the relative performance of sea surface temperature measurements from different types of drifting and moored buoys using satellite-derived reference products. Journal of Geophysical Research, 117, C02029, doi:10.1029/2011JC007472.

Chiodi, A.M. and D.E. Harrison (2006), Summertime subtropical sea surface temperature variability. Geophysical Research Letters, 33, L08601, doi:10.1029/2005GL024524

Collins, C., L. Giovando and K. Abbott-Smith (1975), Comparison of Canadian and Japanese merchant-ship observations of sea-surface temperature in the vicinity of present ocean station P 1927-33. Canadian Journal of Fisheries and Aquatic Science, 32(2), 253-258, doi:10.1139/f75-023.

Compo G.P., P.D. Sardeshmukh, J.S. Whitaker, P. Brohan, P.D. Jones and C. McColl (2013), Independent confirmation of global land warming without the use of station temperatures. Geophysical Research Letters doi: 10.1002/grl.50425

Cummings, J.A. (2005), Operational multivariate ocean data assimilation. Quarterly Journal of the Royal Meteorological Society, 131: 3583–3604. doi: 10.1256/qj.05.105

de Boyer Montégut, C., G. Madec, A.S. Fischer, A. Lazar, and D. Iudicone (2004), Mixed layer depth over the global ocean: an examination of profile data and a profile-based climatology. Journal of Geophysical Research, 109, C12003. doi:10.1029/2004JC002378

Deser C., A.S. Phillips and M.A. Alexander (2010), Twentieth century tropical sea surface temperature trends revisited. Geophysical Research Letters, 37, L10701, doi: 10.1029/2010GL043321.

Dommenget, D. (2007), Evaluating EOF modes against a stochastic null hypothesis. Climate Dynamics, 28(5), 517-531.

Donlon, C., et al. (2007), The Global Ocean Data Assimilation Experiment High-resolution Sea Surface Temperature Pilot Project. Bulletin of the American Meteorological Society, 88, 1197–1213. doi: 10.1175/BAMS-88-8-1197

Emery, W. J., K. Cherkauer, B. Shannon and R.W. Reynolds (1997), Hull-Mounted Sea Surface Temperatures from Ships of Opportunity. Journal of Atmospheric and Oceanic Technology, 14, 1237–1251. doi: http://dx.doi.org/10.1175/1520-0426(1997)014<1237:HMSSTF>2.0.CO;2

Embury, O., C.J. Merchant and G.K. Corlett (2012), A Reprocessing for Climate of Sea Surface Temperature from the Along-Track Scanning Radiometers: Initial validation, accounting for skin and diurnal variability. Remote Sensing of Environment, 116(15) 62-78. DOI:10.1016/j.rse.2011.02.028

Emery, W., D. Baldwin, P. Schlüssel and R. Reynolds (2001), Accuracy of in situ sea surface temperatures used to calibrate infrared satellite measurements. Journal of Geophysical Research, 106(C2), 2387-2405, doi:10.1029/2000JC000246.

Folland, C., R. Reynolds, M. Gordon and D. Parker (1993), A study of six operational sea surface temperature analyses. Journal of Climate, 6(1), 96-113, doi:10.1175/1520-0442(1993)006<0096:ASOSOS>2.0.CO;2.

Folland, C.K. and D.E. Parker (1995), Correction of instrumental biases in historical sea surface temperature data. Quarterly Journal of the Royal Meteorological Society 121: 319-367.

Folland C.K., N.A. Rayner, S. J. Brown, T.M. Smith, S.S.P. Shen, D.E. Parker, I. Macadam, P.D. Jones, R.N. Jones, N. Nicholls and D.M.H. Sexton (2001), Global temperature change and its uncertainties since 1861, Geophysical Research Letters, 28(13), 2621–2624.

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. K. and Salinger, M. J. (1995), Surface temperature trends and variations in New Zealand and the surrounding ocean, 1871–1993. International Journal of Climatology, 15: 1195–1218. doi: 10.1002/joc.3370151103

GCOS (2003), The second report on the adequacy of the Global Observing Systems for Climate in support of the UNFCCC. GCOS–82, WMO Tech. Doc. 1143, 85 pp. [Available online at www.wmo.int/pages/prog/gcos/Publications/gcos-82_2AR.pdf.]

Gilhousen, D.B. (1987), A field evaluation of NDBC moored buoy winds. Journal of Atmospheric and Oceanic Technology, 4(1), 94-104

Gille, S.T. (2012), Diurnal variability of upper ocean temperatures from microwave satellite measurements and Argo profiles. Journal of Geophysical Research, 117, C11027, doi:10.1029/2012JC007883.

Good, S.A., G.K. Corlett, J.J. Remedios, E.J. Noyes, and D.T. Llewellyn-Jones (2007), The global trend in sea surface temperature from 20 years of Advanced Very High Resolution Radiometer data. Journal of Climate, 20(7), 1255-1264 doi:10.1175/JCLI4049.1

Gouretski, V., J.J. Kennedy, T. Boyer, and A. Köhl (2012), Consistent near-surface ocean warming since 1900 in two largely independent observing networks. Geophysical Research Letters, 39, L19606, doi:10.1029/2012GL052975.

Grodsky, S.A., J.A. Carton and H. Liu (2008), Comparison of bulk sea surface and mixed layer temperatures. Journal of Geophysical Research 113, C10026, doi:10.1029/2008JC004871.

Hanawa, K., S. Yasunaka, T. Manabe, and N. Iwasaka (2000), Examination of correction to historical SST data using long-term coastal SST data taken around Japan. Journal of the Meteorological Society of Japan, 78, 187-195.

Hannachi A., I.T. Jolliffe, and D.B. Stephenson (2007), Empirical orthogonal functions and related techniques in atmospheric science: A review. International Journal of Climatology, 27:1119-1152.

Houseago-Stokes, R.E., and P.G. Challenor (2004), Using PPCA to Estimate EOFs in the Presence of Missing Values. J. Atmos. Oceanic Technol., 21, 1471–1480. doi: 10.1175/1520-0426(2004)021<1471:UPTEEI>2.0.CO;2

Hurrell, J.W. and K.E. Trenberth (1999), Global Sea Surface Temperature Analyses: Multiple Problemsand Their Implications for Climate Analysis, Modeling, and Reanalysis. Bulletin of the American Meteorological Society, 80, 2661–2678. doi: http://dx.doi.org/10.1175/1520-0477(1999)080<2661:GSSTAM>2.0.CO;2

Ilin A. and A. Kaplan (2009), Bayesian PCA for Reconstruction of Historical Sea Surface Temperatures. Proceedings of the International Joint Conference on Neural Networks (IJCNN 2009), pp. 1322-1327, Atlanta, USA, 2009.

Ingleby, B. (2010), Factors Affecting Ship and Buoy Data Quality: A Data Assimilation Perspective. Journal of Atmospheric and Oceanic Technology 27:9, 1476-1489

Ishii, M., M. Kimoto, and M. Kachi (2003), Historical Ocean Subsurface Temperature Analysis with Error Estimates. Mon. Wea. Rev., 131, 51–73. doi: 10.1175/1520-0493(2003)131<0051:HOSTAW>2.0.CO;2

Ishii, M., A. Shouji, S. Sugimoto, and T. Matsumoto (2005), Objective analyses of sea-surface temperature and marine meteorological variables for the 20th century using ICOADS and the Kobe collection. International Journal of Climatology, 25(7), 865-879, doi:10.1002/joc.1169.

James, R., and P. Fox (1972), Comparative sea surface temperature measurements in WMO reports on marine science affairs, rep 5, Tech. Rep. 336, WMO.

Jones, G.S., and P.A. Stott (2011), Sensitivity of the attribution of near surface temperature warming to the choice of observational dataset. Geophysical Research Letters, 38, L21702, doi:10.1029/2011GL049324.

Jones, P.D. (1994), Hemispheric Surface Air Temperature Variations: A Reanalysis and an Update to 1993. Journal of Climate, 7, 1794–1802. doi: 10.1175/1520-0442(1994)007<1794:HSATVA>2.0.CO;2

Jones, P.D., T.J. Osborn and K.R. Briffa (1997), Estimating Sampling Errors in Large-Scale Temperature Averages. Journal of Climate, 10, 2548–2568. doi: 10.1175/1520-0442(1997)010<2548:ESEILS>2.0.CO;2

Jones, P.D. and T.M.L. Wigley (2010), Estimation of global temperature trends: What's important and what isn't. Climatic Change, 100 (1). pp. 59-69.

Kaplan, A., M. Cane, Y. Kushnir, A. Clement, M. Blumenthal, and B. Rajagopalan (1998), Analyses of global sea surface temperature 1856-1991. Journal of Geophysical Research, 103(C9), 18,567-18,589, doi:10.1029/97JC01736

Karnauskas, K.B. (2013), Can we distinguish canonical El Niño from Modoki? Geophysical Research Letters, doi: 10.1002/grl.51007

Karspeck, A.R., A. Kaplan, and S.R. Sain, (2012), Bayesian modelling and ensemble reconstruction of mid-scale spatial variability in North Atlantic sea-surface temperatures for 1850-2008. Quarterly Journal of the Royal Meteorological Society, 138, 234-248. doi: 10.1002/qj.900

Kawai, Y. and H. Kawamura, (2000), Study on a Platform Effect in the In Situ Sea Surface Temperature Observations under Weak Wind and Clear Sky Conditions Using Numerical Models. Journal of Atmospheric and Oceanic Technology, 17, 185–196. doi: 10.1175/1520-0426(2000)017<0185:SOAPEI>2.0.CO;2

Kawai, Y. and A. Wada (2007), Diurnal sea surface temperature variation and its impact on the atmosphere and ocean: a review. Journal of Oceanography, 2007, 63: 721-744.

Kennedy, J.J., P. Brohan and S.F.B. Tett (2007), A global climatology of the diurnal variations in sea-surface temperature and implications for MSU temperature trends. Geophysical Research Letters, 34(5), L05712 doi:10.1029/2006GL028920

Kennedy, J.J., R. Smith, and N. Rayner (2011a), Using AATSR data to assess the quality of in situ sea surface temperature observations for climate studies. Remote Sensing of Environment. 116, 79–92 http://dx.doi.org/10.1016/j.rse.2010.11.021

Kennedy, J.J., N.A. Rayner, R.O. Smith, M. Saunby and D.E. Parker (2011b), Reassessing biases and other uncertainties in sea-surface temperature observations since 1850 part 1: measurement and sampling errors. Journal of Geophysical Research, 116, D14103, doi:10.1029/2010JD015218

Kennedy, J.J., N.A. Rayner, R.O. Smith, M. Saunby and D.E. Parker (2011c), Reassessing biases and other uncertainties in sea-surface temperature observations since 1850 part 2: biases and homogenisation. Journal of Geophysical Research, 116, D14104, doi:10.1029/2010JD015220

Kent, E., P. Taylor, B. Truscott, and J. Hopkins (1993), The accuracy of Voluntary Observing Ships' meteorological observations - results of the VSOP-NA. Journal of Atmospheric and Oceanic Technology, 10(4), 591-608, doi:10.1175/1520-0426(1993)010<0591:TAOVOS>2.0.CO;2.

Kent, E.C., P.G. Challenor and P.K. Taylor (1999), A statistical determination of the random observational errors present in voluntary observing ships meteorological reports. Journal of Atmospheric and Oceanic Technology, 16(7), 905-914

Kent, E.C. and D.I. Berry (2005), Quantifying random measurement errors in Voluntary Observing Ships' meteorological observations. International Journal of Climatology, 25(7), 843-856, doi:10.1002/joc.1167

Kent, E., and P. Challenor (2006), Toward estimating climatic trends in SST. Part II: Random errors. Journal of Atmospheric and Oceanic Technology, 23(3), 476-486, doi:10.1175/JTECH1844.1.

Kent, E., and A. Kaplan (2006), Toward estimating climatic trends in SST. Part III: Systematic biases. Journal of Atmospheric and Oceanic Technology, 23(3), 487-500, doi:10.1175/JTECH1845.1.

Kent, E.C., S.D. Woodruff and D.I. Berry (2007), Metadata from WMO Publication No. 47 and an Assessment of Voluntary Observing Ship Observation Heights in ICOADS.

Journal of Atmospheric and Oceanic Technology, 24, (2), 214-234. doi:10.1175/JTECH1949.1

Kent, E., and D. Berry (2008), Assessment of the marine observing system (ASMOS): Final report, Tech. Rep. 32, Natl. Oceanogr. Cent., Southampton, U. K.

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

Kirk, T., and A. Gordon (1952), Comparison of intake and bucket methods for measuring sea temperature. Marine Observer, 22, 33-39.

Knudsen, J. (1966), An experiment in measuring the sea surface temperature for synoptic purposes. Tech. Rep. 12, Det. Norske Meteor. Inst.

Kondrashov, D. and M. Ghil (2006), Spatio-temporal filling of missing points in geophysical data sets, Nonlin. Processes Geophys., 13, 151-159, doi:10.5194/npg-13-151-2006.

Lindau, R. (2003), Errors of Atlantic air-sea fluxes derived from ship observations. Journal of Climate, 16, 783–788.

Lumby, J. (1927), The surface sampler, an apparatus for the collection of samples from the sea surface from ships in motion with a note on surface temperature observations. J. Cons. Perm. Int. Explor. Mer., 2, 332-342.

Luttinen, J. and A. Ilin. (2009), Variational Gaussian-Process Factor Analysis for Modeling Spatio-Temporal Data. In Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS 2009), Vancouver, Canada, 2009.

Luttinen J. and A. Ilin. (2012), Efficient Gaussian Process Inference for Short-Scale Spatio-Temporal Modeling. Accepted to the 15th International Conference on Artificial Intelligence and Statistics (AISTATS 2012).

Lyman, J.M., S.A. Good, V.V. Gouretski, M. Ishii, G.C. Johnson, M.D. Palmer, D.M. Smith and J.K. Willis (2010), Robust warming of the global upper ocean. Nature 465, 334-337 doi:10.1038/nature09043

MacKenzie B.R., and D. Schiedek (2007), Long-term sea surface temperature baselines—time series, spatial covariation and implications for biological processes, Journal of Marine Systems, 68(3–4), pp 405-420 doi: 10.1016/j.jmarsys.2007.01.003.

Matthews, J.B.R. (2013), Comparing historical and modern methods of sea surface temperature measurement – Part 1: Review of methods, field comparisons and dataset adjustments, Ocean Sci., 9, 683-694, doi:10.5194/os-9-683-2013.

Matthews, J.B.R., and Matthews, J.B. (2013), Comparing historical and modern methods of sea surface temperature measurement – Part 2: Field comparison in the central tropical Pacific, Ocean Sci., 9, 695-711, doi:10.5194/os-9-695-2013.

Maul, G.A., A.M. Davis, and J.W. Simmons (2001), Seawater temperature trends at USA Tide Gauge sites. Geophysical Research Letters, 28: 3935–3937. doi: 10.1029/2001GL013458

Mears, C. A., F.J. Wentz, P. Thorne and D. Bernie (2011), Assessing uncertainty in estimates of atmospheric temperature changes from MSU and AMSU using a Monte-Carlo estimation technique. Journal of Geophysical Research, 116, D08112, doi:10.1029/2010JD014954.

Merchant, C., D. Llewellyn-Jones, R. Saunders, N. Rayner, E. Kent, C. Old, D. Berry, A. Birks, T. Blackmore and G. Corlett (2008a), Deriving a sea surface temperature record suitable for climate change research from the along-track scanning radiometers. Advances in Space Research 41 (1), 1–11.

Merchant C.J., P. Le Borgne, A. Marsouin and H. Roquet (2008b), Optimal estimation of sea surface temperature from split-window observations. Remote Sensing of Environment, 112, 2469–2484

Merchant, C.J., O. Embury, N.A. Rayner, D.I. Berry, G. Corlett, K. Lean, K.L. Veal, E.C. Kent, D. Llewellyn-Jones, J.J. Remedios, and R. Saunders (2012), A twenty-year independent record of sea surface temperature for climate from Along Track Scanning Radiometers . Journal of Geophysical Research, 117, C12013, doi:10.1029/2012JC008400.

Merchant, C.J., S. Matthiesen, N.A. Rayner, J.J. Remedios, P.D. Jones, F. Olesen, B. Trewin, P.W. Thorne, R. Auchmann, G.K. Corlett, P.C. Guillevic, and G.C. Hulley (2013), The surface temperatures of the earth: steps towards integrated understanding of variability and change, Geosci. Instrum. Method. Data Syst. Discuss., 3, 305-345, doi:10.5194/gid-3-305-2013.

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

Morrissey, M., and J. Greene (2009), A theoretical framework for the sampling error variance for three-dimensional climate averages of ICOADS monthly ship data. Theoretical and Applied Climatology, 96(3-4), 235-248, doi:10.1007/s00704-008-0027-3

Moyer, K.A., and R.A. Weller (1997), Observations of Surface Forcing from the Subduction Experiment: A Comparison with Global Model Products and Climatological Datasets. Journal of Climate, 10, 2725–2742. doi: http://dx.doi.org/10.1175/1520-0442(1997)010<2725:OOSFFT>2.0.CO;2

Nixon, S.W., S. Granger, B.A. Buckley, M. Lamont, B. Rowell (2004), A one hundred and seventeen year coastal water temperature record from Woods Hole, Massachusetts. Estuaries 27(3), 397-404, doi: 10.1007/BF02803532

O'Carroll, A.G., J.R. Eyre and R.W. Saunders (2008), Three-way error analysis between AATSR, AMSR-E and in situ sea surface temperature observations. Journal of Atmospheric and Oceanic Technology, 25(7), 1197-1207

Palmer, M. D., K. Haines, S.F.B. Tett and T.J. Ansell (2007), Isolating the signal of ocean global warming, Geophysical Research Letters, 34, L23610, doi:10.1029/2007GL031712.

Palmer, M.D. and P. Brohan (2011), Estimating sampling uncertainty in fixed-depth and fixed-isotherm estimates of ocean warming. International Journal of Climatology, 31(7), 980–986, doi: 10.1002/joc.2224

Parker, D.E. (1987), The sensitivity of estimates of trends of global and hemispheric marine temperatures to limitations in geographical coverage. Long Range Forecasting and Climate Research Memo LRFC 12.

Perlroth, I. (1962), Relationship of central pressure of hurricane Esther (1961) and the sea surface temperature field. Tellus, 14: 403–408. doi: 10.1111/j.2153-3490.1962.tb01353.x

Prytherch, J., J.T. Farrar, and R.A. Weller (2013), Moored surface buoy observations of the diurnal warm layer. Journal of Geophysical Research (in press) doi: 10.1002/jgrc.20360

Rayner, N.A., D.E. Parker, E.B. Horton, C.K. Folland, L.V. Alexander, D.P. Rowell, E.C. Kent and A. Kaplan (2003), Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. Journal of Geophysical Research, Vol. 108, No. D14, 4407 10.1029/2002JD002670

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

Rayner, N., et al. (2009), Evaluating climate variability and change from modern and historical SST observations, in Proceedings of OceanObs'09: Sustained Ocean Observations and Information for Society, vol. 2, Venice, Italy, 21-25 September 2009, ESA Publ. WPP-306, edited by J. Hall, D. Harrison, and D. Stammer, ESA, Paris, doi:10.5270/OceanObs09.cwp.71

Reverdin, G., J. Boutin, N. Martin, A. Lourenco, P. Bouruet-Aubertot, A. Lavin, J. Mader, P. Blouch, J. Rolland, F. Gaillard and P. Lazure (2010), Temperature Measurements from Surface Drifters. Journal of Atmospheric and Oceanic Technology, 27, 1403-1409. doi: 10.1175/2010JTECHO741.1

Reynolds, R.W., N.A. Rayner, T.M. Smith, D.C. Stokes and W.Q. Wang (2002), An improved in situ and satellite SST analysis for climate. Journal of Climate, 15(13), 1609-1625

Reynolds, R.W., C.L. Gentemann and G.K. Corlett, (2010), Evaluation of AATSR and TMI Satellite SST Data. Journal of Climate, 23, 152–165. doi: 10.1175/2009JCLI3252.1

Roberts-Jones, J., E.K. Fiedler and M.J. Martin (2012), Daily, Global, High-Resolution SST and Sea Ice Reanalysis for 1985–2007 Using the OSTIA System. Journal of Climate, 25, 6215–6232. doi: 10.1175/JCLI-D-11-00648.1

Roll, H. (1951), Water temperature measurements on deck and in the engine room. Ann. Meteor., 4, 439-443.

Roweis, S. (1998), EM Algorithms for PCA and SPCA. Neural Information Processing Systems 10 (NIPS'97) pp.626-632

Rutherford, S., M.E. Mann, T.L. Delworth, and R.J. Stouffer (2003), Climate Field Reconstruction under Stationary and Nonstationary Forcing. *J. Climate*, **16**, 462–479. doi: 10.1175/1520-0442(2003)016<0462:CFRUSA>2.0.CO;2

Sarachik, E.S., (1984), Large-scale surface heat fluxes. Large-Scale Oceanographic Experiments and Satellites, C Gautier and M. Fieux, Eds., Reidel, 147–165.

Saur, J. (1963), A study of the quality of sea water temperatures reported in the logs of ships' weather observations. Journal of Applied Meteorology, 2(3), 417-425, doi:10.1175/1520-0450(1963)002<0417:ASOTQO>2.0.CO;2.

Schell, I.I. (1959), On a criterion of representativeness of sea-surface data. Bull. Amer. Met. Soc. 40(11) pp 571-574.

Schneider, T. (2001), Analysis of Incomplete Climate Data: Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values. J. Climate, 14, 853–871. doi:10.1175/1520-0442(2001)014<0853:AOICDE>2.0.CO;2

She, J., J.L. Høyer and J. Larsen (2007), Assessment of sea surface temperature observational networks in the Baltic Sea and North Sea. Journal of Marine Systems, 65(1–4), 314–335. doi:10.1016/j.jmarsys.2005.01.004,

Shen, S.S.P., H. Yin and T.M. Smith, (2007), An Estimate of the Sampling Error Variance of the Gridded GHCN Monthly Surface Air Temperature Data. Journal of Climate, 20, 2321-2331. doi: 10.1175/JCLI4121.1

Sheppard, C.R.C. and N.A. Rayner (2002), Utility of the Hadley Centre sea-Ice and Sea Surface Temperature data set (HadISST1) in two widely contrasting coral reef areas. Marine Pollution Bulletin, 44 303 - 308 (0025-326X)

Simmons, A.J., K.W. Willett, P.D. Jones, P.W. Thorne and D. Dee (2010), Low-frequency variations in surface atmospheric humidity, temperature and precipitation: Inferences from reanalyses and monthly gridded observation datasets. Journal of Geophysical Research - Atmospheres.115, D01110, doi:10.1029/2009JD012442.

Smith, D.M., S. Cusack, A.W. Colman, C.K. Folland, G.R. Harris, and J.M. Murphy (2007), Improved Surface Temperature Prediction for the Coming Decade from a Global Climate Model. Science 317 (5839), 796-799. doi:10.1126/science.1139540

Smith, D., C. Mutlow, J. Delderfield, B. Watkins, G. Mason (2012), ATSR infrared radiometric calibration and in-orbit performance. Remote Sensing of Environment 116, 4–16, doi:10.1016/j.rse.2011.01.027

Smith, T., and R. Reynolds (2002), Bias corrections for historical sea surface temperatures based on marine air temperatures. Journal of Climate, 15(1), 73.

Smith, T.M. and R.W. Reynolds, (2003), Extended Reconstruction of Global Sea Surface Temperatures Based on COADS Data (1854-1997). Journal of Climate, 16, 1495-1510.

Smith, T.M., and R.W. Reynolds, (2004), Improved Extended Reconstruction of SST (1854-1997). Journal of Climate, 17, 2466-2477.

Smith, T.M., R.W. Reynolds, (2005), A Global Merged Land-Air-Sea Surface Temperature Reconstruction Based on Historical Observations (1880-1997). Journal of Climate, 18, 2021-2036. doi: 10.1175/JCLI3362.1

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

Stevenson, R.E. (1964), The Influence of a Ship on the Surrounding Air and Water Temperatures. Journal of Applied Meteorology, 3, 115–118. doi: 10.1175/1520-0450(1964)003<0115:TIOASO>2.0.CO;2

Strong, A.E. and E.P. McClain (1984), Improved Ocean Surface Temperatures From Space-Comparisons With Drifting Buoys. Bulletin of the American Meteorological Society, 65, 138–142. doi: 10.1175/1520-0477(1984)065<0138:IOSTFS>2.0.CO;2

Stubbs, M.W. (1965), The standard error of a sea surface temperature as measured using a canvas bucket. The Meteorological magazine 94(1112).

Tabata, S. (1978a), On the accuracy of sea-surface temperatures and salinities observed in the Northeast Pacific Ocean. Atmosphere Ocean, 16(3), 237-247.

Tabata, S. (1978b), Comparison of observations of sea-surface temperatures at ocean staion P and NOAA buoy stations and those made by merchant ships travelling in their vicinities in the northeast Pacific Ocean. Journal of Applied Meteorology, 17(3), 374-385, doi:10.1175/1520-0450(1978)017<0374:COOOSS>2.0.CO;2.

Tauber, G. (1969), The comparative measurements of sea surface temperature in the USSR, Tech. Rep. 103, WMO.

Thompson, D.W.J., J.J. Kennedy, J.M. Wallace and P.D. Jones (2008), A large discontinuity in the mid-twentieth century in observed global-mean surface temperature. Nature 453, 646-649

Thompson, D.W.J., J.M. Wallace, J.J. Kennedy and P.D. Jones (2010), An abrupt drop in Northern Hemisphere sea surface temperature around 1970. Nature 467, 444-447, doi:10.1038/nature09394

Thorne, P.W., D.E. Parker, J.R. Christy and C.A. Mears (2005), Uncertainties in climate trends: Lessons from upper-air temperature records. Bulletin of the American Meteorological Society, 86, 1437–1442.

Thorne, P.W., J.R. Lanzante, T.C. Peterson, D.J. Seidel, K.P. Shine (2011), Tropospheric temperature trends: history of an ongoing controversy. WIREs Climate Change 2(1) 66-88 doi:10.1002/wcc.80

Thorne, P.W., and Coauthors, (2011b), Guiding the Creation of A Comprehensive Surface Temperature Resource for Twenty-First-Century Climate Science. Bull. Amer. Meteor. Soc., 92, ES40–ES47. doi: http://dx.doi.org/10.1175/2011BAMS3124.1

Tokinaga, H., S. Xie, C. Deser, Y. Kosaka and Y.M. Okumura (2012), Slowdown of the Walker circulation driven by tropical Indo-Pacific warming. Nature 491, 439–443. doi:10.1038/nature11576

Tung, K. and J. Zhou, (2010), The Pacific's Response to Surface Heating in 130 Yr of SST: La Niña–like or El Niño–like?. Journal of Atmospheric Science, 67, 2649–2657. doi: 10.1175/2010JAS3510.1

Vecchi G.A., A. Clement, and B.J. Soden (2008), Examining the Tropical Pacific's Response to Global Warming. Eos, Vol. 89, No. 9, 26 February 2008

Venema, V.K.C., O. Mestre, E. Aguilar, I. Auer, J.A. Guijarro, P. Domonkos, G. Vertacnik, T. Szentimrey, P. Stepanek, P. Zahradnicek, J. Viarre, G. Müller-Westermeier, M. Lakatos, C.N. Williams, M.J. Menne, R. Lindau, D. Rasol, E. Rustemeier, K. Kolokythas, T. Marinova, L. Andresen, F. Acquaotta, S. Fratianni, S. Cheval, M. Klancar, M. Brunetti, C. Gruber, M. Prohom Duran, T. Likso, P. Esteban and T. Brandsma (2012), Benchmarking homogenization algorithms for monthly data. Climates of the Past, 8, 89-115, doi:10.5194/cp-8-89-2012

Wahl, E. (1948), Water temperature measurements on deck and in the engine room. Ann. Meteor., 1(7).

Walden, H. (1966), On water temperature measurements aboard merchant vessels (in German). Ocean Dynamics, 19, 21-28, doi:10.1007/BF02321345.

Weare, B.C. (1989), Uncertainties in estimates of surface heat fluxes derived from marine reports over the tropical and subtropical oceans. Tellus A, 41A: 357–370. doi: 10.1111/j.1600-0870.1989.tb00388.x

Weare, B.C. and Strub, P.T. (1981), The significance of sampling biases on calculated monthly mean oceanic surface heat fluxes. Tellus, 33: 211–224. doi: 10.1111/j.2153-3490.1981.tb01745.x

Wilkerson, J.C., and M. D. Earle (1990), A study of differences between environmental reports by ships in the voluntary observing program and measurements from NOAA buoys. Journal of Geophysical Research, 95(C3), 3373–3385, doi:10.1029/JC095iC03p03373.

Wilkinson, C., S.D. Woodruff, P. Brohan, S. Claesson, E. Freeman, F. Koek, S.J. Lubker, C. Marzin and D. Wheeler (2011), Recovery of logbooks and international marine data: the RECLAIM project. International Journal of Climatology, 31(7) 968-979 doi:10.1002/joc.2102

Woodruff, S.D., S.J. Worley, S.J. Lubker, Z. Ji, E. Freeman, D.I. Berry, P. Brohan, E.C. Kent, R.W. Reynolds, S.R. Smith and C. Wilkinson (2011), ICOADS Release 2.5:

extensions and enhancements to the surface marine meteorological archive. International Journal of Climatology. 31(7) 951-967 doi:10.1002/joc.2103

Worley, S.J., S.D. Woodruff, R.W. Reynolds, S.J. Lubker, and N. Lott, 2005: ICOADS Release 2.1 data and products. International Journal of Climatology. (CLIMAR-II Special Issue), 25, 823-842 doi:10.1002/joc.1166

Xu, F., and A. Ignatov (2010), Evaluation of in situ sea surface temperatures for use in the calibration and validation of satellite retrievals. Journal of Geophysical Research, 115, C09022, doi:10.1029/2010JC006129.

Yasunaka, S. and K. Hanawa (2002), Regime shifts found in the Northern Hemisphere SST field. Journal of the Meteorological Society of Japan 80: 119-135.

Yasunaka, S. and K. Hanawa (2011), Intercomparison of historical sea surface temperature datasets. International Journal of Climatology, 31(7) 1056-1073 doi:10.1002/joc.2104

Yu, L., R.A. Weller, and B. Sun, 2004: Improving latent and sensible heat flux estimates for the Atlantic Ocean (1988-1999) by a synthesis approach. J. Climate, 17, 373–393

References	Estimated measurement uncertainty for ship measurements
Stubbs [1965]	0.11±0.01K for canvas bucket measurements from an Ocean Weather Ship
Strong and McLean [1984]	1.8K RMS difference between ship and AVHRR data
Bernstein and Chelton [1985] pg 11620	1.1 K
Sarachik [1984], *Weare* [1989] pg 359	1 K
Wilkerson and Earle [1990] pg 3381	3.5 K
Cummings [2005] Table 1, pg 3592	1.3 K (ERI) 0.6 K (Hull sensor) 1.2 K (bucket)
Kent and Challenor [2006] pg 484	1.2±0.4 K or 1.3±0.3 K depending on how measurements were weighted
Kent et al. [1999] abstract	1.5±0.1 K
Kent and Berry [2005] Table 2 pg 853	1.3±0.1 K and 1.2±0.1 K
Reynolds et al. [2002] pg 1613	1.3 K
Kennedy et al. [2011a] pg 83	1.0 K
Ingleby [2010] Table 10 pg 1487	0.9 K for automatic systems 1.2 K for manual measurements
Kent and Berry [2008] Table 5a pg 11	1.1 K
Xu and Ignatov [2010] pg 16 of 18	1.16 K

Table 1: List of estimates of measurement error uncertainties for ships where random and systematic errors were not dealt with separately.

References	Estimated measurement uncertainty for drifting buoy measurements
Strong and Mclean [1984]	0.6K RMS difference between drifter and AVHRR
Reynolds et al. [2002] pg 1613	0.5 K
Emery et al. [2001] pg 2393	0.3 K
Cummings [2005] Table 1, pg 3592	0.12 K
O'Carroll et al. [2008] abstract	0.23 K
Kent and Berry [2008] Table 5c pg 12	0.67 K
Ingleby [2010] Table 10 pg 1487	0.33 K
Kennedy et al. [2011a] pg 83	0.2-0.4 K
Xu and Ignatov [2010] pg 16 of 18	0.26 K
Merchant et al. [2012] Table 2 pg 8 of 18	0.15-0.19 K

Table 2: List of estimates of measurement error uncertainties for drifting buoys where random and systematic errors were not dealt with separately.

Reference	Estimated measurement uncertainty for moored buoy measurements
Cummings [2005] Table 1, pg 3592	0.05 K
Kent and Berry [2008] Table 5b pg 11	0.4 K

Kennedy et al. [2011a] pg 83	tropical moorings, 0.12 K; all moorings, 0.21 K
Xu and Ignatov [2010] pg 16 of 18	tropical moorings, 0.30 K; coastal moorings, 0.39 K
Gilhousen [1987] Table 6 pg 104	0.22 K

Table 3: List of estimates of measurement error uncertainties for moored buoys where random and systematic errors were not dealt with separately.

Reference	Platform type	Random	Systematic	Notes
Kent and Berry [2008] pg 11 Table 5a	Ship	0.7 K	0.8 K	From comparison with Numerical Weather Prediction fields provided with VOSClim data
Pg 12 Table 5c	Drifter	0.6 K	0.3 K	
Pg 11 Table 5b	Mooring	0.3 K	0.2 K	
Kennedy et al. [2011a, 2011b] pg 86	Ship	0.74 K	0.71 K	From comparison with Along Track Scanning Radiometer SST retrievals
Pg 86	Drifter	0.26 K	0.29 K	
Brasnett [2008] values estimated for present study by	Ship	1.16 K	0.69 K	From comparison with interpolated fields

author				
Xu and Ignatov [2010] values estimated for present study by author	Ship	0.81 K	0.53 K	From comparison with multisensor satellite SST fields
Kennedy et al. [2011a, 2011b] method using *Atkinson et al.* [2013] whitelist	Ship	0.56 K	0.37 K	From comparison with multisensor satellite SST fields
Gilhousen [1987] Table 6 pg 104	Mooring	0.22 K	0.13 K	Comparison of moored buoys

Table 4: List of estimates of measurement error uncertainties for all platforms for studies where the measurement error uncertainty is decomposed into random and systematic components.

Data set	Input data set	Interpolation method	Resolution
ICOADS summaries [*Woodruff et al.,* 2011]	ICOADS 2.5	None	2°x2° monthly
HadSST2 [*Rayner et al., 2006*]	ICOADS 2.1	None	5°x5° monthly

HadSST3 [*Kennedy et al.,* 2011b; *Kennedy et al.*, 2011c]	ICOADS 2.5	None	5°x5° monthly
TOHOKU [*Yasunaka and Hanawa*, 2002]	ICOADS 2.1	None	5°x5° monthly
HadISST1.1 [*Rayner et al.*, 2003]	Met Office Marine Databank and COADS, AVHRR satellite retrievals	Reduced Space Optimal Interpolation	1°x1° monthly
ERSSTv3b [*Smith et al.*, 2008]	ICOADS 2.1	Separate low and high frequency reconstructions. High frequency component based on EOTs	2°x2° monthly
COBE [*Ishii et al.*, 2005]	ICOADS 2.1 and Kobe collection	Optimal interpolation	1°x1° monthly
COBE-2 [*Hirahara et al.,* 2013]	ICOADS 2.5 and Kobe collection, AVHRR satellite retrievals	Multi scale analysis based on EOFs	1°x1° daily and monthly

Kaplan [*Kaplan et al.,* 1998]	Met Office Marine Databank	Reduced Space Optimal Smoothing	5°x5° monthly
NOCS [*Berry and Kent*, 2011]	ICOADS 2.5	Optimal Interpolation	1°x1° daily and monthly
VBPCA [*Ilin and Kaplan*, 2009]	ICOADS 2.5	Variational Bayesian Principal Component Analysis	5°x5° monthly
GPFA [*Luttinen and Ilin,* 2009]	ICOADS 2.5	Gaussian Process Factor Analysis	5°x5° monthly
GP [*Luttinen and Ilin,*2012]	ICOADS 2.5	Gaussian Process	5°x5° monthly

Table 5: List of datasets used and referred to in the review.

Figure Captions

Figure 1: (a) Estimated bias (with respect to the 1961-1990 average) on global average SST anomalies associated with measurement depth as a function of time (upper panel). (b) Global average SST anomaly from the HadSST3 [Kennedy et al. 2011b, 2011c] median before (black) and after (red) the measurement-depth bias has been subtracted. The two red lines reflect different assumptions concerning data that could not be definitively assigned to any particular measurement type. The large dip during World War 2 arises because the majority of observations were ERI measurements.

Figure 2: Time series of upper ocean temperatures (0-30 m) from nine moorings in the Tropical Ocean Atmosphere (TAO) array and the Subduction Array. The mooring and its location are given above each plot. The different coloured lines represent different depths and these are indicated by the legends in each panel. The Subduction Array data are described in *Moyer and Weller* [1997].

Figure 3: Distributions of estimated measurement errors and uncertainties from ships. (a) distributions of systematic measurement errors for all entries (2003-2007) in *Kennedy et al.* [2011a], *Brasnett* [2008], *Berry and Kent* [2008] and *Xu and Ignatov* [2010]. (b) distributions of random measurement error uncertainties (expressed as variances) from the same analyses as in the top left panel and *Atkinson et al.* [2013]. (c) as for top left except each ship now has only a single entry so the analyses are directly comparable. (d) scatter plot showing systematic measurement errors estimated by *Brasnett* [2008] and *Berry and Kent* [2008] showing the good correlation between the estimates.

Figure 4: (a) Estimated global average SST anomaly from HadSST3 [Kennedy et al. 2011b, 2011c] (red) and for subsamples of the HadSST3 dataset reduced to 19th century coverage. The black line is the median of the samples and the blue area gives the range. (b) difference, on an expanded temperature scale, between the global average SST anomaly from the full HadSST3 data set and global averages calculated from the subsamples.

Figure 5: Global average sea-surface temperature anomalies and night marine air temperature anomalies from a range of data sets. (a) Simple gridded SST data sets including ICOADS v2.1 (red), 200 realizations of HadSST3 (pale grey), HadSST2 (dark green), TOHOKU (darker grey), ARC (*Merchant et al.* [2012] lime green) and the COBE-2 dataset sub-sampled to observational coverage (pale blue). (b) 8 Interpolated SST analyses including the COBE-2 dataset (pale blue), HadISST1.1 (gold), ERSSTv3b (orange), VBPCA, GPFA and GP (deep magenta), Kaplan (pink), NOCS (black). (c) shows the series in (a) and (b) combined. (d) NMAT: Ishii et al. (2005, red and blue), MOHMAT4N3 and HadMAT (*Rayner et al.* [2003], pink and orange), *Berry and Kent* [2009] (green), HadNMAT2 (*Kent et al.* [2013], gold).

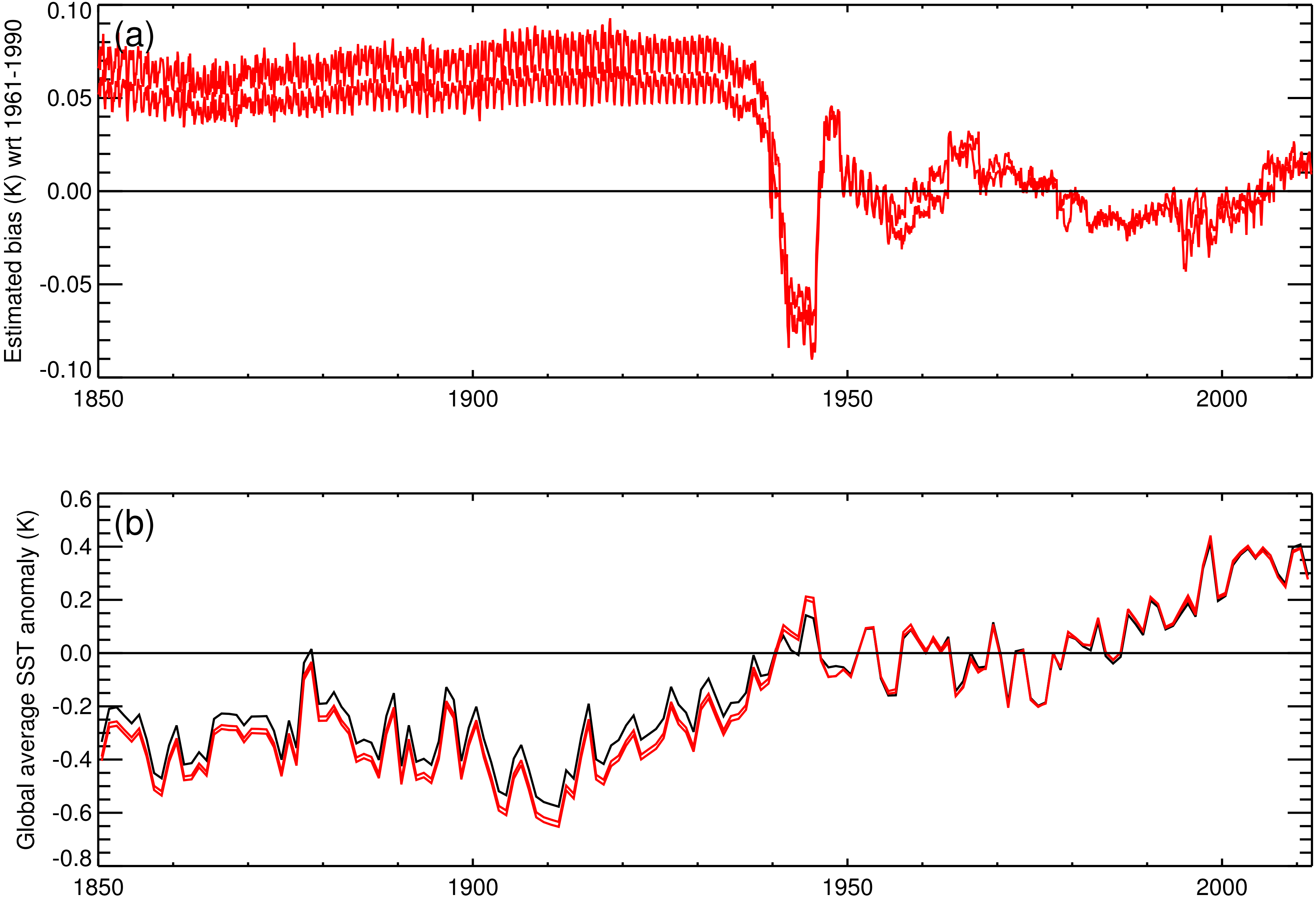
Figure 6: Comparison between COBE-2 (black) and HadSST3 (red) metadata and bias estimates for the period 1920 to 2010. (a) Fraction of buckets assessed as being uninsulated. The two red lines indicate the earliest and latest switchover dates allowed in the generation of the HadSST3 ensemble. (b) Fractional contribution to the global average from buckets, buoys and engine room measurements. The total is less than unity; the remainder are either unknown (in the HadSST3 analysis) or uncategorized (COBE-2). (c) Estimated bias. There are 100 versions of HadSST3 and a single estimate from COBE-2.

Figure 7: Maps showing climatological standard deviation of SST (a, g, m), Structural uncertainty (b, h, n), Sampling uncertainty (c, i, o), measurement uncertainty (d, j, p), bias

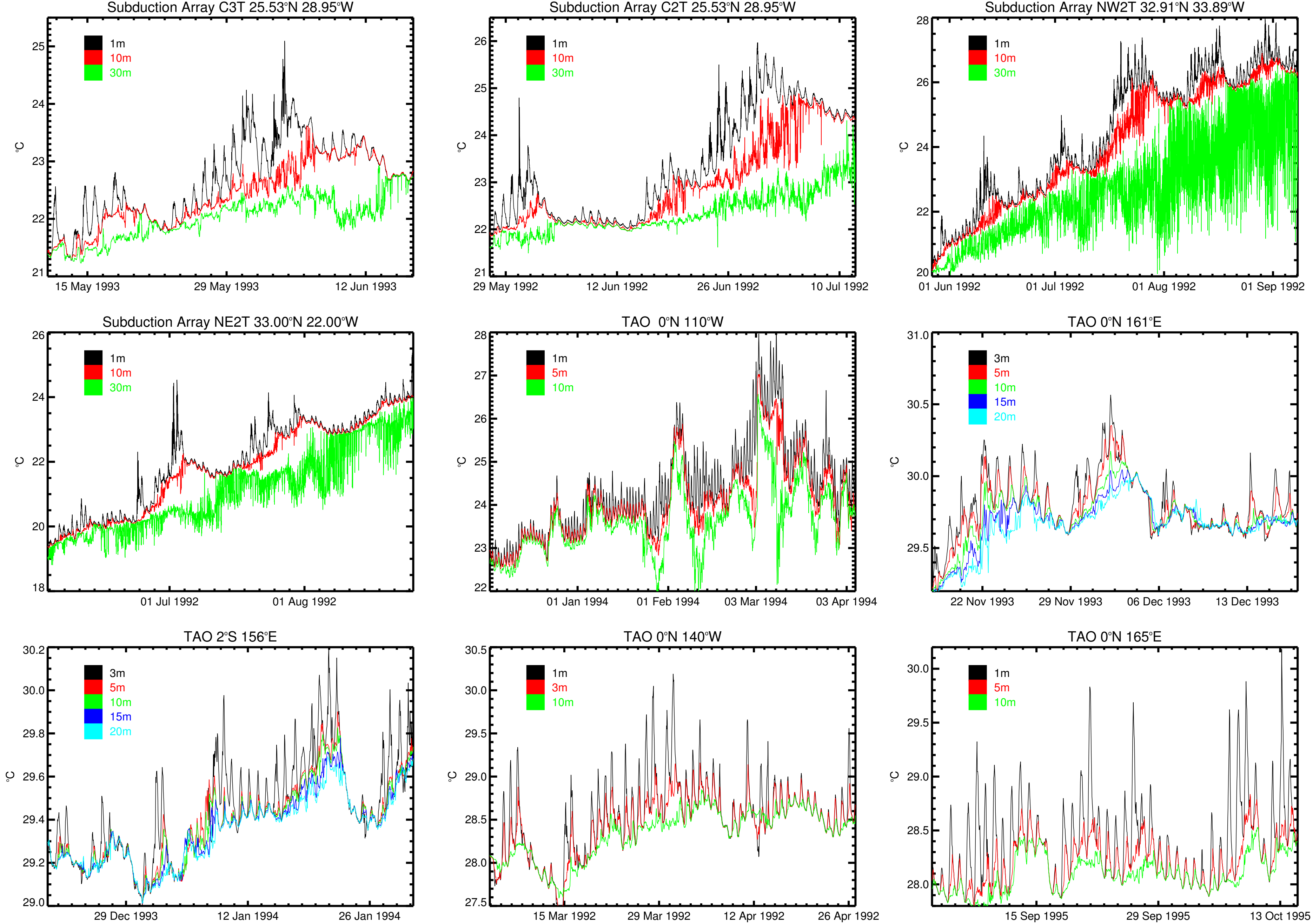
uncertainty (e,k,q) and analysis uncertainty from ERSST (f, l, r). Three months are shown: (a-f) June 1891, (g-l) April 1944 and (m-r) August 2003.

Figure 8: Time series of estimated uncertainties arising from different sources in area-averages: (a) Global annual, (b) Northern hemisphere annual, (c) North Pacific annual, (d) North Atlantic annual and (e) a 5-degree grid box centered on 42.5°W, 27.5°N monthly. Uncertainty components shown are: (pale blue) grid-box sampling uncertainty, (green) uncorrelated measurement uncertainty, (red) correlated measurement uncertainty, (dark blue) parametric bias uncertainty from a 200-member ensemble based on HadSST3, (black) large-scale sampling uncertainty, and (magenta) structural uncertainty estimated by taking the range of the area-average calculated from seven near-globally-complete analyses.

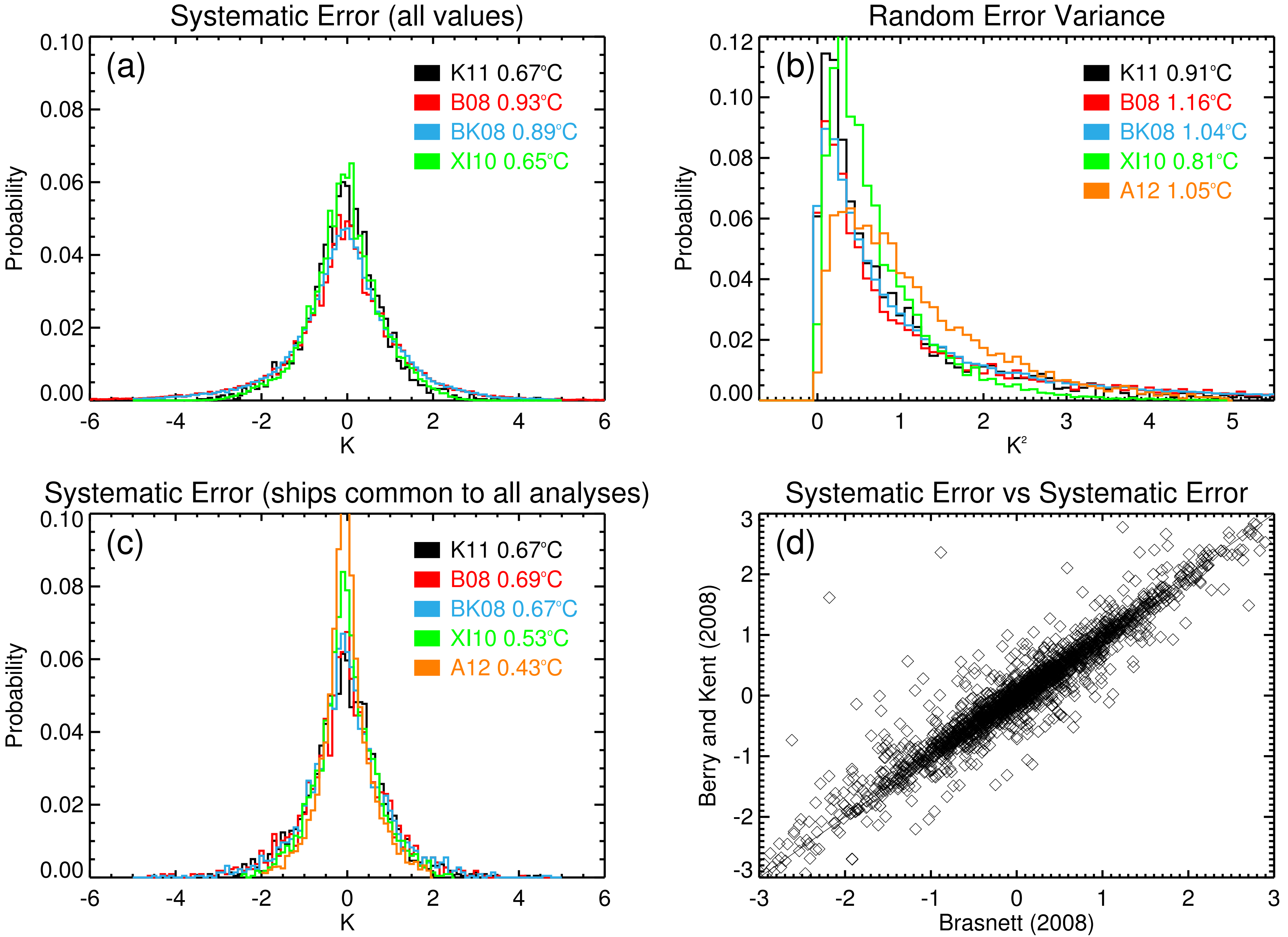
Figure 9: (a) Global, (b) Northern Hemisphere, (c) Southern Hemisphere and (d) Tropical average sea-surface temperature anomalies with estimated 95% confidence range for ERSSTv3b (1880-2012 dark blue line and pale blue shading) and for the HadSST3 based analysis described in section 3.5 (1850-2011 red line and orange and yellow shading). The yellow shading indicates an estimate of the additional structural uncertainty in the HadSST3 series.



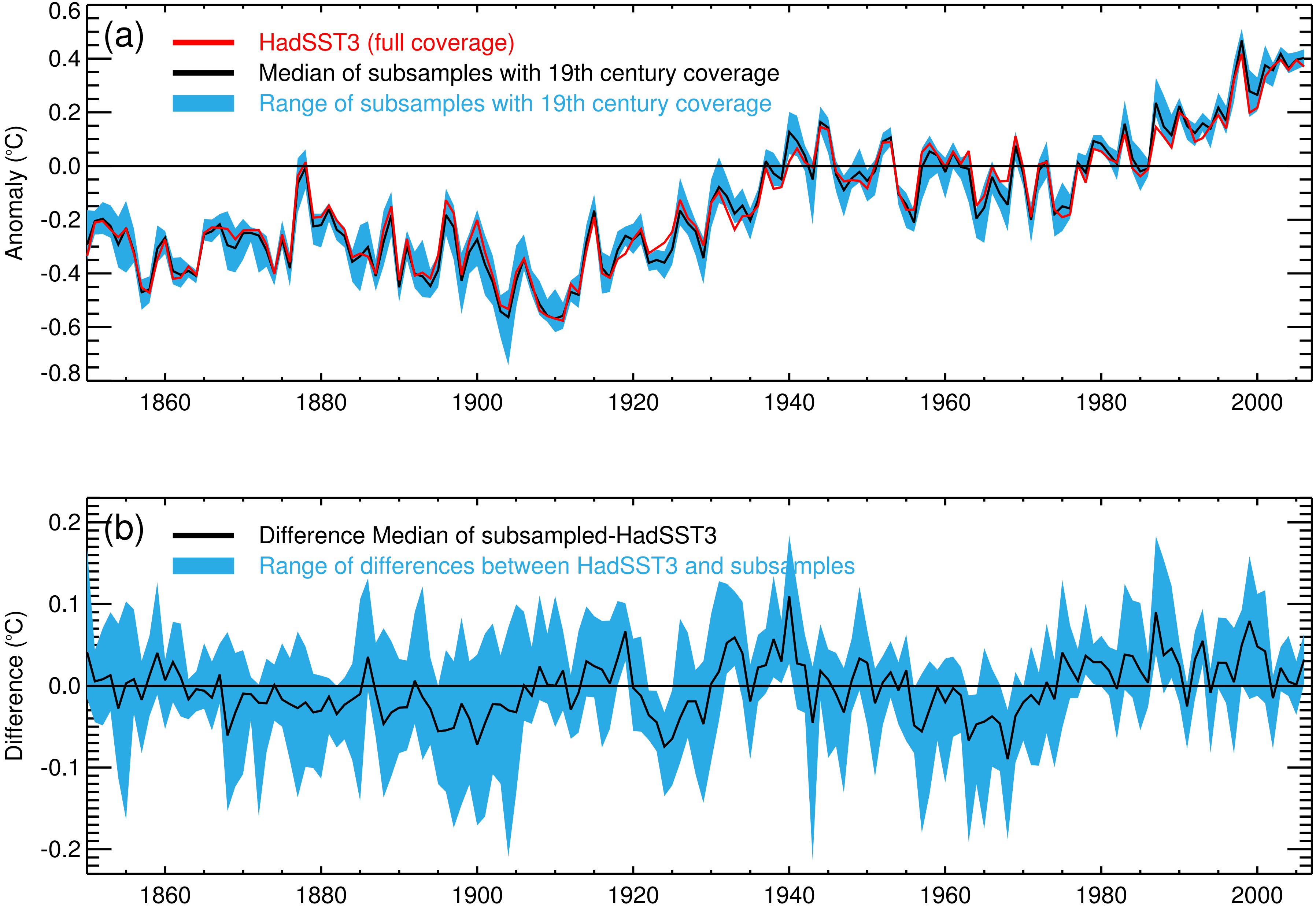
(a)
Estimated bias (K) wrt 1961-1990
0.10
0.05
0.00
-0.05
-0.10
1850
1900
1950
2000
(b)
Global average SST anomaly (K)
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
1850
1900
1950
2000

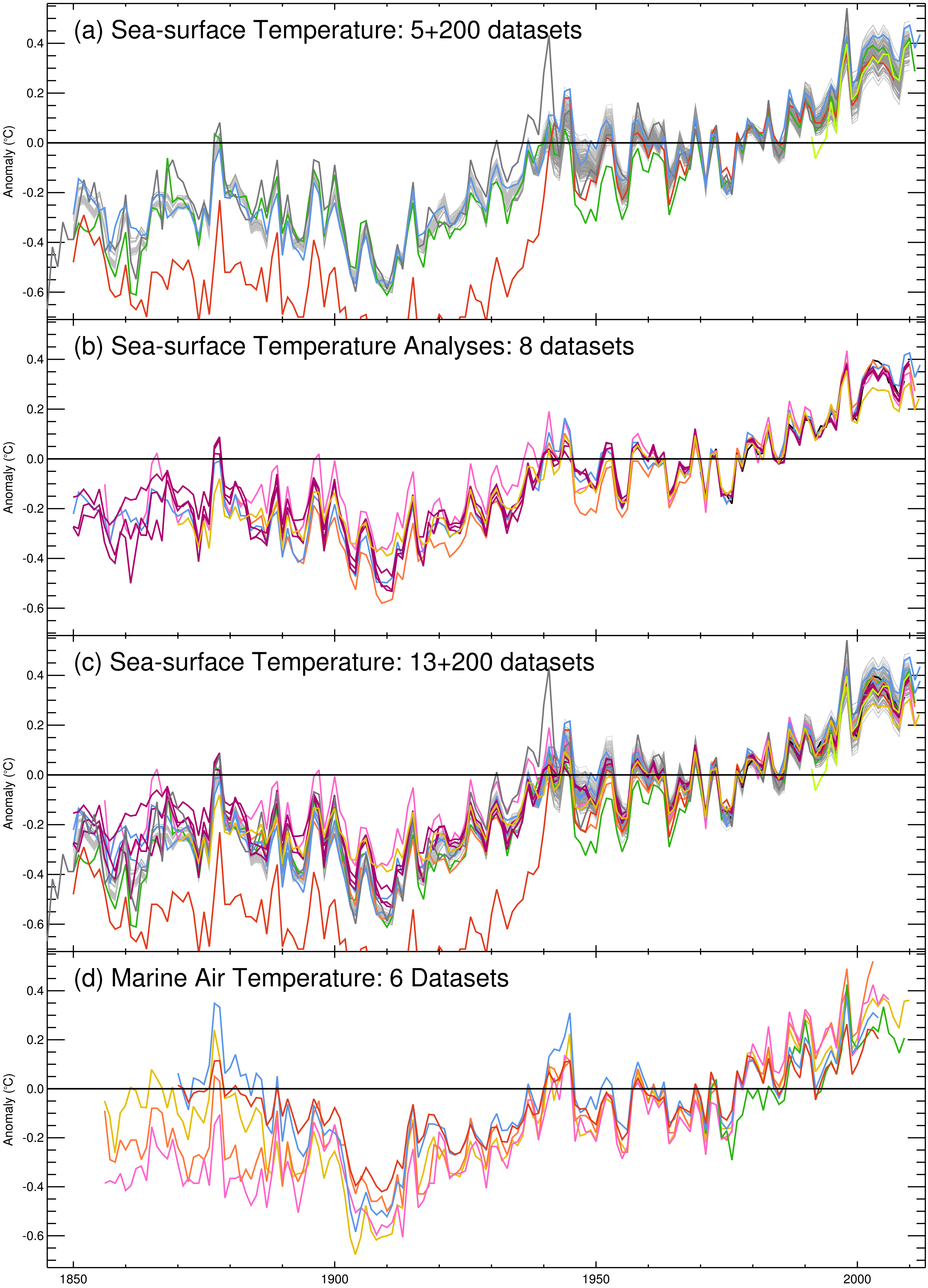


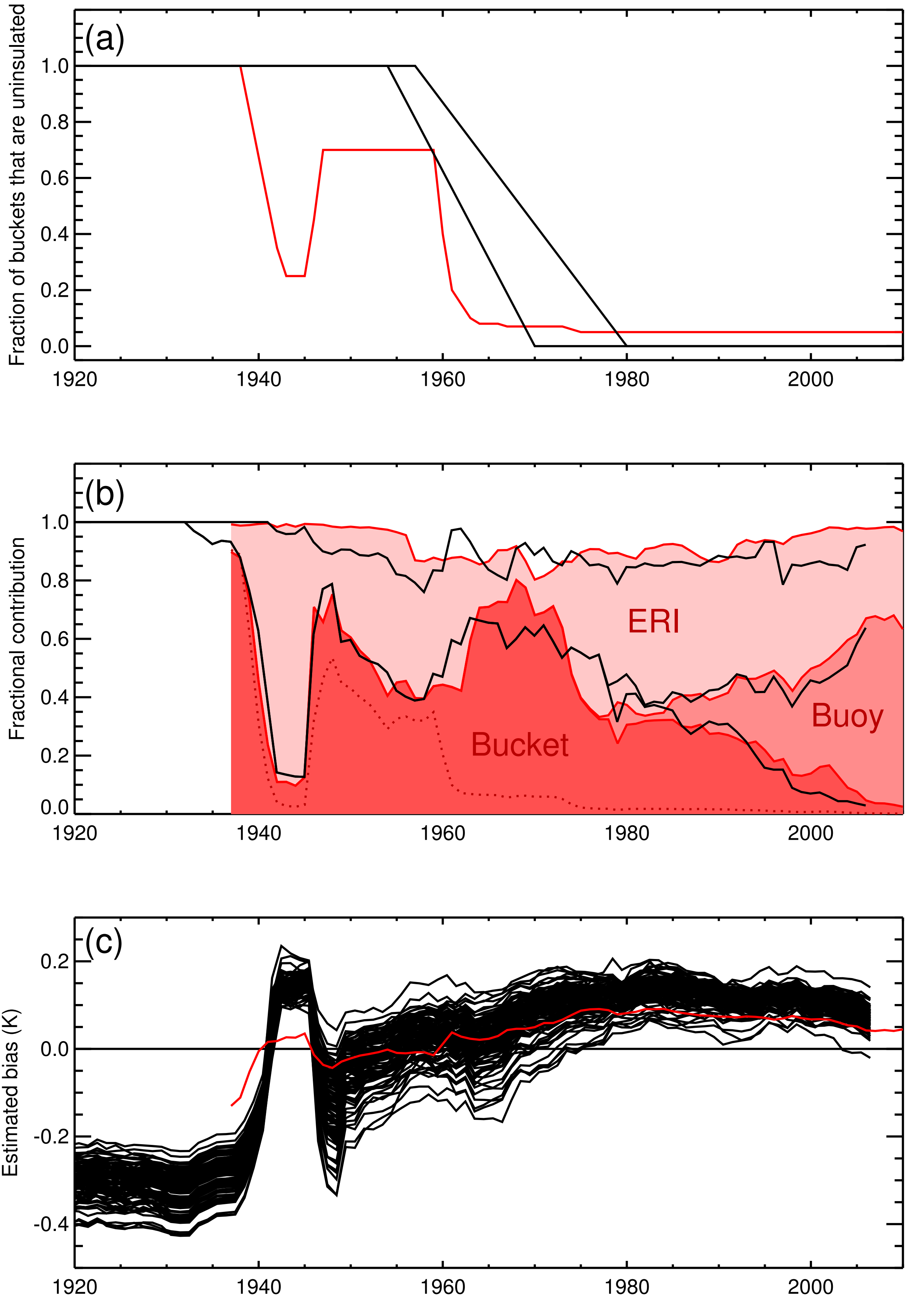
Subduction Array C3T 25.53°N 28.95°W
1m
10m
30m
°C
15 May 1993
29 May 1993
12 Jun 1993
Subduction Array C2T 25.53°N 28.95°W
29 May 1992
12 Jun 1992
26 Jun 1992
10 Jul 1992
Subduction Array NW2T 32.91°N 33.89°W
01 Jun 1992
01 Jul 1992
01 Aug 1992
01 Sep 1992
Subduction Array NE2T 33.00°N 22.00°W
01 Jul 1992
01 Aug 1992
TAO 0°N 110°W
1m
5m
10m
01 Jan 1994
01 Feb 1994
03 Mar 1994
03 Apr 1994
TAO 0°N 161°E
3m
5m
10m
15m
20m
22 Nov 1993
29 Nov 1993
06 Dec 1993
13 Dec 1993
TAO 2°S 156°E
29 Dec 1993
12 Jan 1994
26 Jan 1994
TAO 0°N 140°W
1m
3m
10m
15 Mar 1992
29 Mar 1992
12 Apr 1992
26 Apr 1992
TAO 0°N 165°E
1m
5m
10m
15 Sep 1995
29 Sep 1995
13 Oct 1995



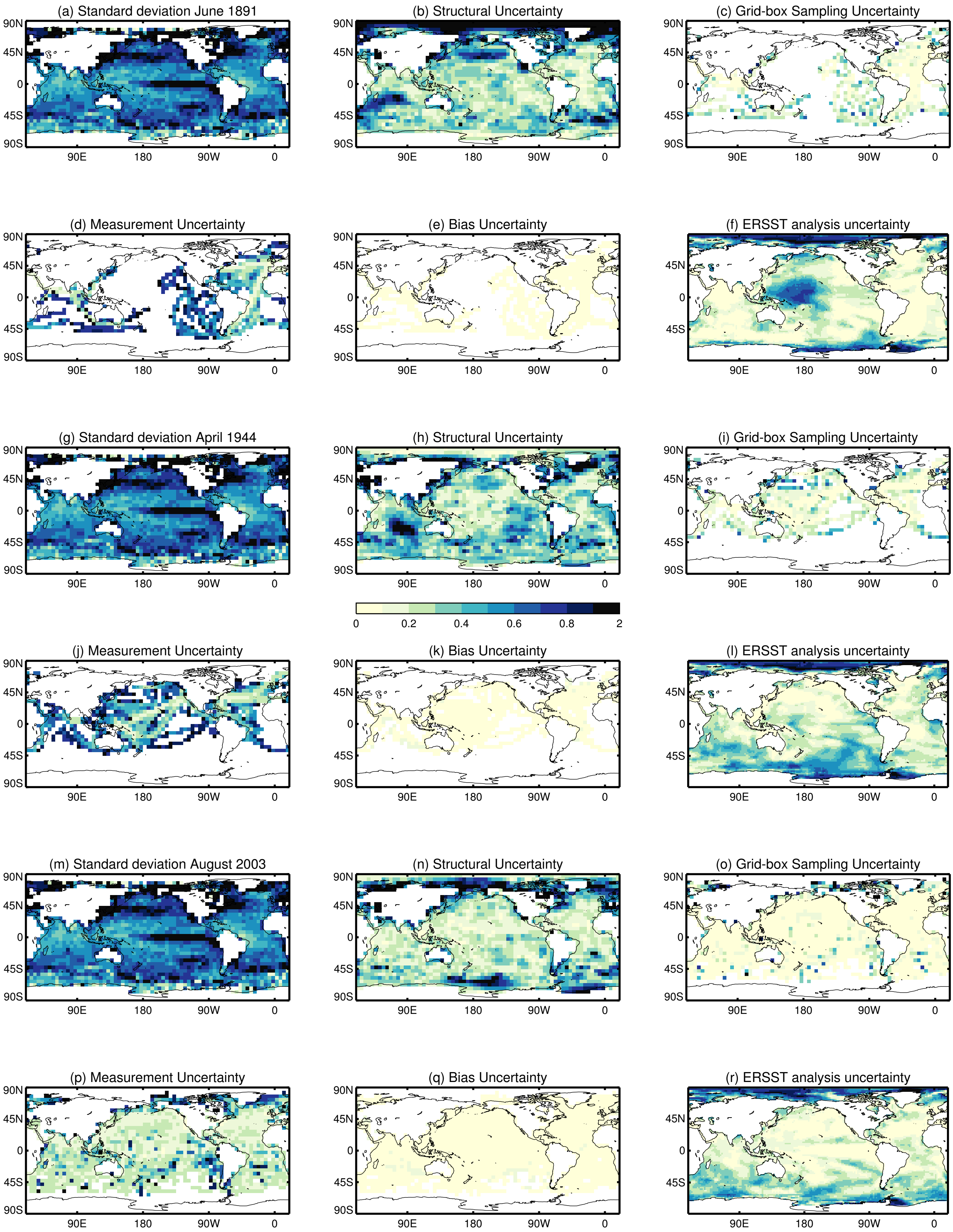
Systematic Error (all values)
(a)
K11 0.67°C
B08 0.93°C
BK08 0.89°C
XI10 0.65°C
Probability
K
Random Error Variance
(b)
K11 0.91°C
B08 1.16°C
BK08 1.04°C
XI10 0.81°C
A12 1.05°C
Probability
K²
Systematic Error (ships common to all analyses)
(c)
K11 0.67°C
B08 0.69°C
BK08 0.67°C
XI10 0.53°C
A12 0.43°C
Probability
K
Systematic Error vs Systematic Error
(d)
Berry and Kent (2008)
Brasnett (2008)

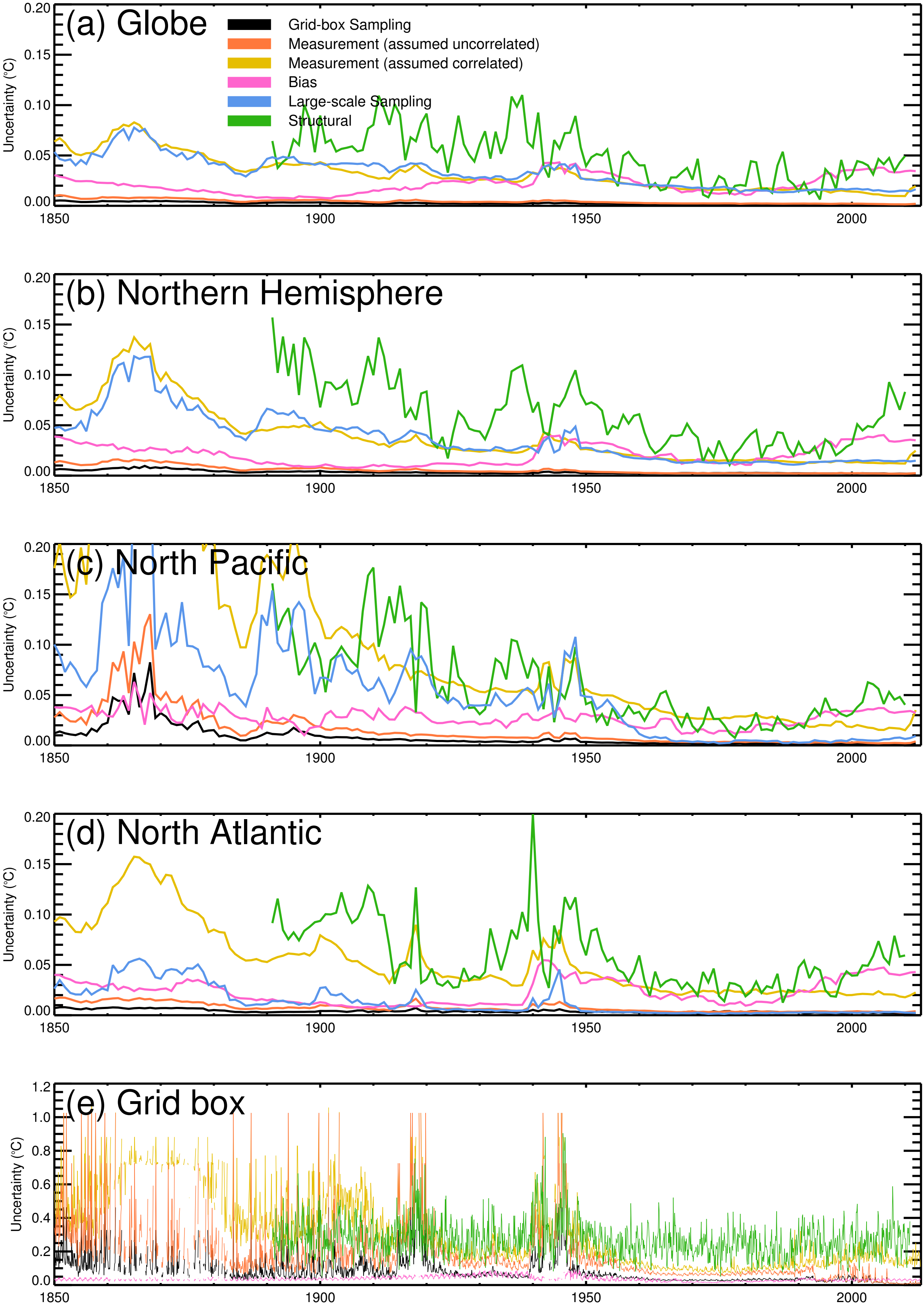

(a)
HadSST3 (full coverage)
Median of subsamples with 19th century coverage
Range of subsamples with 19th century coverage
Anomaly (°C)
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
1860
1880
1900
1920
1940
1960
1980
2000
(b)
Difference Median of subsampled-HadSST3
Range of differences between HadSST3 and subsamples
Difference (°C)
0.2
0.1
0.0
-0.1
-0.2

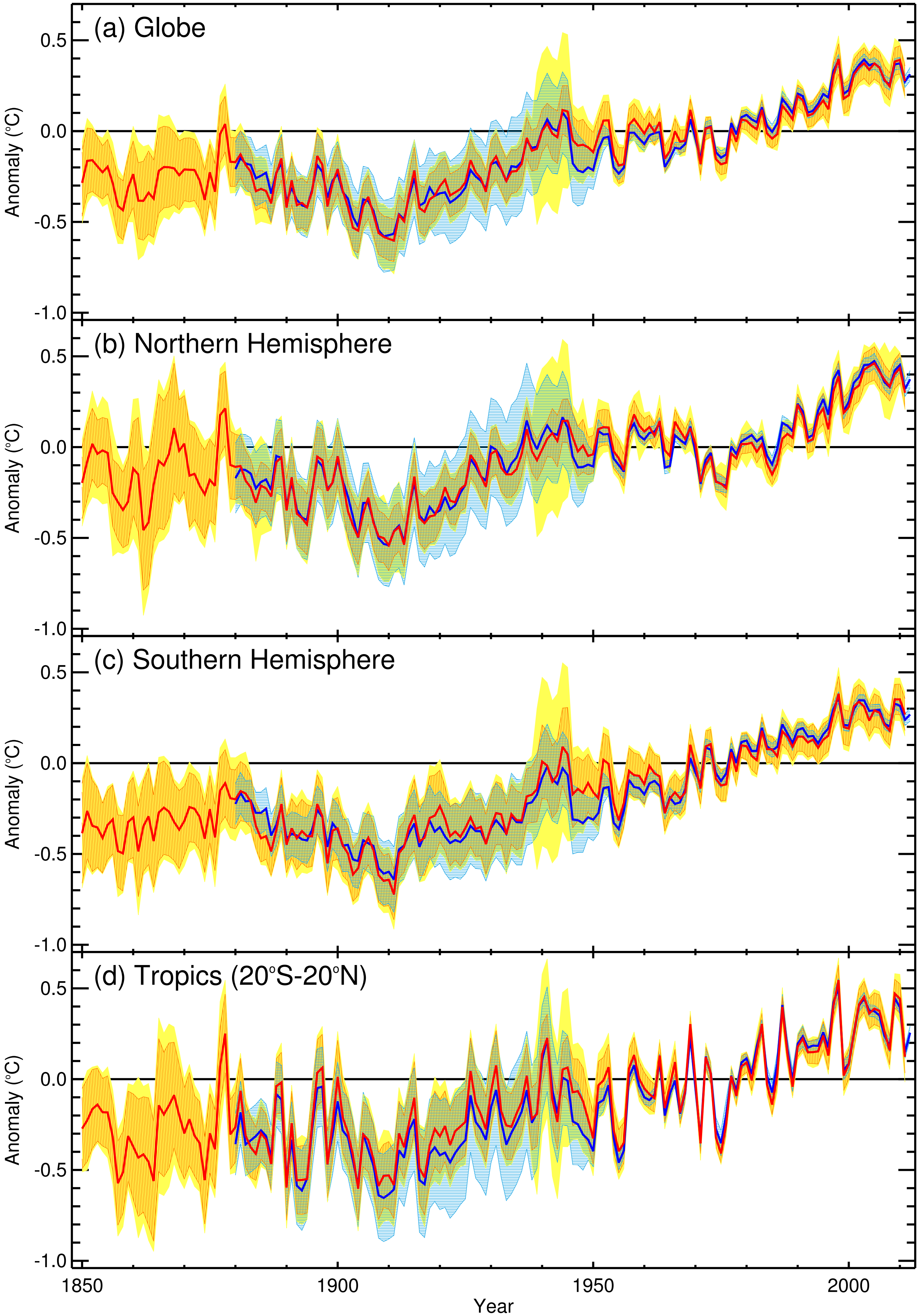

(a) Sea-surface Temperature: 5+200 datasets
(b) Sea-surface Temperature Analyses: 8 datasets
(c) Sea-surface Temperature: 13+200 datasets
(d) Marine Air Temperature: 6 Datasets
Anomaly (°C)
0.4
0.2
0.0
-0.2
-0.4
-0.6
1850
1900
1950
2000



(a)
Fraction of buckets that are uninsulated
1.0
0.8
0.6
0.4
0.2
0.0
1920
1940
1960
1980
2000
(b)
Fractional contribution
ERI
Buoy
Bucket
(c)
Estimated bias (K)
0.2
0.0
-0.2
-0.4


(a) Standard deviation June 1891
(b) Structural Uncertainty
(c) Grid-box Sampling Uncertainty
(d) Measurement Uncertainty
(e) Bias Uncertainty
(f) ERSST analysis uncertainty
(g) Standard deviation April 1944
(h) Structural Uncertainty
(i) Grid-box Sampling Uncertainty
(j) Measurement Uncertainty
(k) Bias Uncertainty
(l) ERSST analysis uncertainty
(m) Standard deviation August 2003
(n) Structural Uncertainty
(o) Grid-box Sampling Uncertainty
(p) Measurement Uncertainty
(q) Bias Uncertainty
(r) ERSST analysis uncertainty
90N
45N
0
45S
90S
90E
180
90W
0
0
0.2
0.4
0.6
0.8
2


(a) Globe
Grid-box Sampling
Measurement (assumed uncorrelated)
Measurement (assumed correlated)
Bias
Large-scale Sampling
Structural
(b) Northern Hemisphere
(c) North Pacific
(d) North Atlantic
(e) Grid box
Uncertainty (°C)
0.00
0.05
0.10
0.15
0.20
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1850
1900
1950
2000



(a) Globe
(b) Northern Hemisphere
(c) Southern Hemisphere
(d) Tropics (20°S-20°N)
Anomaly (°C)
0.5
0.0
-0.5
-1.0
1850
1900
1950
2000
Year