INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **37**: 2233–2247 (2017)
Published online 28 September 2015 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/joc.4492

RMetS
Royal Meteorological Society

A probabilistic approach to ship voyage reconstruction in ICOADS

Giulia Carella,* Elizabeth C. Kent and David I. Berry

National Oceanography Centre, Marine Physics and Ocean Climate, Southampton, UK

ABSTRACT: The International Comprehensive Ocean-Atmosphere Data Set (ICOADS) provides the main archive for surface marine observations for the past approximately 150 years. ICOADS ship identifier (ID) information is often missing or unusable, preventing the linking of reports to an individual ship. A method for the reconstruction of ship voyages in ICOADS is presented, by which groups of reports can be associated with an individual ship or ship track. The method defines a function representing the probability density function (*pdf*) of any particular report being associated with a group of reports. The parameters of the pdf are calculated from the ship data themselves, giving the likely variation of a ship report perpendicular to its overall direction of travel. For groups of reports with ID information, the PDF is used to associate reports without ID information with the known-ID track. Reports without ID information are then clustered together to form the most probable track. Results are shown for the period 1855–1969. Both the percentage of reports associated with tracks and the length of those tracks increase substantially following tracking. Initial validation of the results was performed by visual inspection: the model implementation was then refined to improve the results. Confidence in the tracking is increased by a demonstration that the method clusters together reports with similar sea surface temperature characteristics. Issues in the data were found to be one of the main challenges in implementing the tracking technique. Particular problems encountered included the coarse resolution of some position information; reports that were mispositioned in either space or time; unidentified duplicate reports; and the fragmentation of voyages between different ICOADS acquisition sources. Some of these effects could be ameliorated by pre-processing of ICOADS reports, however a full reprocessing of the historical input sources to ICOADS would be required to make further improvements.

KEY WORDS marine meteorological data; ship data; voyage tracking; ocean climate; climate change; sea surface temperature

Received 1 April 2015; Revised 30 June 2015; Accepted 1 August 2015

1. Introduction

The International Comprehensive Ocean-Atmosphere Data Set (ICOADS, e.g. Woodruff *et al.*, 2011) provides an archive of *in situ* surface marine observations presently starting in 1662, but sparse before about 1850. These observations come from a variety of sources, including ships, buoys and coastal data, and are used to construct gridded analyses that document changes in surface marine conditions. Examples include gridded analyses for sea surface temperature (SST, Smith and Reynolds, 2004; Kennedy *et al.*, 2011b; Hirahara *et al.*, 2014; Huang *et al.*, 2015), air temperature (Kent *et al.*, 2013), wind (Kalnay *et al.*, 1996), pressure (Allan and Ansell, 2006), humidity (Willett *et al.*, 2008) and air–sea fluxes (Berry and Kent, 2009; Berry and Kent, 2011). These gridded analyses are used in climate assessments (IPCC, 2013; Blunden and Arndt, 2014). Of these variables, SST is the focus of most attention and forms the marine component of the global surface temperature record, a primary metric of climate change (IPCC, 2013). *In situ* SST observations are used for monitoring the present climate and for comparison between the present and past climatic variations over the oceans (Stott *et al.*, 2010), as well as for validation of climate models (Sutton *et al.*, 2007; Boer, 2011) and to provide boundary conditions for atmospheric models (Compo *et al.*, 2011; Stickler *et al.*, 2014).

COADS Release 1 (Woodruff *et al.*, 1987) was based on data collections in the form of punched card decks (hereafter decks) that had been obtained by the United States from major maritime nations from the 1940s onwards. Reports were available in a variety of different formats, and not all contained metadata identifying the observing platform or methods. COADS combined archives of marine data from several countries, and it was known that there was substantial duplication of observations between some of the sources. A complicated process of identification, exclusion and compositing of duplicate reports (known as duplicate elimination: dupelim) was developed to address this (Slutz *et al.*, 1985). Dupelim was later extended as the number of ICOADS data sources expanded. A substantial proportion of observations in the current ICOADS Release 2.5 (R2.5) are from these historical archives. More recent observations may contain ship identifiers (hereafter IDs, Kent *et al.*, 2007), and recent data digitisation activities have been careful to retain ship and observational metadata

*Correspondence to: G. Carella, National Oceanography Centre, European Way, Southampton SO14 3ZH, UK. E-mail: giulia.carella@noc.ac.uk

(García-Herrera *et al.*, 2005; Allan *et al.*, 2011; Wilkinson *et al.*, 2011). However, missing (Figure 1) or incorrect platform identifiers are recurrent, particularly for ship observations, throughout ICOADS.

This article addresses the problem of reconstructing ship voyages in the ICOADS archive by associating groups of reports to an individual vessel (hereafter 'ship tracking'). The aim of this article is to explain the method and to show the results for the period 1855–1969. Section 2 details the motivations for this work and explains how ship tracking will enable improved SST bias adjustment, uncertainty estimation, quality control and data assimilation. Section 3 describes the data association technique developed to group observations to give plausible tracks; Section 4 shows the results and Section 5 discusses the conclusions and describes the potential for future improvements.

2. Motivations

2.1. SST bias adjustment

Quantification of biases in historical SSTs is important for estimating global temperature trends (Jones and Wigley, 2010). SST observations from ships form one of the longest instrumental records of surface marine temperature change. However, over the years, several different methods of measuring SST have been used, each with different bias characteristics (James and Fox, 1972; Kent *et al.*, 1993, 2010; Kent and Taylor, 2006; Kennedy *et al.*, 2011b). For historic observations, the measurement practice is almost never known in detail (Folland and Parker, 1995), and therefore integral to the SST bias adjustment is the assignment of measurement methods, ideally to individual ships or reports. Although there has been progress towards understanding the characteristics of historical SST observations (e.g. as reviewed by Kennedy, 2014), we do not yet have a full quantification of their bias and uncertainty. The estimation of systematic biases is critical for climatic decadal predictions (Kennedy *et al.*, 2011b), and uncertainties in long-term trends are expected to be controlled by uncertainties in biases introduced by changes of instrumentation and measurement practices (Jones and Wigley, 2010).

Currently, SST data sets use bias models representing large-scale effects, either based on 5° area average monthly climatological environmental conditions (Folland and Parker, 1995) or on large-scale variations in air–sea temperature difference (Smith and Reynolds, 2002), which is also uncertain (Kent *et al.*, 2013). There are differences between the bias adjustment fields used to date, which limits our confidence, particularly in regional estimates of historical SST (Kennedy *et al.*, 2011a). It is known that changes in observational practice can be rapid and undocumented (Thompson *et al.*, 2008), such changes cannot be captured by large-scale approaches to bias adjustment.

There are two main barriers to finer-scale adjustment of SST. Firstly, there is usually not enough information about how the observations were made; for example, we may know that a bucket was used, but not the type of bucket, or the conditions of its exposure. The second barrier is that many ICOADS reports cannot be confidently assigned to a particular vessel and hence, cautiously, to the same measurement methodology. It is this latter point that we address here, noting that similar arguments can be made for other ICOADS variables such as wind speed or humidity.

2.2. Improved estimation of measurement uncertainty

Recent studies (Kent and Berry, 2008; Kennedy *et al.*, 2011a; Kent *et al.*, 2013) have partitioned measurement uncertainties into random and systematic parts, where the systematic uncertainty represents mean biases for individual platforms (typically a ship or buoy). It is therefore important to know which platforms took measurements in each grid box and how many observations each platform made. However, the present lack of comprehensive ID information hampers the application of such an error model (Kennedy, 2014). Linking observations from the same vessel together, through ship tracking, will aid in applying such an error model and lead to improved estimates of the uncertainty in gridded analysis through better treatment of the uncertainty arising from correlated errors.

2.3. ICOADS quality assurance

There are two main aspects of quality assurance (QA) that will benefit from ship tracking. The first is identifying mispositioned and misassigned data. It is well known that ICOADS contains mispositioned data, particularly in deck 732 (Minobe and Maeda, 2005; Kennedy *et al.*, 2011b). Mispositioned data have also been identified post-1970 for ships with valid ID information, often as a duplicate, or partial duplicate, of a report in the correct location (Kent and Challenor, 2006). Some data are mispositioned in time, possibly as a result of report corruption of the observation or by conversion from local time to GMT with incorrect longitude. For reports with valid ID information, tracking can identify mispositioned reports and perhaps relocate them in position or time. For reports with no valid ID information, any reports that could not be associated with other reports in the tracking process might be down-weighted or excluded from analyses.

The second benefit of grouping reports by observing platform is the identification of platforms that consistently report biased observations as a result of, for example, poor observation practice, miscalibrated instruments or persistent miscoding. This permits the exclusion, or in some cases correction, of observations on a vessel-by-vessel basis.

2.4. Data assimilation

Finally, ship tracking constitutes an important step in the data assimilation process of atmospheric reanalyses. For example, the European Centre for Medium-Range Weather Forecasts (ECMWF) pilot reanalysis of the 20th-century (ERA-20C) uses a platform level adaptive bias correction systems that updates the bias parameters during the assimilation simultaneously with the meteorological variables

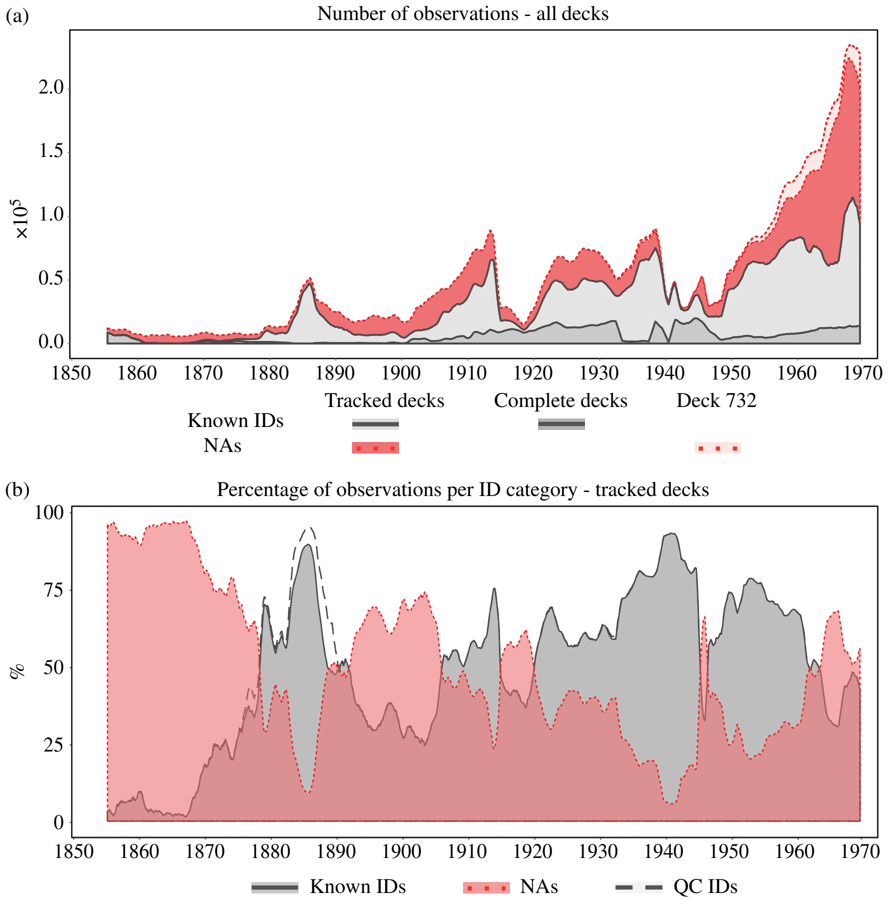

Figure 1. (a) For all the decks in ICOADS, the number of observations with platform type field values of 0–5 (ship data) or missing. The observations have been grouped according to three different deck categories (*tracked decks*, *complete decks*, *deck 732*) and to two different ID categories (*known-ID*, *NAs*). Note that in the *complete decks* category, the figure shows the number of known-IDs only (no NAs are present) while for *deck 732* it shows the number of NAs only (no known-IDs are present). (b) For the *tracked decks* category, the percentage of observations in each of the three different ID categories (*known-ID*, *NAs*, *QC IDs*) the data were grouped into. Both plots show monthly data filtered with a 12-month running mean. [Colour figure can be viewed at wileyonlinelibrary.com].

(Poli *et al.*, 2013). To address the problem of missing platform IDs in ICOADS, a simple tracking algorithm was implemented that split ICOADS observations into plausible subsets based on ship speed constraints combined with any available ID information (Hersbach *et al.*, 2015). Improvements to platform identification, combined with the QA improvements described in Section 2.3, will have clear benefits for atmospheric reanalysis activities.

3. Method

3.1. Data selection, pre-processing and QA applied to ICOADS

We have used observations from ICOADS Release 2.5, using only those reports with platform type (PT) field values of 0–5 (various types of ship data) or missing (unknown PT, which inspection suggests are mostly reports from ships). Data known to originate in decks containing only observations from buoys, fixed platforms or coastal stations (*buoy-only decks*, see Appendix for details) have been excluded. Based on preliminary tracking results, decks that contained complete and reliable ID information with no overlap with observations from other decks (*complete decks*) were not tracked. These decks are typically from modern digitisation efforts or are collections of data from research cruises. The final exclusions were reports from decks 732 and 874 known to have suffered problems with format conversion. Finally, decks that were thought to be unique but with incomplete ID information were tracked within the deck only, while decks that were thought to have common data were tracked together. More information is given in the Appendix.

The observations selected have been divided into two categories based on the availability of ID information, those with an empty ID field (hereafter NAs) and those with extant ID information (hereafter known-IDs). The known-IDs are grouped together and each unique ID assumed to represent a single ship, whilst the NAs remain unassigned. Additionally, a number of the known-IDs are not assigned to a group, either due to a corrupted ID or the use of a generic callsign (e.g. SHIP, PLAT, etc.). Figure 1(a) illustrates the data, separated into the two categories (NAs and known-IDs), for both the decks requiring tracking and those excluded. The two groups of observations, those with known-IDs and NAs, have been treated differently in the tracking analysis (Section 3.2).

Prior to tracking, it has been necessary to pre-process the data. Firstly, the ICOADS IDs have been quality controlled. Within a deck, the IDs conform to a limited number of patterns in terms of combinations of digits and/or characters (see Table A2). Those reports where the ID does not confirm to an expected pattern for the deck were reallocated to the NAs. This helps to prevent the use of invalid ID information in the initial construction of the tracks. For example, without this step, reports with truncated ID information from several different ships would be erroneously assigned to a single ship. In some cases, ID information from different decks contained common sequences of characters or digits and had clearly been derived from the same original information (e.g. logbook number), but one of the decks had appended additional characters or digits to the start or end of the common sequence. Where the link was clear, the common sequence only was retained using rules for particular decks (Appendix A3). Further, ID modification reduced sequences of numeric IDs each representing a single report to their common root. Any IDs modified in these ways will be referred to as QC IDs. Figure 1(b) shows that the ID modification was particularly important for reports from the 1880s. Moreover, for IDs associated with less than four observations per month (QC IDs if appropriate), the report was tracked in the NA category.

Secondly, in addition to QC of the IDs, a duplicate elimination process has been applied. For all pairs of reports at a given time and with similar locations, the available parameters were checked for matching data. Those found to contain matching data, excluding missing elements and location information, were flagged as potential duplicates. The duplicate containing the most complete report or from the deck expected to be of highest quality was then selected. This relaxes the dupelim restrictions applied by ICOADS, which considers only potential duplicates within the same 1° grid box. The largest monthly proportion of reports removed was 3%, but substantially less than 1% was much more typical. Although percentages of identified duplicates were small, the process was judged to be worthwhile as ship tracking inevitably works best when there are fewer choices of nearby reports to consider.

Finally, we considered mispositioned data, which represent a significant obstacle to ship tracking. Generally, mispositioned data are characterized by a wrongly reported position or time. Firstly, for some reports, we identified and corrected potential errors in the reported time variable (see Appendix A3 for details). However, these adjustments have been decided on a qualitative basis only; a full identification of time-shifted reports should use the archive prior to dupelim processing and be performed both within decks and between decks. Similarly, spatially mispositioned data were processed only partially. Speed checks were used to spot isolated mispositioned reports as well as to split known-ID tracks resulting from the aggregation of different ships using the same identifier. Tracks were only split where more than 10% of reports were inconsistent to avoid splitting long tracks containing a few mispositioned reports.

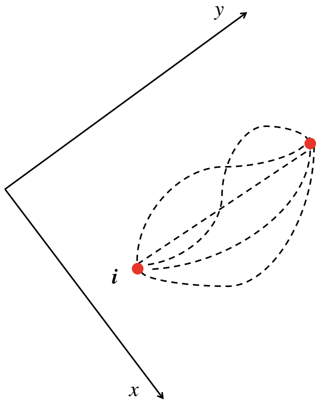

Figure 2. Example of different possible trajectories between (time) consecutive observed positions (i and f) in a known-ID ship track. [Colour figure can be viewed at wileyonlinelibrary.com].

3.2. Model formulation

The tracking model consists of three steps. Firstly, each NA report is tested to see whether it is associated with a gap in a track with known-ID and can therefore be associated with that ID. The second step is the clustering of the residual NAs to give new tracks and the final step is the joining of original and new tracks.

3.2.1. Assignment of reports to existing tracks

Known-ID tracks may be subset into smaller fragments, each defined by the two points closest in time (i and f) and the probability of assigning a N observation to each of these fragments may be modelled as a function of the distance from both i and f. The variance of the expected probability density function (pdf) must increase with the distance from the extremes of each fragment, being maximum in the middle and minimum at i and f. In fact, as shown in Figure 2, from i to f, the ship can move along any of the infinite number of trajectories between the two points, and it is possible to reconstruct its trajectory only within a margin of uncertainty which increases with the distance from the observed positions. The problem then reduces to determine how the pdf variance changes between the two extremes of each track fragment.

A simple model for the evolution of the ship trajectory assumes that the ship moves with a constant speed (the whole trajectory can be always divided in short time intervals to make this a reasonable assumption) and is subject to white noise variations. These variations represent many different processes that might include the effects of the wind, currents and random movements of the ship and will also include a contribution from any inaccuracies in the reported positions perpendicular to the overall direction of travel. For each coordinate, the combined effect of these processes may be modelled as a one-dimensional random walk process. Let x and y be the ship longitude and latitude, respectively, and let the spatial reference system be the one where the ship velocity is parallel to the y-axis. We neglected noise variations in the direction of travel and considered only the component of noise perpendicular to the direction of the ship velocity (y). Hence, if v is the mean speed of the known-ID ship along the *i-f* segment, the probability for a ship with initial position (x_i, y_i) to be at position (x_k, y_k), after time $\delta t \equiv |y_k - y_i| / v$, is a Gaussian

$$P\left(x_k,\ y_k \mid x_i,\ y_i\right) = \frac{1}{\sqrt{4\pi D\ \delta t}}\ e^{-\left(\frac{(x_k - x_i)^2}{4D\ \delta t}\right)} \tag{1}$$

where D can be considered to be a diffusion coefficient, linked to the average ship perpendicular displacement from the line connecting i and f (Ibe, 2013). Equation (1) represents the probability of the ship being at position (x_k, y_k) given its (known-ID) track and can therefore also be used to calculate the probability of an NA report, in this location, being associated with that known-ID track (we call this the assignment pdf). The diffusion coefficient D can then be calculated for every known-ID ship as the variance of the assignment pdf written as a function of the new variable $z_k = \left(x_k - x_i\right) / \sqrt{2\delta t}$

$$D \equiv \sigma^2 = \frac{1}{n-1} \sum_{k=1}^{n} z_k^2 \tag{2}$$

D is in fact a measure of how well the ship trajectory, for small intervals of time, can be approximated by uniform motion: ships with lower D will tend to move in a straighter line than ships with larger D. Knowing D we can write the assignment pdf for an NA report at position $k \equiv (x_k, y_k)$ as

$$P_k = \begin{cases} \frac{1}{\sqrt{4\pi D \left|y_k - y_i\right| / \mathrm{v}}}\ e^{-\left(\frac{(x_k - x_i)^2}{4\mathrm{D}\left|y_k - y_i\right| / v}\right)} & \text{if } y_k \leq y_m \\ \frac{1}{\sqrt{4\pi D \left|y_f - y_k\right| / v}}\ e^{-\left(\frac{\left(x_k - x_f\right)^2}{4D\left|y_f - y_k\right| / v}\right)} & \text{if } y_k > y_m \end{cases} \tag{3}$$

the equation being symmetric around the intermediate point of the *i-f* segment y_m (see Figure 3(a)).

Equation (3) is used to select the best NA candidate (i.e. the one with highest probability) in the assignment process. For cases of reports with some ID information but that were classified as NAs (as described in Section 3.1), the similarity of the available ID information was taken into account (see Section 3.2.3). To eventually accept or reject an NA assignment, some additional basic temporal and spatial constraints are also applied: no ship can record twice at the same time and, for a given known-ID, δt must be no smaller than the observed minimum time gap between two subsequent observations. Moreover, at any time, the ship speed cannot exceed a maximum threshold: this is typically set at 160 km h^{-1}. A large threshold for the ship speed is required as many ICOADS positions have 1° resolution and often tracks are characterized by several reports at the same position followed by one degree 'jumps'.

3.2.2. *Clustering of reports not already assigned to a track*

NA reports not assigned to any known-ID track are then clustered together. Starting from an NA report, representing the initial point of a track, the clustering pdf is modelled similarly to that of the assignment model, the only difference being that, as the final point of each track is unknown, the pdf variance will increase as the distance from the initial NA point increases (see Figure 3(b)). Moreover, when successively clustering observations together, the ship course and speed at the starting point of each track are typically unknown, while in the assignment of NA reports to known-ID tracks the ship velocity was derived for each *i-f* segment. A few ICOADS reports contain information on the ship course and speed, but more usually this information is missing. As a first approximation, a typical ship speed can be computed from the known-ID tracks for each different month. The best guess for the ship course must be instead determined for each individual case. The adopted method is illustrated in Figure 4: the clustering technique relies on the idea that the best guess for the ship course is the direction of the track formed by the combination of n observations, in an appropriate neighbourhood of the starting NA report, which forms the straightest line out of all possible n-point tracks. For a given period of time, the neighbourhood radius is calculated as the mean ship speed derived from all the ships with known-IDs times the length of the chosen period. In particular, recalling that the chosen minimum number of observations per track is four, the clustering procedure is implemented firstly over 4 days and then repeated over 6 days in order not to exclude ships reporting daily and with incomplete reporting sequences.

From the ICOADS speed-course information, when available, or from their computed values, the clustering pdf is calculated for every competing (i.e. taken at the same time) NA observation in the neighbourhood. Keeping track of the assignments, this process is then repeated for each NA observation. As before, the acceptance of a track is then subjected to additional constraints: observations belonging to each track must be temporally ordered, the direction of the track must be unambiguously defined (no tracks with reversals, i.e. changes of direction in both coordinates, are accepted) and the previously defined maximum threshold for the ship speed is adopted.

3.2.3. *Joining of new and existing tracks*

The last step of the model consists of joining tracks. In fact, for tracks obtained via clustering (Section 3.2.2), so far, only observations within a fixed interval of time have been grouped and the maximum temporal coverage for these tracks is either 4 or 6 days. However, the joining step must be also applied to known-ID tracks: in fact, often two tracks with different IDs belong to the same ship and need to be joined. An example is where the ID is derived from a logbook page number and sequential pages need to be joined. The approach to joining tracks is similar to that taken to assign and cluster the observations. There are two cases, the first where there is a long gap in the track of a

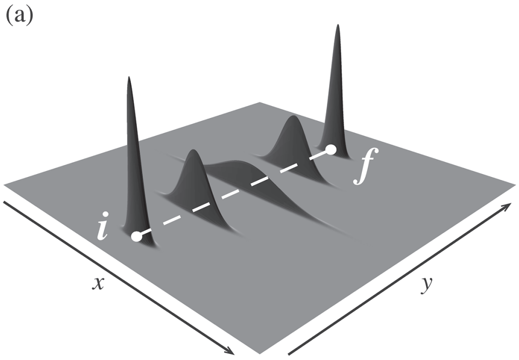

Figure 3. (a) The assignment pdf is shown as a function of the distance from the extreme points (i and f) of the fragment of a known-ID track, defined by two of the points in the track which are closest in time. (b) The clustering pdf is shown as a function of the distance from the starting NA observation (i). Note that in both plots the chosen spatial reference system is the one where the ship velocity is parallel to the y-axis.

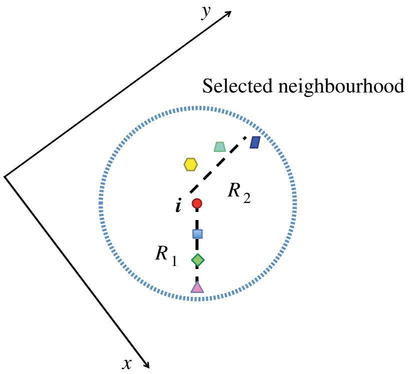

Figure 4. Selection method for the determination of the ship course in the clustering step. Among all the possible competitive tracks starting from an initial NA observation (i), the one selected (R_1) is the one characterized by the straightest line out of all possible n-point tracks. [Colour figure can be viewed at wileyonlinelibrary.com].

known-ID ship, and the second where we try to join tracks constructed from NA reports or from ships with different IDs. The first case uses the assignment pdf with speed and direction determined from the pair of observations spanning the gap. In the second case, we refer to the first (in time) track as the target track and all subsequent tracks as candidates for joining. The clustering pdf is calculated from the speed and direction determined from the target track from the last day of reports or, for ships observing daily, from the last four days. The assignment or clustering pdf is then derived for each point of all the tracks that are potential candidates for joining to the target track. To join whole tracks together, it is necessary to calculate a joint probability for all the points in a candidate track. In order to discriminate between competitive tracks, we introduced the variable

$$L_\perp = \frac{\sum_k P_k \cdot L_{\perp_k}}{\sum_k P_k} \tag{4}$$

where $L_{\perp_k}$ is the perpendicular distance of the k-th observation of a candidate track from the direction line of the target ship. The best guess is then the candidate track that minimizes the distance from the target ship direction line averaged over the single point probabilities: observations with higher P_k are down-weighted, giving more weight to those reports closer in time to the end of the target track. For the case of known-ID candidate tracks, this approach has also the advantage of down-weighting any observation at a large distance away from the main line of the (candidate) track which may be mispositioned.

Some of the additional temporal and spatial constraints used before also apply to this step. No tracks with reversals are accepted and the ship speed cannot exceed 160 km h^{-1} (for reports with coarse resolution positions) or a calculated maximum speed based on the mean and standard deviation of the reports (for reports with higher precision positions). Moreover, the joining operation must be transitive, i.e. if ship A = ship B and ship B = ship C then ship C = ship A: this problem is solved using the union-find algorithm (e.g. Cormen *et al.*, 2001). Additional checks on the similarity of the recorded variable types are also implemented. Specifically, the joining is accepted only for tracks reporting at least three of the same variable types between selected ICOADS variables (country code, sea surface temperature, sea level pressure, air temperature, total cloud cover, wind speed, wind speed indicator, wave direction, visibility, present weather, past weather). As a further constraint, for each potential candidate, the differences between the ID strings were computed (Damerau, 1964; Levenshtein, 1966) and, where competitive candidates were found, the one with more similar ID was chosen.

3.2.4. *Estimating the quality of the derived ship tracks*

One of the best methods to check the results of the tracking analysis is visual inspection. However, the amount of data precludes this except for a few examples. The quantification of the track uncertainty may be used for model validation purposes. To estimate the uncertainty associated with each track, we adopted an ensemble approach. Each possible ‘candidate’ in the assignment, clustering and joining step can be considered as one of the possible N outcomes of a probabilistic process: the track uncertainty must then be an increasing function of the number of possible outcomes (i.e. of possible competitive observations in the assignment, clustering and joining processes).

It is well known that entropy and information can be considered as measures of uncertainty. In conventional information theory (Shannon, 1948), entropy measures the amount of uncertainty of a random variable with a certain number of outcomes, each with probabilities P_i, as

$$u_i = -\sum_i P_i \, log\left(P_i\right) \quad (5)$$

The logarithmic base remains arbitrary, but it is natural to choose base 2 and to measure the amount of uncertainty in bits (Robinson, 2008). The ensemble maximum uncertainty is obtained assuming that all the outcomes are equally probable, $P_i = 1/N$: for a track of m points, the total uncertainty can then be estimated as

$$u_{\text{track}} = \sum_{i=1}^{m} u_{i_{\text{assignment}}} + u_{i_{\text{clustering}}} + u_{i_{\text{joining}}} \quad (6)$$

While the first two terms in the sum are determined separately for each observation, the last one is calculated from the number of competitive tracks in the joining step and the resulting value applied to all the observations in the track. We refer to u_{track} as the 'track quality indicator' (TQI hereafter). The TQI is zero when the track is unambiguous and increases with the number of choices made during the track construction. The TQI is particularly useful when comparing tracks, and a smaller value (especially zero) means that the track is likely to be more reliable. However, a track with a large TQI may indeed be correct, if the appropriate decision is made at every step, and equally a track with TQI of 1 may be wrong if the single choice taken during its construction was not made correctly. This is because the TQI records only the number of choices and does not contain information about how clear-cut any particular choice might have been.

4. Results

The effectiveness of the tracking model presented in this article varies over time, but overall the method works well both in terms of increasing the proportion of reports associated with known-IDs and of increasing the overall length of the tracks.

Generally, the impact of single NA assignments is small: the largest monthly proportion of reports assigned was 2%, but substantially less than 1% was much more typical. Although these percentages were small, this step is important as the remaining NAs are subsequently assumed to form independent tracks and not missing observations from known-ID tracks. In particular, the percentage of assignments increases significantly after the 1950s, when the number of NA reports increases significantly (Figure 1(a)). On the other hand, as shown in Figure 5, the total fraction of observations with known-IDs is greatly increased after tracking (Figure 5(b)), compared to the ICOADS original record (Figure 5(a)). In particular, the impact of the tracking is particularly clear during the 1860s, when the percentage of observations in ICOADS with known-ID is less than 10%, rising to more than 75% after tracking. Note that in Figure 5, we included both the observations from the *tracked decks* and the *complete decks* categories, in order to show all the final 'usable' tracks. Figure 5(a) and (b) also show the breakdown of the fraction of observations according to the length of the track (i.e. the number of points per track). Overall, not only observations are assigned or clustered together, increasing the percentage of tracked observations up to almost 90% for most of the record, but also there is an increase in the number of reports associated with long (more than 50 observations per month) and medium (between 50 and 10 observations per month) tracks. Moreover, observations with unusable ID information, i.e. tracks with invalid or generic IDs, are processed and the final identified IDs are all unique and characterized by a minimum of four reports per track per month.

Generally, visual inspection of multiple cases proved to be very useful to test the model results as well as to improve the method and to refine its details. Figure 6 shows some selected examples illustrating the results of the various model steps on the data. Figure 6(a) and (b) illustrates examples of the adopted pre-processing criteria. In particular, Figure 6(a) shows the case of a known-ID track (104906), which, during pre-processing, was merged with another known-ID (04906) from a different deck. Figure 6(b) shows an example of tracks originally formed by individual reports with sequences of unique numeric IDs, identified and clustered prior to tracking. The remaining panels of Figure 6 show examples of the typical results of applying the assignment, clustering and joining steps. Specifically, Figure 6(c) shows the case of NA observations assigned to a known-ID track (13560) which is then joined to another known-ID track (31313560) containing a common sequence not identified during the pre-processing. Figure 6(d) illustrates the case of NA observations clustered together to create a new track. Finally, Figure 6(e) and (f) shows examples of tracks formed by joining different IDs: the match may arise not only between tracks with formerly known-IDs, as in Figure 6(f), but also between new and known-ID tracks (Figure 6(e)).

Overall, from the analysis of individual cases much can be learned about ICOADS reports and their sources. For instance, tracks belonging to different decks and originally classified by different IDs may be joined together: the tracks in Figure 6(a) belong, respectively, to deck 201 and 194, while in Figure 6(e) the track made by formerly NA reports belonging to deck 155 is joined to known-ID tracks originating from deck 720. Figure 6(e) also provides an example of a case where observations from decks with different rounding of position information are clustered in the same track: in deck 155 latitude and longitude are approximated to 0.5° whilst the position reports in deck 720 are rounded to whole degrees.

In addition to specific cases, the comparison between the original ICOADS record and that resulting from the tracking analysis also demonstrates the effectiveness of

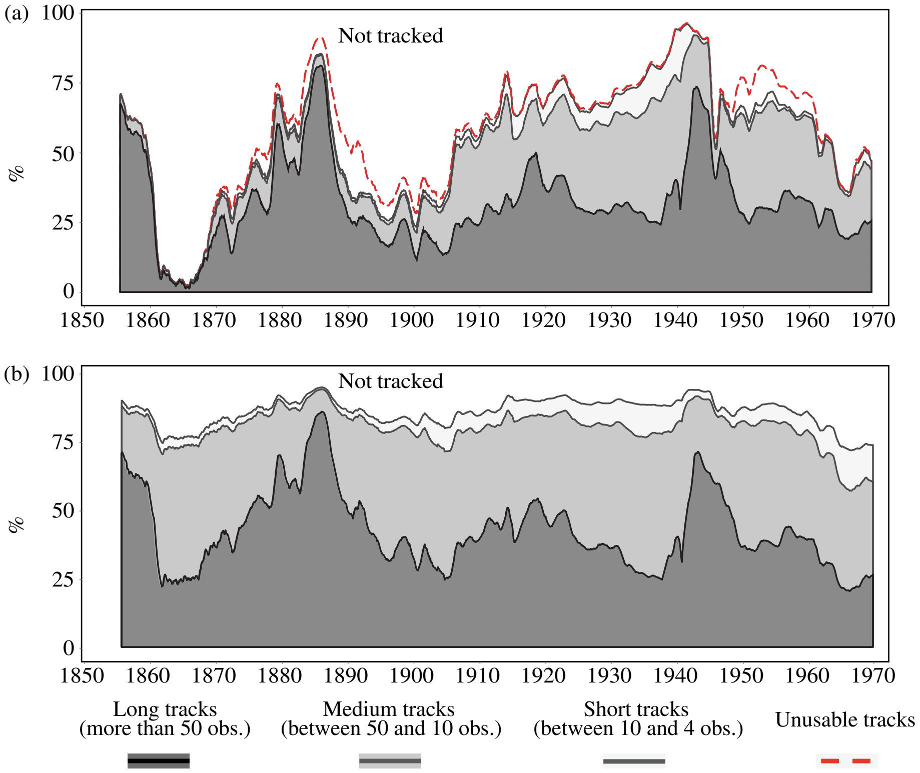

Figure 5. Percentage of observations with known-IDs before (a) and after (b) the tracking analysis. In both figures, the sum of the grey shaded areas identifies all the reports with a valid known-ID. Also shown is the percentage of reports in *long tracks* (more than 50 observations per month) represented by the dark grey shaded area, the fraction of *medium tracks* (between 50 and 10 observations per month), specified by the medium grey shaded area and the percentage of *short tracks* (between 10 and 4 observations per month) and *unusable tracks,* identified respectively by the light grey shaded area and the area between the solid grey line and the black dashed one. Note that only decks from the *tracked* and *complete decks* categories have been included. Both plots show monthly data filtered with a 12-month running mean. [Colour figure can be viewed at wileyonlinelibrary.com].

the method. Figure 7 shows the known-ID tracks and the NA reports before (left) and after (right) applying the tracking method for three different periods: not only do new tracks appear but also IDs characterized by 'odd' patterns are rearranged to form new, more consistent, tracks. For example, in Figure 7(a) and (b), the east Pacific is characterized in the original ICOADS record by several 'zig zag' tracks which correspond to truncated IDs (e.g. a four digit ID when the expected format was eight digits) representing a mixture of different ships. Figure 7(a) also exhibits some examples of mispositioned data remaining in the final record: in fact, as mentioned before, in order not to fragment long tracks with few mispositioned reports, mispositioned data were not always removed, as appears in some tracks in the west Pacific characterized by 'sudden' jumps.

In order to test the assignment of IDs through tracking, we performed an analysis of variance (R Core Team, 2015) of SST within 20° monthly grid boxes, calculating the percentage of variance explained by partitioning the observations between IDs. The location of each observation within the 20° box is the factor that explains the most variance of the SST observations. Using the ID as an additional factor is equivalent to assuming each ship has a constant SST bias, and that accounting for these biases would improve the consistency of the data. Figure 8 shows the percentage of variance that can be explained using different assignments of ID. When the pre-tracking known-IDs are used to group the observations typically 10–30% of the variance can be explained, varying with the proportion of observations assigned to a known-ID compared to the number of NAs (see Figure 1(b)). In contrast, when the tracked-IDs are used the variance explained is larger, and more consistent over time, typically between 20 and 30% over the period. For periods where the majority of observations have a known-ID, such as the late 1880s, the variance explained by the tracked-IDs and known-IDs is similar. When a random clustering of observations with a number of groups similar to the tracked-IDs is used, the variance explained is typically between 5 and 10%. The increase in the variance explained by the tracked-IDs, compared to known-IDs, gives confidence that the tracking process is working, with observations with similar properties clustered together. The much smaller percentage of variance explained for the randomly clustered data gives confidence that the improvement is not due to changes in the number of degrees of freedom.

Figure 9 shows the percentage of observations for different ranges of the TQI, computed as in Equation (6). The quality indicator varies over time, but tracks during the periods between about 1890 and 1910 and during World War II are particularly low quality (high TQI), as NA reports are concentrated together giving a larger

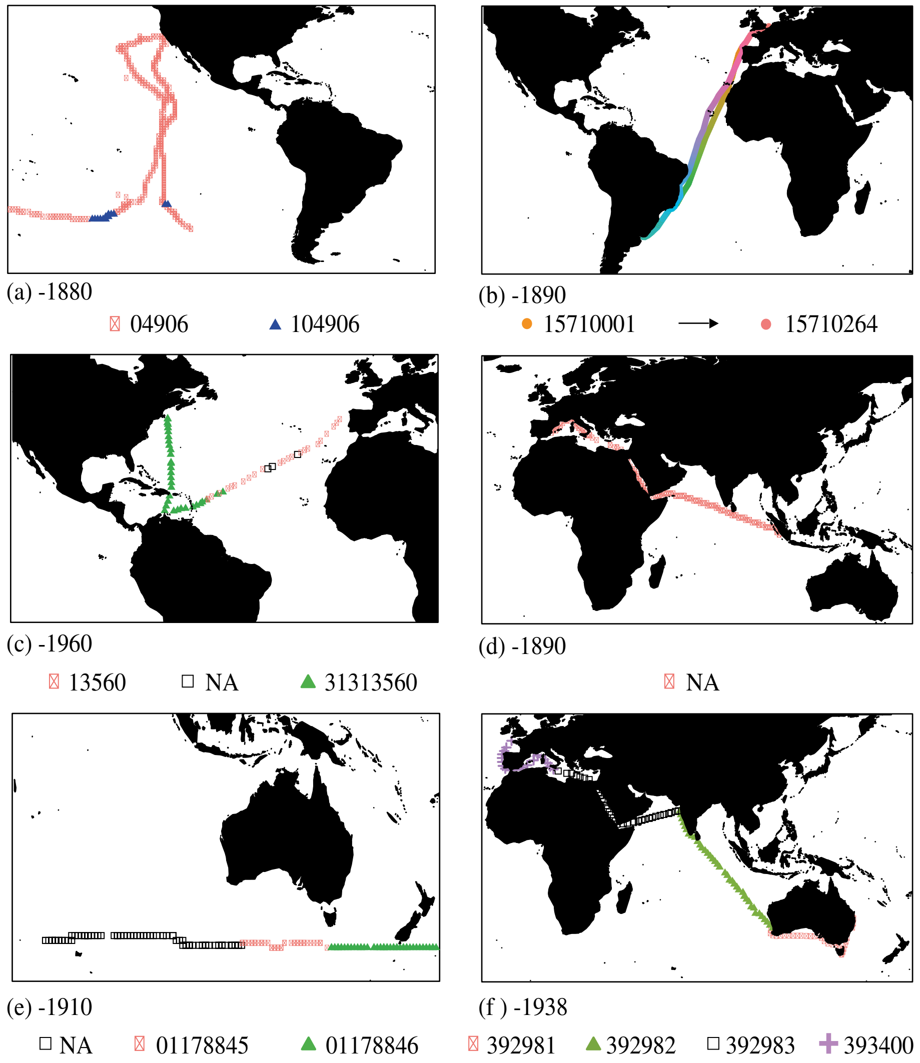

Figure 6. (a) ID 104906 was converted to 04906, dropping the first digit to make the ID consistent with those in a different deck (QC ID). (b) Sequentially ordered IDs (represented by dots of sequentially ordered colours), originally assigned to only one report, clustered together. (c) NA observations assigned to a known-ID track (13560), and then joined to another known-ID track (31313560). (d) NA observations clustered together to create a new track. (e)–(f) Examples of tracks formed by joining different IDs. [Colour figure can be viewed at wileyonlinelibrary.com].

number of choices for any assignment, clustering or joining. Different ranges can be computed (tracks not modified by the tracking analysis must have $u_{\text{track}} = 0$) and can be explored separately. Figure 10 shows an example for 1870 of the tracks with a TQI of zero and those with larger values ($u_{\text{track}} \geq 2$) separately. Tracks with a zero TQI are those that are unchanged by the processing, or where reports were assigned, clustered or track segments joined with no competing reports. Comparing Figures 10(a) and 7(a) shows that some of the tracks with a TQI of zero are new tracks. The tracks which may be of lower quality (Figure 10(b)) are concentrated in the major shipping lanes for that period, which is also the region where NA reports are concentrated (compare with Figure 7(a)).

In order to explore the record after the tracking analysis, we created different indicators describing the track characteristics, as the track speed, its temporal coverage and a measure of the difference between the strings of any combined ID. These flags may be used to understand the quality of the derived record but can also help to describe the evolution of the marine observing system and its changes. Figure 11 shows the 12-month running mean ship speed overlaid on a density plot of the speed distribution. In the early record, before 1900, many ships have distributions containing either zero or large speed due to the rounding of locations to 1° resolution in some decks. Changes to shipping during World War II are also evident, with a clear decrease in the ship speed. During the war, the percentage of ships characterized by a low speed (less than 5 km h^{-1}) or almost stationary in their position exceeds 40%, however the drop in the mean ship speed is not so dramatic due to the presence of some high speeds of 50 km h^{-1} or more (off the scale of the plot). The overall reduction in speed is expected from the convoy system adopted to protect merchant ships as the convoy had to assemble and then move at the speed of the slowest ship (Burn, 1998). The SST bias correction model (Folland and Parker, 1995) assumes a

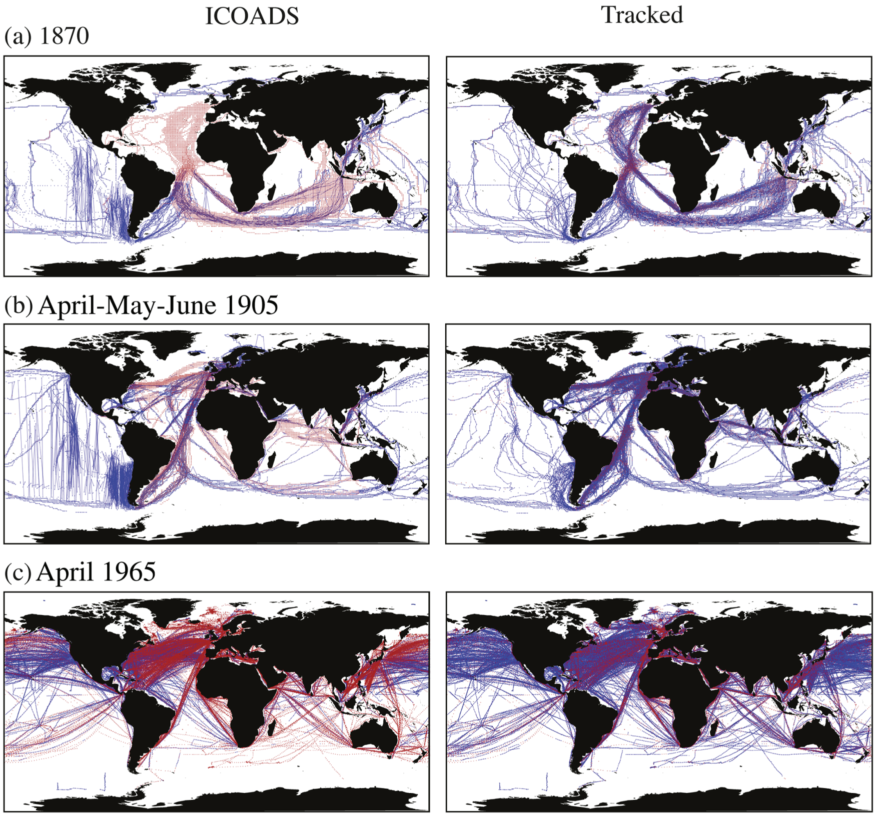

Figure 7. Comparison between the original ICOADS record and the record after the tracking analysis (observations in the *tracked decks* category only). Known-IDs (blue lines and points) and NAs observations (red points) are shown for the original ICOADS record (left) and the record after the tracking analysis (right) for three different periods.

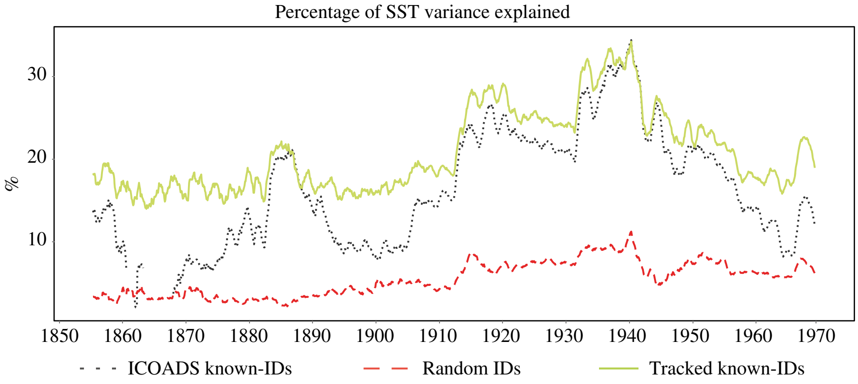

Figure 8. Percentage of SST variance explained (observations in the *tracked decks* category only). Shown is the percentage of variance for the SST that can be explained by partitioning the observations using the known-IDs in the original ICOADS record (black dotted line), the tracked known-IDs (green solid line) and using random clustered observations (red dashed line). The plot shows monthly data filtered with a 12-month running mean. [Colour figure can be viewed at wileyonlinelibrary.com].

linear increase in the ship speed from 4 to 7 m s^{-1} (approximately 14–25 km h^{-1}) over the period 1850–1940. Ship speeds derived from our analysis show an increase of similar magnitude, but the change is not linear.

5. Summary and the potential for future improvements

ICOADS ship ID information is often missing or unusable, preventing the linking of reports to an individual ship. In this study, we used a probabilistic approach to reconstruct ship voyages that groups observations together to give plausible ship tracks. The increased proportion of reports associated with known-IDs and the increased overall length of the tracks illustrates the efficacy of the method. Validation was initially by visual inspection of the tracks and statistics indicating track quality were calculated.

Issues in the data (such as duplicates, mispositioned reports and rounding of position information) were found to be one of the main challenges in implementing the

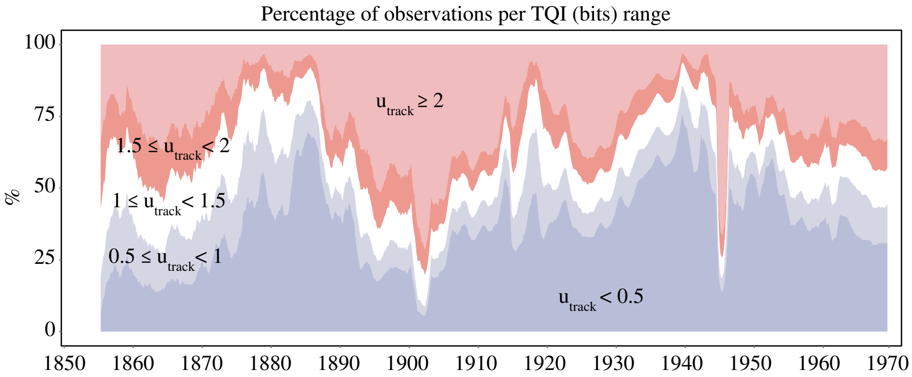

Figure 9. Percentage of observations for different TQI (bits) ranges (observations in the *tracked decks* category only). Shown is the percentage of observations for five different ranges of the track quality indicator, with the quality of the track decreasing from the bottom ($u_{track} < 0.5$) to the top ($u_{track} \geq 2.0$). Note that tracks not modified by the tracking analysis have $u_{track} = 0$. The plot shows monthly data filtered with a 12-month running mean. [Colour figure can be viewed at wileyonlinelibrary.com].

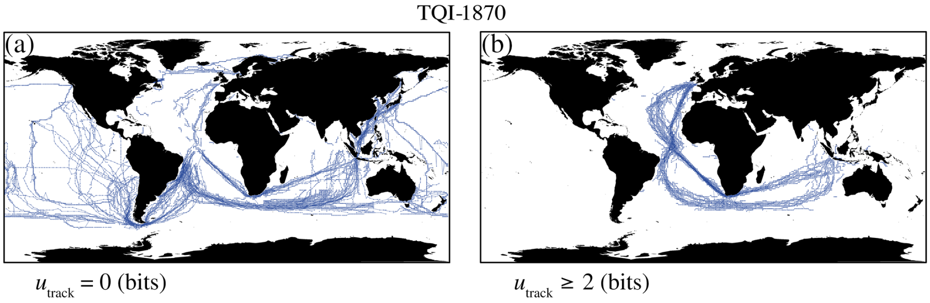

Figure 10. Example of tracks with a TQI (bits) of zero and tracks with larger values (observations in the *tracked decks* category only). Shown for 1870 are the tracked observations with a track quality indicator of zero (a) and those with larger values of u_{track} (b). [Colour figure can be viewed at wileyonlinelibrary.com].

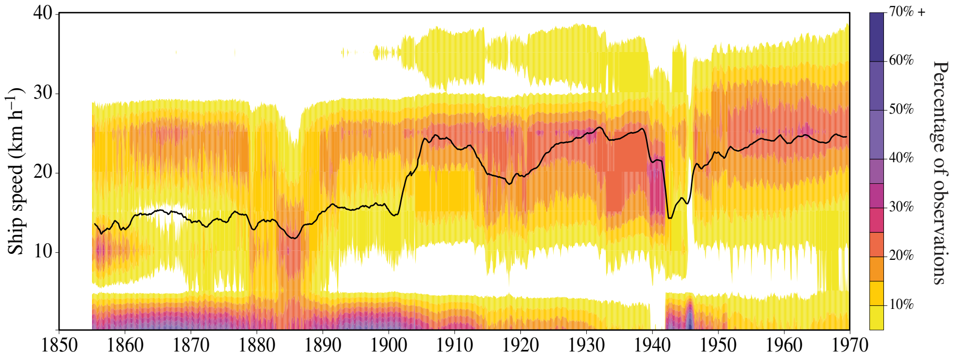

Figure 11. Ship speed (km h^{-1}) (observations in the *tracked decks* category only). The mean ship speed (black line) is presented overlaid on a density plot of the speed distribution. The plot shows monthly data filtered with a 12-month running mean. [Colour figure can be viewed at wileyonlinelibrary.com].

tracking algorithm. There is much to be gained from the identification of duplicates and relocation of mispositioned data, as ship tracking inevitably works best when observations are correctly located and unique. Particularly, in the early record before 1900, many tracks are characterized by truncated positions (longitude and/or latitude rounded to the closest degree) or by 'dead reckoning', where a known position is advanced by means of recorded heading, speed and time. Therefore, even in the absence of corrupted position information, some IDs may show 'jumps' in their tracks. Many of the reports with rounded position information are contained in legacy decks (Freeman *et al.*, 2015; personal communication) with no ID information, which are then fragmented by the ICOADS dupelim processing.

Presently, the tracking algorithm does not account for the presence of land. It is possible therefore that the constructed tracks may require the ship to cross land. This was not a problem that was identified as prevalent from the visual inspection of results, due to the constraints applied to the process, including speed and similarity of reports. However, improving this aspect of the tracking is clearly desirable.

Mispositioned data were not fully handled by the tracking algorithm. For known-ID tracks, mispositioned data were identified but not relocated. Interpolation techniques could be used to correct, as a first approximation, the errors in the positions of some reports. Mispositioned reports with no ID information cannot be easily relocated. Positional inaccuracy, including some gross mispositioning of reports, affects the quality of the record and may impact the quality of any derived climate analysis.

Even though results are encouraging and show that the implemented data association method works well, future effort is needed to elucidate the reasons of some of the data issues present in ICOADS and to eventually correct them. ICOADS has long been the focus of the main improvements in the understanding of marine climatology (Kennedy, 2014). Alongside ongoing efforts to identify and digitize new data, a critical reprocessing of ICOADS legacy data is needed to provide a more reliable baseline for the understanding of the future global climate.

Acknowledgements

This work was funded by a grant from the Natural Environment Research Council (NE/J020788/1). The authors thank the anonymous reviewers for their help in improving this article as wells as the ICOADS team for useful advice and providing the data. The authors thank Hans Hersbach at ECMWF for helpful comments and discussion.

Appendix

This appendix describes the criteria applied in the pre-processing step.

Appendix A1. Selection of decks for analysis

Table A1 shows the decks excluded from our analysis. The excluded decks can be grouped into three main categories:

1. *Buoy-only decks*: the observations coming from these decks have missing platform type PT (e.g. PT = 0–5 for ship types) and must be excluded explicitly.
2. *Complete decks*: unique data with full ID information thought to be from unique sources. While the majority of these decks were not tracked because of their good data quality, some have inconsistent positions due to dead reckoning.
3. *Decks 732 and 874*: observations from deck 732 between 1958 and 1974 were identified as being incorrectly located and 17 areas of 5° or blocks of 5° areas were found artificially warm or cold relative to neighbouring areas and relative to other observations within the area (Kennedy *et al.*, 2011b). Reports from deck 874 [US e-logbook software package, Shipboard Environmental Acquisition System (SEAS)] will be excluded from the next ICOADS Release as they were found to be unreliable (Freeman *et al.*, 2015; personal communication). We did not therefore include this deck in our tracking.

Table A1. List of ICOADS decks, containing data with PT = 0–5 or missing, that were excluded from tracking analysis.

Deck	Description	Category
143	Pacific Marine Environmental Laboratory (PMEL) Buoys	1
144	TAO/TRITON and PIRATA Buoys (from PMEL and JAMSTEC)	1
145	PMEL Equatorial Moorings and Island Stations	1
239	British Navy (HM) Ships	2
245	Royal Navy Ships Logs (keyed by 2007)	2
246	Antarctic Expeditions: Printed/Published (Met. Office)	2
247	Atmospheric Circ. Reconstructions over the Earth (ACRE) Data	2
701	US Maury Collection	2
702	Norwegian Logbook Collection	2
707	US Merchant Marine Collection (1912-46), 700 series	2
714	Canadian Integrated Science Data Mgmt. (ISDM) Buoys	1
730	Climatological Database for the World's Oceans (CLIWOC)	2
731	Russian S.O. Marakov Collection	2
732	Russian Marine Met. Data Set (MARMET)	3
734	Arctic Drift Stations	2
736	Byrd Antarctic Expedition (keyed by Hollings Scholars)	2
740	Research Vessel (R/V) Data Quality Evaluated by FSU/COAPS	2
761	Japanese Whaling Ship Data (CDMP/MIT digitization)	2
762	Japanese Kobe Collection Data (keyed after decks 118-119)	2
780	NODC/OCL World Ocean Database (WOD)	2
793–795	NCEP BUFR GTS	1
874	Shipboard Environmental (Data) Acquisition System (SEAS)	3
883	US National Data Buoy Center (NDBC)	1
900	Australian	2

The excluded decks are in three different categories: *buoy-only decks with missing PT* (1), *complete deck*s (2) and *decks 732 and 874* (3).

Note that for the analysed period (1855–1969) the number of reports in category 1 is negligible and is not shown in Figure 5(a). Finally, to be added to the excluded observations, are those without day or hour information from any deck. Prior to 1855, there are a substantial number of reports with missing time information, while after 1855, over 90% of reports, and by 1858 over 99% of reports, have full time information.

Appendix A2. Flagging of invalid ID types

Table A2 shows the possible ID types for each deck. If the ID was not of the expected format, it was flagged as invalid and treated as NA in the tracking analysis. The extant ID information was however used in the similarity testing of ID information used to choose amongst competing reports or tracks. For example, a truncated version of an ID would be matched in preference to an ID of different format.

Table A2. ID types per deck.

ID type	Decks
Ship name	704, 897
Callsign	926, 927
N	188
NN	128
NNN	116, 128, 196, 197, 926, 927
NNNN	116, 117, 128, 186, 187, 197, 667, 735, 926, 927
NNNNN	116, 189, 194, 195, 201, 202, 203, 206, 207, 209, 210, 211, 213–215, 215, 218, 221, 224–227, 229, 230, 233, 234, 254, 255, 735, 926, 927
NNNNNN	118, 119, 189, 194, 197, 216, 254, 926, 928
NNNNNNN	197, 928
NNNNNNNN	192, 215, 720, 902
-NNN	128, 927
-NNNN	116, 195
ANN	128, 197, 849
ANNN	128, 197, 889
ANNNNN	197
ANNNNNN	197
NNNNA	762
NNNSNNNN & ANNSNNNN, ANNSNNNN & NNSNNNN & NNSSNNNN & NNNSSNNN & ANNSSNNN	184
NNNNANNN	192
NSNNNN	194
US Journals ID (AANNNNN & AANNNNNN & AAANNNNN & AANNNNNA)	705
OWS ID (NNNNA, starting C7 or 4Y)	896

The expected ID types are listed for each deck in the *tracked decks category* only (see Table A1). Key: N = [0–9]; A = [A–Z, a–z]; S = space; note '-', 'C7' and '4Y' represent their specific characters.

Appendix A3. Pre-processing of IDs and time information

Further pre-processing criteria were adopted to correct the ID or the time information.

1. Sequential IDs in deck 720 before 1891, originally assigned to only one report, were clustered together.
2. Based on preliminary results from the tracking analysis, the IDs from some decks were altered to match formats in other decks that were seen to have common data. In particular, we removed the first digit in six digits IDs from deck 194, while for deck 701 additional information from ICOADS supplementary material was appended to make unique IDs from the same ship name.
3. Reports from deck 201 before 1899 taken at the GMT midnight were moved one day before the reported date.

Appendix A4. Constraints on tracking jointly reports from different decks

Finally, Table A3 lists the constraints we adopted on tracking jointly reports from different decks in the *tracked decks* category. Decks were classified according to different types. Data from decks that were thought to be unique but did not have complete ID information (as opposed to the *complete decks* in Table A1, unique and with complete ID information) were only tracked within the deck. On the other hand, decks that were though to have common data were tracked jointly and data coming from these decks were allowed to be assigned, clustered and joined together.

Table A3. Constraints on tracking jointly reports from different decks in the *tracked decks* category.

Deck types	Decks	Action
Decks with substantial ID information (few missing or invalid), that could be unique	117, 118, 119, 187, 188, 195, 229, 667, 704, 705, 706	Tracked within deck
Decks containing data from single ship with missing ID	897	Tracked within deck
Decks with missing or partial ID information, that could be unique	666, 899	Tracked within deck
Decks that may have common data	110, 116, 128, 150, 151, 152, 155, 156, 184, 185, 189, 192, 193, 194, 196, 197, 201, 202, 203, 204, 205, 206, 207, 209, 210, 211, 213, 214, 215, 216, 218, 221, 223, 224, 226, 227, 230, 233, 234, 254, 255, 281, 555, 700, 720, 735, 749, 792, 849, 850, 888, 889, 892, 896, 898, 901, 902, 926, 927, 928, 999	Tracked together
Ice stations with some common data	186, 733	Tracked together

The table shows the decks that, according to their deck type, were tracked together and those that were tracked within the deck only.

References

Allan R, Ansell T. 2006. A new globally complete monthly historical gridded mean sea level pressure dataset (HadSLP2): 1850–2004. *J. Clim.* **19**: 5816–5842, doi: 10.1175/JCLI3937.1.

Allan R, Brohan P, Compo GP, Stone R, Luterbacher J, Brönnimann S. 2011. The international Atmospheric Circulation Reconstructions over the Earth (ACRE) initiative. *Bull. Am. Meteorol. Soc.* **92**: 1421–1425, doi: 10.1175/2011BAMS3218.1.

Berry DI, Kent EC. 2009. A new air-sea interaction gridded dataset from ICOADS with uncertainty estimates. *Bull. Am. Meteorol. Soc.* **90**: 645–656, doi: 10.1175/2008BAMS2639.1.

Berry DI, Kent EC. 2011. Air-sea fluxes from ICOADS: the construction of a new gridded dataset with uncertainty estimates. *Int. J. Climatol.* **31**: 987–1001, doi: 10.1002/joc.2059.

Blunden J, Arndt DS. 2014. State of the climate in 2013. *Bull. Am. Meteorol. Soc.* **95**: S1–S279, doi: 10.1175/2014BAMSStateoftheClimate.1.

Boer GJ. 2011. The ratio of land to ocean temperature change under global warming. *Clim. Dyn.* **37**: 2253–2270, doi: 10.1007/s00382-011-1112-3.

Burn A. 1998. *Fighting Captain: The Story of Frederic John Walker RN, CB, DSO and the Battle of the Atlantic*. Pen & Sword Books Ltd: Barnsley, UK, 224 pp. ISBN: 9781844154395.

Compo GP, Whitaker JS, Sardeshmukh PD, Matsui N, Allan RJ, Yin X, Gleason BE, Vose RS, Rutledge G, Bessemoulin P, Brönnimann S, Brunet M, Crouthamel RI, Grant AN, Groisman PY, Jones PD, Kruk MC, Kruger AC, Marshall GJ, Maugeri M, Mok HY, Nordli Ø, Ross TF, Trigo RM, Wang XL, Woodruff SD, Worley SJ. 2011. The twentieth century reanalysis project. *Q. J. R. Meteorol. Soc.* **137**: 1–28, doi: 10.1002/qj.776.

Cormen TH, Leiserson CE, Stein C. 2001. *Chapter 21: Introduction to Algorithms*, 2nd edn, MIT Press and McGraw-Hill (ed). The MIT Press: Cambridge, MA and London. 498–524. ISBN: 02620 32937.

Damerau FJ. 1964. A technique for computer detection and correction of spelling errors. *Commun. ACM* **7**(3): 171–176, doi: 10.1145/363958.363994.

Folland CK, Parker DE. 1995. Correction of instrumental biases in historical sea-surface temperature data. *Q. J. R. Meteorol. Soc.* **121**: 319–367, doi: 10.1002/qj.49712152206.

García-Herrera R, Können GP, Wheeler D, Prieto MR, Jones PD, Koek FB. 2005. CLIWOC: a climatological database for the world's oceans 1750–1854. *Clim. Change* **73**: 1–12, doi: 10.1007/s10584-005-6952-6.

Hersbach H, Poli P, Dee D. 2015. The observation feedback archive for the ICOADS and ISPD data sets, ECMWF ERA Report Series, ECMWF, Shinfield Park, Reading, UK, 18 pp.

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

Huang B, Banzon VF, Freeman E, Lawrimore J, Liu W, Peterson TC, Smith TM, Thorne PW, Woodruff SD, Zhang H. 2015. Extended reconstructed sea surface temperature version 4 (ERSST.v4). Part I: upgrades and intercomparisons. *J. Clim.* **28**: 911–930, doi: 10.1175/JCLI-D-14-00006.1.

Ibe OC. 2013. *Chapter 4: Elements of Random Walk and Diffusion Processes*. John Wiley & Sons, 50–52. ISBN: 978-1-118-61809-7.

IPCC. 2013. Summary for policymakers. In *Climate Change 2013: The Physical Science Basis*, Stocker TF, Qin D, Plattner G-K, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, Midgley PM (eds). Cambridge University Press: Cambridge, UK and New York, NY.

James RW, Fox PT. 1972. Comparative sea surface temperature measurements. Reports on Marine Science Affairs Report No. 5, WMO: Geneva, 27 pp.

Jones PD, Wigley PML. 2010. Estimation of global temperature trends: what's important and what isn't. *Clim. Change* **100**: 59–69, doi: 10.1007/s10584-010-9836-3.

Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Leetmaa A, Reynolds R, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo KC, Ropelewski C, Wang J, Jenne R, Joseph D. 1996. The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.* **77**: 437–471, doi: 10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

Kennedy JJ. 2014. A review of uncertainty in in situ measurements and data sets of sea surface temperature. *Rev. Geophys.* **52**(1): 1–32, doi: 10.1002/2013RG000434.

Kennedy JJ, Rayner NA, Smith RO, Parker DE, Saunby M. 2011a. Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 1. Measurement and sampling uncertainties. *J. Geophys. Res. Atmos.* **116**: D14103, doi: 10.1029/2010jd015218.

Kennedy JJ, Rayner NA, Smith RO, Parker DE, Saunby M. 2011b. Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization. *J. Geophys. Res. Atmos.* **116**: D14104, doi: 10.1029/2010jd015220.

Kent EC, Berry DI. 2008. Assessment of the Marine Observing System (ASMOS): Final Report, NOCS Research and Consultancy Report No. 32, 55 pp.

Kent EC, Challenor PG. 2006. Toward estimating climatic trends in SST. Part II: random errors. *J. Atmos. Oceanic Technol.* **23**: 476–486, doi: 10.1175/JTECH1844.1.

Kent EC, Taylor PK. 2006. Toward estimating climatic trends in SST. Part I: methods of measurement. *J. Atmos. Oceanic Technol.* **23**: 464–475, doi: 10.1175/jtech1843.1.

Kent EC, Taylor PK, Truscott BS, Hopkins JS. 1993. The accuracy of voluntary observing ships' meteorological observations – results of the VSOP-NA. *J. Atmos. Oceanic Technol.* **10**: 591–608, doi: 10.1175/1520-0426(1993)010<0591:taovos>2.0.co;2.

Kent EC, Woodruff S, Berry DI. 2007. Metadata from WMO publication no. 47 and an assessment of voluntary observing ship observation heights in ICOADS. *J. Atmos. Oceanic Technol.* **24**(2): 214–234, doi: 10.1175/JTECH1949.1.

Kent EC, Kennedy JJ, Berry DI, Smith RO. 2010. Effects of instrumentation changes on sea surface temperature measured *in situ*. *WIREs Clim. Change* **1**: 718–728, doi: 10.1002/wcc.55.

Kent EC, Rayner NA, Berry DI, Saunby M, Moat BI, Kennedy JJ, Parker DE. 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.

Levenshtein V. 1966. Binary codes capable of correcting deletions, insertions, and reversal. *Sov. Phys.Dokl.* **10**(8): 707–710.

Minobe S, Maeda A. 2005. A 1° monthly gridded sea-surface temperature dataset compiled from ICOADS from 1850 to 2002 and Northern Hemisphere frontal variability. *Int. J. Climatol.* **25**(7): 881–894, doi: 10.1002/joc.1170.

Poli P, Hersbach H, Tan D, Dee D, Thépaut JN, Simmons A, Peubey C, Laloyaux P, Komori T, Berrisford P, Dragani R, Trémolet Y, Hólm E, Bonavita M, Isaksen L, Fisher M. 2013. The data assimilation system and initial performance evaluation of the ECMWF pilot reanalysis of the 20th-century assimilating surface observations only (ERA-20C), ERA Report Series No. 14, 59 pp.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna. http://www.R-project.org/ (accessed 25 June 2015).

Robinson DW. 2008. Entropy and uncertainty. *Entropy* **10**: 493–506, doi: 10.3390/e10040493.

Shannon CE. 1948. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**: 379–423, doi: 10.1002/j.1538-7305.1948.tb01338.x.

Slutz RJ, Lubker SJ, Hiscox JD, Woodruff SD, Jenne RL, Joseph DH, Steurer PM, Elms JD. 1985. Comprehensive ocean-atmosphere data set; release 1. NOAA Environmental Research Laboratories, Climate Research Program, Boulder, CO, 268 pp (NTIS PB86-105723).

Smith TM, Reynolds RW. 2002. Bias corrections for historical sea surface temperatures based on marine air temperatures. *J. Clim.* **15**: 73–87, doi: 10.1175/1520-0442(2002)015<0073:bcfhss>2.0.co;2.

Smith TM, Reynolds RW. 2004. Improved extended reconstruction of SST (1854-1997). *J. Clim.* **17**: 2466–2477, doi: 10.1175/1520 0442(2004)017<2466:IEROS>2.0.CO;2.

Stickler A, Brönnimann S, Valente MA, Bethke J, Sterin A, Jourdain S, Roucaute E, Vasquez MV, Reyes DA, Allan R, Dee D. 2014. ERA-CLIM: historical surface and upper-air data for future reanalyses. *Bull. Am. Meteorol. Soc.* **95**: 1419–1430, doi: 10.1175/bams-d-13-00147.1.

Stott PA, Gillett NP, Hegerl GC, Karoly DJ, Stone DA, Zhang X, Zwiers F. 2010. Detection and attribution of climate change: a regional perspective. *WIREs Clim. Change* **1**: 192–211, doi: 10.1002/wcc.34.

Sutton RT, Dong B, Gregory JM. 2007. Land/sea warming ratio in response to climate change: IPCC AR4 model results and comparison with observations. *Geophys. Res. Lett.* **34**: L02701, doi: 10.1029/2006gl028164.

Thompson DWJ, Kennedy JJ, Wallace JM, Jones PD. 2008. A large discontinuity in the mid-twentieth century in observed global-mean surface temperature. *Nature* **453**: 646–649, doi: 10.1038/nature06982.

Wilkinson C, Woodruff SD, Brohan P, Claesson S, Freeman E, Koek F, Lubker SJ, Marzin C, Wheeler D. 2011. Recovery of logbooks and international marine data: the RECLAIM project. *Int. J. Climatol.* **31**: 968–979, doi: 10.1002/joc.2102.

Willett KW, Jones PD, Gillett NP, Thorne PW. 2008. Recent changes in surface humidity: development of the HadCRUH dataset. *J. Clim.* **21**: 5364–5383, doi: 10.1175/2008JCLI2274.1.

Woodruff SD, Slutz RJ, Jenne RL, Steurer PM. 1987. A comprehensive ocean-atmosphere data set. *Bull. Am. Meteorol. Soc.* **68**: 1239–1250, doi: 10.1175/1520-0477(1987)068<1239:ACOADS>2.0.CO;2.

Woodruff SD, Worley SJ, Lubker SJ, Ji Z, Freeman EJ, Berry DI, Brohan P, Kent EC, Reynolds RW, Smith SR, Wilkinson C. 2011. ICOADS Release 2.5: extensions and enhancements to the surface marine meteorological archive. *Int. J. Climatol.* **31**(7): 951–967, doi: 10.1002/joc.2103.