



Copernicus Climate Change Service Global Land and Marine Observations Database

Documentation for marine duplicate identification and linking of platform identifiers

Issued by: NUIM / Peter Thorne

Date: 25/11/2019

Ref: C3S_D311a_Lot2.????

Official reference number service contract: 2017/C3S_311a_Lot2_NUIM/SC2

Contributors

NATIONAL OCEANOGRAPHY CENTRE (NOC)

1. Elizabeth C. Kent
2. David I. Berry
3. Irene Pérez González
4. Richard Cornes

MET OFFICE UK

1. John Kennedy



Table of Contents

1. Introduction	6
2. Background	7
2.1 Marine observational data sources	7
2.2 Data selection	7
2.3 Preprocessing of IDs	11
2.3.1 ID replacements	11
2.4 Preprocessing of data record for duplicate identification	15
2.4.1 Data that fails QC	15
2.4.2 Adding information used in duplicate identification processing.	15
2.5 Correcting time errors	21
2.5.1 Incorrect dates and times	21
2.5.2 File format with corrected dates and times	22
3. Identifying duplicates	23
3.1 Why do we have duplicates?	23
3.2 Approach to Duplicate Identification	23
3.2.1 Summary of ICOADS duplicate identification	23
3.2.2 Summary of approach taken here	23
3.3 Duplicate Record Identification Procedure	24
3.3.1 Finding potential duplicate pairs	24
3.3.2 Identifying duplicate groups and selecting the "best" duplicate	29
3.3.3 Output data file format	32
3.4 Duplicate Records by ID	33
3.4.1 Linking IDs	33
3.4.2 Forming ship tracks for linked IDs	35
4. Results	36
4.1 Duplicate identification and ID linking results	36
5. Summary	41
References	42



Annex: Software used in processing	44
Base R (version 3.5.1),	44
jsonlite 1.6	44
stringdist 0.9.5.1	44
lubridate 1.7.4	44
data.table 1.12.2	44
igraph 1.2.4.1	44
maps 3.3.0	44
local package versions of MO climatological check and track check	44



Executive Summary

The C3S 311a Lot 2 (Global Land and Marine Observations Database) service is concerned with the provision of globally distributed land and marine surface meteorological records from the national archives. The service includes inventorying of, and brokering access to, data sources, their harmonization (via conversion to a Common Data Model (CDM), merging, and quality assurance) and their provision via the Copernicus Climate Change Service (C3S) Climate Data Store (CDS).

This document describes processing applied to homogenise platform identifier information and to identify duplicate observations in the International Comprehensive Ocean-Atmosphere Data Set (ICOADS) marine data source.

Version	Release Date	Release notes
1.0	????	Initial version describing platform identifier homogenisation and duplicate identification processing for first full release



1. Introduction

The Copernicus Climate Change Service (C3S) Global Land and Marine Observations Database service provides brokered access to global historical holdings of surface meteorological observations. It builds upon existing national, regional and global efforts to create an augmented set of quality assured holdings that can be used to create a multitude of datasets, products and services. This document contains all relevant information to enable a user to discover, understand, and use the marine data holdings appropriately. The document is ordered as follows:

- Section 2 provides background information including data selection and preprocessing.
- Section 3 described the duplicate identification
- Section 4 presents diagnostics
- The Annex to this document lists the software packages used in the processing

Feedback on this document can be provided via the C3S service desk facility.

This document supplements information in other C3S 311a Lot2 publications, in particular the following documents:

- The marine inventory {C3S_D311a_Lot2.1.1.1_201708_Preliminary_Marine_Inventory_v1}
- The annex to the marine inventory {C3S_D311a_Lot2.1.1.1_201708 Preliminary_Marine_Inventory_Annex_I v1}
- Third version of Marine User Guide {C3S_D311a_Lot2.3.4.4-2019_201910_Third version_Marine_User_Guide_v1}
- The Common Data Model for in-situ data holdings
https://github.com/glamod/common_data_model/blob/master/cdm_latest.pdf
- The C3S 311a Lot 2 Technical service document { C3S_311a_Lot2.3.4.2-2019_202003_Third version of Technical_Service_Document.v1}
- John's QC software (Kennedy et al. 2017)

These documents shall be hosted and discoverable on the service website once instigated. In the interim they can be provided upon request by Ms. Corinne Voces (Corinne.voces@mu.ie).



2. Background

2.1 Marine observational data sources

Marine observations within the C3S 311a Lot 2 service have been sourced from the International Comprehensive Ocean-Atmosphere Data Set (ICOADS) Release 3.0 (Freeman et al. 2017). Further information is available in the marine user guide.

2.2 Data selection

Reports were processed for the period 1850 to 2018, however only data from 1950 to 2010 have presently been made available. This document therefore contains information not directly relevant to the first full release, but will be available in future releases and is included for completeness.

Data were selected to include ship data only, and to exclude specialist ship data sources, such as research vessels. Data selection rules are presented in Table 1.

Data were required to have day of month present, and valid (e.g. 30 Feb not used), but data with missing hour were included for certain DCK between 1850 and 1872.

Table 1. Data Selection criteria

Rule	Action
PT missing & DCK = 128, 150, 151, 152, 155, 156, 192, 201, 255, 701, 721, 875, 897, 899	PT = 0 (ship)
PT missing & DCK = 714	PT = 7 (drifting buoy)
PT missing & DCK = 797	PT = 13 (C-MAN)
PT missing & DCK = 896 & ID contains PLAT or RIG	PT = 15 (platform)
PT missing & DCK = 896 & ID contains SHIP	PT = 0 (ship)
PT missing & DCK = 896, 883 & ID starts with number	PT = 6 (moored buoy)
PT missing & DCK = 896 & ID contains 4Y or C7	PT = 2 (OWS)
IRF = 2 except DCK = 732	Use ICOADS IRF flag but retain DCK 732 regional exclusions
PT = 7	Exclude drifters
PT = 6	Exclude moored buoys
PT = 13	Exclude C-MAN
PT = 14	Exclude other coastal or island stations
PT = 16	Exclude tide gauges



Rule	Action
PT = 18	Exclude profiling floats
PT = 19	Exclude undulating oceanographic recorders
PT = 20	Exclude pinnepeds
PT = 21	Exclude gliders
DCK = 795, 995	Exclude C-MAN
DCK = 735, 740, 780, 782	Exclude research vessels
DCK = 793, 794, 993, 994	Exclude moored buoys
ID = PLAT, BUOY, RIGG, BOUY	Exclude non-ship data
ID missing & DCK 700	Exclude drifters
ID = 5 digits & DCK = 700 & SID = 147 & PT = 5	Exclude buoys
ID = 5 digits & DCK = 892 & SID = 29 & PT = 5	Exclude buoys
ID contains TEST but /= CONTEST	Exclude suspect data
Invalid day	Exclude data with invalid dates
Hour missing	Except for DCK = 246, 701, 721
At least one present of SST, SLP, AT, W, D, WW, N, WBT, DPT, VV, RH, WH, NH, W1, OSV	Exclude reports with no selected variables.

Figures 1 and 2 illustrate some of the important features of the ICOADS Total Files (Freeman et al. 2017). Figure 1 (top panel) shows the total number of observations per month, coloured according to selection criteria. The largest amount of data is shown in orange and has been excluded by platform type. Figure 2 splits the data by observation type, showing that the period from 2000 onwards is dominated by fixed platforms, and moored and drifting buoys which we exclude from our analysis. ICOADS suffered a loss of drifter data in its near real time stream, after 2014, which is clear in the lower panel of Figure 2 and should be remedied during late 2019 or 2020.

The second panel in Figure 1 reduces the y-axis range to focus in on the selected observations (shown in green and pink). Observations shown in pink have been flagged as duplicates by ICOADS and are not included in the final files typically served to users (Freeman et al. 2017). The third panel of Figure 1 shows the same data as a fraction of the total number of observations in each month. Data with missing hour (shown in brown) are prevalent at the start of the period and if excluded preclude



analysis before about 1855. A smaller proportion of observations have missing hour later in the record, these are excluded from the processing.

The lower panel in Figure 1 focuses in on the reasons for exclusion of data. Early in the record a small amount of data is excluded, in some months none. The black line shows the number of reports excluded ($\div 100k$) showing that exclusions reach about 1000 per month in around 1950 and are typically much smaller than this before 1950. Platform type exclusions (orange) before about 1970 are from ice stations or oceanographic data, after about 1970 the bulk are fixed platforms, buoys and drifters. Exclusions by the IRF flag are shown in green, these are data judged by ICOADS to not be of suitable quality. Exclusions by ID (e.g. PLAT, RIGG, TEST) are shown in cream. A tiny proportion of data is excluded because it has no information on day of month and in in the 19th Century and during World War 1 and 2 a significant proportion of the small amount of data excluded has no information in the selected variable fields (Table 1).

Figure 1: Time series of numbers (top and middle panels) and proportions (lower panel) of reports from ICOADS Release 3.0 Total Files showing effect of data selection

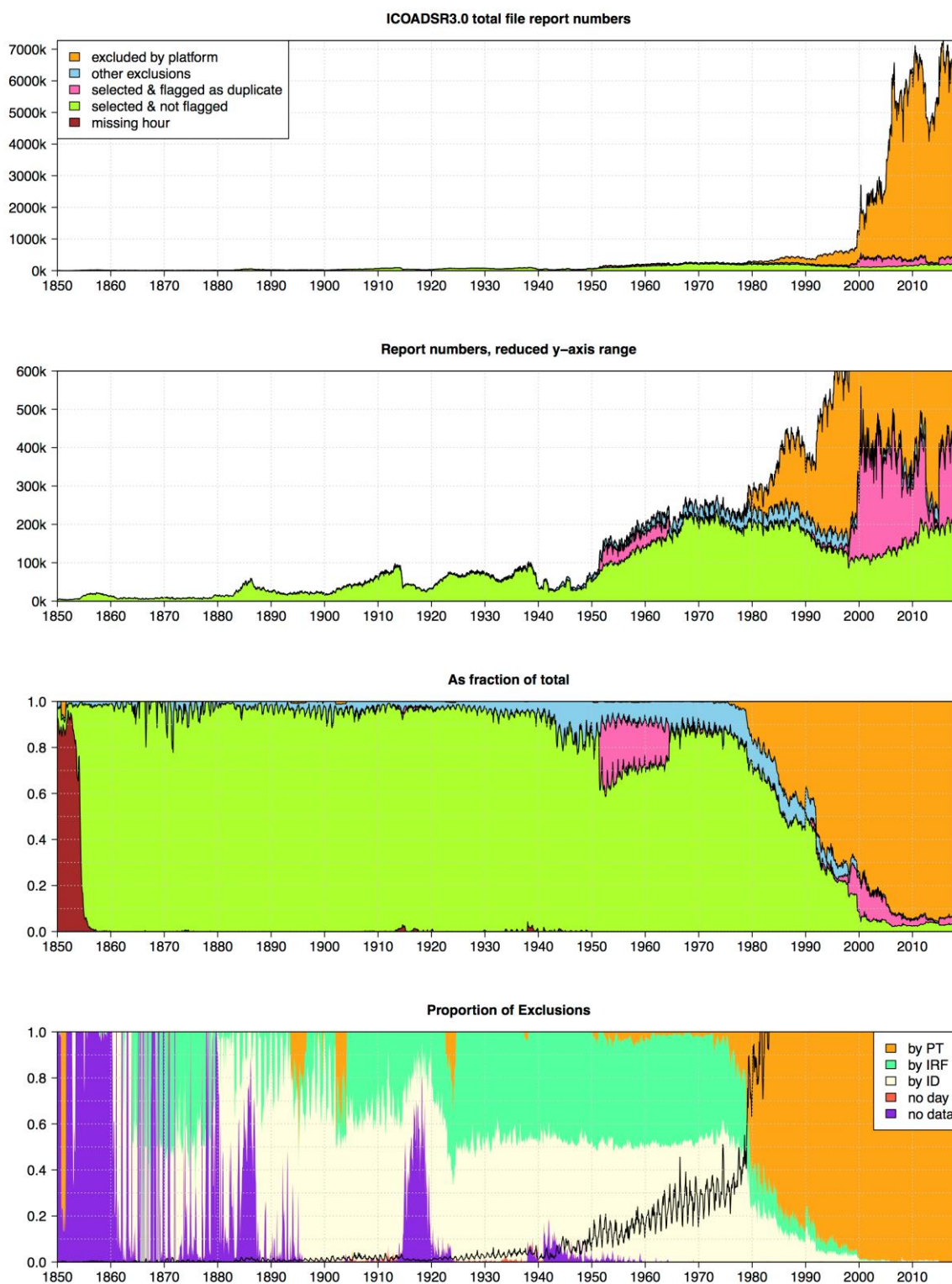
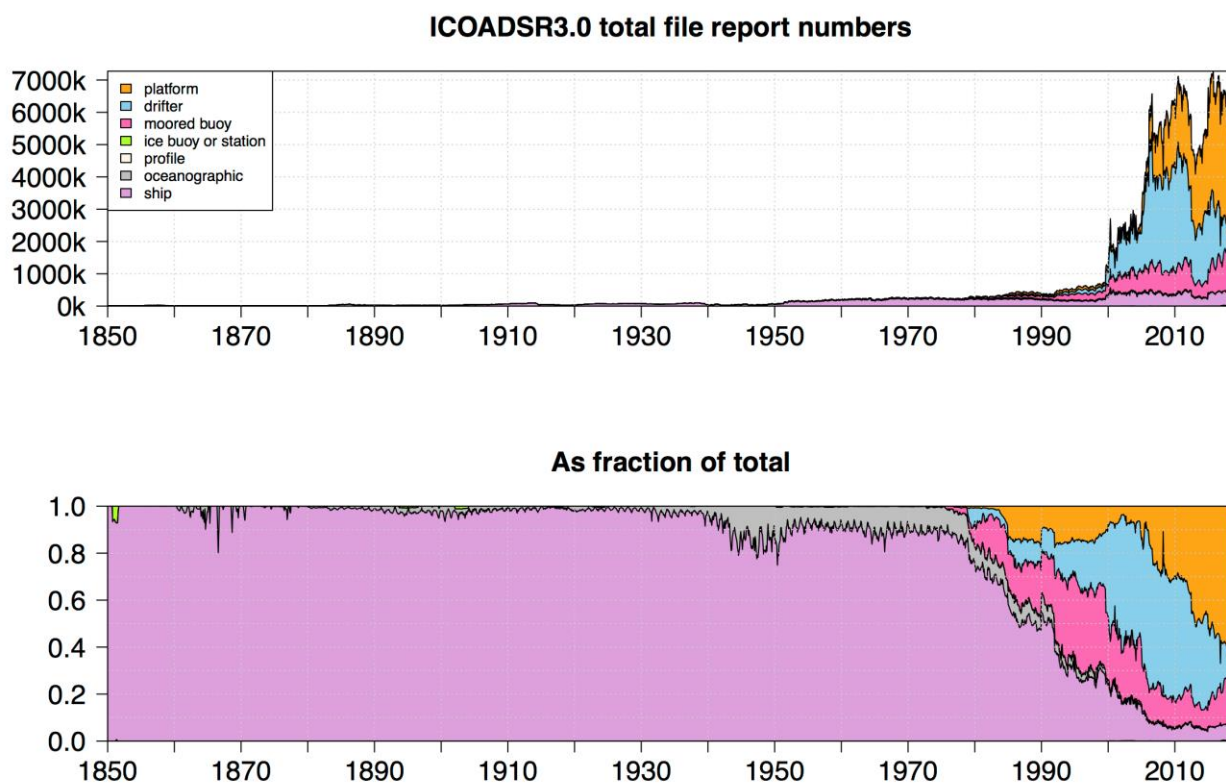


Figure 2: Time series of numbers (top panel) and proportions (lower panel) of reports from ICOADS Release 3.0 Total Files by type of platform



2.3 Preprocessing of IDs

2.3.1 ID replacements

ID corrections

For the period 1878 to 1894 some minor changes to the ship names from DCK 704 have been made to correct e.g. typos. Similarly, for DCK 701 (1867-1899), and 711 (1889-1899) some ship names have been corrected.

For the period 1663 to 1860 CLIWOC logbook IDs from DCK 730 have been converted to ship names using information from the project (<https://projects.knmi.nl/cliwoc/download/shiplogbookid21.htm>).

For the period 1663 to 1863 ship names from the US Maury collection DCK 701 have been extended using information from http://icoads.noaa.gov/software/transpec/maury/mauri_out. Also data with missing ID from DCK 701 have been split into voyages by inspection. Ship names from the German Maury collection (DCK 721) have been extended where they overlap with names from US Maury (DCK 701). Where names were the same across DCK 701 and 721 and it was not clear that the ships were



the same, the DCK number has been appended (AUSTRALIA, JAMESTOWN, SWORDFISH, ANN MARIA, ASHBURTON).

In DCK 555 (1966-1973) North Pole and South Pole station IDs have been corrected by prepending "N" or "S" depending on latitude.

ID reformatting

Manual corrections to 2 IDs from DCK 187 (1946-1956) to make them conform to the expected format.

For the period 1953 to 1961 IDs from DCK 184 were truncated to remove the first digit which indicated the ocean region and the IDs reformatted to match the expected form.

Between 1962 and 1963 the ID "Eltanin" was added to DCK 897 which contains only data from that ship and had ID missing.

Between 1957 and 1961 a small number of IDs from DCK 902 were reformatted to match the expected format by prepending a "3" to truncated IDs.

Between 1930 and 1961 IDs for DCK 118 and 119 a small number of IDs were reformatted to match the expected format by inserting a 2-digit year.

For DCK 720 and SID 135 8-character IDs represented a single report, there were truncated to the first 4 digits and "-SEQ" appended.

ID homogenisation

IDs in DCKs 194, 201, 202, 203 and 227 all derive from the same 5-digit ship identifiers. Leading digits were removed where needed.

Some ship IDs that are callsigns need to be reformatted to enable linking of data from the same ship across DCKs and also linking to metadata information in WMO Publication No. 47 (Pub. 47, Kent et al. 2007, Freeman et al. 2011). Where an ID was identified as a callsign or other identifier listed in Pub. 47 other IDs containing the same character string were identified and leading digits were removed to homogenise the callsigns across DCKs.

ID linking

Several DCK have IDs that indicate a logbook, sheet or other block of data that can be linked together to form ship tracks.

DCKs 705, 706 and 707 have IDs that are a 2-character country code followed by a 6-digit number, each unique combination joining about 10 records. Ship names are available in the supplemental data (<https://icoads.noaa.gov/e-doc/other/transpec/usmm1912-46/>) but these are inconsistently formatted and contain many typos. The ship names were corrected manually where possible, this was hampered by not having a list of the names of the ships that were digitised, so is far from perfect. A new ID was formed from the character code from the old ID plus the ship name from the supplemental data, homogenised allowing joins between records with similar supplemental ship names that formed a plausible track using the Met Office QC scheme tracking code (MOQC track check).



Eight-digit IDs from DCKs 192, 215 and 720 (not SID 135) represents log sheets that were joined sequentially where they formed a ship track. Sometimes data from new ship started on the log sheet started from another ship. Where such a discontinuity was identified the ID linking was restarted from that sheet number.

Six-digit IDs from DCK 216 were similarly joined.

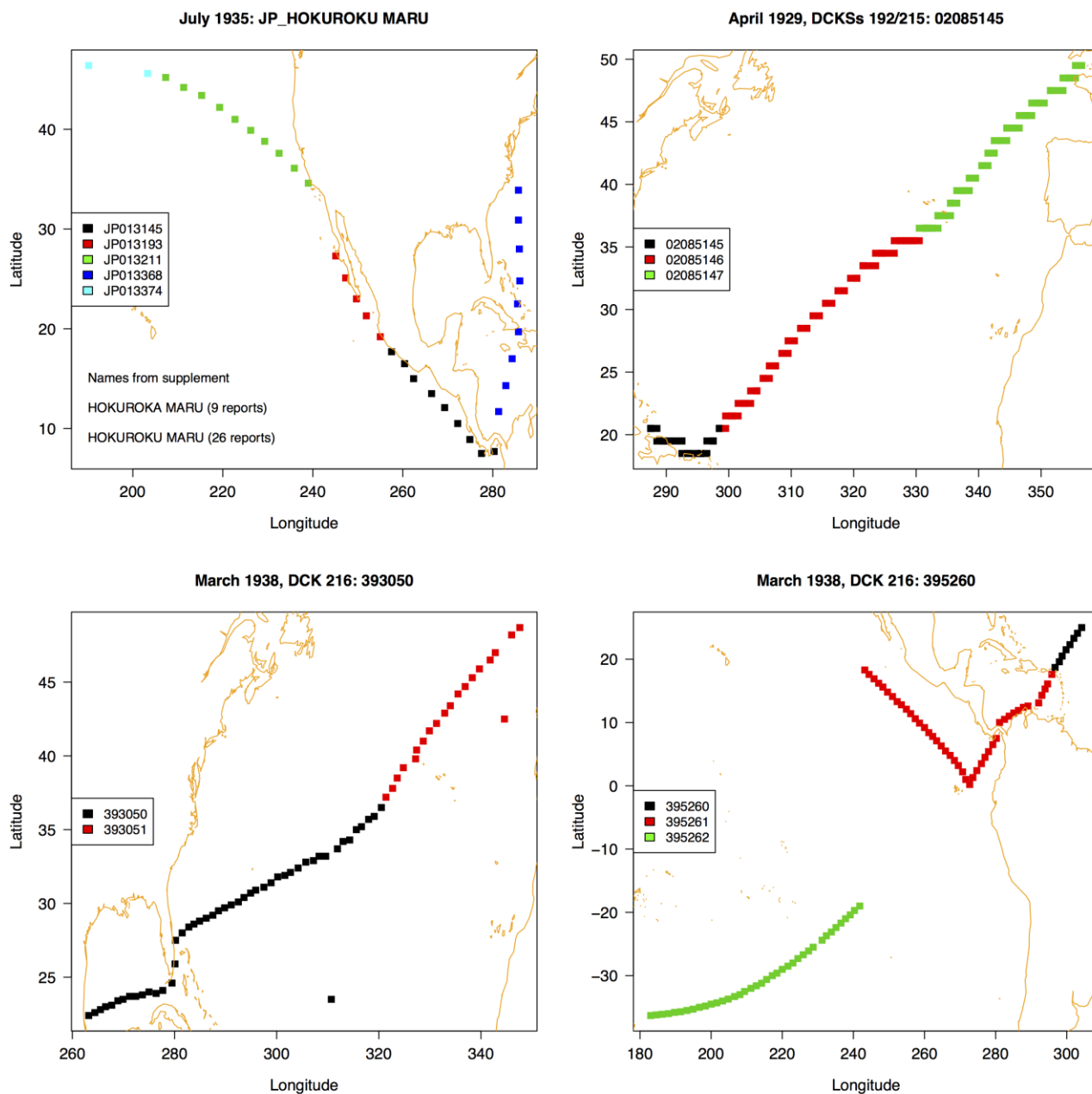
Figure 3 illustrates sample output from the ID linking procedure. Top left is from DCK 705 during July 1935 and showing data from Japan, originally split across 5 different IDs. The ship name from the supplement shows 2 similar variants and the assigned ID is the most frequently occurring of these 2 ship names, with JP prepended to avoid mixing ships with the same name from different countries. Whilst in this case the association is rather clear, there are cases where the ship names from the supplement are rather more varied and it is likely that improvements to the linking for DCKs 705, 706 and 707 can be made. There is no list available giving the names of the ships that were included in this digitisation, if located that would be extremely helpful in linking and correctly naming the ships. We note that data from DCKs 705, 706 and 707 is duplicated with data from other DCK (see Table 8).

DCKs 192, 215 and 720 all contain data with 8-digit IDs that were made by German ships. DCK 192 was ingested in the original release of COADS (Slutz et al. 1985), DCK 215 originates from the UK Marine Data Bank (Worley et al. 2005) and contains data from several other national archives and DCK 720 contains data from the German Weather Service that has been more recently reprocessed or digitised (Freeman et al. 2017). The top right panel of Figure 3 shows data from 3 original IDs in April 1929 which are numerically sequential log sheet numbers used in the German data system. The assigned ID is the first in the numerical sequence.

The lower panels in Figure 3 show data from DCK 216, also from the UK Marine Data Bank. Here 5-digit IDs from March 1938 are joined sequentially and assigned a new ID of the first in the sequence. These plots also illustrate problems with the data that are found in DCK 216, but also more widely in ICOADS. In the lower left panel 2 reports have been mispositioned in latitude. These data will be flagged by the MOQC track check, but future extensions to processing might allow correction. The lower right plot shows a more problematic issue, a systematic miscoding of hemisphere that is only identified when the sequential IDs are linked. In this case the track check on the joined ID will not correctly identify which data are mis located.



Figure 3: Examples of ID joining, see text for description



An excerpt from a pipe-delimited year-month output file from the IDs change processing follows. The report UID is followed by the report ID then a flag with value 1 if the ID has been changed, 0 if it remains the same.

ICOADS-30-0Y0HJK|32024|0

ICOADS-30-0Y0HJL|14 00117|1



ICOADS-30-0Y0HJM|14 00370|1

ICOADS-30-0Y0HJV||0

ICOADS-30-0Y0HJX|613744|0

ICOADS-30-0Y0HJY|611224|0

2.4 Preprocessing of data record for duplicate identification

2.4.1 Data that fails QC

Each of the variables to which the Met Office QC (Kennedy et al. 2017) is applied (SST, air temperature, sea level pressure, wet bulb temperature, dewpoint temperature, relative humidity) are duplicated with ".old" appended. The original variable is then set missing and is used in the duplicate identification processing. This helps to avoid a failure to identify duplicate records where a particular element has been corrupted. Only the climatological QC check is applied.

2.4.2 Adding information used in duplicate identification processing.

Based on the ICOADS documentation, and earlier source documentation, the IDs have been classified as to their "ID type" (logbook number, ship number, etc) and validity. Where an ITU callsign is found the country for that callsign is also identified.

Flags are added to indicate whether the report comes from a DCK containing Global Telecommunication Systems (GTS) data or Delayed Mode (DM) data from the International Data Exchange DCKs (Table 2). DM data is expected to be of higher quality than GTS data so when there is a match between them, the DM data is preferred. This information is also used in the selection of which humidity variable is included in the duplicate identification procedure (Table 3).

Table 2. Definition of Global Telecommunication Systems (GTS) and related Delayed Mode (DM) DCKs.

Type	ICOADS DCK
GTS	555, 700, 792, 793, 794, 795, 796, 797, 888, 992, 993, 994, 995, 996, 997
DM	926, 927, 928

Table 3. Procedure for excluding derived humidity variables from duplicate identification processing

Type	ICOADS DCK
DPT	GTS or DM DCK and DPTI = 1 or 3 and either WBT or RH present
WBT	GTS or DM DCK and WBTI = 1 or 3 and either DPT or RH present
RH	GTS or DM DCK and RHI = 3 or 4 and either WBT or DPT present



Type	ICOADS DCK
DCK=117	Remove DPT and RH

In the duplicate identification procedure priorities are assigned to data from each DCK (Table 4) and that with the highest priority is selected. The data expected to be of best quality is assigned a priority of 1, data with larger priority numbers will be flagged as the worst duplicate if identified as potential matches. These priority values are based on those from ICOADS (see <https://icoads.noaa.gov/e-doc/other/> for details).

Table 4. Assigned DCK priorities

Priority	ICOADS DCK
1	143, 144, 145, 146, 201, 202, 203, 204, 205, 206, 207, 209, 210, 211, 213, 214, 216, 218, 226, 227, 246, 247, 248, 249, 667, 701, 702, 704, 705, 706, 707, 710, 711, 715, 721, 730, 731, 734, 735, 736, 750, 781, 761, 762, 780, 876, 877, 878, 879, 880, 881, 882, 883, 891, 927
2	110, 116, 118, 119, 128, 143, 184, 185, 187, 188, 189, 192, 193, 194, 195, 196, 197, 245, 666, 703, 709, 714, 720, 740, 874, 875, 897, 898, 899, 900, 902, 926
3	117, 186, 233, 234, 254, 792, 793, 294, 700, 708, 782, 792, 793, 794, 795, 795, 797, 892, 893, 896, 901, 928, 992, 993, 994, 995, 996, 997
4	150, 151, 152, 155, 156, 221, 223, 224, 229, 230, 239, 733, 849, 850, 889
5	281, 255, 555, 749, 888
6	215, 732, 999

An estimate of the precision of each variable per DCK and sometimes by year and SID is required to set the tolerances for allowing a match in the duplicate identification procedure (Table 5). The default value to use is indicated by DCK=0, SID=0. When testing whether 2 reports match, the difference between the estimates of each variable are compared to the expected tolerance from Table 5. For example, if a DCK has positions rounded to whole degrees and matches to another DCK where position is available to 0.1 degree, then the precision value is 1. These precision values are likely to be updated in future versions of the processing.



Table 5. Precision values, per variable, per DCK and SID. For DCK, SID and year 0 is the default value

DCK	SID	yr.s	yr.e	lat	lon	sst	at	dpt	slp	w	d	vv	ww	n	w1
0	0	0	0	1	1	1	1	1	1	1	1	0	0	0	0
110	0	0	0	1	1	0.1	0.1	0.1	0.1	0.1	20	0	0	0	0
116	0	1945	1948	1	1	0.3	0.3	1	0.1	0.1	10	0	0	0	0
117	0	1949	1963	1	1	0.1	0.1	0.3	0.1	0.1	25	0	0	0	0
118	0	0	0	0.1	0.1	1	1	NA	0.1	0.1	13	0	0	0	0
119	0	0	0	0.1	0.1	1	1	1	0.1	0.1	10	0	0	0	0
128	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	10	0	0	0	0
150	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	10	0	0	0	0
151	0	1862	1957	1	1	0.1	0.1	0.1	0.1	0.1	10	0	0	0	0
151	0	1958	1960	0.1	0.1	0.1	0.1	0.1	0.1	0.1	10	0	0	0	0
152	0	0	0	1	1	0.3	0.3	0.3	0.1	0.1	10	0	0	0	0
155	0	1861	1942	1	1	0.1	0.1	0.1	0.1	0.1	25	0	0	0	0
155	0	1943	1960	0.1	0.1	0.1	0.1	0.1	0.1	0.1	25	0	0	0	0
156	0	0	0	1	1	0.1	0.1	0.1	0.1	0.1	23	0	0	0	0
184	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	13	0	0	0	0
185	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	10	0	0	0	0
186	0	0	0	0.1	0.1	1	0.1	0.1	0.1	0.1	10	0	0	0	0
187	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	10	0	0	0	0
188	0	0	0	0.1	0.1	0.1	0.1	NA	0.1	0.1	13	0	0	0	0
189	0	0	0	0.1	0.1	0.1	0.1	NA	0.1	0.1	10	0	0	0	0
192	0	0	0	1	1	0.1	0.1	0.1	0.1	0.3	23	0	0	0	0
193	0	0	0	1	1	0.1	0.1	NA	0.1	0.3	23	0	0	0	0
194	0	1856	1929	1	1	0.1	0.1	0.1	0.1	0.3	13	0	0	0	0
194	0	1930	1955	0.1	0.1	0.1	0.1	0.1	0.1	0.3	13	0	0	0	0



DCK	SID	yr.s	yr.e	lat	lon	sst	at	dpt	slp	w	d	vv	ww	n	w1
195	0	0	0	1	1	0.1	0.1	0.1	0.1	0.1	10	0	0	0	0
196	0	0	0	1	1	0.1	0.1	0.1	0.1	0.3	10	0	0	0	0
197	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.3	13	0	0	0	0
201	0	0	0	1	1	0.1	0.1	0.1	0.1	0.3	13	0	0	0	0
202	0	0	0	1	1	0.1	0.1	0.1	0.1	0.3	13	0	0	0	0
203	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.3	13	0	0	0	0
204	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.3	13	0	0	0	0
205	0	0	0	0.3	0.3	0.1	1	NA	NA	0.3	13	0	0	0	0
206	0	0	0	0.3	0.3	0.1	0.1	0.1	0.1	0.3	10	0	0	0	0
207	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.3	13	0	0	0	0
209	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.3	13	0	0	0	0
210	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.3	10	0	0	0	0
211	0	0	0	0.1	0.1	0.1	0.1	NA	1	0.1	10	0	0	0	0
213	0	0	0	0.1	0.1	0.1	0.1	NA	0.1	0.1	10	0	0	0	0
214	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	10	0	0	0	0
215	0	0	0	1	1	0.1	0.1	NA	0.1	0.3	13	0	0	0	0
216	0	0	0	0.1	0.1	0.3	0.3	NA	0.1	0.3	13	0	0	0	0
218	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	10	0	0	0	0
221	0	0	0	0.1	0.1	0.1	NA	NA	NA	0.1	10	0	0	0	0
223	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.3	10	0	0	0	0
224	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.3	10	0	0	0	0
226	0	0	0	0.3	0.3	0.1	0.1	0.1	0.1	0.3	10	0	0	0	0
227	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.3	13	0	0	0	0
229	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.3	10	0	0	0	0
230	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.3	10	0	0	0	0



DCK	SID	yr.s	yr.e	lat	lon	sst	at	dpt	slp	w	d	vv	ww	n	w1
233	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.3	10	0	0	0	0
234	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.3	10	0	0	0	0
239	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.3	10	0	0	0	0
245	0	0	0	0.5	0.5	0.1	0.1	0.1	0.1	0.3	23	0	0	0	0
246	0	0	0	0.5	0.5	0.1	0.1	0.1	0.1	0.3	23	0	0	0	0
247	0	0	0	0.5	0.5	0.1	0.1	0.1	0.1	NA	NA	0	0	0	0
249	0	0	0	0.5	0.5	0.1	0.1	0.1	0.1	0.3	13	0	0	0	0
254	0	1860	1948	1	1	0.1	0.1	0.1	0.1	0.3	23	0	0	0	0
254	0	1949	1994	0.1	0.1	0.1	0.1	0.1	0.1	0.3	10	0	0	0	0
255	0	1857	1946	1	1	0.1	0.1	0.1	0.1	0.3	13	0	0	0	0
255	0	1947	1994	0.1	0.1	0.1	0.1	0.1	0.1	0.3	13	0	0	0	0
281	0	0	0	1	1	0.1	0.1	0.1	0.1	0.3	13	0	0	0	0
555	0	1966	1969	0.1	0.1	0.1	1	1	0.1	0.1	10	0	0	0	0
555	0	1970	1973	0.1	0.1	0.1	1	1	0.2	0.1	10	0	0	0	0
666	0	0	0	0.1	0.1	0.1	0.1	1	0.1	0.3	10	0	0	0	0
667	0	0	0	0.1	0.1	0.1	NA	NA	NA	0.3	23	0	0	0	0
700	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.3	10	0	0	0	0
701	0	0	0	0.5	0.5	0.3	0.3	NA	0.3	0.3	23	0	0	0	0
702	0	0	0	0.5	0.5	0.3	0.3	NA	0.3	0.3	23	0	0	0	0
703	0	0	0	1	1	0.3	0.3	0.3	0.3	0.3	23	0	0	0	0
704	0	0	0	1	1	0.3	0.3	0.3	0.3	0.3	23	0	0	0	0
705	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.3	23	0	0	0	0
706	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.3	23	0	0	0	0
707	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.3	23	0	0	0	0
708	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.3	10	0	0	0	0



DCK	SID	yr.s	yr.e	lat	lon	sst	at	dpt	slp	w	d	vv	ww	n	w1
709	0	0	0	0.1	0.1	0.3	0.3	0.3	0.1	0.3	10	0	0	0	0
710	0	0	0	0.5	0.5	0.3	0.3	NA	0.1	NA	NA	0	0	0	0
711	0	0	0	0.5	0.5	NA	0.3	NA	0.1	0.3	NA	0	0	0	0
720	0	1868	1949	0.1	0.1	0.1	0.1	0.1	0.1	0.3	23	0	0	0	0
720	0	1950	1988	0.1	0.1	0.1	0.1	0.1	0.1	0.3	20	0	0	0	0
721	0	0	0	0.5	0.5	0.3	0.3	NA	0.1	0.3	23	0	0	0	0
730	0	0	0	0.5	0.5	0.3	0.3	NA	0.1	0.3	23	0	0	0	0
731	0	0	0	0.5	0.5	0.3	0.3	NA	NA	0.3	23	0	0	0	0
732	0	1888	1934	0.1	0.1	0.3	0.3	0.1	0.3	1	10	0	0	0	0
732	0	1935	1938	0.1	0.1	0.3	0.3	1	0.3	1	10	0	0	0	0
732	0	1939	1995	0.1	0.1	0.3	0.3	0.1	0.3	1	10	0	0	0	0
733	0	0	0	0.1	0.1	NA	0.1	0.1	0.1	1	23	0	0	0	0
734	0	0	0	0.5	0.5	0.1	0.1	0.1	0.1	0.3	13	0	0	0	0
735	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	1	10	0	0	0	0
736	0	0	0	0.5	0.5	NA	NA	0.1	NA	NA	NA	0	0	0	0
740	0	0	0	0.01	0.01	0.1	0.1	0.1	0.1	0.1	1	0	0	0	0
749	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	1	10	0	0	0	0
750	0	0	0	0.1	0.1	0.1	NA	NA	NA	NA	NA	0	0	0	0
761	0	0	0	0.5	0.5	0.1	0.1	NA	0.1	0.3	23	0	0	0	0
762	0	0	0	0.1	0.1	0.3	0.3	0.3	0.1	0.3	23	0	0	0	0
780	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.3	10	0	0	0	0
781	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	10	0	0	0	0
782	0	0	0	0.01	0.01	0.1	NA	NA	NA	NA	10	0	0	0	0
792	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	10	0	0	0	0
849	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.3	10	0	0	0	0



DCK	SID	yr.s	yr.e	lat	lon	sst	at	dpt	slp	w	d	vv	ww	n	w1
850	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.3	10	0	0	0	0
874	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.3	10	0	0	0	0
875	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.3	10	0	0	0	0
888	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.3	10	0	0	0	0
889	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.3	10	0	0	0	0
892	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.3	10	0	0	0	0
896	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.3	10	0	0	0	0
897	0	0	0	0.1	0.1	NA	0.1	0.1	0.1	0.3	10	0	0	0	0
898	0	0	0	0.1	0.1	0.1	1	1	0.1	0.3	10	0	0	0	0
926	0	0	0	0.1	0.1	1	1	1	0.1	0.1	1	0	0	0	0
927	0	0	0	0.1	0.1	1	1	1	0.1	0.1	1	0	0	0	0
928	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	10	0	0	0	0
992	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	10	0	0	0	0
999	0	0	0	0.1	0.1	0.1	0.1	1	0.1	NA	10	0	0	0	0

2.5 Correcting time errors

2.5.1 Incorrect dates and times

It has been noticed that some DCK have incorrect dates or times. This largely arises from confusion over historical definitions of the marine day and conversions between local time and UTC. Consequently, some DCK-level corrections were made to observations either at local midnight or midnight UTC where it was clear that there were systematic time misplacement of data. This could be discerned from 'jumps' in the tracks of ships with known IDs (Table 6). There are likely to be similar problems for DCK where there are no IDs present, this needs further investigation.

Table 6. Time adjustments

Time	ICOADS DCK	Year range	Action
0 UTC	189	1948-1949	Add one day



Time	ICOADS DCK	Year range	Action
0 UTC	201	to 1899	Add one day
0 UTC	207	1946-1949	Add one day
0 UTC	209	from 1957	Add one day
0 local	215	all	Subtract one day
0 UTC	216	1936-1939	Add one day
0 UTC	707	1913-1939	Add one day

2.5.2 File format with corrected dates and times

Changes to dates, times and positions (not yet implemented) are recorded in pipe-delimited year-month files with contents: UID, datetime, datetime flag (0 if unchanged or 1 if changed), latitude, latitude flag, longitude, longitude flag. An excerpt from a sample file is shown below.

```
ICOADS-30-0LBYZL|1939-04-01 00:00:00+00:00|0|67.5|0|11.5|0
ICOADS-30-0LBYZM|1939-04-01 00:00:00+00:00|0|68.5|0|12.5|0
ICOADS-30-0LBYZN|1939-03-31 00:00:00+00:00|1|52.2|0|-25.9|0
ICOADS-30-0LBYZO|1939-03-31 00:00:00+00:00|1|50|0|-15.8|0
ICOADS-30-0LBYZP|1939-04-01 00:00:00+00:00|0|59.5|0|-2.5|0
```



3. Identifying duplicates

3.1 Why do we have duplicates?

ICOADS is made up of data from many different sources, there are several hundred combinations of the DCK and SID flags that indicate the origin of the data. Typically, DCK indicates the type of data (e.g. US Navy ships; Japanese Whaling Fleet) and SID provides more information about the data system or format (e.g. International Maritime Meteorological (IMM) Data; NCEP BUFR GTS: Operational Tanks: Converted from Original Message). Sometimes a single DCK is associated with a single SID, sometimes a single DCK will contain several SID and vice versa. Historically archives of marine data have been maintained by individual nations, and often these were shared so that the same observations appear in the archives of several nations. Truncated formats often did not contain sufficient information to identify the observations made by a particular ship or platform, and these compact formats sometimes converted or encoded data in different ways. For example, many observations do not have an identifier linking to the ship or platform, and for those that do have such identifiers they may be different between data sources. The main types of duplicates are:

- Observations historically shared among national archives, likely to have different formats, precision, conversions and metadata.
- Re-ingestion of the same data more than once.
- Data from near real time sources that can be replaced with higher quality delayed mode data.
- Re-digitisation of logbooks, newer data likely to have higher precision, more metadata etc.
- Planned redundancy, for example the ingestion of several near real time data streams.

3.2 Approach to Duplicate Identification

3.2.1 Summary of ICOADS duplicate identification

The ICOADS duplicate identification procedure (duplicate elimination, dupelim) is based on comparison of individual reports falling in the same 1° latitude/longitude grid box (described in a series of webpages, <https://icoads.noaa.gov/e-doc/other/>). Seven weather elements are checked: wind speed, visibility, present weather, past weather, sea level pressure, air temperature, and sea surface temperature. Dates and times must match but in some circumstances a "cross" of the day or hour is allowed. The ID is checked for an exact match after standard substitution of erroneous characters. There are a range of different allowances made to account for some known differences between DCK. Any duplicates found are flagged, the flag indicates which is the best duplicate and provides some information about the extent of the match. No information linking the reports identified as duplicates is available. From Release 2.5 all data are flagged and carried through to the "total" file, used as the input for processing here. In earlier Releases only the preferred duplicates were carried through the processing. Typically for these early Releases, the full data are still available, but in earlier ICOADS binary formats, or in a range of ascii formats.

3.2.2 Summary of approach taken here

The pre-processing steps described in Section 2 improve the homogeneity of the ID information across and within the different DCKs, corrects some pervasive miscoding in time information, and



appends information required for duplicate selection to each report, including an estimate of the precision of the elements of the data record, the number of extant elements (after application of QC), and the priority assigned to the DCK. Potential report pairs are identified based on position and time, with some constraints on the closeness of variables within the report.

The pairing process is conducted 4 times, once requiring a full match in date and time, and subsequently allowing each of hour, day or month to be one unit different. Mismatches of a single hour typically arise from differences between data sources in conversion from local time to UTC. Mismatches of a single day are typically related to associating either midnight local or midnight UTC to the wrong date. Day or hour mismatches of ± 1 are allowed, although day mismatches were not allowed for DCK 192, 194, 197, 201, 702, 704, 720, 721, 762 and hour mismatches not allowed for DCK 117. Month mismatches are usually associated with GTS reports where year and month are not reported, examples found typically have the incorrect month assignment one month earlier than the correct assignment.

The candidate pairs are then selected according to the number of matching elements and the precision of the elements in each report, the DCKs and a comparison of the IDs. Whilst there is no compositing of duplicate reports (for example we do not reinstate any variables unavailable in the preferred data source), a preferred ID is assigned for each pair. For example, when a GTS and DM report are paired as duplicate records, the DM report is preferred, but the preferred ID might be from the GTS report if that contains a callsign.

The next stage clusters together all of the pairs with common reports and selects one report as the best duplicate. Each report in the cluster is flagged with its duplicate status.

In this pairing process IDs are linked, and this information is used to associate a preferred ID with all reports with IDs that have been linked in the duplicate identification. The results are checked manually and the processing iterated to remove obvious cases where IDs should not have been linked. For example, an ID that has been truncated might associate with a longer ID for a single report, but actually be associated with several different longer IDs so a bulk association is not appropriate.

In the final step the data associated with an ID is checked to flag and select duplicate times, using the MOQC track check or the assigned DCK priorities to choose between and flag time duplicates.

3.3 Duplicate Record Identification Procedure

3.3.1 Finding potential duplicate pairs

The potential data pairs are identified by first indexing the reports by year, month, day, hour rounded to an integer, rounded latitude and longitude. The indexed reports are only considered as potential duplicate pairs when the differences between extant variables is within the tolerance limits shown in Table 7. Pairs are considered with a full date time match (nearest hour), and ± 1 hour, 1 day and 1 month. When the hour, day or month is mismatched at least 4 variables from the following list must be present: SST, SLP, air temperature, dewpoint temperature, wind speed, wind direction, cloud cover, present weather, past weather, and visibility. A candidate pair list is generated by appending the difference between each variable for the paired reports, a flag to say whether that difference is



within the tolerance (Table 5), and the number of extant elements in each report in the pair. Because some DCK contain reports at different precision, a check is made that a variable compared at tolerance 1 is indeed rounded.

Table 7. Tolerances used in pair selection

Variable	Tolerance
Latitude/longitude	0.51 °
wind speed	2 ms ⁻¹
wind direction	13 °
SST, air temperature, dewpoint temperature	1.01 °
SLP	1.01 mb

A flag indicating whether an ID match is allowed is then added. Generic IDs (e.g. blank, "SHIP", "MASKSTID") are allowed to match within a DCK. Table 8 contains the information used to decide whether IDs in a pair are an allowed match. These criteria have been developed by inspection of the paired IDs and are therefore likely to be approximate. Damerau–Levenshtein (DL) distance is the number of insertions, deletions and swaps necessary to convert one string to another (van der Loo M, 2014). A substring is where one ID is contained within the other. Italics represents the “ID type”.

Table 8. ID match criteria. Matches of reports where the IDs meet the criteria listed below are allowed.

DCK	ID
Within any DCK	blank, SHIP, MASKSTID
116, 117, 218	any ID to blank
150, 151, 152, 155, 156, 192, 193, 215, 720, 901	any ID to blank
128, 254, 720	any ID to blank
187, 196, 197, 229, 230, 720, 732	any ID to blank
227, 246, 732	any ID to blank
761, 898	any ID to blank
204, 245	any ID to blank



DCK	ID
230, 254	any ID to blank
128, 230	any ID to blank
195, 281	any ID to blank
192, 193, 194, 201, 202, 706, 732	any ID to blank
194, 201, 202, 203, 207, 221, 223, 227, 233, 239, 254, 926	substring; or 1 digit DCK 194 ID
254, 926	2-5 characters of 254 match 3-6 characters of 926
194, 927	"7- " in 194 with "00" in 227
194 with 207 or 227	3-6 characters of 194 match 2-5 characters of 207 or 227
194 with 194, 201, 202, 203, 207, 227	DL = 1 or substring of length at least 3 and number of occurrences of one of the IDs = 1
194 with 201, 203, 207, 227	substring
194, 201	DL <= 2 and one of the IDs classed as invalid
184, 209	characters 5-8 of 184 with 2-5 or 209
555, 733	add "N" at start of 555 match of characters 2-4 555 ID is SHIP
733 with 849, 888	849, 888 ID is SHIP
733 with 888, 892	888, 892 ID has DL <= 1 with ROBB 888, 892 ID is EMIO, UYAJ, UFRE
186, 733	186 has 4 digit ID, 733 is <i>north_pole_station</i>
750, 888	888 ID is SHIP
781 with 128, 735, 849, 888, 926, 927	781 is AAAA with callsign
927 with 230, 720	927 ID is SHIP
213, 902	213 is characters 4-8 of 902



DCK	ID
926 with 888, 892	888, 892 is characters 4-8 of 926
892, 926	892 is characters 1-4 of 926
117	any match to <i>invalid</i> ID
116, 117	<p><i>id_over_X</i> to <i>id_minus</i>, match of characters 2-4</p> <p>characters 3-4 of 116 with 117 and 116 is <i>osv_onstation</i></p> <p>match of characters 1-3 and 116 is <i>osv_onstation</i></p> <p>match characters 2-3 of 116 with 1-2 of 117 and 116 is <i>osv_noship</i></p> <p>match of characters 1-4 and 116 is <i>other</i></p> <p>match of - at start of 116 with 5 at start of 117</p> <p>prepend 5 to 117</p> <p>prepend - to 116</p> <p>1 digit ID in each</p> <p>match of start of 116 with 2 character 117</p> <p>within or between 116 and 117 when DL <=2 when one ID has 3 or fewer occurrences</p> <p>116 missing to extant 117</p> <p>116 is <i>osv_onstation</i> and characters 3-4 are 00 with 117 ID of length 4</p> <p>substring, one ID has <= 4 occurrences, the other >= 10 occurrences</p>
117	<p>prepend "-" to one of the IDs</p> <p>DL = 1 if 3 or fewer occurrences of one ID</p>
116, 116	22014, 22004
116, 226	<i>osv_noship</i> to <i>ows_logbook</i>
117, 218	<p>prepend 0 to 117 and 218 is <i>us_ows_folio</i></p> <p>characters 1-3 of 117 with 2-4 of 218 is <i>us_ows_folio</i></p>
117, 128	both 4 characters in length and match of characters 1-2 in 117 with 3-4 in 128



DCK	ID
192, 215, 720	match blank ID match characters 1-4 with 4 character ID allow letter as 5th character in 192 in 8 character ID one ID invalid and not <i>id_5digit_pership</i> and not containing 0000 and DL<=2 or substring DL<=2 and one ID has <= 3 occurrences and other has >= 8
192, 215, 254, 720	one is 5 character ID, the other is not
246	"PQP PTMNI" to "PORQUOIP"
762	2617A to 26174
128, 233, 254, 255, 555, 700, 708, 709, 732, 735, 749, 781, 792, 849, 874, 875, 888, 889, 892, 926, 927, 992, 993, 995, 999 hereafter "call.dcks"	subset one is <i>invalid</i> and DL <= 2 DL <= 2 and one has a single occurrence and the other at least 3 one has a single occurrence and the other at least 20
call.dcks, 850	SHIP, MASKSTID or AAAA to anything
call.dcks, 896	SHIP to OWS
128, 555	128 platform type = 3 and 555 ID starts 4Y
128, 230 with 555	ship number to call sign if at least 3 occurrences
128, 230 with 555, 720	matches with blank ID
Any	match when 0 replaced with O; O/O I/J UU/VV U/V with DL=1 WZC/WCZ
892, 896	replace C7O/C7M QR/C7R
992	replace XP42/MP42
700 with 792, 992	BBXX removed from start of ID



DCK	ID
711, 201	> 3 occurrences
720, 734	> 3 occurrences
246, 720	TERRANOVA to ID starting 610426
193 with 705, 706, 707	705, 706, 707 starting NL or DN
118, 762 with 705, 706, 707	705, 706, 707 starting JP
203 with 705, 706, 707	705, 706, 707 starting UK
705, 706, 707	with matching characters 1-2 of original ID
703, 927	927 ID starts 05 and has >=5 occurrences

3.3.2 Identifying duplicate groups and selecting the "best" duplicate

The duplicate pairs are then combined into groups. By this stage each pair will match in date and time (or have a permitted mismatch), and each extant variable will be within the tolerances listed in Table 7 and the IDs match as listed in Table 8. There can be many versions of the same report in ICOADS, so all reports that are associated across different pairs are grouped.

Each group of possible duplicates is then assessed. In this process it is important to account for known differences between DCKs that are not captured in the precision information in Table 5. An example is a systematic difference in coding of weather information, the allowed mismatches are listed in Table 10.

The next step is to exclude reports that are expected to be of lower quality from the duplicate groups before assessment of the differences between the higher quality reports. The reports from the lower quality groups are flagged as being worst duplicates (Table 10). Where there is a mix of data from DM and GTS DCKs (Table 11), the data from the GTS is automatically flagged as the worst duplicate. Further whittling down of DM reports keeps only the report with the most recent format version flag. Where the duplicate group contains reports from multiple DCK, data from certain DCK are excluded (Table 12). This reduces the risk that duplicates are not identified because of coding differences between DCK and the inclusion of reports from sources that are known to be less reliable.

If more than one report remains for assessment after application of these rules, the remaining reports are compared according to Table 13. If differences exceed these limits the reports in the group are not flagged as duplicates.

If the group of reports meets the matching criteria then the reports are ordered according to Table 14, with ties in DCK priority being broken with delayed mode format version, then the number of extant elements and so on down the list. The final tie-breaker is the ICOADS duplicate status, so that



if there is no other difference in the expected quality of the reports, there is no change to the preferred report from ICOADS.

Table 9. Allowed mismatches

Category	Allowance
Mismatches of weather codes in DCK 194 & 201	53/50 81/52 80/52 90/89 63/60
Mismatches of wind speeds in DCK 194 & 201	9.3/9.8
Mismatches of wind speeds in DCK 128 & 254	9.3/9.8
Allow match to floored (rounded down to integer) sea level pressure in DCK 720	
Allow match to wind speed = 1 m/s in DCK 720	
DCK 116, 117, 227, 194, 720 only compare wet bulb temperature, not other humidity variables	
DCK 117, SID 140	Cloud and weather codes, clouds and visibility do not match 116 or other 117 SID
DCK 116, 117	Allow for differences in coding of wind direction, 10 degree vs 16 point

Table 10. Duplicate flags

Flag value	Meaning	Note
0	Unique observation, no known duplicates	
1	Best duplicate	
2	Duplicate	This has been used to identify reports thought to be duplicates but where no selection was made, e.g. because there were unique variables present in each report



Flag value	Meaning	Note
3	Worst duplicate	
4	Unchecked	Reports that have not been through the duplicate identification process

Table 11. Classification of DCK as Delayed Mode (DM) or GTS

DCK type	DCKs included
DM	926, 927, 928
GTS	555, 700, 792, 793, 794, 795, 796, 797, 888, 992, 993, 994, 995, 996, 997

Table 12. Selection of duplicate status within groups

Reports in duplicate group	Action
Delayed mode reports in duplicate groups	Flag GTS report(s) as worst duplicate(s)
If multiple IMMT version numbers	Flag delayed mode report(s) with older IMMT number as worst duplicate(s)
If DCK other than 700 are present	Flag DCK 700 report(s) as worst duplicate(s)
If DCK other than 792 are present	Flag DCK 792 report(s) as worst duplicate(s)
If DCK 117 and either 116 or 128 present	Flag DCK 117 report(s) as worst duplicate(s)
If multiple SID for DCK 117 present, including SID 142	Flag DCK 117 with SID != 142 as worst duplicate(s)
If DCK 733 and 186 present	Flag DCK 733 report(s) as worst duplicate(s)



Table 13. Report match criteria after date/time, location and ID matches allowed.

Number and type of matches	Allow match
Variables in match assessment, if extant : SST, SLP, AT, DPT, WBT , RH, W , D ,WW , N, VV , WH, NH ,W1	
4 or more matches, 0 or 1 mismatches, one ID invalid	Yes
4 or more matches, 0 mismatches	Yes
4 or more matches, 0 or 1 mismatches, match between DCK 192 & 254 or 193	Yes
Any mismatch and missing ID	No
More than one mismatch and missing or invalid ID	No
Mean absolute difference across all variables in match assessment < 0.01	Yes
Mean absolute difference across all variables in match assessment >= 0.01	No

Table 14. Selection of best duplicate

Ranking criteria
Highest priority DCK
Delayed mode format version (if any)
Number of elements in list: SST, SLP, AT, W, D, WW, N, WBT, DPT, VV, RH, WH, NH, W1, OSV
Number of elements in list sst, SLP, AT, W, D, WBT, DPT, RH
Presence of C1M, indicating match to callsign in Pub 47
ICOADS duplicate status,DUPS

3.3.3 Output data file format

The results of the duplicate flagging are output in year-month pipe-delimited files. The first column is ICOADS UID, the second duplicate flag status (Table 11) then in the third column a list of the other UIDs in the duplicate group (comma separated within curly braces).



ICOADS-30-D30K3W|1|{ICOADS-30-D30K3V}
ICOADS-30-D30K3V|3|{ICOADS-30-D30K3W}
ICOADS-30-D31APM|1|{ICOADS-30-D31APK,ICOADS-30-D31APL}
ICOADS-30-D31APK|3|{ICOADS-30-D31APM,ICOADS-30-D31APL}
ICOADS-30-D31APL|3|{ICOADS-30-D31APM,ICOADS-30-D31APK}
ICOADS-30-0S9WKD|1|{ICOADS-30-D30SXM,ICOADS-30-D30SXX,ICOADS-30-D30SXL}
ICOADS-30-D30SXM|3|{ICOADS-30-0S9WKD,ICOADS-30-D30SXX,ICOADS-30-D30SXL}
ICOADS-30-D30SXX|3|{ICOADS-30-0S9WKD,ICOADS-30-D30SXM,ICOADS-30-D30SXL}
ICOADS-30-D30SXL|3|{ICOADS-30-0S9WKD,ICOADS-30-D30SXM,ICOADS-30-D30SXX}
ICOADS-30-D31VJA|0|{}
ICOADS-30-D31VJB|0|{}

3.4 Duplicate Records by ID

3.4.1 Linking IDs

Once the date/time/location parameter value duplicates have been identified and flagged, the next step in the processing considers together the data that have associated IDs. Sometimes the link between IDs can be used to homogenise the IDs beyond the individual pairs, sometimes the link is specific to a particular pair of reports, particularly if one of the matched IDs is generic. Table 15 details the DCKs that are assigned to particular groups where matches of IDs are not expected outside these groupings. ID matches are therefore only considered within-group. At the end of the processing the suffix “_gN” is appended to the IDs, where N is the group number from Table 15. Group 0, the “all other data” grouping, does not have this information appended.

Table 15. Group assignments by DCK and ID. The group is appended to the ID to avoid mixing data with the same ID from different types of data.

Group	DCK Selection	Comment
1	118, 119, 762, (705, 706, 707 & ID starts with JP)	Japanese data
2	116, 117, 195	US, including US Navy
3	229, 239	UK Navy
4	205, 211, 213, 221, 218, 902	UK
5	128, 298, 230, 254, 255, 926, 927 and at least 1 character	callsigns



Group	DCK Selection	Comment
6	900	Australian data
7	667	Inter-American Tropical Tuna Commission
8	186, 733	Ice stations
9	188	Norwegian Antarctic Whaling Factory Ships
10	226	UK OWS
0	All other data	

Table 16. ID linking criteria

Criterion	Allow Match
Different group (Table 16)	No
Generic ID	No
Callsign to id.class = <i>IMMPC, id_4digit, id_3digit, selected_logbook, AU_ship_number, HMS, id_minus, id_7digit_pership, id_5digit, MARID, IATTC_number, IMM_3dig_shipno, numeric, id_dash_3dig_CHECK, IMM_4dig_shipno, IMM_6dig_shipno, id_4dig_CHECK, north_pole_station</i>	Yes
Numeric ID in DCK 254	No
ID starts C7	No
ID of length 1 in DCK 194	No
IDs both length 2	No
DCK 116, mix of ship and OWS id.class: <i>ship_number, osv_onstation</i>	No
6 character IDs starting SHIP in DCK 700	No
DCK 720, SID 135	No
For all matches, at least one ID in pair must have at least 6 occurrences	

Table 17. ID preferences

Choice criteria, choices are sequential so the last met will be selected
Most occurrences
Not DCK 700
Not invalid
Callsign, length ≥ 4
Valid ITU
C1M present
DCK 118, 202 or 203 over 705, 706, 707
6 character IDs starting SHIP in DCK 700
DCK 720, SID 135
For all matches, at least one ID in pair must have at least 6 occurrences

3.4.2 Forming ship tracks for linked IDs

The linked IDs are then checked using the MOQC track check, and for time duplicates. Reports that fail the track check are flagged as a worst duplicate. Where positions are similar the best duplicate is selected by dck priority and number of elements as before.



4. Results

4.1 Duplicate identification and ID linking results

Figure 4 shows timeseries of the numbers and proportions of duplicates identified in the selected data from ICOADS R3.0. The duplicate numbers from the ICOADS dupelim are also shown. The top panel shows the numbers of reports passing the new duplicate check (beige), the numbers identified by dupelim (blue) and those identified in the new duplicate procedure (cyan). For most of the period the new approach identifies more duplicates (compare panels 2 and 3 in Figure 4, also shown in lower 2 panels as a difference).

Comparison with dupelim output is complicated as the total files only contain the full original record for data sources ingested from Release 2.5 (Woodruff et al. 2011). For data where the worst duplicates are still available there is no information about which reports are linked. Comparison of the different procedures is therefore by inspection of individual cases, which is timeconsuming and requires further work.

Early in the record duplicate data are mostly identified within-DCK, with a few additional duplicates between some known common DCK groups (152/155/156/193, 194/201, 192/215/720, 701,721). This is illustrated in Figure 5 which illustrates the relationships between DCK (around the edge of each circular plot) using a chord plot of the duplicates found. The thickness of the chord plot lines is \log_{10} of the number of duplicates found. The colour indicates the best duplicate, in the first panel of Figure 5 all of the linkages to DCK 151-156 do not match to the colours of those DCKs. For example, links to DCK 156 (blue) are in brown and pink showing these data are worst duplicates in comparison with DCK 910 and 193 respectively. The proportion of duplicates identified remains low until the 1930s, with a range of DCK being linked (Figure 5).

In the 1930s a large number of duplicates are found mainly within DCK 192. Another source of duplicates not identified by dupelim is with the DCKs 705/706/707 and several national archive DCKs: 193, 118, 762. Post World War 2 the number of duplicates rises as a new source ingested for Release 3, DCK 117, is heavily duplicated with existing data from 117, but also overlaps with 116 and also 218. Figure 5 shows that the relationship between DCKs post World War 2 to the end of the 1950s is complicated.

From the 1965 to 1972 the new procedure typically identifies fewer duplicates than dupelim. The reasons for this need investigation – either dupelim is overflagging, or the new procedure needs to be improved to catch the missing duplicates. One example was found was a report with ID “SHIP” and PT = 5 (ship) was identified by dupelim as a duplicate with 2 other reports with PT = 6 (moored buoy) – these buoy reports were not considered here, so the duplicate was missed. It’s not clear whether this is a common occurrence or not.

From the 1973 to 2013 the new procedure identifies more duplicates than dupelim, up to about 3%, with the exception of 1999 and 2000. Figure 4 shows a large jump in the number of reports, and also the fraction of duplicates toward the end of 1999. Most reports have ID information during this period so it is likely that the cause of the difference between the methods is due to the different ways that IDs are treated, which method performs better in this period is not presently clear.



It seems likely that from 2014 the new procedure needs improving dupelim identifies more than 5% more duplicates, mostly with generic masked IDs. Most of the reports from DCK 792 with ID “MASKSTID” are likely to be duplicates with DCK 992 – more investigation is required.



Figure 4: Duplicate identification timeseries

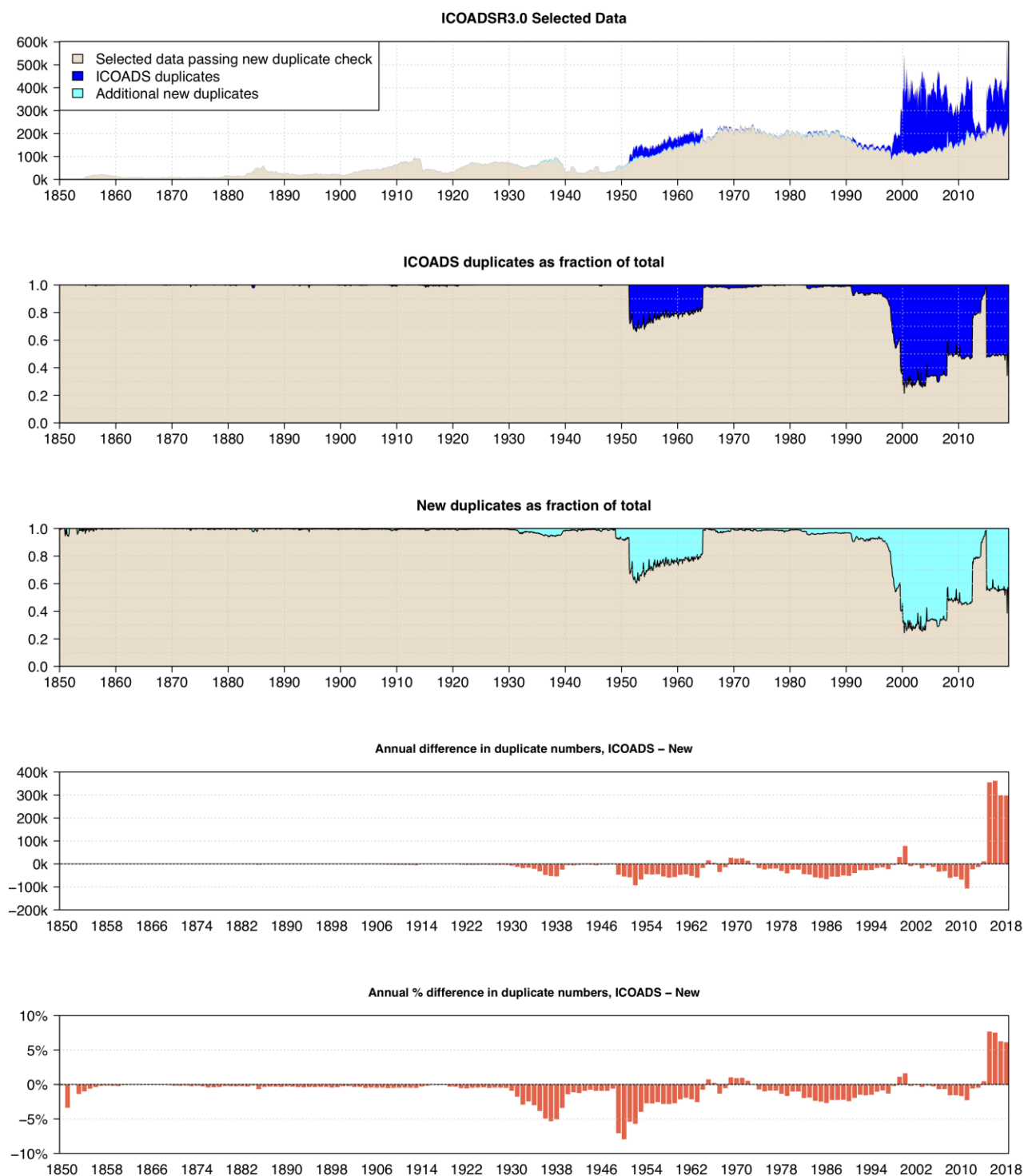




Figure 5: Chord diagrams of duplicates, 1850 to 1959

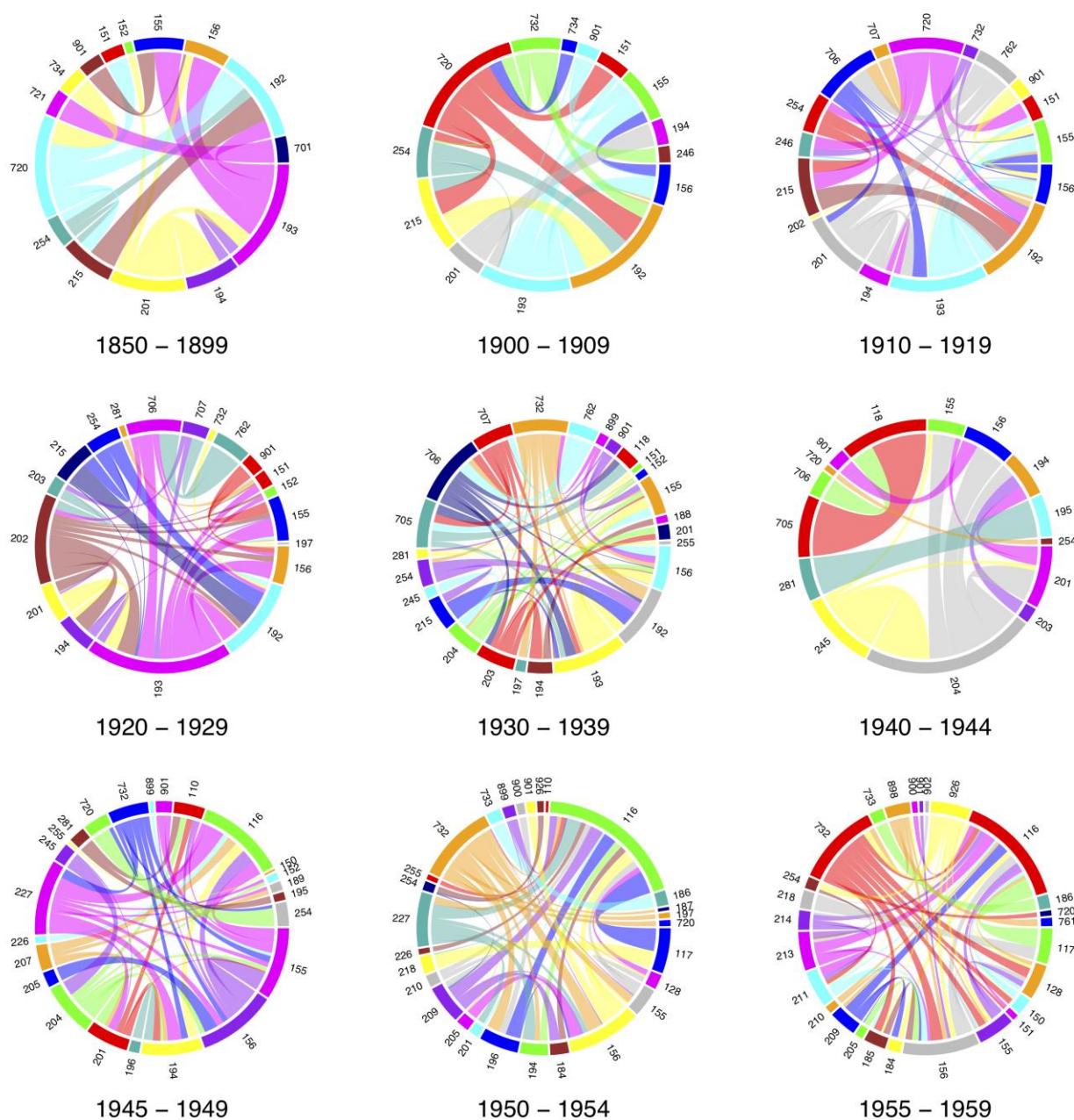
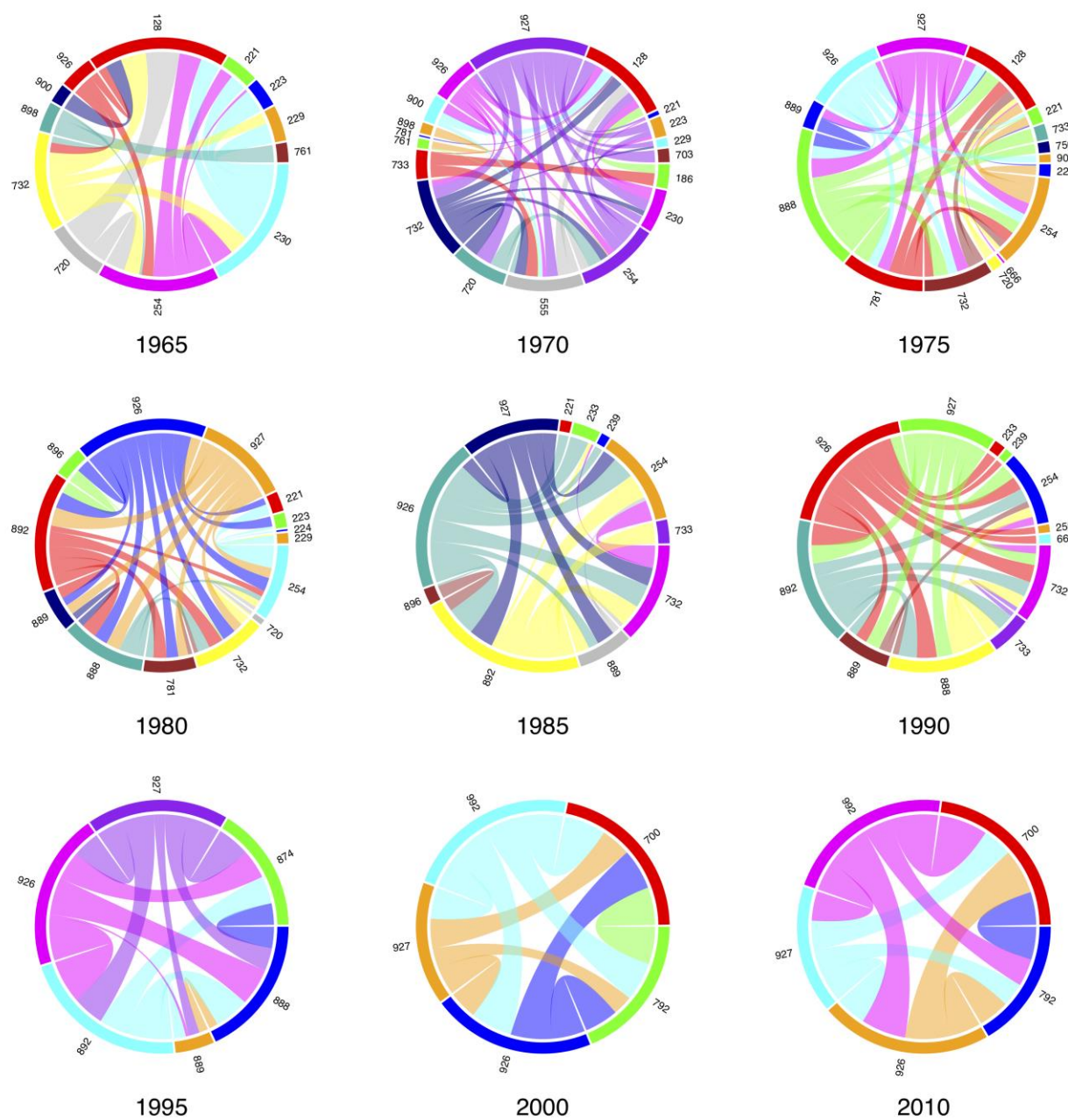




Figure 5, continued: Chord diagrams of duplicates, 1960 onwards





5. Summary

A new procedure is described to identify and flag duplicates within ICOADS Release 3.0. Whilst conceptually quite simple, the implementation becomes complex due to differences between data sources that often require case-by-case consideration.

Typically the method flags more reports as duplicates than the ICOADS dupelim procedure. From 2014 the method may need improvement as dupelim consistently flags about 5% more duplicates.

One advantage over dupelim is that the information linking all the suspected duplicates together is retained, allowing inspection of the performance of the procedure. The absence of this information for dupelim makes comparison difficult. This difficulty is compounded as only the preferred duplicates were carried through ICOADS processing for Releases prior to 2.5, this means large volumes of duplicates are not available for training, testing or comparison.



References

{C3S_D311a_Lot2.1.1.1_201708 Preliminary_Marine_Inventory_v1}. Available upon request from the service manager Ms. Corinne Voces (Corinne.voces@mu.ie).

{C3S_D311a_Lot2.1.1.1_201708 Preliminary_Marine_Inventory_Annex_I v1}. Available upon request from the service manager Ms. Corinne Voces (Corinne.voces@mu.ie).

{C3S_D311a_Lot2.3.4.4-2019_201910_Third version_Marine_User_Guide_v1}. Available upon request from the service manager Ms. Corinne Voces (Corinne.voces@mu.ie).

{C3S_D311a_Lot2.2.1.1_201708_Initial_specification_for_CDM_v1}. Available upon request from the service manager Ms. Corinne Voces (Corinne.voces@mu.ie).

{C3S_D311a_Lot2.3.2.1_201712_Specification_of_test_data_delivery_service. v1}. Available upon request from the service manager Ms. Corinne Voces (Corinne.voces@mu.ie).

{C3S_D311a_Lot2.3.5.1_201709_Agreed_user_model_v1}. Available upon request from the service manager Ms. Corinne Voces (Corinne.voces@mu.ie).

Freeman E, Woodruff SD, Worley SJ, Lubker SJ, Kent EC, Angel WE, Berry DI, Brohan P, Eastman R, Gates L, Gloeden W, Ji Z, Lawrimore J, Rayner NA, Rosenhagen G, Smith SR. 2017. [ICOADS Release 3.0: A Major Update to the Historical Marine Climate Record](#), *International Journal of Climatology*, 37, 2211–2232. doi: [10.1002/joc.4775](https://doi.org/10.1002/joc.4775).

Kennedy, J. J., C. Atkinson and K. Willett, 2017: Marine Data System Quality Control, UK Met Office technical note, 23pp.

Kent, Elizabeth C.; Woodruff, Scott D.; Berry, David I. 2007. [Metadata from WMO Publication No. 47 and an Assessment of Voluntary Observing Ships Observation Heights in ICOADS](#), *Journal of Atmospheric and Oceanic Technology*, 24 (2). 214-234. doi:[10.1175/JTECH1949.1](https://doi.org/10.1175/JTECH1949.1).

Slutz, R.J., S.J. Lubker, J.D. Hiscox, S.D. Woodruff, R.L. Jenne, D.H. Joseph, P.M. Steurer, and J.D. Elms, 1985: [Comprehensive Ocean-Atmosphere Data Set; Release 1](#). NOAA Environmental Research Laboratories, Climate Research Program, Boulder, CO, 268 pp. (NTIS PB86-105723).

Smith SR, Freeman E, Lubker SJ, Woodruff SD, Worley SJ, Angel WE, Berry DI, Brohan P, Ji Z, Kent EC. 2016. The International Maritime Meteorological Archive (IMMA) Format. <http://icoads.noaa.gov/e-doc/imma/R3.0-imma1.pdf> (17 December 2018).

Woodruff, S.D., R.J. Slutz, R.L. Jenne, and P.M. Steurer, 1987: [A comprehensive ocean-atmosphere data set](#). *Bull. Amer. Meteor. Soc.*, **68**, 1239-1250.

Woodruff, S.D., S.J. Worley, S.J. Lubker, Z. Ji, J.E. Freeman, D.I. Berry, P. Brohan, E.C. Kent, R.W. Reynolds, S.R. Smith and C. Wilkinson, 2011: ICOADS Release 2.5 and Data Characteristics, *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. *Int. J. Climatol.* (*CLIMAR-II Special Issue*), **25**, 823-842 ([doi:10.1002/joc.1166](https://doi.org/10.1002/joc.1166)).



Annex: Software used in processing

Base R (version 3.5.1),

Code runs in R (version 3.5.1), with several packages required:

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

R included packages (all 3.5.1): stats, graphics, grDevices, utils, datasets, methods, base

jsonlite 1.6

Jeroen Ooms (2014). The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. arXiv:1403.2805 [stat.CO] URL <https://arxiv.org/abs/1403.2805>.

stringdist 0.9.5.1

van der Loo M (2014). "The stringdist package for approximate string matching." The R Journal, 6, 111-122. URL: <https://CRAN.R-project.org/package=stringdist>

lubridate 1.7.4

Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL <http://www.jstatsoft.org/v40/i03/>.

data.table 1.12.2

Matt Dowle and Arun Srinivasan (2019). data.table: Extension of `data.frame`. R package version 1.12.2. <https://CRAN.R-project.org/package=data.table>

igraph 1.2.4.1

Csardi G, Nepusz T: The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006. <http://igraph.org>

maps 3.3.0

Original S code by Richard A. Becker, Allan R. Wilks. R version by Ray Brownrigg. Enhancements by Thomas P Minka and Alex Deckmyn. (2018). maps: Draw Geographical Maps. R, package version 3.3.0. <https://CRAN.R-project.org/package=maps>

local package versions of MO climatological check and track check

Converted to R by Richard Cornes (NOC) (note, code conversion performed under a different project, code remains to be uploaded to git).



ECMWF - Shinfield Park, Reading RG2 9AX, UK

Contact: info@copernicus-climate.eu