Received: 17 August 2020 | Accepted: 5 October 2020

DOI: 10.1002/gdj3.109

DATA SERVICES PAPER

Geoscience Data Journal | RMetS | WILEY

Progress towards a holistic land and marine surface meteorological database and a call for additional contributions

Simon Noone[1] | **Chris Atkinson**[2] | **David I. Berry**[3] | **Robert J. H. Dunn**[2] | **Eric Freeman**[4] | **Irene Perez Gonzalez**[3] | **John J. Kennedy**[2] | **Elizabeth C. Kent**[3] | **Anthony Kettle**[1] | **Shelley McNeill**[4] | **Matthew Menne**[4] | **Ag Stephens**[5] | **Peter W. Thorne**[1] | **William Tucker**[5] | **Corinne Voces**[1] | **Kate M. Willett**[2]

[1]Department of Geography, Irish Climate Analysis and Research UnitS (ICARUS), Maynooth University, Kildare, Ireland

[2]Met Office, Exeter, UK

[3]National Oceanography Centre, Southampton, UK

[4]NOAA National Centers for Environmental Information (NCEI), Asheville, NC, USA

[5]Science and Technology Facilities Council, Swindon, UK

Correspondence
Simon Noone, Department of Geography, Irish Climate Analysis and Research UnitS (ICARUS), Maynooth University, Kildare, Ireland.
Email: simon.noone@mu.ie

Funding information
The Service is funded by Copernicus C3S under contract C3S 311a Lot 2 led by Maynooth University. The marine processing uses software and methods developed by NOC under NERC funding from grants NE/J020788/1, NE/S015647/2, NE/R015953/1. Link to Copernicus C3S Climate Data Store https://cds.climate.copernicus.eu/#!/home

Abstract

This paper outlines progress of the Copernicus Climate Change Service's (C3S) Global Land and Marine Observations Database service in securing data sources and introduces the data upload component. We present details of land and marine data holdings inventoried, highlighting priority needs in terms of periods, regions and Essential Climate Variables (ECVs) where additional data could bring most benefit. These holdings are being iteratively merged and integrated to best meet user needs and are served to the user via the Copernicus Climate Data Store (CDS). The secure Data Upload Server enables any data provider to share additional data and metadata with the service. We outline the process for registering as a data provider and how data sets are prioritized for integration. We encourage all data owners to share their data with the C3S service via our Data Upload Server. All unique and relevant data acquired or submitted will be also archived at the NOAA National Centers for Environmental Information World Data Center for Meteorology, Asheville, North Carolina, USA and used in their database curation efforts which are being jointly developed.

KEYWORDS

climate data, climate services, land, marine, meteorological observations

1 | INTRODUCTION

Historical observational climate records are crucial in understanding climatic variability, extreme past weather and climate events and allowing us to make informed decisions for better societal adaptation to climate change (e.g., Kennedy *et al*., 2010; Shapiro *et al*., 2010; Noone *et al*., 2017; Murphy *et al*., 2017; Thorne *et al*., 2018). Historical observations are also a key component to derive reanalysis products (e.g., Compo *et al*., 2011; Dee *et al*., 2011) and evaluate climate models (e.g., Barker *et al*., 2004; Brunet *et al*., 2013; Flato *et al*., 2013; Prohom *et al*., 2015; Wilby, 2016).

The management of both marine and land historical data sets has been highly fragmented, leading to diverse data holdings held by multiple institutions. Consequently, it is necessary to confront the challenges of a plethora of distinct data formats; gross duplication of records with differing identifiers, names; and in many cases varying geo-location information. Within available land and marine data holdings, there are greatly differing levels of completeness, data quality checks and data processing applied. There are further issues with limited data discovery metadata and sometimes a distinct lack of traceability to the underlying original data source (Thorne *et al*., 2018).

There have been many efforts to produce land observation data holdings, but these have been essential climate variable (ECV) and/or timescale (observations available hourly, daily and monthly) specific. In most cases, these data holdings are also regionally or nationally specific with only a few providing truly global coverage. Examples of truly global holdings include Cram *et al*.(2015) who produced The International Surface Pressure Databank (ISPD) the world's largest collection of hourly and synoptic global surface and sea level pressure observations. Similarly, the International Surface Temperature Initiative (ISTI) produced a set of monthly temperature data holdings (Rennie *et al*., 2014). Others have produced global precipitation data holdings such as the Global Precipitation Climatology Centre (GPCC) and the Global Precipitation Climatology Project (GPCP) (Adler *et al*., 2003; Schneider *et al*., 2013). These efforts tend to concentrate either upon single ECVs, single observational timescales, or both. Yet, many applications wish to consider changes across a range of ECVs and/or a range of timescales. For example, drought analyses typically require both temperature and rainfall data (Banimahd and Khalili , 2013; Saeidipour *et al*., 2019; Jin *et al*., 2020).

Recognizing this, there have been efforts to merge land data holdings with multiple ECVs derived from various existing data holdings. At a regional level, the European Climate Assessment and data set (ECA&D) (Klein Tank *et al*., 2002; Klok and Klein-Tank., 2009) contains multiple ECVs for stations in 65 countries across Europe, Middle east and the Mediterranean. Van den Besselaar *et al*.(2015) produced the International Climate Assessment and Dataset (ICA&D) which offers science-based services to help the gathering of observations for archiving, quality control and homogeneity checks, facilitating climate extremes analysis in different parts of the world (e.g., Latin America and south east/central Asia). At a global level, Menne *et al*.(2012) created the Global Historical Climatology Network-Daily database (GHCNd) of merged daily observations. However, these data sets are typically limited in geographical coverage, temporal integration and in most cases only provide a subset of variables. The current approach to data management makes it difficult for users to obtain the full benefits of the available historical land meteorological holdings. On the other hand, the International Comprehensive Ocean-Atmosphere Data Set (ICOADS, 2016; https://icoads.noaa.gov) provides access to surface marine data in a consolidated holding. For over 30 years, the community has pooled efforts to create a single repository for these data containing all reported variables collated together. However, there are issues with the underlying data due to past data management practices which have not been revisited. For example, *among other things*, recoverable mispositioned data (i.e., data that can be corrected) have been discarded when apparently over land; undetected duplicates due to merging of different sources with different levels of numerical precision and station ID formats; and corrupted station IDs due to early real-time data sharing and digitization errors.

The C3S Global Land and Marine Observations Database (hereafter, the Service) is part of the Copernicus Climate Change Service (C3S; https://climate.copernicus.eu) making climate data and information more easily accessible to support adaptation and mitigation policies of the European Union and the wider global community. The Service, in collaboration with NOAA's National Centers for Environmental Information (NCEI), aims to provide comprehensive access

Action A2	Land database
Action	Set up a framework for an integrated land database which includes all atmospheric and surface ECVs and across all reporting timescales
Benefit	Centralized archive for all parameters. Facilitates QC among elements, identifying gaps in the data, efficient gathering and provision of rescued historical data, integrated analysis and monitoring of ECVs. Supports climate assessments, extremes, etc. Standardized formats and metadata
Who	NCEI and contributing centres
Time frame	Framework agreed by 2018
Performance Indicator	Report progress annually to AOPC
Annual cost	US$ 100 000-1 million

TABLE 1 Action A2 extracted from the Global Climate Observing System's latest Implementation Plan (GCOS, 2018) (https://gcos.wmo.int/en/home)

to the global archives of surface meteorological observations made over land and oceans through a common interface and data model, integrated across timescales and ECV's. In doing so, the Service is actively fulfilling several actions called for in the Global Climate Observing System's latest Implementation Plan (GCOS, 2018 –https://gcos.wmo.int/en/home) most notably Action A2, see Table 1 for details).

The data in the Service will be hugely important for producing climate service products such as reanalysis. Reanalyses systematically produce data sets at regular intervals over long periods of time for climate monitoring and research (Parker, 2016; Hersbach *et al*., 2020). Reanalysis products are heavily reliant on observations from marine, land and atmospheric monitoring networks (for more information see https://www.ecmwf.int/en/research/climate-reanalysis).

There are large quantities of both land and marine data that are already digitized but not yet collectively archived. Many marine data sources are well-managed at the national, regional or observing network level, but for climate applications, the observations need to be integrated and harmonized internationally and across platform types. One obvious example where climate quality observations are not routinely available for climate applications are underway observations from research vessels and ships of opportunity (Smith *et al*., 2019). It is also important to identify the original data sources for ICOADS as many reports were not included in the public releases and are not presently available (Kent *et al*., 2019). The land-based station component requires a huge effort to locate and acquire all known available meteorological data and metadata at multiple timescales (e.g., sub-daily, daily and monthly). These holdings need to be acquired, inventoried, merged and integrated across ECVs to meet the needs of climate service users.

The purpose of this paper is threefold: a) to provide details of the data secured by the Service to date; b) to outline data integration priorities; and c) to introduce the data upload component of the Service. Any data owner can provide data for potential inclusion in subsequent database releases. This data contribution may arise from organizations such as National Meteorological and Hydrological Services (NMHS) or other public or private entities, or via an individual or group undertaking, for example a data rescue effort (e.g., ACRE; https://met-acre.net). The Service has already received historical observations from several data rescue projects and has started to integrate data from 1900 to 1910 Met Office Daily Weather Reports as part of the citizen science project Weather Rescue (https://weatherrescue.org) into the current database release. The Copernicus C3S Data Rescue Service facilitates and coordinates the rescue of weather and climate data from around the world. The online portal collects and shares information on past, current and planned data rescue projects, whilst promoting best practice and standards for all aspects of the data rescue process (https://datarescue.climate.copernicus.eu/).

The remainder of this paper is set out as follows: Section 2 provides details of all data inventoried as at April 2020 along with temporal, spatial and ECV priorities for the Service. Section 3 provides details on how data can be contributed to the Service and how a data set is prioritized for integration and subsequently served via the CDS. Section 4 presents a summary.

2 | OVERVIEW OF AVAILABLE OBSERVATIONS

2.1 | Land-based holdings

As of April 2020, a total of 319 land-based sources have been acquired and the data inventoried. These sources comprise of 107,894 sub-daily station series, 173,850 daily station series and 186,015 monthly station series, with many sources containing data at multiple timescales. Many stations will exist in multiple sources so the true number of unique stations is lower than these counts. The following ECVs are generally available across all three timescales: precipitation, temperature, sea level pressure, humidity, snow and wind measurements. However, these may not all be available at each station or in each source. For example, many sources arise from the International Surface Pressure Databank (Cram *et al*., 2015) and the International Surface Temperature Initiative databank (Rennie *et al.,*2014) and thus consist of mainly pressure and temperature data, respectively. Data for other parameters may exist, but historically the governance of land in situ holdings has encouraged fracturing of data holdings (see Section 1). GCOS therefore instigated Action A2 and the community responded, first with a white paper (Thorne *et al*., 2018), and then with this collaboration between C3S and NOAA NCEI. Full details of both land and marine inventories can be found at: https://datadeposit.climate.copernicus.eu/inventories/

2.1.1 | Sub-daily stations inventoried

The Service has already significantly improved the state of land-based inventories of sub-daily timestep (mainly hourly) observations. The Service began in mid-2017 and in the first year inventoried 13 sub-daily data sources (814 stations). In 2018, the Service added 38 new sources (22,805 stations) and in 2019 a further 72 sources (75,064 stations). As at April 2020, the sub-daily station inventory consists of 126 data sources and 107,894 stations, in 200 different countries, territories and dominions, although many of these stations are duplicated.

One important addition to the sub-daily inventory since 2017 was the re-issue of the United States Air Force (USAF) sub-daily data holdings to NOAA/NCEI under a data-sharing agreement and archived as the Data Set Index (DSI) 9966-03 at NCEI. The USAF sub-daily data has formed the basis of the International Surface Database (https://www.ncdc.noaa.gov/isd). We extracted four main global 'platform' types, referring to the preferred identifier used to index the USAF data (see Table 2). These four platform types (AFWA, WMO, C-MAN and ICAO) shown in Table 2 index the vast majority of long-term station observations consisting of FM12 (SYNOP Report from a fixed land station), FM15 (METAR Aviation routine weather report), Surface Airways Observations etc. The USAF holdings in totality hold in excess of 150 fields pertaining to observed variables but for the vast majority of stations very many of these are either perpetually missing or only sporadically filled. We have initially retained nine principal variables (Temperature, Precipitation, Wind, Humidity, Sea Level pressure, Surface pressure, Snow, Cloud Ceiling, Visibility and Hail size) from the four extracted platform types. Future work may expand the selection of fields to be processed and served.

The USAF data include the Coastal-Marine Automated Network (C-MAN) network of stations located on lighthouses, at capes and beaches, on nearshore islands, and on offshore platforms. These stations are almost exclusively US and US territory data. This source is also partially present in ICOADS and thus also in the marine data holdings. C-MANs have not been used to date in the land data processing pending resolution of how best to manage and quality control these data between the land and marine domains.

The USAF source also contains several tens of thousands of additional observations which are mostly available post 1990s. These include private and cooperatively owned and operated networks of weather stations called Mesonets. These Mesonets consist of usually closely spaced stations, usually within a 30 km radius, that report meteorological data frequently (Mahmood *et al*., 2017). The ambition is to include as many of these as possible, but they have not yet been inventoried as longer records have so far been prioritized. Substantial work would also be required to better understand the usage conditions that pertain to these data. The sub-daily inventory includes 70 sources obtained from the input sources to the International Surface Pressure Databank (ISPD, Cram *et al*., 2015) which are almost exclusively sea level pressure and station level pressure data. Two more data sets 'Tape Deck TD-13 and TD-14', from the National Centre for Atmospheric Research (NCAR; Colorado, USA) data archive, have also been added. TD-13 consists of 10,851 global stations and TD-14 has 300 stations located across the USA containing temperature, precipitation, humidity, wind and pressure observations.

We have also acquired sub-daily data from the NOAA Climate Database Modernization Program (CDMP). The CDMP funded the imaging and keying of nearly 56 million climate data observations (https://www.ncdc.noaa.gov/climate-information/research-programs/climate-database-modernization-program). The CDMP digitized in situ surface observations from the 19th century, primarily 1820–1892 data across the United States, mainly collected by the Smithsonian Institute and the U.S. Army Signal Service. However, some surface pressure data from this period are already included in the ISPD. The CDMP project also digitized daily and sub-daily observations collected by the U.S. Weather Bureau from 1892 to about 1950. Many of the daily records have already been incorporated into GHCNd. The Service has added over 500 CDMP stations located across the United States to the inventory with sub-daily observations of wind, snow, temperature, water vapour and pressure from 1892 to 1997 which are now archived at NOAA NCEI as DSI 3,850, 3,851 & 3,853.

2.1.2 | Daily stations inventoried

Since 2017, the Service has added inventories of 53 sources (10,890 stations) to the existing 91 data sources (162,892 stations) acquired and curated as part of GHCNd (Menne *et al*., 2012). This brings the current total to 144 data sources and 173,782 stations in 250 different countries, territories and dominions worldwide.

TABLE 2 Summary of the four main platform types extracted from the USAF sub-daily data

Source identifier	Source name	No. of stations	Principal variables available across sources	Data start and end years
220	(WMO) World Meteorological Organisation WMO/WIGOS	12,167	Temperature, Precipitation, Wind, Humidity, Sea Level Pressure, Surface pressure Snow, Cloud Ceiling, Visibility, Hail size.	1901–2019
221	(AFWA) Air Force Weather Agency	2,706		1926–2019
222	(C-MAN) Coastal-Marine Automated Network	551		1942–2019
223	(ICAO) International Civil Aviation Organisation	7,282		1901–2019

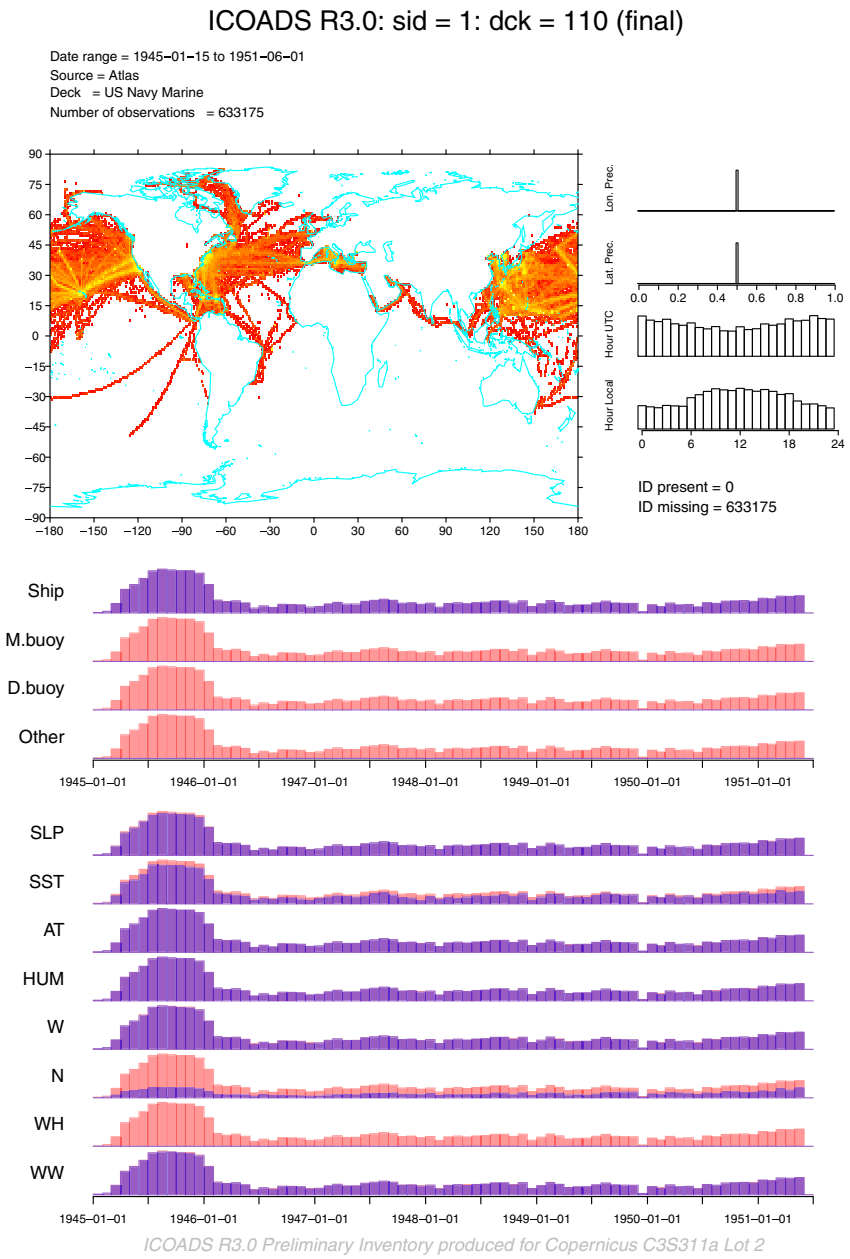

FIGURE 1 Inventory summary of sample ICOADS source (SID = 1, DCK = 110). Map shows locations of points as a heat map and the reporting precision of the location and time information to the right (in this example the latitude and longitude are both reported at the centre of a 1° box, and reports are hourly. Often data sources report at fixed hours in either local time or UTC). Report dates are shown as bar charts below. Purple shading in the bars shows presence of information, over-plotted on pink bars indicating the total number of reports, in various categories. Bars show observation type (ship, moored buoys, drifting buoy or other) and the presence of selected ECVs (surface pressure, sea surface temperature, air temperature, humidity, wind speed/direction, wave height and weather code). Figure developed using R software (R Core Team, 2019)

The following ECVs are generally available at daily resolution:

- Temperature
- Precipitation
- Sunshine Hours
- Mean Sea Level Pressure
- Wind measurements
- Water Vapour measurements
- Snow measurements

The new additions since 2017 have provided many more daily stations in important regions such as South America, Mexico, Hawaii and the Arctic. Newly added data sources for countries including Mexico, Suriname, Chile, American Samoa and Venezuela have increased both the station density and also the available ECVs from just temperature to now include precipitation, humidity and wind observations.

2.1.3 | Monthly stations inventoried

The Service has focussed on adding daily and sub-daily sources but has secured 55 sources for a total of 186,015 contributing series. These cover over 200 different countries, territories and dominions worldwide. This includes the addition of the Global Summary of the Month (which is derived from the GHCNd data set).

2.1.4 | Sources acquired but not inventoried yet and potential new data sources

The Service has acquired and is working to inventory, additional data sources exceeding 800 GB in volume. We are also working closely with the C3S Data Rescue Service (https://data-rescue.copernicus-climate.eu/) to ensure all rescued climate data is uploaded to the Service via the Data Uploads Server (Section 3). We will assess as many data sources as possible for inclusion in the data integration process over the term of the Service contract. However, due to the sheer volume of data and the varying source formats, assessment is very labour-intensive and will likely need to continue for many years.

The recently concluded EUMETNET (https://www.eumetnet.eu), European Environment Agency (EEA) and Copernicus data-sharing agreement (https://insitu.copernicus.eu/news/eea-and-eumetnet-sign-public-duty-license-agreement-for-data-provision-to-copernicus) offers an opportunity to gather data from European NMHS. Working with other C3S services and the EEA a data request has been made to the NMHSs resulting in indications of willingness to share data from several European NMHSs and Institutes and provision of some data. The data-sharing agreement with EUMETNET is to share with Copernicus services to at a minimum derive products but also many members have agreed to more open data policies. A visit to NCAR by Service team members during 2019 identified 82 different data sources varying from small country data sets to large global data sets that will be a useful addition to the land data inventory. The team at NCAR agreed to share all the archived data with the Service.

2.2 | Marine data sources

The marine component of the Service has initially been derived from the International Comprehensive Ocean-Atmosphere Data Set (ICOADS; https://icoads.noaa.gov). The ICOADS Release 3.0 (Freeman *et al*., 2017) is the most complete archive of surface marine observations taking input from historical collections, global data centres and near-real-time data systems). The Service builds on ICOADS Release 3.0.0 and the near-real-time updates denoted as Release 3.0.1.

To document the composition of the data sources underlying ICOADS, the observations have been inventoried according to the original data deck (named after the punch card decks the observations were originally stored on, for example Smith *et al*., 2016; Freeman *et al*., 2017) and the data source, noting that some decks contained observations from multiple sources. For example, the 'Deutscher Wetterdienst (DWD) Marine Meteorological Archive' deck contains data from 5 different sources separated by data type (e.g., modern lightships vs 19th century merchant vessels) and digitization project. For each combination of deck and source identifier, a summary of the observations available has been produced (see example in Figure 1). These provide a summary of where the observations were apparently taken (including any mis-located observations), the type of reporting platform (e.g., ship, moored buoy), the spatial reporting resolution, dates of observation, reporting times and the availability of relevant ECVs through time. These summary figures and tabulated discovery metadata are available as part of the Service via the CDS catalogue entry.

As part of the service ICOADS reports have been converted to use a consistent data model and vocabulary, common across the land and marine domains. QC flags have been applied using the UK Met Office QC suite (e.g., see Kennedy *et al*., 2019, Rayner et al., 2006; https://github.com/ET-NCMP/MarineQC). ICOADS contains many observations derived from the same original data sources, but past data management means that the reports are no longer exact duplicates in many cases. The identification and removal of these duplicates has been handled by ICOADS by the comparison of observations falling in the same 1° gridbox (Slutz *et al*., 1985). Prior to ICOADS Release 2.5 (Woodruff *et al*., 2011), reports identified as inferior duplicates were excluded from any further processing and are now only available in archaic formats. Processing for the Service starts with the 'total' files that include all reports that passed the ICOADS duplicate procedure prior to Release 2.5 and all reports from Release 2.5 onward.

The approach used to identify and flag duplicate reports in the Service is based on methods developed to identify reports from the same ship or platform when platform identification information (IDs) is missing, ambiguous or inconsistent (Carella *et al*., 2017). The availability of IDs is important for QC, the development of bias adjustments (Kent *et al*., 2017) and the calculation of uncertainty estimates for gridded data products (Kennedy *et al*., 2019; Kent *et al*., 2019b). Duplicates are identified by similarities in report dates, times, position and content (Kent *et al*., 2019b) in a similar manner to Slutz *et al*.(1985). The retention and consistency of ID information are prioritized throughout the processing, and after checking and homogenizing where necessary (Kent *et al*., 2019a), a further step ensures that there is only a single report at each time associated with a particular platform. Information on the full duplicate group is retained to allow evaluation of differences between data from different sources (e.g., Chan *et al*., 2019). Figure 2 shows sample output from the duplicate identification process for 1920 and 1960 for duplicate reports only. The coloured bands link ICOADS data decks containing duplicate reports, the colour of the band indicates the source of the preferred duplicate and the width of the band indicates the number of associated reports.

Although the marine processing has only so far been applied to the data already ingested into ICOADS, the same

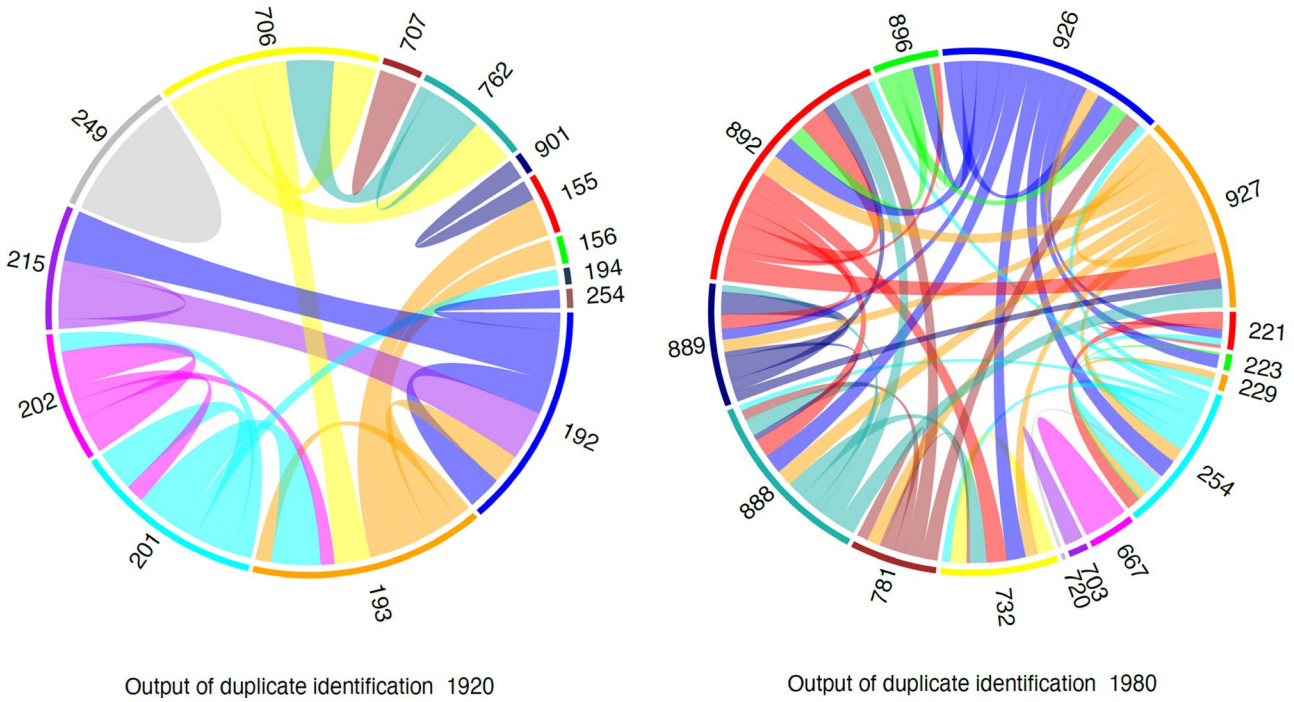

FIGURE 2 Summary information from output of duplicate identification for sample years of 1920 (left) and 1980 (right). Connections between ICOADS DCK numbers and colours around the edge show where reports in different DCK s have been associated as duplicates, coloured by the preferred DCK (e.g., in 1920 for all duplicate pairs containing reports from DCK 155 and 156 these sources are judged to be inferior). The widths are scaled by log10 of the number of report pairs. Plots produced using software package circlize (Gu, 2014). These are shown here as an illustration; more information on the meaning of the DCK designations can be found in Smith *et al*.(2016)

procedure is readily applicable to other data sources, for example from data rescue, exchange or reprocessing. Reprocessing the original ICOADS data sources to recover the rejected duplicate reports will be a future priority as comparison of different versions of the same observations from different data sources will enable a more comprehensive evaluation of data quality and quantification of uncertainty. Newly rescued data may also have been previously ingested into ICOADS in an inferior form (e.g., a subset of records or parameters, or with uninformative IDs, or poorly defined observational metadata), so identifying and flagging reports that should no longer be included in processing will become increasingly important.

2.3 | Land data coverage, data inventoried and priorities for integration

The spatial coverage of available stations is, as expected, densest in the most populated and economically developed areas of the globe. Remoteness and hostility of the environment make sustained and routine monitoring in many regions challenging and costly. Figure 3 shows the spatial coverage of all land stations inventoried and operational with at least one ECV available for specific time slices over time. All three timescales show limited data coverage outside of the United States, Europe and Australia prior to 1940. Currently, some regions have limited data at all timescales, including Sub-Saharan and Central Africa and parts of the Middle East, Siberia, South America, central Australia, Antarctica and Greenland. In some cases, no data exist but in many cases this highlights likely issues around data recovery or data-sharing policy.

For some regions, data are not available at every timescale. Specifically, at monthly and daily timescales, more coverage is required in parts of Western Asia. Monthly coverage is also sparse across South-Central Asia and Siberia. The sub-daily timescale has the smallest number of stations but there are some regions such as northern Africa where station coverage is greatest for the sub-daily timescale. This is the case also for south-central Asia and Siberia. In addition, most of the monthly station records presently are limited to temperature or precipitation only. Therefore, a priority for the Service is to acquire more ECVs at the monthly timescale across all regions. However, we will be able to fill some of the data gaps in monthly data coverage by deriving monthly averages from the available sub-daily and daily stations. Such monthly values will be given priority as they ensure consistency of provided data holdings across timescales by construction.

In addition to widespread and dense spatial coverage, long and continuous records are especially important for climate change analysis. The temporal coverage of the inventory could be improved across all timescales. Table 3 presents the average number of inventoried stations open for 50-year periods from 1750 to 1999 and a 20-year period 2000–2019 with observations for temperature, precipitation, sea level pressure and

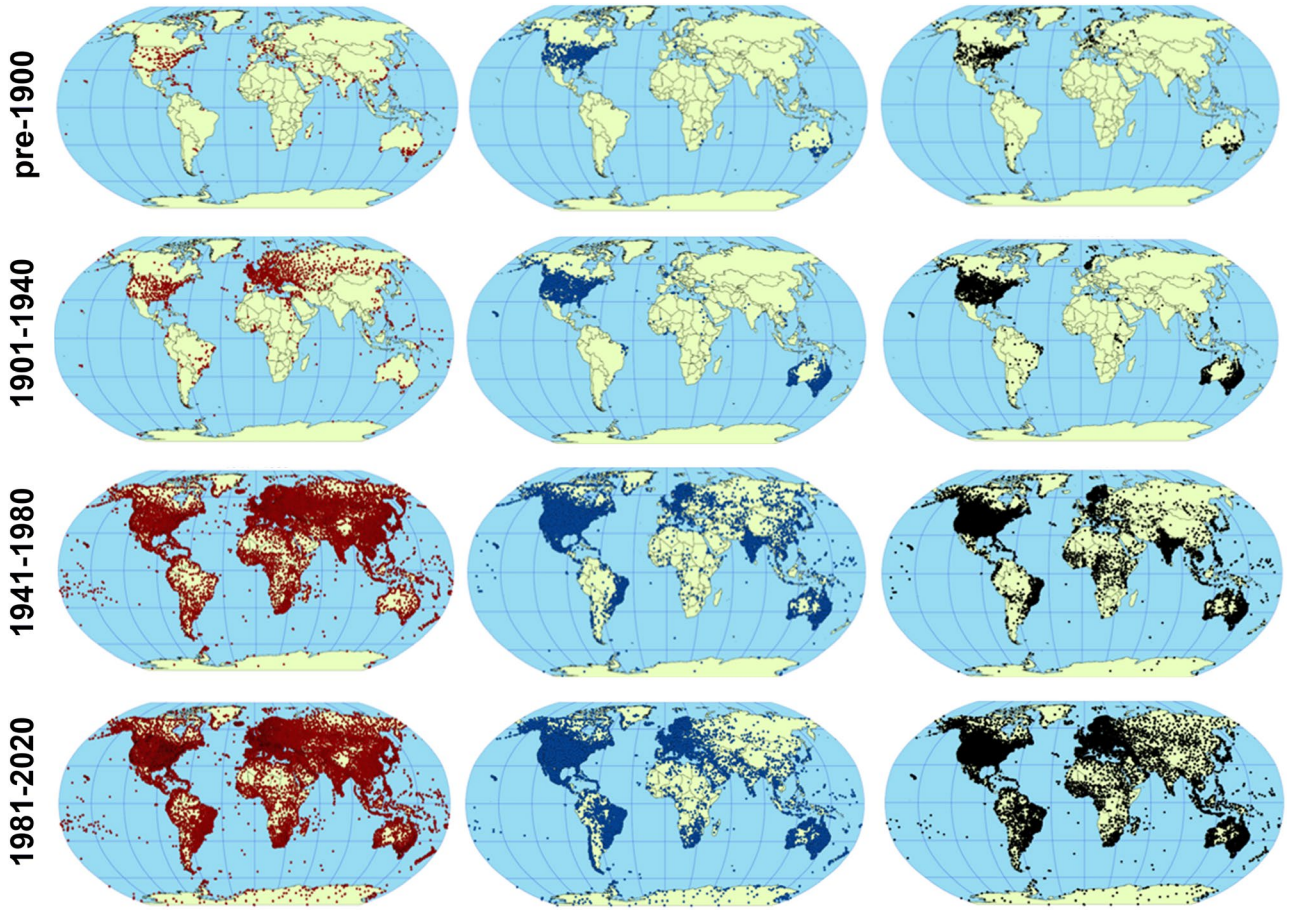

FIGURE 3 Location of land-based stations inventoried at each timescale operational with at least one ECV during specific time slice periods. Left panel: sub-daily stations [red dots], centre panel: daily stations [blue dots] and right panel: monthly stations [black dots]

snow. Figures 4 and 5 use the logarithmic scale to show the average number of stations inventoried for each year from 1750 to 2019 for the same four ECV's. The period 1750–1799 across all timescales shows limited data availability for all selected ECVs. The period 1800–1849 also has sparse coverage with an average of 248 stations reporting monthly temperature. It is known that many more data potentially exist prior to 1850 than are presently held (Brönnimann *et al*., 2019). During 1850–1899, there are on average 3,291 monthly stations with temperature data, nearly 2,000 monthly and daily stations with precipitation data and a few with snow observations (Table 3).

We see a marked increase in available stations with temperature, precipitation and snow observations at both daily and monthly timescale during 1900–1949. There is also an increase in sub-daily and monthly stations with sea level pressure and sub-daily temperature stations during the same period. However, there are limited daily sea level pressure stations open during 1900–1949, but we have 2,637 sub-daily stations which can be used to calculate the average daily sea level pressure for this period if required. The number of stations open on average each year between 1900 and 1949 clearly shows a priority requirement for more data predominantly at the sub-daily timescale during this period. These records may exist given the availability of daily and monthly stations over the same period. Daily and monthly statistics are for many parameters aggregated from the individual sub-daily observations so a daily or monthly report de facto implies that sub-daily observations must have been made. As expected, station counts increase substantially for the period (1950–1999) across all ECVs and timescales. However, there is still limited availability of sub-daily stations with snow data. Conversely, sea level pressure observations are most prolific at the sub-daily scale. However, in the last 20 years (2000–2019), we do see a general drop off in the number of stations at all timescales, although less so for sub-daily data. This is due to factors such as time lag of data getting into the archives, the closure of some networks, and reduced funding.

2.4 | Marine data coverage and priorities for integration

The marine observations included as part of the Service have been derived from the ICOADS Release 3.0 Total files (Freeman *et al*., 2017), with each record containing a single weather report from a ship, drifting or moored buoy, or

TABLE 3 Average number of inventoried stations open each year during 50-year periods from 1750 to 1999 and a 20-year period from 2000 to 2019 for four different variables. The number of stations at each timescale reflects what has been inventoried so far and there may be stations with data across more timescales (e.g., daily stations with month data)

Variable	Timescale	1750–1799 (50 years)	1800–1849 (50 years)	1850–1899 (50 years)	1900–1949 (50 years)	1950–1999 (50 years)	2000–2019 (20 years)
Temperature	Sub-daily	1	4	34	1,801	11,899	11,197
	Daily	6	28	894	7,650	33,453	28,769
	Monthly	64	248	3,291	22,401	51,893	32,981
Precipitation	Sub-daily	1	3	19	204	7,465	10,581
	Daily	2	27	1,966	19,074	45,948	42,171
	Monthly	9	36	1,925	19,448	36,102	34,425
Sea level pressure	Sub-daily	2	13	134	2,637	20,507	19,999
	Daily	1	14	129	148	3,776	3,707
	Monthly	1	6	270	2,405	4,403	2,310
Snow	SUB-daily	0	2	1	0	1,245	2,557
	Daily	0	12	721	7,185	14,500	18,792
	Monthly	0	0	506	6,442	11,983	17,608

other marine platform. These weather reports are typically instantaneous or short temporal averages (10 min) made at frequencies ranging from daily or once per 4-hr watch for the early data to either six-hourly or hourly for the more recent reports. Depending on the observing platform, each report will contain observations of one or more ECVs, typically

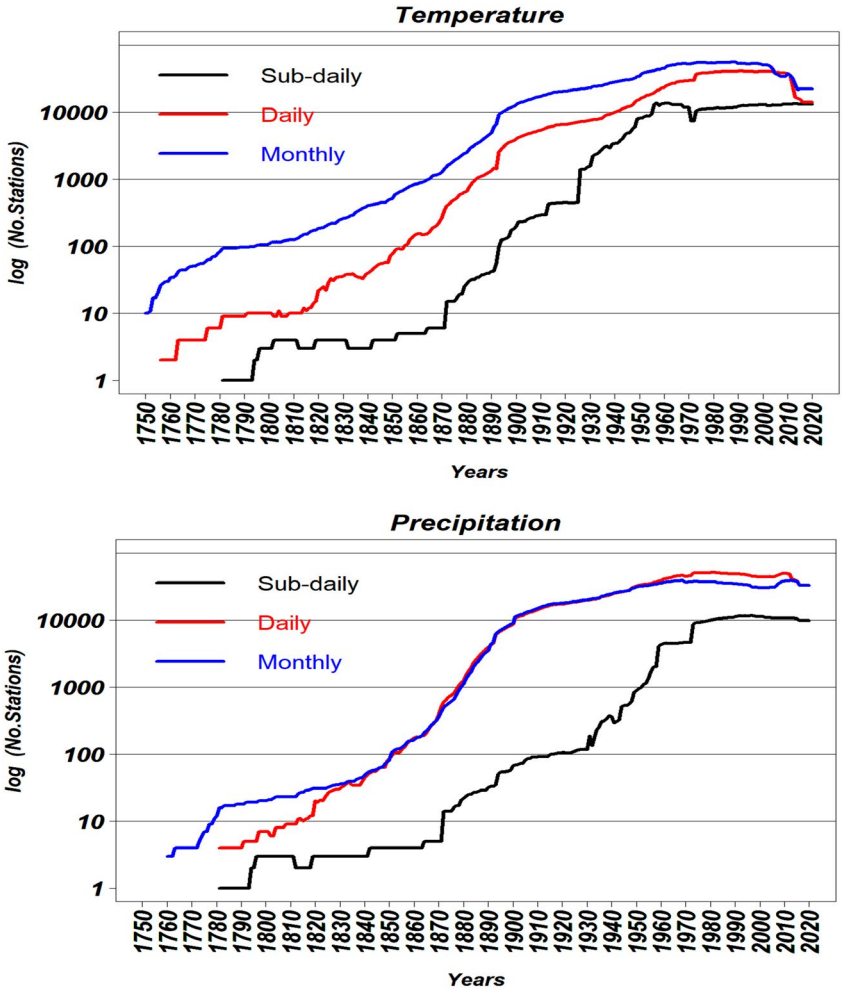

FIGURE 4 Number of land-based stations operational from 1750 to 2020 at each timescale for temperature and precipitation plotted using a logarithmic scale to account for the orders of magnitude changes over the period of record in data availability (black line = sub-daily stations, red line = daily stations and blue line = monthly stations)

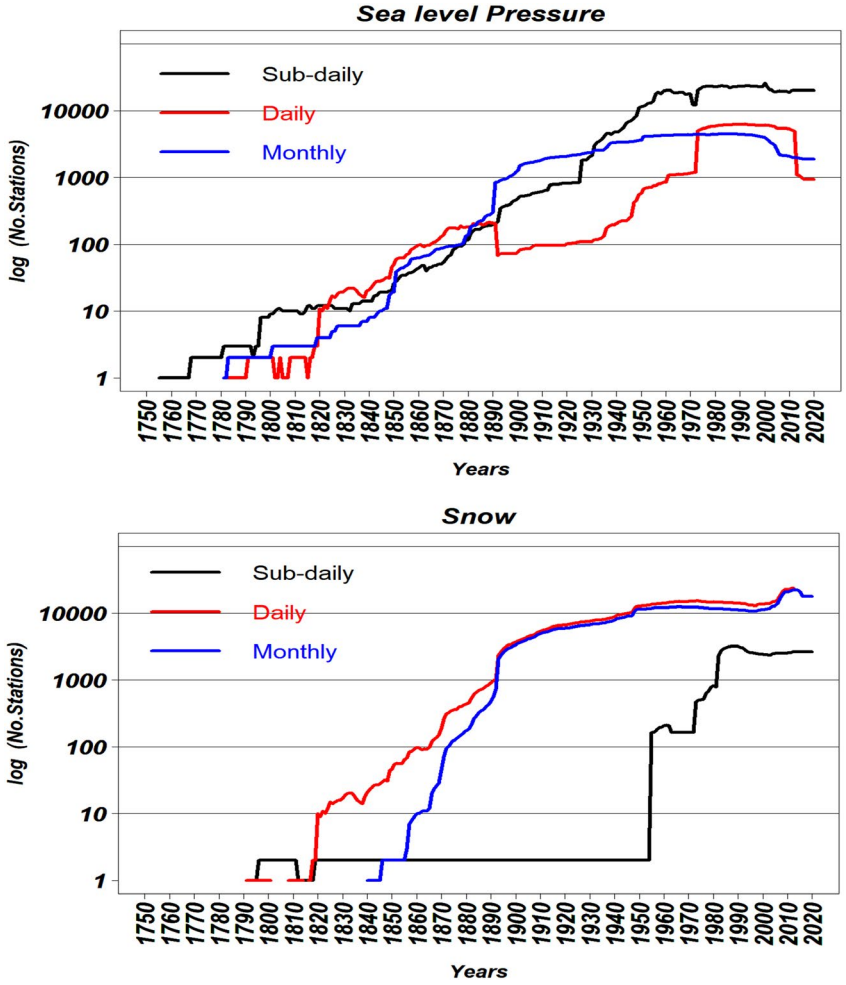

FIGURE 5 same as Figure 4 but with sea level pressure and snow (black line = sub-daily stations, red line = daily stations and blue line = monthly stations)

air temperature, near surface sea temperature, humidity (wet bulb temperature, dew point temperature or relative humidity), wind speed and direction and mean sea level pressure for ships. Some ships report coded weather and cloud observations and visually estimated wave parameters. Moored buoys and other fixed platforms typically report several ECVs whilst drifting buoys typically only report mean sea level pressure and/or sea surface temperature.

Figure 6 shows the fraction of months containing more than 5 weather reports per 5° grid cell from ICOADS independent of reported ECV for different 30-year periods between 1850 and 2014 (noting the shorter final period). Grid cells containing 5 or fewer observations per month have not been counted in order to exclude erroneous observations due to position errors. During the first period (1850–1879), the observations are clustered along the primary shipping routes between Europe and South East Asia, and with some travel to North and South America. The impact of the opening of the Suez (1869) and Panama (1914) canals can be seen, with increased shipping through the Arabian Sea and across the Bay of Bengal towards South East Asia and Australia and between Central America and Australia, respectively. Since 1970, the majority of regions have been sampled with almost monthly frequency. However, it should be noted that since the early 1990s this sampling has been driven by an increasing number of drifting buoys reporting a limited subset of ECVs.

Figure 7 shows the number of months sampled with more than five reports per month between 1850 and present for any report and for selected ECVs (air temperature, sea surface temperature, humidity, sea level pressure and wind speed). As with Figure 6, the concentration along the major shipping routes is clearly visible, with close to continuous monthly sampling since 1850 over some routes. The poorly sampled high latitudes and the tropical and south east regions of the Pacific are also clearly visible. With the exception of humidity, sampling across ECVs is fairly uniform with a similar spatial pattern in the number of months sampled across ECVs. Humidity was not routinely recorded/observed until the early to mid-20th Century and is still less-commonly reported than other ECVs, leading to fewer months overall with data but with a similar spatial distribution.

Figure 8 shows time series of the number of observations per month and number of 5 × 5 grid cells with data per month for any report and for the selected ECVs. With the exception of humidity, there is a general increase from

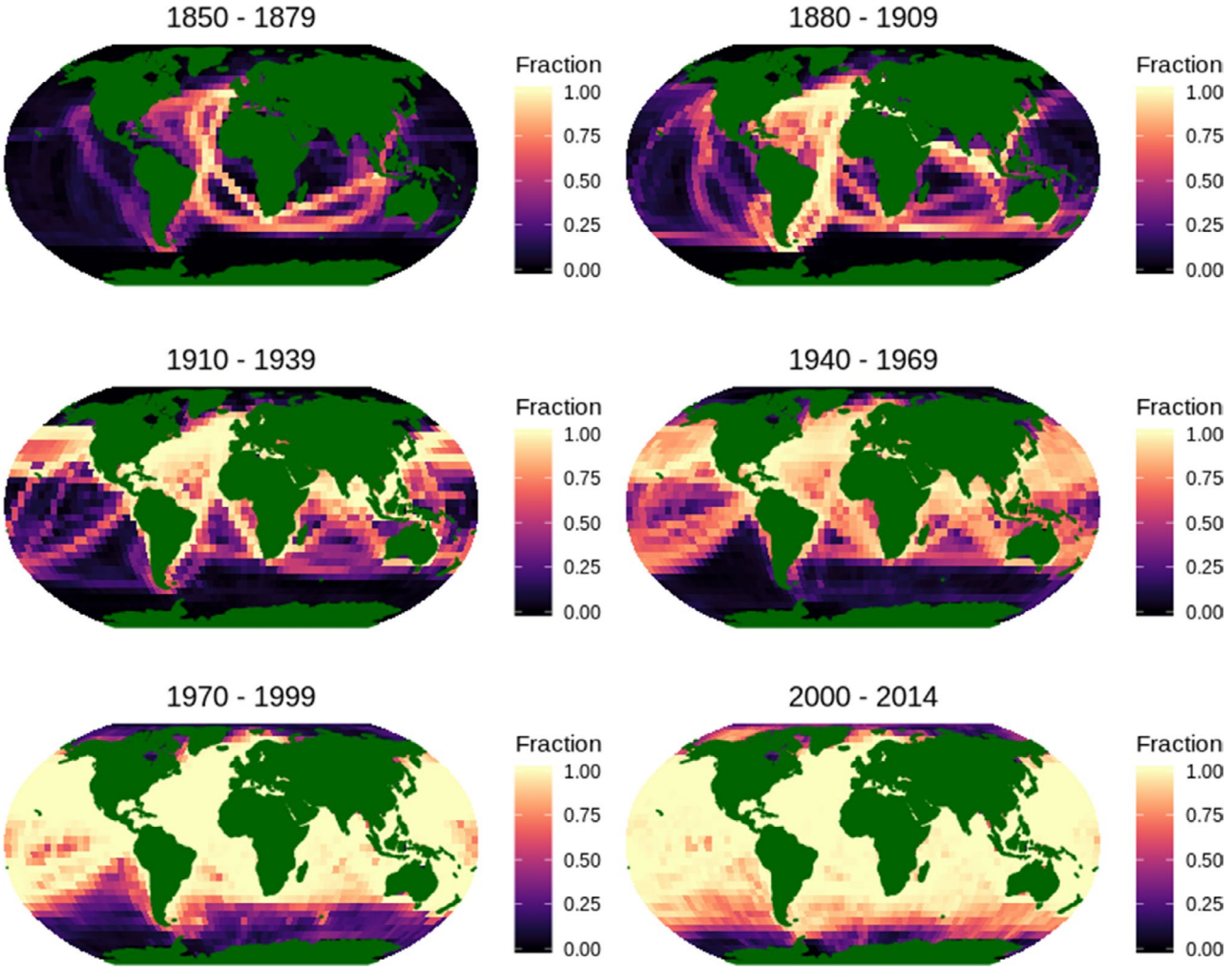

FIGURE 6 Fraction of months containing more than 5 marine weather reports per 5° grid cell for the indicated 30-year periods

tens of observations per month in the mid 1800s, increasing to the 100,000s in the 1970 and 1980s. The peak around 2000 is due to several factors, including the inclusion of coastal stations with higher reporting frequency and a move to automatic weather stations (AWS) on board ships. The second world war is clearly visible with a notable drop in the number of observations. There are few humidity observations early in the record but from ~1910 onwards the number of observations is only slightly lower than the other ECVs. When plotted as the number of 5 × 5 grid cells with more than 5 observations, there is a similar increase in coverage in time up to the 1960s–1970s for all ECVs shown. From the 1970s onwards, the areal coverage for air temperature, humidity and wind speed plateaus and then decreases over the last ~20 years. This is consistent with the decline in the in situ marine component of the Global Climate Observing System previously reported (e.g., Berry and Kent, 2017, Kent *et al*., 2006, Woodruff *et al*., 2011) and due to an increasing frequency of observations from AWS but concentrated over a smaller region of the ocean. The areal sampling for sea surface temperature and sea level pressure starts to increase again from the 1990s onwards, with the increase more notable for sea surface temperature. This increase is primarily driven by the expansion of the drifting buoy network and can also be seen in the total number of reports plot.

2.5 | Data rescue priorities

Many more land observations are needed for all ECVs at all timescales prior to 1900. Sub-daily observations were made in many regions of the world even as far back as the late 18th Century (Brönnimann *et al*., 2019) giving the potential to extend the historical record. Africa is a particular challenge but considerable progress has recently been made and a Copernicus Climate Change Service contract is now underway to rescue from unstable microfiche and film a wealth of sub-Sahara data that otherwise will be lost forever as the original imaging led to the disposal of many of the original paper records. Extension of the marine record is needed to fill data gaps outside the main shipping lanes, especially

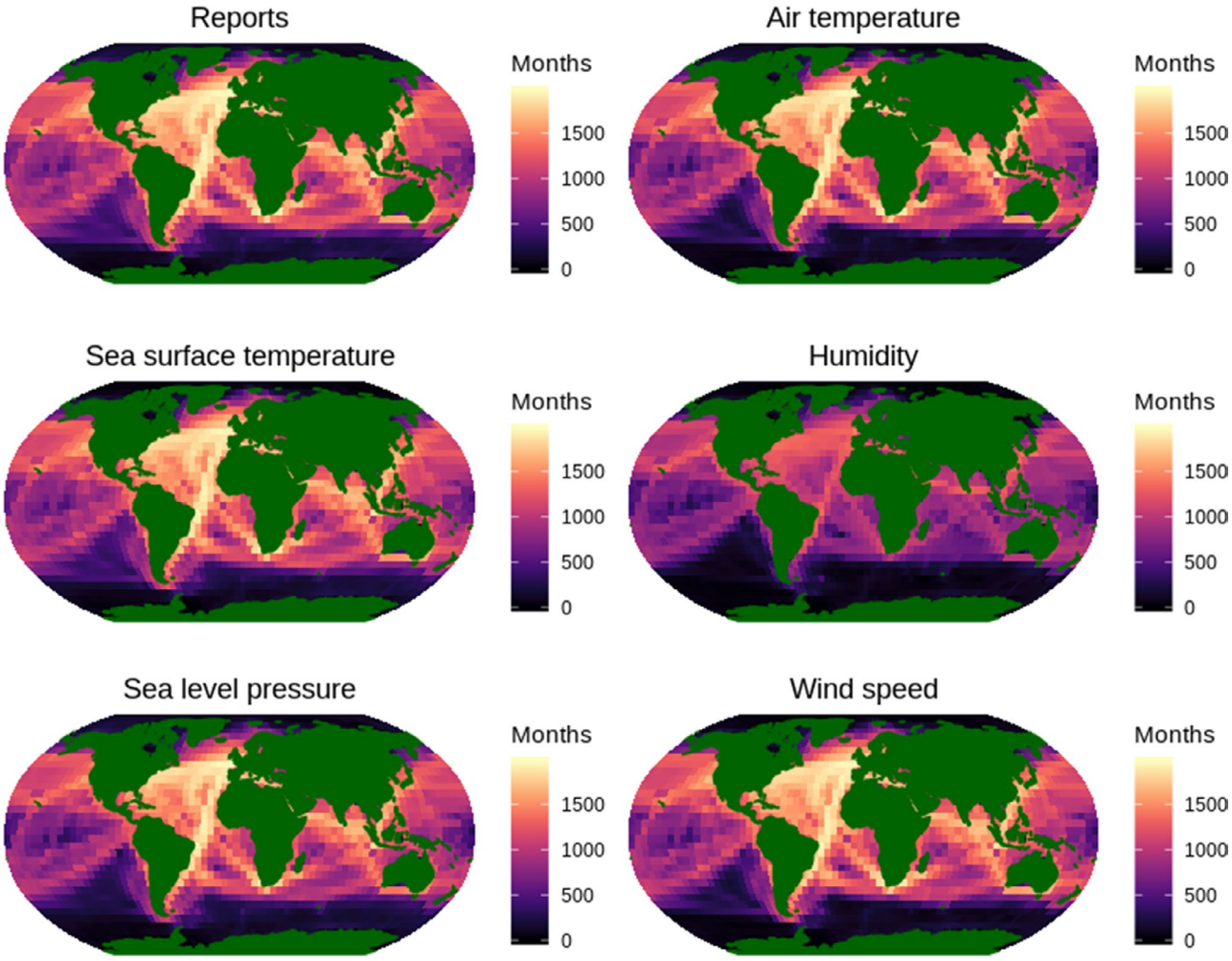

FIGURE 7 Number of months sampled between 1850 and 2014 (max 1980) with at least one marine data report/observation per 5° grid cell of: (a) any ECV; (b) air temperature; (c) sea surface temperature; (d) humidity (dew point temperature, wet bulb temperature or relative humidity); (e) sea level pressure; and (f) wind speed

prior to 1950. Regions that are particularly sparse include the polar oceans, most of the southern hemisphere and the central Pacific. Additional historical data, across all ECVs, are critical to provide long, high quality, reanalysis products (Slivinski *et al*., 2019) and climate data records (Thorne *et al*., 2018; Kent *et al*., 2019b).

Meeting these needs will require large-scale digitization efforts, which are supported by the C3S Data Rescue Service, and recognize the immense contributions already made by groups such as Atmospheric Circulations Reconstructions over the Earth (ACRE; met-acre.net), the WMO Expert Team Data Rescue (http://www.wmo.int/pages/prog/wcp/ccl/opace/opace1/ET-DARE-1-2.php) and MEDARE/I-DARE data rescue initiatives (https://www.idare-portal.org/data/medare). In addition, there have been several very successful citizen science initiatives around data rescue such as oldweather.org and weatherrescue.org. The latter resulted in crowd-sourcing efforts of 16,000 volunteers transcribing 65,000 sheets of historical rainfall data from 1860 to 1960 via Zooniverse during the spring of 2020. There is also a potential to integrate data rescue into formal educational settings (Ryan *et al*., 2018). Given the volumes of data involved, it is likely that a range of approaches will need to be pursued on a sustained basis.

3 | DATA UPLOAD SERVER

There are several methods for data to enter the Service: downloaded from known data repositories; received through personal contact; requested from NMHSs; 'pulled' near-real time from the WMO Information system WIS (https://www.wmo.int/pages/prog/www/WIS/overview.html) or periodically from NOAA/NCEI data streams; and directly contributed from any data owner.

To promote and encourage data provider contributions, we have instigated the Data Upload Server (https://datadeposit.climate.copernicus.eu/home/) where data providers can upload and share their data. This Service is hosted on the JASMIN computer platform in the UK (www.jasmin.

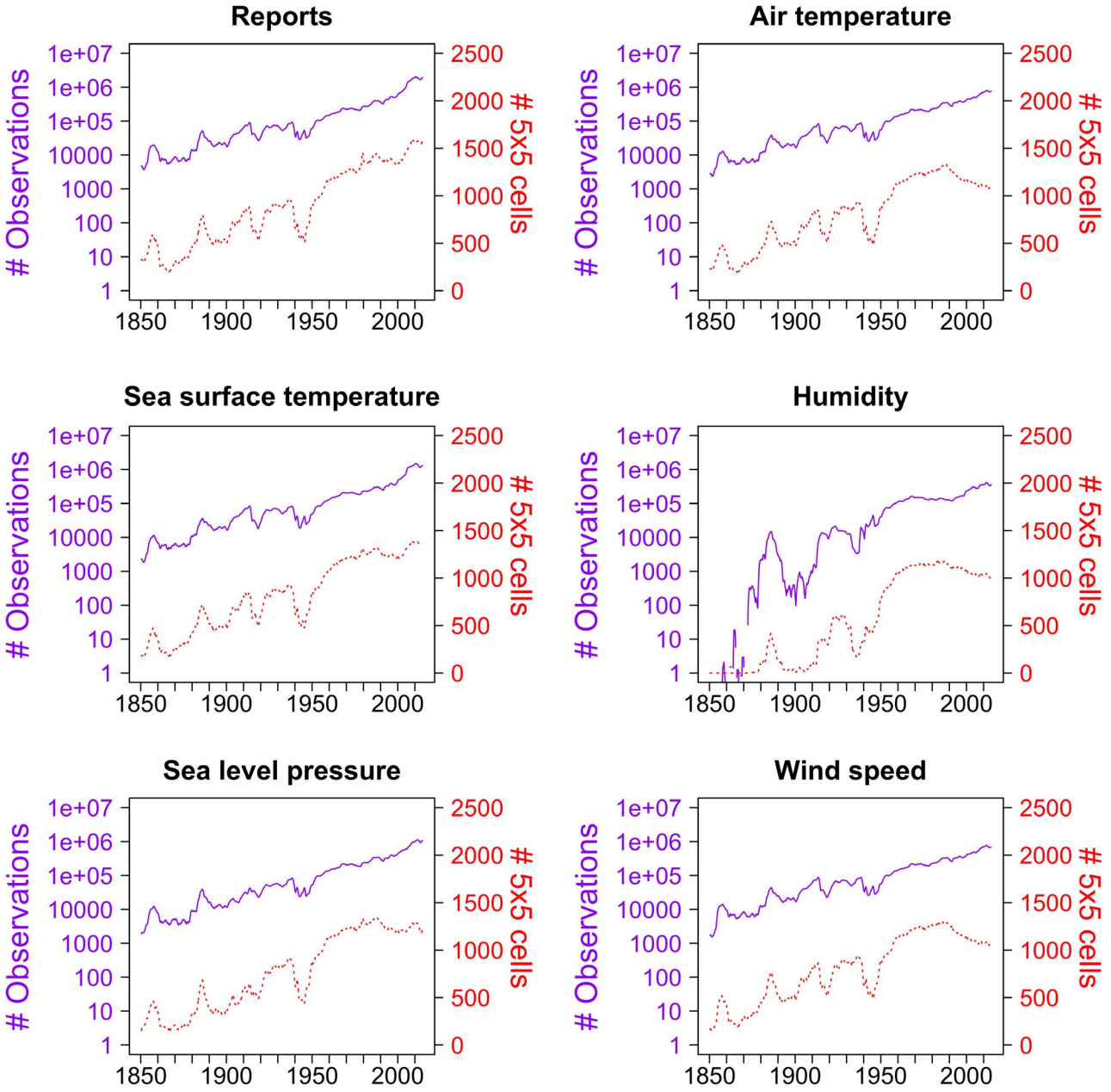

FIGURE 8 Number of marine data reports/observations per month (purple lines/left axis) and number of 5° grid cells with more than five observations (red dotted lines/right axis) of: (a) any weather report; (b) air temperature; (c) sea surface temperature; (d) humidity (dew point temperature, wet bulb temperature or relative humidity); (e) sea level pressure; and (f) wind speed. For clarity, a 12-month running mean filter has been applied to the lines

ac.uk) managed by the Science and Technology Facilities Council (STFC) based at the Centre for Environmental Data Analysis (CEDA; www.ceda.ac.uk). The Data Upload Server is a dedicated disc area enabling data upload by either user-initiated push or Service-initiated pull. Data can be safely checked and inventoried in this area to ensure against viruses, malware and data corruption prior to integration into the C3S database on the JASMIN servers. The Data Upload Server enables data providers to share their data and metadata with the Service via a web form, Rsync or FTP method.

Any relevant data can be provided in any format. However, for land data submissions, we have been working closely with the C3S Data Rescue service (https://data-rescue.copernicus-climate.eu/) who have developed a standard Station Exchange Format (SEF) and an R code package to read and write the SEF (documentation, code and examples at: https://github.com/C3S-Data-Rescue-Lot1-WP3/SEF/wiki. We encourage, but do not require, data providers to convert their data to the standard SEF to expedite the data integration process. Nevertheless, we still require the data in the original format in addition, to allow us to identify and problems with

the conversion to SEF. We also require all available supporting metadata and documentation including information on any quality control checks and any known historical station or instrumental changes or moves. We will endeavour to incorporate all source QC flags with our internal QC flags and provide these to the end users. A link to step by step instructions on how a data provider can register for an account and upload data can be found at the top of the following page: https://datadeposit.climate.copernicus.eu/home/.

TABLE 4 Prioritization matrix for processing of acquired land data sets

Data policy	Discoverability	Spatial coverage	Temporal coverage/ECV type	Format
Open access	Complete set of collection-level discovery metadata and minimal granular metadata	Data set contains stations located in sparse data region or data sets derived from National Meteorological/ Hydrological Service	Data sets contains large volume of stations that increase temporal coverage on existing holdings and/or increase ECV coverage	Standard Exchange Format (SEF) as read/write code has been developed.
Non-commercial	Minimal/basic metadata	Data set has global/ regional spatial coverage	Data set contains medium volume of stations that increase temporal coverage on existing holdings and/or increase ECV coverage	Standards based machine readable (e.g., CSV, ASCII JSON, XML)
Other	Limited data set information and needs some investigation	Data set has local/ national spatial coverage	Data set contains small volume of stations that increase temporal coverage on existing holdings and/or increase ECV coverage	Basic machine readable (formatted binary data)
Unknown (withheld and once known it will be treated accordingly) or Restricted (which will be withheld until policy changes)	By personal contact only; Data set information not discoverable	Data set/Stations already exist in the Service	No increase in existing temporal coverage or ECV type.	Non-machine readable (Untapped data; Obsolete media)

Note: Green shading indicates urgent priority; amber shading indicates a high/medium priority; and red indicates the lowest priority.

TABLE 5 Prioritization matrix for ingest of new marine data sources

Maturity level	Data policy	Traceability	Location	Date/time
Best quality	Open	Imaged logbook page available and traceable to archive	Location to tenths degree or greater precision	Time of observation to nearest hour or greater precision
Data likely to be of good quality	Non-commercial	Traceable to logbook in archive (national or other) or well-documented translation	Location to whole degrees	Other documented information, for example watch number
Quality and completeness of data and metadata uncertain, investigation needed	Other	Some information available	Location accuracy uncertain	Date of observation
Data with little or no accompanying documentation or metadata, uncertain provenance, quality and completeness uncertain and likely to be low	Unknown	Other/unknown	Location accuracy known to be poor	Other

Note: Green shading indicates the highest priority; amber shading medium priority; and red the lowest.

3.1 | Prioritizing uploaded data for integration into the service for land-based data sets

Once data have been uploaded via the Data Upload Server the team will evaluate a data source and prioritize integration into the Service based on the criteria guideline outlined in the priority matrix (Table 4). There are five main aspects of a data source that need to be considered when prioritizing it for inclusion in the process through to the CDS. Data usage policy is the most important aspect when prioritizing a data set. Sharing and onward data usage policy relating to land surface meteorological data is complex and highly variable depending on the owning country or entity, and in most cases is not clearly associated with the data itself, requiring case by case detailed investigation. Historical changes to network designations, historical bilateral agreements, political boundaries, data archives and policy practices further complicate the issue. This means that for most data sources, there is no immediately clear data policy information. For sources where provenance is not presently satisfactorily attained, further investigations will be conducted to try and obtain documentation. Where only emails from original data sources were available in the metadata, we have sent information requests and in many cases are awaiting a response. Out of an abundance of caution, as all the data currently obtained is freely available from public-facing repositories the default approach of the Service is to determine data to be restricted to non-commercial usage unless and until a more open policy can be ascertained.

The Service will give highest priority to an open data policy data source even if all the other aspects are in lower priority criteria. The second highest priority will be given to non-commercial data policy data sets and finally data sets with other types of data policies. When a source data policy is unclear or unknown, the team will withhold the data set from the integration process until the data policy is confirmed. In addition, if a data source is acquired by the Service and the data policy is restricted then it may be accepted but then withheld in case the data policy changes in future. Discoverability relates to the amount of detailed metadata that is made available with the data source. There is a minimum requirement for a basic set of source metadata (e.g., station locations, station names, source name and source data policy) and will be an important consideration in the prioritization process.

The Service aims to provide full traceability of the holdings served to users back to the original data source. It is the aim of the Service to eventually include all original source processing, quality control flags, homogenization checks and other metadata, which will be made available to the end user of the Service. The Service is primarily looking for data that is of proven quality, extends the coverage of data spatially, temporally or by ECV, which is well-documented and in a format that is easily readable. All unique and relevant data acquired by the Service or submitted via the Data Upload Server will be archived for posterity at the NOAA NCEI World Data Center for Meteorology, Asheville, North Carolina, USA (https://www.ncdc.noaa.gov/wdcmet).

3.2 | Prioritizing uploaded data into the service for marine-based data sets

Marine data uploaded to the Data Upload Server for integration into Service will be prioritized according to the matrix shown in Table 5. Those sources that meet all the criteria in

Format	Platform metadata	Observation metadata	Documentation	Completeness
CSV compatible with the Common Data Model (CDM)	Comprehensive platform metadata	Comprehensive metadata, including observing practices and instrumentation	Comprehensive documentation, including observing practices and instrumentation	Extensive range of ECVs available
IMMA		Instrument heights available (e.g., barometer height)	Reasonable documentation	Limited range of ECVs available
Machine readable (e.g., CSV, JSON, XML)	Ship name, callsign or other identifier available	Documentation, including units and code tables for parameters	Limited documentation	Single ECV
Unstructured/free format	Identifiers missing, ambiguous or undocumented	Metadata missing, ambiguous or undocumented	No documentation	No ECVs

green will be given highest priority, with priority decreasing through the green, amber and red categories. Temporally, new sources in the period prior to 1950 add the greatest value, especially when those new sources are high quality, well-documented and extend the range of ECVs available in otherwise data-sparse areas. Similarly, those sources that sample the full diurnal cycle will add more value than a source with one observation per day. Isolated voyages in otherwise data-sparse periods, whilst potentially of historical interest, are unlikely to be of great value on their own and may be given lower priority for integration into the Service. Open data sources add the greatest value whilst those that have usage restrictions have less value and will similarly be given a lower priority.

4 | SUMMARY

This paper has provided details on the climate data that has been acquired and inventoried up until April 2020 as part of the activities of the C3S Global Land and Marine Observations Database Service. This Service meets the need for integration and harmonization of the disparate sources of observational in situ land and marine meteorological data that have developed over many decades. For the land-based holdings, as of April 2020, 107,894 sub-daily, 173,850 daily and 186,015 monthly station series have been inventoried (still with gross duplication). Marine observations have been ingested from ICOADS and converted to a common data format whilst homogenizing platform identifiers and other metadata to improve the consistency of the marine climate record. Quality control flags have been applied, and a new approach to the identification, flagging and prioritization of duplicate reports implemented.

This paper has also highlighted spatial and temporal data gaps in our current holdings across a number of ECVs. These have driven a set of priority rankings for the conversion, ingestion and processing of acquired data sources based on data policy; the value of the data source in extending spatial or temporal coverage for each ECV; the quality of the data, metadata and documentation and the ease with which the data and metadata can be read and converted to a standard format.

The C3S Global Land and Marine Observations Database in collaboration with NOAA NCEI is a work in progress and although a huge body of work has been completed there is much more work to do. The database will provide a comprehensive and fully integrated archive of land and marine surface meteorological observations. The first data release is expected to be available to users by the end of 2020 or early 2021 via the C3S Climate Data Store (https://cds.climate.copernicus.eu/#!/home). This improves the availability of climate data by producing the first climate database to serve global climate observations consistently over land and ocean, across different timescales, and for a wide range of ECVs. There is a growing need for traceability of observations to original source and international standards, given the importance of historical climate assessments in informing policy around weather and climate change (Thorne *et al*., 2018). The Service has therefore ensured that each observation is fully traceable to the original data source and includes all available provenance as well any citation requirements. Eventually, each report will contain all the available corresponding observational metadata.

We envisage that this database will continue to develop over time, growing with periodic user uploads through the Data Upload Server including from data rescue, periodic discovery and ingest by the Service team and incorporation of near-real time streams. Enhancement of the existing database could come from discovery of supporting information relating to origin, instrumentation, observing environment and practices and improvements to processing methodology, including improvements to quality flagging, duplicate identification and data or metadata merging. This Service provides verifiable climate data to scientists, policymakers and other users and will secure the archiving and access of these data for future generations.

If you have climate data, we encourage you to share it with the Service by visiting our Data Upload Server: https://datadeposit.climate.copernicus.eu/home/

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

AUTHOR CONTRIBUTIONS

SN wrote the original draft of the paper; SN, DB and EK provided the figures; SN, DB, EK, PT, RD, MM and contributed full sections of text; all other authors reviewed and commented on the paper.

ORCID

Simon Noone https://orcid.org/0000-0003-1661-1423
Robert J. H. Dunn https://orcid.org/0000-0003-2469-5989
Eric Freeman https://orcid.org/0000-0001-9654-5109
John J. Kennedy https://orcid.org/0000-0002-6841-7289
Elizabeth C. Kent https://orcid.org/0000-0002-6209-4247
Peter W. Thorne https://orcid.org/0000-0003-0485-9798

REFERENCES

Adler, R.F., Huffman, G.J., Chang, A., Ferraro, R., Xie, P., Janowiak, J. *et al*. (2003) The version-2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979–present). *Journal of Hydrometeorology*, 4, 1147–1167.

Banimahd, S.A. & Khalili, D. (2013) Factors influencing Markov chains predictability characteristics, utilizing SPI, RDI, EDI and SPEI drought indices in different climatic zones. *Water Resources Management*, 27(11), 3911–3928.

Barker, P.A., Wilby, R.L. & Borrows, J. (2004) A 200-year precipitation index for the central English Lake District. *Hydrological Sciences Journal*, 49(5), 769–785.

Berry, D.I. and Kent, E.C. 2017 'Assessing the health of the in situ global surface marine climate observing system', *Int. J. Climatol.* 37: 2248–2259. https://doi.org/10.1002/joc.4914

Brönnimann, S., Allan, R., Ashcroft, L., Baer, S., Barriendos, M., Brázdil, R. *et al*. (2019) Unlocking pre-1850 instrumental meteorological records: a global inventory. *Bulletin of the American Meteorological Society*, 100, ES389–ES413. https://doi.org/10.1175/BAMS-D-19-0040.1

Brunet, M., Jones, P.D., Jourdain, S., Efthymiadis, D., Kerrouche, M. & Boroneant, C. (2013) Data sources for rescuing the rich heritage of Mediterranean historical surface climate data. *Geoscience Data Journal*, 1, 61–73. https://doi.org/10.1002/gdj3.4

Carella, G., Kent, E.C. & Berry, D.I. (2017) A probabilistic approach to ship voyage reconstruction in ICOADS. *International Journal of Climatology*, 37, 2233–2247. https://doi.org/10.1002/joc.4492

Chan, D., Kent, E.C., Berry, D.I. & Huybers, P. (2019) Correcting datasets leads to more homogeneous early 20th century sea surface warming. *Nature*, 571, 393–397. https://doi.org/10.1038/s41586-019-1349-2

Compo, G.P., Whitaker, J.S., Sardeshmukh, P.D., Matsui, N., Allan, R.J., Yin, X. *et al*. (2011) The Twentieth Century Reanalysis Project. *Quarterly Journal of the Royal Meteorological Society*, 137, 1–28. https://doi.org/10.1002/qj.776

Cram, T.A., Compo, G.P., Yin, X., Allan, R.J., McColl, C., Vose, R.S. *et al*. (2015) The International Surface Pressure Databank version 2. *Geoscience Data Journal*, 2, 31–46. https://doi.org/10.1002/gdj3.25

Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S. *et al*. (2011). The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137(656), 553–597. https://doi.org/10.1002/qj.82

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S.C., Collins, W.*et al*. (2013) Evaluation of climate models. In: Stocker, T.F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V. & Midgley, P.M. (Eds.) *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Freeman, E., Woodruff, S.D., Worley, S.J., Lubker, S.J., Kent, E.C., Angel, W.E. *et al*. (2017) ICOADS release 3.0: A major update to the historical marine climate record. *International Journal of Climatology*, 37, 2211–2232. https://doi.org/10.1002/joc.4775

GCOS, 2018 Global Climate Observations System, webpage available at: https://gcos.wmo.int/en/home, [accessed 11/11/2020].

Gu, Z. (2014) Circlize implements and enhances circular visualization in R. *Bioinformatics*, 30, 2811–2812.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J. *et al*. (2020) The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*. 146, 1999–2049. https://doi.org/10.1002/qj.3803

ICOADS (2016) *ICOADS Release 3.0 Quality Control (QC) and Related Processing* (DRAFT, 14 July 2016). Available at: http://icoads.noaa.gov/e-doc/R3.0-stat_trim.pdf [Accessed 5 July 2017]

Jin, X., Qiang, H., Zhao, L., Jiang, S., Cui, N., Cao, Y. *et al*. (2020) SPEI-based analysis of spatio-temporal variation characteristics for annual and seasonal drought in the Zoige Wetland, Southwest China from 1961 to 2016. *Theoretical and Applied Climatology*, 139(1), 711–725.

Kennedy, J.J., Rayner, N.A., Atkinson, C.P. & Killick, R.E. (2019) An ensemble data set of sea-surface temperature change from 1850: the Met Office Hadley Centre HadSST.4.0.0.0 data set. *Journal of Geophysical Research: Atmospheres*, 124, 7719–7763https://doi.org/10.1029/2018JD029867

Kennedy, J.J., Thorne, P.W., Peterson, T.C., Ruedy, R., Stott, P.A., Parker, D.E. *et al*. (2010) How do we know the world has warmed? *Bulletin of the American Meteorological Society*, 91(7), S26–S27.

Kent, E.C., Berry, D.I., Pérez González, I. & Cornes, R. (2019a) *Copernicus Climate Change Service Global Land and Marine Observations Database, Documentation for marine duplicate identification and linking of platform identifiers*. C3S_D311a_Lot2_NUIM/SC2_DUP_NMAT.

Kent, E.C., Kennedy, J.J., Smith, T.M., Hirahara, S., Huang, B., Kaplan, A. *et al*. (2017) A call for new approaches to quantifying biases in observations of sea-surface temperature. *BAMS*, 98, 1601–1616. https://doi.org/10.1175/BAMS-D-15-00251.1

Kent, E.C., Rayner, N.A., Berry, D.I., Eastman, R., Grigorieva, V., Huang, B. *et al*. (2019b) Observing requirements for long-term climate records at the ocean surface. *Frontiers in Marine Science*, 6, 441. https://doi.org/10.3389/fmars.2019.00441

Klein Tank, A.M., Wijngaard, J.B., Können, G.P., Böhm, R., Demarée, G., Gocheva, A. *et al*. (2002) Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. *International Journal of Climatology*, 22, 1441–1453. https://doi.org/10.1002/joc.773

Klok, E.J. & Klein-Tank, A.M. (2009) Updated and extended European dataset of daily climate observations. *International Journal of Climatology*, 29, 1182–1191. https://doi.org/10.1002/joc.1779

Mahmood, R., Boyles, R., Brinson, K., Fiebrich, C., Foster, S., Hubbard, K. *et al*. (2017) Mesonets: Mesoscale weather and climate observations for the United States. *Bulletin of the American Meteorological Society*, 98, 1349–1361. https://doi.org/10.1175/BAMS-D-15-00258.1

Menne, M.J., Durre, I., Vose, R.S., Gleason, B.E. & Houston, T.G. (2012) An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology*, 29, 897–910. https://doi.org/10.1175/JTECH-D-11-00103.1

Murphy, C., Broderick, C., Burt, T.P., Curley, M., Duffy, C., Hall, J. *et al*. (2017) A 305-year continuous monthly rainfall series for the Island of Ireland (1711–2016). *Climate of the Past Discussions*, 14, 413–440. https://doi.org/10.5194/cp-14-413-2018

Noone, S., Broderick, C., Duffy, C., Matthews, T., Wilby, R. & Murphy, C. (2017) A 250-year drought catalogue for the island of Ireland (1765–2015). *International Journal of Climatology*, 37, 239–254. https://doi.org/10.1002/joc.4999

Parker, W.S. (2016) Reanalyses and observations: What's the difference? *Bulletin of the American Meteorological Society*, 97, 1565–1572. https://doi.org/10.1175/BAMS-D-14-00226.1

Prohom, M., Barriendos, M. & Sanchez-Lorenzo, A. (2015) Reconstruction and homogenization of the longest instrumental precipitation series in the Iberian Peninsula (Barcelona, 1786-2014).

International Journal of Climatology, 36(8), 3072–3087. https://doi.org/10.1002/joc.4537

R Core Team (2019) *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/

Rayner, N. A., Brohan, P., Parker, D. E., Folland, C. K., Kennedy, J. J., Vanicek, M., Ansell, T. J. and Tett, S. F. B., (2006), 'Improved analyses of changes and uncertainties in sea surface temperature measured in situ since the mid-nineteenth century: The HadSST2 dataset', *Journal of Climate* 19(3), 446–469 https://doi.org/10.1175/JCLI3637.1

Rennie, J.J., Lawrimore, J.H., Gleason, B.E., Thorne, P.W., Morice, C.P., Menne, M.J. *et al*. (2014) The international surface temperature initiative global land surface databank: monthly temperature data release description and methods. *Geoscience Data Journal*, 1, 75–102. https://doi.org/10.1002/gdj3.8

Ryan, C., Duffy, C., Broderick, C., Thorne, P.W., Curley, M., Walsh, S. *et al*. (2018) 527 Integrating data rescue into the classroom. *Bulletin of the American Meteorological Society*. 1757–1764. https://doi.org/10.1175/BAMS-528D-17-0147.1

Saeidipour, M., Radmanesh, F., Eslamian, S. & Sharifi, M.R. (2019) Regionalization analysis of SPI and SPEI drought indices for Karoon basin. *JWSS-Isfahan University of Technology*, 23(2), 397–415.

Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Ziese, M. & Rudolf, B. (2013) GPCC's new land surface precipitation climatology based on quality-controlled in situ data and its role in quantifying the global water cycle. *Theoretical and Applied Climatology*, 115, 15–40. https://doi.org/10.1007/s00704-013-0860-x

Shapiro, M., Shukla, J., Brunet, G., Nobre, C., Béland, M., Dole, R. *et al*. (2010) An earth-system prediction initiative for the twenty-first century. *Bulletin of the American Meteorological Society*, 91, 1377–1388. https://doi.org/10.1175/2010BAMS2944.1

Slivinski, L.C., Compo, G.P., Whitaker, J.S., Sardeshmukh, P.D., Giese, B.S., McColl, C. *et al*. (2019) Towards a more reliable historical reanalysis: Improvements for version 3 of the Twentieth Century Reanalysis system. *Quarterly Journal Royal Meteorological Society*, 145, 2876–2908. https://doi.org/10.1002/qj.3598

Slutz, R.J., Lubker, S.J., Hiscox, J.D., Woodruff, S.D., Jenne, R.L., Joseph, D.H. *et al*. (1985) *Comprehensive Ocean-Atmosphere Data Set; Release 1*. NOAA Environmental Research Laboratories, Climate Research Program, Boulder, CO, 268 pp. (NTIS PB86-105723).

Smith, S.R., Alory, G., Andersson, A., Asher, W., Baker, A., Berry, D.I. *et al*. (2019) Ship-based contributions to global ocean, weather, and climate observing systems. *Frontiers in Marine Science*, 6, 434. https://doi.org/10.3389/fmars.2019.00434

Smith, S.R., Freeman, E., Lubker, S.J., Woodruff, S.D., Worley, S.J., Angel, W.E. *et al*. (2016) *The International Maritime Meteorological Archive (IMMA) format*, 3. Available at: http://icoads.noaa.gov/e-doc/imma

Thorne, P.W., Allan, R., Ashcroft, L., Brohan, P., Dunn, R., Menne, M. *et al*. (2018) Towards an integrated set of surface meteorological observations for climate science and applications. *Bulletin of the American Meteorological Society*, 98, 2689–2702. https://doi.org/10.1175/BAMS-D-16-0165.1

Van Den Besselaar, E.J., Klein-Tank, A.M., Van Der Schrier, G., Abass, M.S., Baddour, O., Van Engelen, A.F. *et al*. (2015) International climate assessment & dataset: climate services across borders. *Bulletin of the American Meteorological Society*, 96, 16–21. https://doi.org/10.1175/BAMS-D-13-00249.1

Wilby, R.L. (2016) When and where might climate change be detectable in UK river flows? *Geophysical Research Letters*, 33(19), https://doi.org/10.1029/2006GL027552

Woodruff, S.D., Worley, S.J., Lubker, S.J., Ji, Z., Freeman, J.E., Berry, D.I. *et al*. (2011) ICOADS release 2.5 and data characteristics. *International Journal of Climatology*, 31(7), 951–967. https://doi.org/10.1002/joc.2103

How to cite this article: Noone S, Atkinson C, Berry DI, et al. Progress towards a holistic land and marine surface meteorological database and a call for additional contributions. *Geosci Data J*. 2020;00: 1–18. https://doi.org/10.1002/gdj3.109