ECMWF COPERNICUS REPORT

Copernicus Climate Change Service

Best Practice Guidelines for Climate Data and Metadata Formatting, Quality Control and Submission

C3S_311a_Lot1_Met Office – Collection and Processing of In Situ Observations - Data Rescue

Issued by: University of Bern / Brönnimann

Date: 09/12/2020

Ref: C3S_DC3S311a_Lot1.3.4.2_2020_BestPracticeGuidelines_Part2_v6.docx

Official reference number service contract: 2020/C3S_311a_Lot1_Met Office/SC3

To cite this report:
Brunet, Manola; Brugnara, Yuri; Noone, Simon; Stephens, Ag; Valente, Maria Antónia; Ventura, Clara; Jones, Phil; Gilabert, Alba; Brönnimann, Stefan; Luterbacher, Juerg; Allan, Rob; Brohan, Philip; Compo, Gilbert P., 2020: Best Practice Guidelines for Climate Data and Metadata Formatting, Quality Control and Submission of the Copernicus Climate Change Service Data Rescue Service [Ref: C3S_DC3S311a_Lot1.3.4.2_2020_v4 - contract: 2020/C3S_311a_Lot1_Met Office/SC3].
DOI: 10.24381/kctk-8j22, https://doi.org/10.24381/kctk-8j22

This document has been produced in the context of the Copernicus Climate Change Service (C3S).
The activities leading to these results have been contracted by the European Centre for Medium-Range Weather Forecasts, operator of C3S on behalf of the European Union (Delegation Agreement signed on 11/11/2014). All information in this document is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose.
The user thereof uses the information at its sole risk and liability. For the avoidance of all doubts, the European Commission and the European Centre for Medium-Range Weather Forecasts has no liability in respect of this document, which is merely representing the authors view.

Contributors

UNIVERSITY ROVIRA I VIRGILI

1. Manola Brunet

UBERN

1. Stefan Brönnimann
2. Yuri Brugnara

NUIM

1. Simon Noone

STFC

1. Ag Stephens

FCIÊNCIAS.ID

1. Maria Antónia Valente
2. Clara Ventura

UEA

1. Phil Jones

FURV

1. Alba Gilabert

JLU

1. Jürg Luterbacher

MET OFFICE

1. Rob Allan
2. Philip Brohan

UNIVERSITY OF COLORADO

1. Gilbert P. Compo

Table of Contents

1. Introduction

1.1 Scope

Despite our atmosphere being routinely and regularly monitored since the inception of the meteorological instrumental era, the available archive of historical and current instrumental data is still limited. This hampers our capacity to understand fully and respond properly to hazards arising from climate variability or from anthropogenic influences on our climate (Brunet and Jones, 2011). Many climate products and services require enhanced climate data availability and accessibility at either global or finer scales. Historical global climate reanalysis requires more extensive input data, particularly in regions with sparse data, and more early instrumental observations. Regional reanalyses of higher spatial resolution need denser climate networks to be used as input (e.g., Compo et al., 2011; Slivinski et al., 2019). Climate change attribution studies, looking especially but not only at extreme event attribution, also need longer observational time-series at hourly or higher temporal resolution to assess the potentially unprecedented character of any event or underlying climate trend. Moreover, improved knowledge and characterization of weather and climate extremes and their potentially harmful socio-economic impacts will benefit from an enhanced archive of climatic observations. “Climate data” here refers to long observation series at the highest possible temporal resolution.

There is growing awareness among scientific and operational (e.g., National Meteorological and Hydrological Services, NMHS) organizations of the need to recover both past and present meteorological and climate-relevant observations. This has recently engendered intensified international efforts to conduct and coordinate data rescue (DARE) activities. For example, the community efforts coordinated by the international Atmospheric Circulation Reconstructions over the Earth initiative (ACRE, Allan et al., 2011a; 2011b; 2016) (http://www.met-acre.net/), and the World Meteorological Organization (WMO)/Global Framework for Climate Services (GFCS) I-DARE portal (https://www.idare-portal.org/) both aim at coordinating and accelerating DARE efforts worldwide. In addition, current activities are being undertaken by the Copernicus Climate Change Service (C3S) Data Rescue Service (https://datarescue.climate.copernicus.eu/) to provide DARE practitioners with enhanced guidance, tools, software, and best practice guidelines to ensure more efficient planning, development and co-ordination of their projects.

These Best Practice Guidelines for Climate Data and Metadata Formatting, Quality Control and Submission (henceforward BPG2) build upon, follow up and complement the first Best Practice Guidelines for Climate Data Rescue (BPG1) document coordinated by Wilkinson et al. (2019). While the latter was focused on facilitating more efficient planning of DARE projects by guiding the archive work, the scanning of data sheets, and the digitization of the data from the imaged sheets, BPG2 is intended to facilitate the work on metadata inventorying, formatting observations and metadata, and to ensure that both are quality controlled and submitted through the Global Land and Marine

Observations Database (GLAMOD) and the C3S Historic In Situ Upper Air Database to other global data centres.

In summary, BPG2 aims to facilitate and guide DARE practitioners undertaking small or large projects, to carry out all the remaining components integrating the value chain of DARE; namely, metadata inventorying, formatting metadata and observations and ensuring their quality control, submission, and consolidation.

1.2 Aims and objectives: Setting the scene

Planning and executing any DARE project in the most efficient way, be it large or small, requires DARE's practitioners to follow a set of procedures. These include discovering undigitized data sources, imaging them for duplication and preservation of the original data sources, keying the observations and associated metadata they contain, and then formatting, Quality Controlling (QC), and finally submitting the original and QC'ed observations and metadata. Fig. 1 shows a simplified scheme of the main processes to be followed by any DARE project, in order to facilitate its successful planning and realization.

The first two components of the DARE workflow (Fig. 1), addressed in BPG1, include the description and type of documents and data sources that contain relevant climate data and metadata, along with their different data holders. BPG1 provided guidance for more efficient searching, locating and inventorying of the data sources identified, either in physical or imaged archives, and focused on the imaging and digitization processes, either for marine or terrestrial data sources. All this was accompanied with full technical details on these steps. The last two main components of any DARE exercise, the observations and metadata formatting and their QC, along with data submission and consolidation, are addressed in BPG2. Here, the focus is specifically on the recovery of station-based observations and their derived historical time-series, except for the climate metadata software tool stlocationqc", which can also be applied to land surface, marine and upper-air data in fixed platforms datasets. In short, while the C3S BPG1 was mainly focused on guiding the initial steps for a plan to set up a DARE project, namely imaging and digitizing the data and metadata contained in the relevant data sources identified, BPG2 continues the DARE workflow to address the formatting, quality control, submission, and consolidation of the digitized observations and metadata.

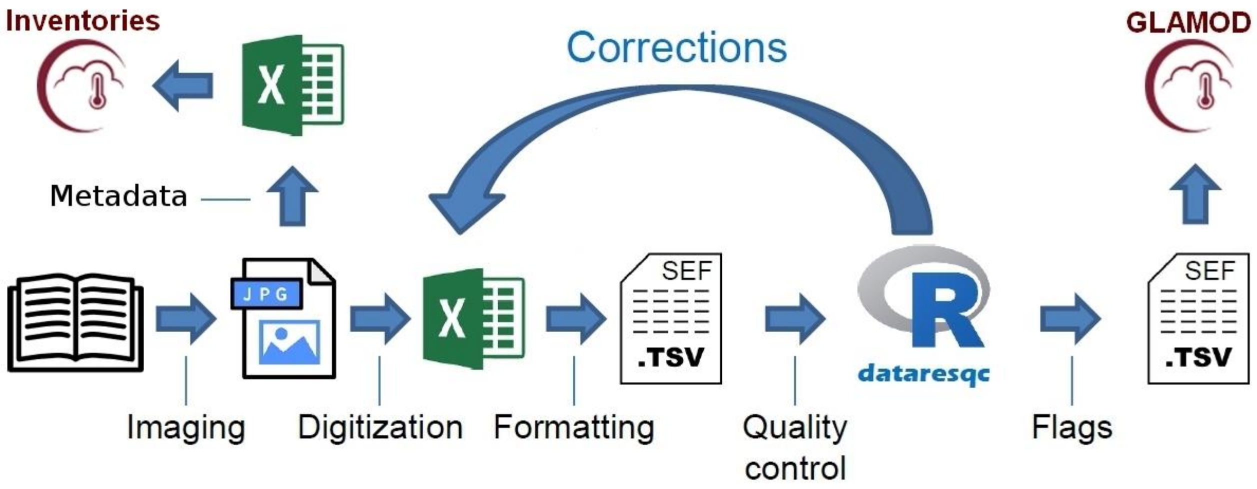


Fig. 1. Workflow scheme illustrating all the steps to be taken for an efficient and traceable climate data and metadata transference from unusable formats to machine-readable data and metadata.

Once meteorological observations have been recovered and digitized, and before they can be used confidently in any climate assessment, application, product, or service delivery, there is a need to make sure the climate data series are in a usable format and each and every one of the observations that are contained in the derived time-series are true meteorological observations. Therefore, it is essential to subject any observation and derived climate time-series to a QC procedure and to give the data series a recognizable and easily used format. In addition, the climate time-series derived from the temporal collection of meteorological observations for any climatic variable measured at any observing site must be accompanied by their corresponding metadata. Metadata must inform about when, where, how, and by whom each observation has been taken or collected. Station metadata, also called station histories, are as important as the observational data and this information must also be recovered, analyzed and made accessible to guide the QC and homogenization exercises to be undertaken. Complete metadata will greatly facilitate the application of homogenization tests, since the metadata will confirm or not the veracity of any break points detected by the statistical test applied and will help to validate/reject them in a more confident way. BPG2 does not address the homogenization issue, but points to further insights that can be gained by consulting the recent WMO guidance on homogenization (Venema et al., 2020). Reliable metadata are necessary to guarantee that the end user has all the information about the circumstances in which data has been recorded, compiled and transferred (Aguilar et al., 2003).

A climate data QC exercise consists of applying statistical tests (e.g., to check for gross and coding errors, internal consistency, temporal and spatial consistency, physical, and climatological limits or tolerance tests) to detect potential errors in the data series that may have been introduced by mistakes in any of the steps. The goal here is to detect, validate or reject any suspicious values or sequences of values. These are flagged by the tests applied. It is now standard, either at the station level or data centres processing data, that real-time or near real-time QC is carried out, ensuring that current data are reasonably free of mistakes and all values are true observations. The QC procedure generates confidence in these data. This does not ensure, though, that historical data series, as well as the data recovered, are of high enough quality to be used confidently in any climate study. Numerous studies have indicated that additional QC is required (WMO/CIMO, 2014).

The objective of BPG2 then, is to facilitate the C3S Data Rescue Service's application tools and guide the remaining processes necessary in any integrated DARE effort. In this regard, this Service has contributed to the veracity of the remaining DARE work by defining and developing several strategies for climate data and metadata formatting. This includes their QC checks and facilitating their submission and consolidation into global data repositories. Therefore, BPG2 has as its main objectives:

i. To guide the usage of the C3S Data Rescue Service metadata inventory for land surface observations, including the tools for submitting metadata,
ii. To describe and apply the metadata QC tools for station locations,
iii. To provide insights on the application of the Station Exchange Format (SEF),
iv. To give an overview on the QC tests and software,
v. To facilitate the submission of climate data to GLAMOD

All the above will be illustrated through an example in Sect. 3, using the Zurich (Switzerland) pressure record taken by J. J. Scheuchzer during the period 1718-1730.

In addition to this introduction, BPG2 is structured as follows: the first section provides insights for generating metadata and their QC, followed by a section on climate data formatting, QC, and data submission and consolidation. The third section provides a step by step example of formatting and quality control, using the Zurich (Switzerland) air pressure record.

2. General Guidelines

2.1. Generating Metadata and their Quality Control (QC)

2.1.1. What are climate metadata?

Metadata are information about data and provide knowledge about data, e.g., how, where, when and by whom information was recorded, gathered, transmitted and managed (Aguilar et al., 2003). This should include station or platform identifiers, geographical location, data owner or manager, and a description of the site and its surrounding area with its local topography and encompassing land use and coverage of land stations. Further, it should include instrumental details and exposures, observational schedules and practices, the meteorological variables measured, the observing times for each station, start and end dates of observations, maintenance procedures and results, any correction, conversion or adjustment applied to the measurements, and information on quality control (QC) and homogenization results.

Ideally, the metadata should include a complete history of the station or observing platform, including the dates and details of all changes undergone during its lifetime, inspections, any interruption to operation, and the possible eventual closure, with all this information managed by a computerized database that enables updating and use (WMO/CIMO, 2014). In short, the elements of a metadata database following the WIGOS[1] metadata standards should include specific information on the observed variable, purpose of observations, station/platform, environment, instruments and methods of observation, sampling, data processing and reporting, data quality, ownership, data policy and contact (WMO/WIGOS, 2017).

Metadata are useful information that are needed to properly guide data usage, and its management and stewardship, as many different climate assessments, applications, products and services require knowledge of the conditions under which observations were taken. Most of the NMHS central databanks run near-real-time QC exercises on observations before they are archived. However, this does not guarantee the observations taken in the past (historical series) have been subjected to QC procedures. Therefore, it is vital to make sure that historical climate time-series are subjected to QC examination prior to being used in further analyses. In addition, there is the need to subject these data to a homogeneity test and to a homogenization exercise, if required. This exercise will greatly benefit from reliable and complete metadata, which will assist with the validation of the detected breakpoints in any homogeneity examination. Complete metadata are of great value for both data users and data providers, particularly when dealing with historical data. Both QC and homogenization are key exercises to ensure climate time series are only composed of true observations and represent only variations and trends forced by weather and climate factors.

[1] WMO Integrated Global Observing System

2.1.2. Metadata minimum requirements and best practices

It is desirable to obtain a fully complete metadata database, but this is difficult to achieve, since there are many different types of potentially relevant information about the data. Many factors influence the measurements, and some of these may not even be known at this time. Thus, in the absence of all possible sources of metadata or the impossibility of recovering them all in data rescue exercises, there are minimum requirements that should be gathered and recovered and standards that are considered as metadata best practices.

Following the WMO guidelines on metadata (Aguilar et al., 2003), minimum requirements should be considered such as: information on station/platform identifiers, the geographical and locational data (e.g., geographical coordinates, elevation) and data processing (e.g., elemental units, special codes, calculations, algorithms). While best practices require more complete metadata such as:

- Stations/platform identifiers (e.g., start/end date of station, type of station, organization responsible for data),
- Geographical data (e.g., topographical information, location),
- Local environment (e.g., local land use/land cover, obstacles, soil type),
- Instrumentation maintenance (e.g., type of instruments, instrument sheltering and mounting, instrument calibration results and inspection results),
- Observing practices (e.g., observing times and schedules, corrections),
- Data processing (e.g., methods and algorithms),
- Historical events and data transmission (e.g., formats, reporting period).

A much more detailed and widely used metadata list is provided in the WIGOS guide on metadata standards (WMO/WIGOS, 2017), to which these metadata guidelines have followed to some extent (see Table 2.1 in Sect. 2.1.5).

2.1.3. Metadata consolidation

Data rescuers can benefit from having a centralized single metadata archive. Many projects define their priorities for data rescue without having full knowledge of what data have been already discovered by other groups. There is sometimes duplication of effort, and common data sets may end up being digitized twice. If the metadata that is collected in these data rescue efforts can be conveyed to a single archive, there will be the opportunity to consolidate the metadata.

One of the objectives of the C3S Data Rescue Service is to consolidate metadata of past, current and planned data rescue projects, through the Metadata Inventories (https://datarescue.climate.copernicus.eu/inventories). These inventories provide a tool to upload metadata, using a pre-determined format (in downloadable inventory templates) that is described in Sect. 2.1.5.1 of this document.

The Metadata Inventories will allow users (1) to discover what metadata exist for duplicated records and if they have discrepancies, (2) to verify if a dataset has already been rescued, and (3) to look for regions with sparse data observations. In short, the Metadata Inventories contribute to improving the quality of data rescue efforts by making them more efficient (e.g., avoiding duplications or discovering that what was apparently a duplication was in fact a new record).

Gathering the metadata in a single storage system, with a common format, will take some time, as it is necessary to encourage users to convert their metadata inventories into formatted inventories that can be uploaded to the C3S Data Rescue Service. It might require users to incorporate extra metadata that they have ignored previously but are considered essential to include in a consolidated metadata system. On the other hand, it will give users more confidence in the metadata included in a consolidated system, as they can cross check information and contact other metadata providers.

Including metadata in a consolidated and centralized system requires and encourages providers to agree to share the information with the community at large. This also gives more visibility to each metadata set, and the data rescue work performed by all groups, as well as increasing the potential of collaboration between groups.

2.1.4. Rationale for the C3S Data Rescue Service climate metadata QC

The C3S Data Rescue Service inventories adhere to metadata standards established in a previous version of the WIGOS metadata standards (WMO/WIGOS, 2015), but these guidelines are adapted to the needs of the C3S Data Rescue Service inventories and their end users. We consider here not only DARE practitioners but also the public at large.

The C3S Data Rescue Service Guidelines for inventory metadata standards and formats (Valente, 2019) details the writing and formatting of metadata inventories for land surface, upper-air and marine data worldwide. Here, we provide insights behind our approach for metadata QC, using the development of the software stlocationqc, which is described in depth in the next sub-section.

The stlocationqc software is designed for quality controlling large metadata bases that are of extreme interest not only for DARE activities worldwide, but also for global observational-based databases. The software provides an introduction, format and access to metadata, ensuring locations of land-surface stations are correct (stlocationqc can also be applied to QC metadata from upper-air fixed or moving platform data and marine observations). Station/platform locations are metadata of vital importance to ensure data provenance and avoid locational mistakes that can cause issues when interpreting the data. For small data rescue projects, stlocationqc can be used to check the coordinates (i.e., the software checks whether they are consistent with the designated country).

2.1.5. Metadata QC tools developed by the C3S Data Rescue Service

The C3S Data Rescue Service provides a Metadata Database with inventories for land surface observations, as well as upper air fixed and moving platforms plus marine observations. In this document, the focus falls on the Land Surface Metadata Inventory. For other inventories, the reader is referred to the C3S Data Rescue Service Portal (https://datarescue.climate.copernicus.eu/) or the Guidelines for Metadata Inventories Standards and Formats (https://datarescue.climate.copernicus.eu/met). Users can search the inventories, plot the stations' location files on a global Earth map and download the search results as CSV (tables) and JPEG (plot images) files. QC tools for metadata are also available, as well as metadata submission tools. These will be described in the following three subsections.

2.1.5.1. How to write metadata inventories of land surface series for the C3S Data Rescue Service

Table 1 presents the columns that form the Land Surface Observations Metadata Inventory. The Inventory has been constructed to show the metadata for one variable per row. Table 1 is taken directly from the Guidelines for Metadata Inventories Standards and Formats. The C3S Data Rescue Service supplies templates, which are blank Excel tables (in CSV format) containing just the headers (column designation), ready to be filled in by users (https://datarescue.climate.copernicus.eu/met). Any metadata from the Land Surface Inventory (e.g., for a given country or city) can be obtained by performing a parameter search at https://datarescue.climate.copernicus.eu/lso and then clicking the button "Get CSV". Looking at examples taken from this inventory will help users to fill in the templates.

As each row contains the metadata for one variable, it is advisable to start the inventory by filling the "Variable Name" column. The Variable Name is tabled (Table 2) and follows the standards agreed in the Common Data Model for the C3S Collection and Processing of In Situ Observations Services (Thorne, 2017).

Table 1. Metadata information for Land Surface Observations included in the inventories (Mandatory Elements)*

Land Surface Inventory Fields		Fill Options			Examples
Column (#60)	**Description**	**List Options**	**General / Other Options**	**Input Messages / Error Alerts**	**ERA-CLIM**
Type of Inventory (ID)*	Surface (01)	Surface (01)		Stop: Choose from the list. No other option allowed.	Surface (01)
Unique_metadata_record_ID*	Type of Inventory (ID) followed by inventory entry number (e.g., 010000160).	(Will be automatically generated in future versions)		Stop: Type 01 followed by the inventory entry number.	1000260
Project Title	Common Project Title in I-DARE database.		Free text; blank		ERA-CLIM

Project Status	State if the Project has ended, is ongoing, on hold, planned, postponed or other. For the "other" option, please specify details in the comments column.	Ended; Ongoing; Onhold; Planned; Postponed; Other	Free text; blank	Stop: Select a Project Status from the list or leave blank. If necessary, provide additional information in the "Comments" column.	Ongoing
Archive*	Institution, DARE initiative or person owning/holding the archive's documents.		Free text; NA		Insituto Dom Luiz;
Archive Link/Contact*	Link to data owner/holder website and/or e-mail contact.		Free text; hyperlink; NA		Insituto Dom Luiz/Maria Antónia Valente mavalente@fc.ul.pt
Archive Reference	Archive document identifier, if exists. Books in libraries usually are catalogued and have a reference.		Free text; blank		
Collection Name	Data collection name (e.g., ERA-CLIM2).		Free text; blank		ERA-CLIM
Document Title*	Title as indicated on the front cover or box (e.g., Lisbon Geophysical Institute Publications). If it doesn't have a title, describe briefly the document focused on its content.		Free text; NA		Lisbon Geophysical Institute Publications;
Document Type*	State if it's a manuscript, printed publication, digital database, chart, map, microfilm, microfiche or other.	Manuscript; Printed publication; Digital database; Printed publication and digital database; Microfilm; Microfiche; Chart; Map; Other; NA		Stop: Select a Document Type from the list. If necessary, select NA and provide additional information in the "Comments" column.	Printed publication; Manuscript
Document Description	Indicate (when available), separated by semicolon: the general description of document; the format; the language; any additional material included like barograms, charts, etc.		Metadata semicolon separated; blank		Meteorological records book; A4 format; Portuguese; description of various meteorological instruments, units, scales and methods used.
Observer name	Observer's name as indicated in the document.		Free text; blank		

Type of Access*	Type of access.	Public; Partially public (WMO resolution 40); Restricted; Other; NA		Stop: Select a Type of Access from the list. If necessary, select NA and provide additional information in the "Comments" column.	Public
Document Imaged*	This field indicates if and for which years the data have been imaged	Full set; Not imaged; Unnecessary - digital native format; NA	Year or interval of years of imaged data in the format YYYY or YYYY-YYYY		1906-1939
Document Keyed*	This fields indicates if, how and which years have been digitized	Full set; Full set by typing - original units; Full set by typing - units converted to SI; Full set with OCR; Digital native format and units; Digital native format - units converted to SI; Not keyed; NA	Year or interval of years of digitized data in the format YYYY or YYYY-YYYY		1906-1939
Keyed Data Processed/QC*	Type the interval of years of the digitized data that have been processed/quality controlled or select an option from the list.	Full set; Not quality controlled; NA	Year or interval of years of quality controlled data in the format YYYY or YYYY-YYYY		1906-1916
Data Homogenized	Type the interval of years of the digitized data that have been homogenized or select an option from the list.	Full set; Not homogenized; NA	Year or interval of years of homogenized data in the format YYYY or YYYY-YYYY		Not homogenized
Data Merged	Type the interval of years of the digitized data that have been merged into global databases or select an option from the list.	Full set; Not merged; NA	Year or interval of years of merged data in the format YYYY or YYYY-YYYY into Database Name		1956 supplied to ISPD

Comments on State of Data Rescue	Provide additional details on State of Data Rescue. If known, should be indicated the track of data rescue, i.e, the name and/or contact of the institution(s) that imaged, digitized, quality controlled, homogenized and merged the data		Free text; blank		Imaged, digitized and quality controlled by FFCUL - Fundação da Faculdade de Ciências da Universidade de Lisboa (http://www.fciencias-id.pt/)
Imaged Data/Metadata Link	Link for data/metadata images, if available.		Free text; hyperlink; blank		http://sign.fc.ul.pt/anais.html
Data Provider	Name of digitized data provider, which can be different from the data owner (e.g., CHUAN, IGRA).		Free text; blank		Instituto Dom Luiz
Data Provider Link/Contact	Website and/or e-mail of digitized data provider.		Free text; hyperlink; blank		mavalente@fc.ul.pt
Data Series in Published Databank Citation	Indicate Databank Publication DOI if the digitized data has been published.		Free text; hyperlink; blank		
Platform Type	Classical manual, Automatic Weather Station (AWS), synoptic network, local network, resulting from historical observations campaign or other.	Classical manual; Classical manual/Synoptic network; Automatic Weather Station (AWS); Synoptic network; Local network; Resulting from historical observations campaign; Other	Free text; blank	Stop: Select from the list or leave blank. If necessary, provide additional information in the "Comments" column.	Classical Manual/Synoptic Network
Station Name*	Station name at time of observations. (can be in English and/or in any of these original languages: Spanish, Portuguese, French and German)		Free text; NA		Beja
Current Station Name	Current station name if still active. (can be in English and/or in any of these original languages: Spanish, Portuguese, French and German)		Free text; blank		Beja
Country*	Station location's current country as stated on "List of Countries" spreadsheet.	Table - Country; NA		Stop: Select a country from the list or select NA and provide additional information in the	Portugal

				"Comments" column.	
Original Country/Region	Country or autonomous region at the time of observations (e.g., Mozambique).		Free text; blank		Portugal
State or Province	If applicable, indicate the state or province where the station is/was located or as stated on document.		Free text; blank		
City/Town/Village	Current station location at local level.		Free text; blank		Beja
Original City/Town/Village	Old name of station location at local level and at time of observations.		Free text; blank		Beja
Latitude *	Latitude from -90° to 90° with precision at least 0.001°, the format being a real number with 6 or more characters) (e.g., -65.565; 40.30373).		Decimal in the interval [-90, +90]; NA	Information: Type latitude in decimal degrees [-90, +90]. If unavailable, type NA and provide additional information in the "Comments" column.	38.017
Longitude*	Longitude from -180° to 180°, Greenwich at 0°, with precision at least 0.001°, the format being a real number with 7 or more characters) (e.g., -125.565; 60.6055).		Decimal in the interval [-180, +180]; NA	Information: Type longitude in decimal degrees [-180, +180]. If unavailable, type NA and provide additional information in the "Comments" column.	-7.883
Altitude	Altitude in meters above sea level (masl) with precision at least 0.1m (e.g., 5.0).		Decimal less than 8850; blank	Stop: Type a decimal less than 8850] meters or leave blank. If necessary, provide additional information in the "Comments" column.	284.0
Original Latitude Units	Latitude in original units.		Free text; blank		38°1' N
Original Longitude Units	Longitude in original units.		Free text; blank		7°53' W
Original Altitude Units	Altitude in original units.		Free text; blank		284 m
Local Gravity	Recorded local gravity with units at time of observations.		Free text; blank		

Original Location/Relocation	State whether the series was observed at the first station location or is a relocation.	Original location; Relocation	Free text; blank		Original location
Additional location references	Any additional location references such as the address of the place, the name of the building, descriptions of the surrounding building, landscape, relief (local environment, and other.		Free text; blank		
WMO ID	Station WMO identifier in the GCOS - current or original number (e.g., 85767).		Text with 5 to 6 characters; blank	Stop: Type the WMO ID (set of 5-6 digits) or leave blank. If necessary, provide additional information in the "Comments" column.	85620
WMO Region*	WMO region in which the station is located according to the map: https://cpdb.wmo.int/	Africa (1); Asia (2); South America (3); North, Central America and the Caribbean (4); South West Pacific (5); Europe (6); Antarctica (7); NA		Stop: Select a WMO Region from the list or select NA and provide additional information in the "Comments" column.	Europe (6)
Original_collection_record_ID	Station unique record identifier in the original collection inventory, if exists (e.g., 149 - ERACLIM2 Portugal collection).		Free text; blank		135
Network1_name	National or regional network name 1st level (e.g., Dirección Meteorológica de Chile network).		Free text; blank		Portuguese_IPMA_network
Network1_ID	Station number in Network1 (e.g., 39008).		Free text; blank		562
Network2_name	National or regional network name 2st level.		Free text; blank		ERA-CLIM_unique_record_id
Network2_ID	Station number in Network2.		Free text; blank		20
WIGOS Station Identifier	State the WIGOS Station Identifier. WIGOS identifier, if it exists, according to https://oscar.wmo.int/surface/#/).		Free text; blank		
Start Station Date	Date when station started originally the observations		Date in the format "YYYY-MM-DD"; blank	Input Message: YYYY-MM-DD	1897-01-01

End Station Date	Date when station stopped completely the observations. If the station is still active leave blank.		Date in the format "YYYY-MM-DD"; blank	Input Message: YYYY-MM-DD	2004-03-31
Variable Name*	Name of the observed variable as stated on spreadsheet "List of Variables" – Table 2.2 column "Variable Name".	Table - Variable Name; NA		Stop: Select a Variable from the list. If doesn't exist, select NA and provide the name and other details in the "Other Observations" column.	daily_maximum_air_temperature
Units*	Variable units if the original data units have been converted			Stop: If not converted, select NA and provide additional information in the "Comments" column.	C
Original Units*	Variable's original units, if they were not converted.		Free text; NA		C
Variable Instrument	Type of measuring instrument(s) used (can be more than one if changes have occurred). It can also be a visual observation, value obtained by calculation tables, estimated, computed or other.	Table - Variable Instrument	Free text; blank	Stop: Select an Instrument from the list or leave blank. If it isn't on the list, provide the name it in the "Comments" column.	Mercury Thermometer
Variable Instrument Make and Number	Make and/or number of variable's instrument, if stated.		Free text; blank		Negretti and Zambra
Corrections/Conversions	Gravity correction, pressure reduced to 0°C, conversion coefficients, other.		Free text; blank		
Sources of Inhomogeneity	Change in instruments, observing procedures, hours, calculation tables, standards, events at the station.		Free text; blank		
Start Record Date*	Start date of the variable series		Date in the format YYYY-MM-DD; NA	Input Message: YYYY-MM-DD or NA	1906-01-01
End Record Date*	End date of the variable series		Date in the format YYYY-MM-DD; NA	Input Message: YYYY-MM-DD or NA	1919-12-31
Time Resolution*	State the frequency of variable observations in the format	From X to Y times Z", where X and Y are integers, and Z is a choice between "hourly"/"daily"/"weekly"/"monthly"/"yearly"/";	Time resolution in the format: From X to Y times Z	Stop: Type the Time Resolution in the format "From X to Y times Z", select from the list or select NA and provide additional information in the	From 1 to 4 times daily

		Infrequently; Irregular; No observations; NA		"Comments" column.	
Observation Times	Actual time of regular observations in Local Time/UTC/MST- Mountain Time Zone (USA)/Other.		Times in the format HH:MM Local Time/UTC/MST /Other; blank		9:00 Local Time (-37 m GMT)
Time Reference Meridian	State the reference meridian for time of observations		Free text; blank		GMT
Time Gaps	State the time gaps for the selected variable from years to days		Time gaps in the format YYYY to YYYY, YYYY-MM to YYYY-MM or YYYY-MM-DD to YYYY-MM-DD; blank		1950-04-01 to 1950-05-31
Estimated Station Days	Number of days (integer) with observations, discounting gaps.		Positive integer; blank	Stop: Type a positive integer value or leave it blank. No other option allowed.	5113
Additional Variable Metadata	Any information considered relevant		Free text; blank		
Notes on (Severe) Weather Events	Indication of (severe) weather events (e.g., hurricanes, floods, auroras) optionally with dates and extreme values. .		Free text; blank		

Other Observations	Other observed variables, not on the "List of Variables" table. Duplicate the information for the station		Free text; blank		The original document may contain hourly values of cloud cover, wind direction and speed, visibility, present and past weather, pressure, temperature, clouds type, pressure tendency, precipitation (twice daily), maximum and minimum temperature, wet bulb and relative humidity. Sometimes these values are missing.
Comments	Any information considered relevant that does not fit into another column. Photos of stations can be uploaded to the C3S Data Rescue Projects section (https://datarescue.climate.copernicus.eu/projects) or to the I-DARE portal https://www.idare-portal.org/) and the link to these photos can be added		Free text; blank		Interior region station in an Air Base.

Certain inventory entries have been made mandatory and need to be filled in (write “NA” if there is no information available) and uploaded to the Metadata Inventory. These are marked with an asterisk (*) in Table 1, and the automated QC process (see Sect. 2.3) applied to the inventories detects whether these entries are missing or not. If any of these entries are missing, the validation process supplies a list of error/warning messages to the user.

Other types of more generic metadata inventories are being considered to be included in this Service for data collections that do not have precise indications on available variables, exact station locations and other generic information.

Table 2 presents the variable names currently being used in the C3S Data Rescue Service metadata inventories. For entries with no metadata, a user should either enter an “NA” if the field is mandatory or leave blank. The Variable Name must be filled in this detailed metadata inventory.

Table 2. List of variables and abbreviations to use in the C3S Data Rescue Service metadata inventories

Variable group	Domain	Abbreviation	Variable name	Description / Notes
aerosols		aaod	aerosol_absorption_optical_depth	Vertical column integral of spectral aerosol absorption coefficient
aerosols		acb	aerosol_column_burden	2D field of the column burden of condensed particles in the atmosphere
aerosols		adc	aerosol_dust_concentration	3-D field of concentration of dust or sand in the atmosphere
aerosols		aer	aerosol_effective_radius	3D field of mean aerosol particle size, defined as the ratio of the third and second moments of the number size distribution of aerosol particles. Requested in the troposphere (assumed height: 12 km) and as columnar average
aerosols		aec	aerosol_extinction_coefficient	3D field of spectral volumetric extinction cross-section of aerosol particles.
aerosols		ammr	aerosol_mass_mixing_ratio	3D field of the mass mixing ratio of condensed particles in the atmosphere
aerosols		aod	aerosol_optical_depth	Effective depth of the aerosol column from the viewpoint of radiation propagation
aerosols		asmf	aerosol_species_mole_fraction	3D field of the mole fraction of condensed-phase chemical species (e.g., sulphate, nitrate, ammonium, elemental carbon, organic carbon), in the atmosphere
aerosols		astcb	aerosol_species_ total_column_burden	2D field of the total column burden concentration of condensed-phase chemical species (e.g., sulphate, nitrate, ammonium, elemental carbon, organic carbon), in the atmosphere
aerosols		at	aerosol_type	Selection, out of a pre-defined set of aerosol classes, that best fits an input data set (observed or modeled). The pre-defined set of aerosol classes includes specification of the particle composition, mixing state, complex refractive index, and shape as a function of particle size. The definition of aerosol type includes specification of all the classes as well as the algorithm used to choose the best fit to the input data.
aerosols		ava	aerosol_volcanic_ash	3D field of mass mixing ratio of volcanic ash
aerosols		avat	total_column_aerosol_volcanic_ash	Field of total column mass of volcanic ash
aerosols		ac	air_conductivity	TBD

albedo		bsa	blue_ice_and_snow_albedo	TBD
albedo		bir	blue_ice_bidirectional_reflectance	TBD
albedo		cga	clean_glacier_ice_albedo	TBD
albedo		dga	dirty_glacier_ice_albedo	TBD
albedo		esa	earth_surface_albedo	Hemispherically integrated reflectance of the Earth surface in the range 0.4 - 0.7 micro-m
albedo		sbr	snow_bidirectional_reflectance	TBD
cloud	atmospheric upper air	hb	cloud_base_height	Cloud base height
cloud	atmospheric upper air	h	cloud_base_lowest_height	Height above surface of the base of the lowest cloud seen (coded 0-9)
cloud	atmospheric upper air	n	cloud_cover	Total amount of clouds
cloud	atmospheric upper air	c	cloud_genus	Genus of cloud (0 - Cirrus … 9 - Cumulus-Nimbus)
cloud	atmospheric upper air	hs	cloud_genus_base_height	Height of base of cloud whose genus is c
cloud	atmospheric upper air	ch	high_cloud_type	Type of high clouds (coded number according to WMO SYNOP standards 1-9; 0 – no high clouds)
cloud	atmospheric upper air	cl	low_cloud_type	Type of low clouds (coded number according to WMO SYNOP standards 1-9; 0 – no low clouds)
cloud	atmospheric upper air	nh	lowest_cloud_amount	Low or (if low clouds do not exist) middle cloud amount
cloud	atmospheric upper air	cm	middle_cloud_type	Type of middle clouds (coded number according to WMO SYNOP standards 1-9; 0 – no middle clouds)
composition	atmospheric		BrO	
composition	atmospheric		C10H16 (3-Carene)	
composition	atmospheric		C10H16 (alfapinene)	
composition	atmospheric		C10H16 (betapinene)	
composition	atmospheric		C10H16 (Limonene)	
composition	atmospheric		C2H2	
composition	atmospheric		C2H5OH	
composition	atmospheric		C2H6	
composition	atmospheric		C2H6S	
composition	atmospheric		C3H6O	
composition	atmospheric		C4H10 (Methylpropane)	
composition	atmospheric		C4H10 (n-butane)	
composition	atmospheric		C5H12 (2-Methylbutane)	
composition	atmospheric		C5H12 (n-Pentane)	
composition	atmospheric		C5H8	
composition	atmospheric		C6H6	
composition	atmospheric		C7H8	
composition	atmospheric		CFC-11	
composition	atmospheric		CFC-12	
composition	atmospheric		CH3CN	
composition	atmospheric		CH3OH	
composition	atmospheric		CH4	
composition	atmospheric		ClO	
composition	atmospheric		ClONO2	
composition	atmospheric		CO	
composition	atmospheric		CO2	

composition	atmospheric		COS	
composition	atmospheric		H2O	
composition	atmospheric		HCHO	
composition	atmospheric		HCHO (Total Column)	
composition	atmospheric		HCl	
composition	atmospheric		HDO	
composition	atmospheric		HNO3	
composition	atmospheric		N2O	
composition	atmospheric		N2O5	
composition	atmospheric		NO	
composition	atmospheric		NO2	
composition	atmospheric		NO2 (Total column)	
composition	atmospheric		O3	
composition	atmospheric		O3 (Total column)	
composition	atmospheric		OH	
composition	atmospheric		PAN	
composition	atmospheric		PSC occurrence	
composition	atmospheric		SF6	
composition	atmospheric		SO2	
composition	atmospheric		SO2 (Total column)	
evaporation	atmospheric surface	eee	evaporation	From evaporimeter inside Stephenson shelter (Piche)
evaporation	atmospheric surface	ev	evaporation	From tank/pan (evaporation-precipitation; if precipitation>evaporation than ev<0)
evaporation	atmospheric surface	pev	potential_evapotranspiration	Quantity of water evaporated from the soil and plants when the ground is at its natural moisture content
evaporation	atmospheric surface	rev	real_evapotranspiration	TBD
humidity	atmospheric surface	ah	absolute_humidity	Measure of water vapour (moisture) in the air, regardless of temperature
humidity	atmospheric	dep_dew	dew_point_depression	Dew point depression is also called dew point deficit. It is the amount by which the air temperature exceeds its dew point temperature.
humidity	atmospheric surface; upper air	td, t_dew	dew_point_temperature	Dew point temperature is the temperature at which a parcel of air reaches saturation with respect to liquid water upon being cooled at constant pressure and specific humidity.
humidity	atmospheric surface; upper air	ibt	ice_bulb_temperature	TBD
humidity	atmospheric surface; upper air	rh	relative_humidity	The amount of water vapour present in air expressed as a percentage of the amount needed for saturation at the same temperature.
humidity	atmospheric surface; upper air	q	specific_humidity	Specific means per unit mass. Specific humidity is the mass fraction of water vapour in (moist) air.

humidity	atmospheric surface; upper air	e	water_vapour_pressure	Partial pressure of water vapour in any gas mixture in equilibrium with solid or liquid water
humidity	atmospheric surface; upper air	tb, t_wet	wet_bulb_temperature	Lowest temperature to which air can be cooled by the evaporation of water into the air at a constant pressure
ice		ddd	ice_thickness	Thickness of the ice sheet. It is related to sea-ice elevation and ice density
precipitation	atmospheric surface	rr	accumulated_precipitation	Accumulated precipitation over a specified period
precipitation	atmospheric surface	fs	fresh_snow	TBD
precipitation	atmospheric surface	ht	hydrometeor_type	3D field of the predominant form of condensed water in a volume of free atmosphere, including liquid cloud, rain, ice crystals, snow, graupel and hail. (This variable replaces "precipitation type").
precipitation	atmospheric surface	rrls	precipitation	Precipitation (liquid or solid)
precipitation	atmospheric surface	rril	precipitation_instensity_liquid	Precipitation intensity at surface (liquid or solid)
precipitation	atmospheric surface	rris	precipitation_intensity_solid	Precipitation intensity at surface (solid)
precipitation	atmospheric surface	rrt	precipitation_type	Liquid, snow, hail, fog (coded field)
precipitation	atmospheric surface	nr	rainy_days	Number of days with rain above a certain threshold
precipitation	atmospheric surface	sc	snow_cover	Fraction of a given area which is covered by snow
precipitation	atmospheric surface	sd	snow_depth	Vertical distance from the snow surface to the underlying surface (ground, glacier ice or sea ice).
precipitation	atmospheric surface	sst	snow_status	Wet | dry
precipitation	atmospheric surface	sw	snow_water_equivalent	Surface snow amount
pressure	atmospheric surface	atb	adjunct_temperature_barometer	Temperature of the adjunct thermometer to the barometer to reduce pressure to 0ºC
pressure	atmospheric surface	p	air_pressure	Pressure of air column at specified height
pressure	atmospheric	mslp	air_pressure_at_sea_level	sea_level means mean sea level, which is close to the geoid in sea areas. Air pressure at sea level is the quantity often abbreviated as MSLP or PMSL.
pressure	atmospheric surface	ppp	pressure_tendency	Pressure tendency
pressure	atmospheric surface	a	pressure_tendency_characteristic	Characteristic of pressure tendency (used in synoptic maps – coded value 0-8)
radiation	atmospheric	dr	diffuse_radiation	TBD

radiation	atmospheric	dlwie	downward_longwave_irradiance_ at_earth_surface	Flux density of radiation emitted by the gases, aerosols and clouds of the atmosphere to the Earth's surface
radiation	atmospheric	dswie	downward_shortwave_irradiance_ at_earth_surface	Flux density of the solar radiation at the Earth surface
radiation	atmospheric	dswit	downward_shortwave_irradiance_ at_toa	Flux density of the solar radiation at the top of the atmosphere
radiation	atmospheric	eswr	earth_surface_shortwave_ bidirectional_reflectance	Reflectance of the Earth surface as a function of the viewing angle and the illumination angle in the range 0.4-0.7 micro m. The distribution of this variable is represented by the Bidirectional Reflectance Distribution Function (BRDF)
radiation	atmospheric	fapar	fraction_of_absorbed_par	Fraction of PAR absorbed by vegetation (land or marine) for photosynthesis processes (generally around the 'red')
radiation	atmospheric	gr	global_radiation	TBD
radiation	atmospheric	lwe	longwave_earth_surface_emissivity	TBD
radiation	atmospheric	lr	longwave_radiation	TBD
radiation	atmospheric	mor	meteorological_optical_range	Meteorological optical range at surface
radiation	atmospheric	par	photosynthetically_active_radiation	Flux of downwelling photons of wavelength 0.4-0.7 micro m
radiation	atmospheric	swcr	shortwave_cloud_reflectance	Reflectance of the solar radiation from clouds
radiation	atmospheric	sr	shortwave_radiation	TBD
radiation	atmospheric	sgf	solar_gamma_ray_flux	Radiative flux integrated over the gamma-ray domain.
radiation	atmospheric	suf	solar_UV_flux	Integrated UV flux over the solar disk.
radiation	atmospheric	svf	solar_VIS_flux	Integrated VIS flux over the solar disk
radiation	atmospheric	sxf	solar_X_ray_flux	Integrated X-ray flux over the solar disk
radiation	atmospheric surface	ss	sunshine_duration	Number of hours of sunshine
radiation	atmospheric	ulwie	upward_longwave_irradiance_ at_Earth_surface	Flux density of terrestrial radiation emitted by the Earth surface
radiation	atmospheric	ulwit	upward_longwave_irradiance_ at_TOA	Flux density of terrestrial radiation emitted by the Earth surface and the gases, aerosols and clouds ot the atmosphere at the top of the atmosphere
radiation	atmospheric	uswit	upward_shortwave_irradiance_ at_TOA	Flux density of solar radiation, reflected by the Earth surface and atmosphere, emitted to space at the top of the atmosphere
radiation	atmospheric	usrt	upward_spectral_radiance_ at_TOA	Upward radiant power measured at the top of the atmosphere per area unit, per solid angle, and per wavelength interval. Spectral range 0.2-200 micro m.
salinity	oceanic	sal	salinity	Ocean salinity (PSU)

temperature	atmospheric surface, upper air	Ta; t_air	air_temperature	Air temperature is the bulk temperature of the air, not the surface (skin) temperature.
temperature	atmospheric surface	Tx	daily_maximum_air_temperature	TBD
temperature	atmospheric surface	Txs	daily_maximum_air_temperature_ with_direct_sun_exposure	TBD
temperature	atmospheric surface	TGs	daily_maximum_grass_temperature	Grass maximum thermometer is 5 cm above ground
temperature	atmospheric surface	Tn	daily_minimum_air_temperature	TBD
temperature	atmospheric surface	Tns	daily_minimum_air_temperature_ with_direct_sun_exposure	TBD
temperature	atmospheric surface	TGn	daily_minimum_grass_temperature	Grass minimum thermometer is 5 cm above ground
temperature	atmospheric surface	days_frost	days_with ground_frost	Number of days with frost
temperature	atmospheric surface	t_snow	snow_temperature	TBD
temperature	atmospheric sub-surface	Ts	soil_temperature	Temperature below surface level at indicated depth
temperature	oceanic	t_water	water_temperature	Water (sea, river, lake) temperature at depth indicated, includes SST
visibility	atmospheric surface	vv	horizontal_visibility_in_air	The visibility is the distance at which something can be seen; measured on land or on sea platforms
weather		ld	lightning_detection	Detection of the time and location (latitude, longitude) of lightning events. Accuracy expressed in terms of Hit Rate and False Alarm Rate, which requires predetermination of a specific distance and time tolerance.
weather		ls	lightning_duration	TBD
weather		lhd	lightning_horizontal_distance	TBD
weather	atmospheric surface	w1	past_weather_1	Past weather 1 - most extreme phenomenon (used in synoptic maps)
weather	atmospheric surface	w2	past_weather_2	Past weather 2 - most frequent phenomenon (used in synoptic maps)
weather	atmospheric surface	ww	present_weather	Present weather (used in synoptic maps)
weather		tld	Total_lightning_density	Total number of detected flashes in the corresponding time interval and the space unit. The space unit (grid box) should be equal to the horizontal resolution and the accumulation time to the observing cycle
wind	atmospheric surface, upper air	u	eastward_wind_speed	Eastward indicates a vector component which is positive when directed eastward (negative westward). Wind is defined as a two-dimensional (horizontal) air velocity vector, with no vertical

				component. (Vertical motion in the atmosphere has the standard name upward_air_velocity.)
wind	atmospheric surface, upper air	v	northward_wind_speed	Northward indicates a vector component which is positive when directed northward (negative southward).
wind	atmospheric surface, upper air	dd	wind_from_direction	Direction from which the wind is blowing
wind	atmospheric surface, upper air	w	wind_speed	Speed is the magnitude of velocity. Wind is defined as a two-dimensional (horizontal) air velocity vector, with no vertical component. (Vertical motion in the atmosphere has the standard name upward_air_velocity.) The wind speed is the magnitude of the wind velocity.
wind	atmospheric surface	fx	wind_speed_of_gust	A gust is a sudden brief period of high wind speed.
wind	atmospheric surface	fm	wind_speed_max	Maximum observed wind speed over a specified period.
precipitation	Atmospheric surface	pwc	Precipitable_water_column	TBD
pressure	Upper Air	TropH	Tropopause_height	TBD
temperature	Upper Air	TropT	Tropopause_temperature	TBD
pressure	Upper Air	TropP	Tropopause_pressure	TBD
temperature	Upper Air	TropPT	Tropopause_potential_temperature	TBD
temperature	Atmospheric surface	FrostT	Frost_point_temperature	TBD
pressure	Atmospheric surface; upper air	gph	Geopotential_height	Height of a standard or significant pressure level in meters
pressure	Atmospheric surface	gdm	Geopotential_height_decameters	Height of standard or significant pressure level in decameters
temperature	Atmospheric surface; upper air	temp_vertgrad	Vertical_gradient_of_temperature	Vertical variation of temperature
temperature	Atmospheric surface	ptemp_vertgrad	Vertical_gradient_of_ potential_temperature	Vertical variation of potential temperature
temperature	Atmospheric surface; upper air	ept	Equivalent_potential_temperature	Temperature a parcel of air would reach if all the water vapour in the parcel were to condense, releasing its latent heat, and the parcel was brought adiabatically to a standard reference pressure, usually 1000 hPa
wind	Atmospheric surface	rs_vertspeed	Vertical_speed_of_radiosonde	Vertical speed of radiosonde ascent
humidity	Atmospheric; upper air		water_vapour_mixing_ratio	Volume mixing ratio (mol/mol) of water vapour calculated using Hyland and Wexler (1983)
humidity	Atmospheric surface		air_relative_humidity_effective_vertical_resolution	Resolution (defined by 1 / cut_off frequency) of the relative_humidity in terms of time
pressure	Atmospheric; upper air		altitude	Altitude

temperature	Atmospheric surface		air_temperature	Air temperature (from profile measurement)
humidity	Atmospheric surface; upper air		air_dewpoint	Dewpoint measurement (from profile measurement)
humidity	Atmospheric surface		relative_humidity	Relative humidity (from profile measurement)
wind	Atmospheric surface; upper air		eastward_wind_speed	Eastward wind speed (from profile measurement)
wind	Atmospheric surface		northward_wind_speed	Northward wind speed (from profile measurement)
radiation	Atmospheric surface; upper air		solar_zenith_angle	solar zenith angle

2.1.5.2. Metadata QC tool for station location: The R Package “stlocationqc”

The R (https://www.r-project.org/) package stlocationqc results from the need to perform quality control over extensive lists of station coordinates in as much of an automated way as possible. It can be applied to land surface and upper air fixed platforms but can also be applied to fixed marine platforms. In the first case, the tool determines the fixed station current country name, in the second case the geographic sea name.

The software was developed to integrate the set of metadata QC tools produced by the C3S Data Rescue Service, which runs in R and the latest version is available from: https://datarescue.climate.copernicus.eu/st_metadata-quality-control. Testing this tool with various station coordinate lists has been essential for issues related to detection and general improvement of the tool.

Lists of stations belonging to inventories of large databases like ISPD – International Surface Pressure Databank or GHCN – Global Historical Climatology Network often do not have a column with the descriptive country name. However, geographical coordinates, altitude, station name and the WMO ID are provided as geospatial references. In this case, the goal is to assign stations to the current country name corresponding to their location, to fill the “Country” column when adapting those inventories to standardized inventory formats (e.g., the C3S Data Rescue Service inventories format – see Sect. 2.1.5.1). For points located at sea, corresponding to ships, buoys or platforms, a geographical sea name is also assigned.

Several examples and tests, already performed with land surface stations only, proved that in many cases it is not possible to assign country names to all coordinate points. Some are only resolved by assigning a sea name. Plotting those points on satellite imagery of the Earth shows that most of them, besides being located at sea, are concentrated near to region/continent shorelines, which suggests poor positional precision in those coastal/island stations. This fact also suggests that the same imprecision problem might occur with some inland stations. However, the stlocationqc R package focuses on positioning errors at country (and sea) level and positional imprecision is addressed only for coastal and border stations.

The stlocationqc tool determines the location in the political country borders by sequentially running a maximum of six functions. Two additional functions exist to download the required spatial datasets that provide the World's country/sea names and boundaries (Fig. 2). These essential datasets are downloaded from Natural Earth (https://www.naturalearthdata.com/about/terms-of-use/) when the user runs the tool for the first time. Considering the occasional unavailability of the Natural Earth's site, copies of these spatial datasets are permanently stored in the \data-raw directory of the package, on GitHub.

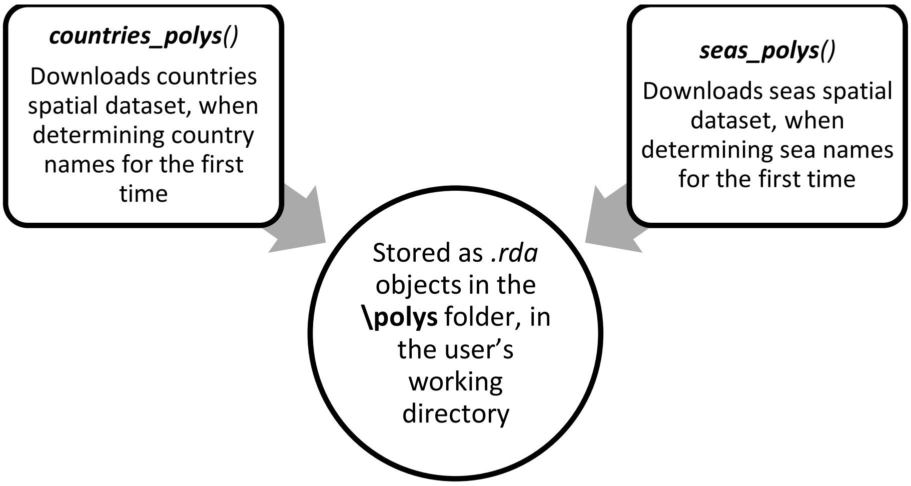

countries_polys()
Downloads countries spatial dataset, when determining country names for the first time
seas_polys()
Downloads seas spatial dataset, when determining sea names for the first time
Stored as *.rda* objects in the **\polys** folder, in the user's working directory

Fig. 2. Functions for downloading the spatial datasets.

The method of assigning names is an exclusion process which consists of four sequential steps:

1. Test the validity of the coordinates, excluding impossible values (latitude outside the [-90, +90] interval, identifying if longitude is in the [0, 360] interval or in the [-180, +180] interval and exclude values outside them, altitude outside Earth's surface heights) and setting the longitude in the range [-180, +180] if not originally in that interval
2. Try to assign country names to all the valid coordinate points
3. Evaluate the coastal stations precision and assign more country names
4. Take only those points for which it is not possible to assign a country name and try to assign them a sea name.

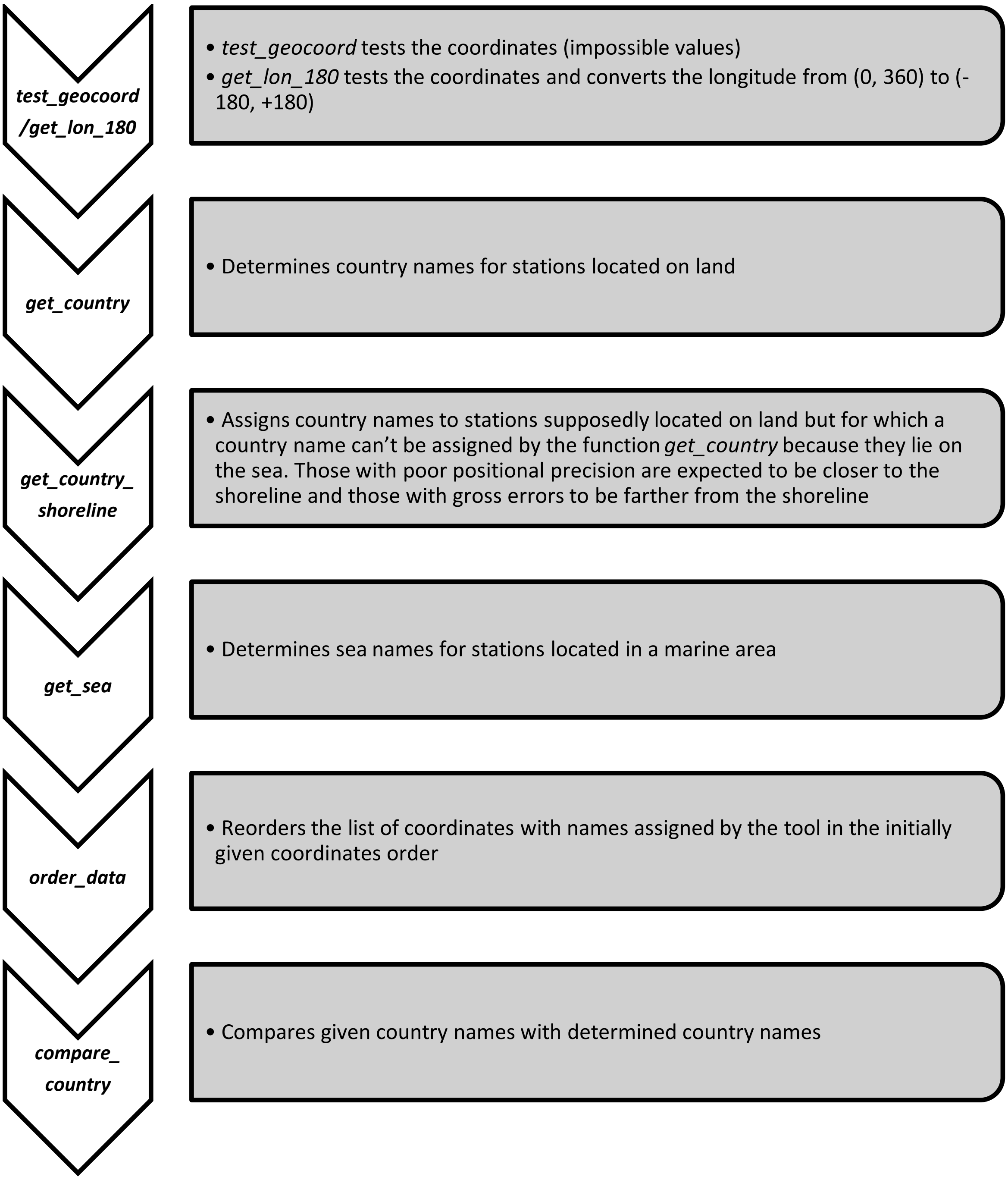
The list of coordinates and the names returned are presented in the output file following the original coordinates' order, ready to incorporate into the inventory. Concerning the third step of the method, this was created for stations supposed to be located on land (e.g., a list of land stations only) and for which it is not possible to assign a country name. The examination of several examples led to the conclusion that these stations are most likely near the shoreline and have poor coordinate precision, which results in them being positioned at sea. The function which evaluates the tolerance addresses that issue, allowing the user to set a tolerance x for those stations and returning the country name of those near the shoreline with an error up to x meters towards the sea. What the function does is to create a buffer zone with a radius of x meters around each point lying on the sea and checks if the buffer overlaps some country polygon. The minimum positional precision required for a fixed station has not been set. However, it makes sense that for stations that are still active, the required precision should be at least 0.001 decimal degrees (111.32 m at the equator) and that for historical stations closed long ago a greater tolerance should be given.

The functions that compose the package and the instructions for its use are described in detail in the package documentation, which can be accessed, for example, by writing the command help(package = “stlocationqc”) in the RStudio console. A summary of the stlocationqc functions which run sequentially is presented in Fig. 3.

The undefined minimum positional precision for fixed stations has its advantages. It allows the tool to be applied both to more precise and less precise sources and in an iterative way which allows the sorting of the stations according to the positional error. The iterative method consists of running the get_country_shoreline function several times, always starting with a smaller tolerance, and successively increasing the tolerance to apply to the previous output of unnamed points (e.g., 400, 1.000, 5.000, 10.000 m), until no unnamed points are left for the largest tolerance. The tolerance given is added to each output file name, and all (or almost all) the points lying at sea should be sorted according to the rank of tolerances given. Sorting the stations allows the user to distinguish those with imprecise coordinates from those completely misplaced.

The disadvantages are that the iterative process is semi-automatic and that the function get_country_shoreline should be applied with care always starting with small tolerances, i.e., tolerances of several tens of meters. A tolerance of several kilometres should never be given at the start, as this can result in a wrong output for stations with smaller precision error. Other situations to keep in mind are those for small island nations (e.g., Nauru, Tuvalu) and some which have neighbours (e.g., Monaco). Other cases are countries that are not so small, but the marine area between them is small, like Morocco and Spain, or countries that are relatively narrow like Lebanon. For all those examples and many others, a large tolerance given to coastal stations can result in a wrong solution.

The function get_sea assigns the sea name to coordinate points located in a water body for which a country name cannot be assigned by get_country and also by get_country_shoreline after a maximum tolerance has been considered. The function intends to locate land stations with gross positional errors or situations where the points correspond to marine records made by ships, buoys or platforms. Therefore, the function might be unnecessary if the sample does not contain marine fixed stations and the coastal stations have good precision. Also, the function compare_country is unnecessary if there are no country names to compare.



- *test_geocoord* tests the coordinates (impossible values)
- *get_lon_180* tests the coordinates and converts the longitude from (0, 360) to (-180, +180)

- Determines country names for stations located on land

- Assigns country names to stations supposedly located on land but for which a country name can't be assigned by the function *get_country* because they lie on the sea. Those with poor positional precision are expected to be closer to the shoreline and those with gross errors to be farther from the shoreline

- Determines sea names for stations located in a marine area

- Reorders the list of coordinates with names assigned by the tool in the initially given coordinates order

- Compares given country names with determined country names

Fig. 3. Functions which run sequentially.

Besides assigning geographic names, another issue addressed by the tool concerns the country names themselves. Unlike the ISPD inventory, other metadata inventories already have a column with the country names, and in these cases, it is important to know if each name is correct. For this purpose, a function which establishes the comparison between the given and the returned names was created. It also aims to ensure the uniformity of country names across all the fixed station inventories to be transformed to the C3S Data Rescue Service inventory format. The function uses standard country names from the list of English names provided by the countries spatial datasets downloaded from https://www.naturalearthdata.com/downloads/. Also, the geographical sea names returned are the ones provided by the seas spatial datasets. The function which compares the country names evaluates the following (other future checks can be added):

- Has a country name been provided by the user?
- Is the provided country name the same as that found by the tool?
- Is the country name provided equal to that of the sovereignty (rather than that found by the tool)?
- Is the country name provided in any of these other languages: Spanish, Portuguese, French and German?
- Is the country name provided given in upper case?
- Is there a partial match (entire word) between the country name provided and that found by the tool (English or in any of these other languages: Spanish, Portuguese, French and German)
- Is there a partial match (three letters at beginning or end of word) between the country name provided and that found by the tool?
- Are the coordinates provided over the sea?

The C3S Data Rescue Service portal also performs metadata QC to submitted inventories. However the portal doesn't use directly the R Package "stlocationqc", but a python function that performs similar tests. It has to be noted that there should be consistency between the country name validations executed by the C3S Data Rescue Service portal when fixed station inventories are submitted, and the standard country names used by stlocationqc when performing QC on a desktop. The reason is that the portal validations are also based on a countries' spatial dataset from Natural Earth.

Two example datasets are included in the package: the ISPD stations inventory containing land and marine stations and without given country names to compare with; the ERACLIM Upper-air stations inventory with land stations only and given country names to be compared with the tool results.

Some known issues which result in an incomplete or wrong solution are related to a spatial dataset's coverage and precision. Hence it has been taken as a preferred option to download these datasets directly from the source (Natural Earth) rather than using a static copy of it. This allows the tool to benefit from the updates/corrections made periodically by Natural Earth.

Concerning the get_country_shoreline function, the default tolerance is set to 500 m. This, however, is an empirical value, justified by the need to consider imperfections in countries' and seas' polygons used by the tool which do not match perfectly. This shortcoming leads to gaps in the global coverage of the spatial dataset used. Consequently, coastal points falling in one of those gaps would be unnamed but using the 500 m tolerance has been shown to solve the coverage problem (at least for the coordinates tested so far).

In what concerns the compare_country function, the output categorized as “different country name” can present false positive errors. This is due to poor precision of the country polygons datasets. In some regions the polygon lines do not match the country borders precisely, which results in a wrong solution by the tool. An example of this situation was found in the ERA-CLIM Upper-air inventory: a station belonging to French Guiana (France) was found to be located by the tool in Brazil. Besides being an error that does not occur very often or for many points, this is the most serious issue detected, and a solution will be added in future versions.

Future functions to add to this tool will evaluate the consistency between the country name, WMO ID and WMO Region in the inventories and address the altimetric QC of land surface stations.
The first stable version of the software (1.0.0) was released in 2019. Fig. 4 represents the workflow scheme, the tool concept, and how the various functions relate to each other.

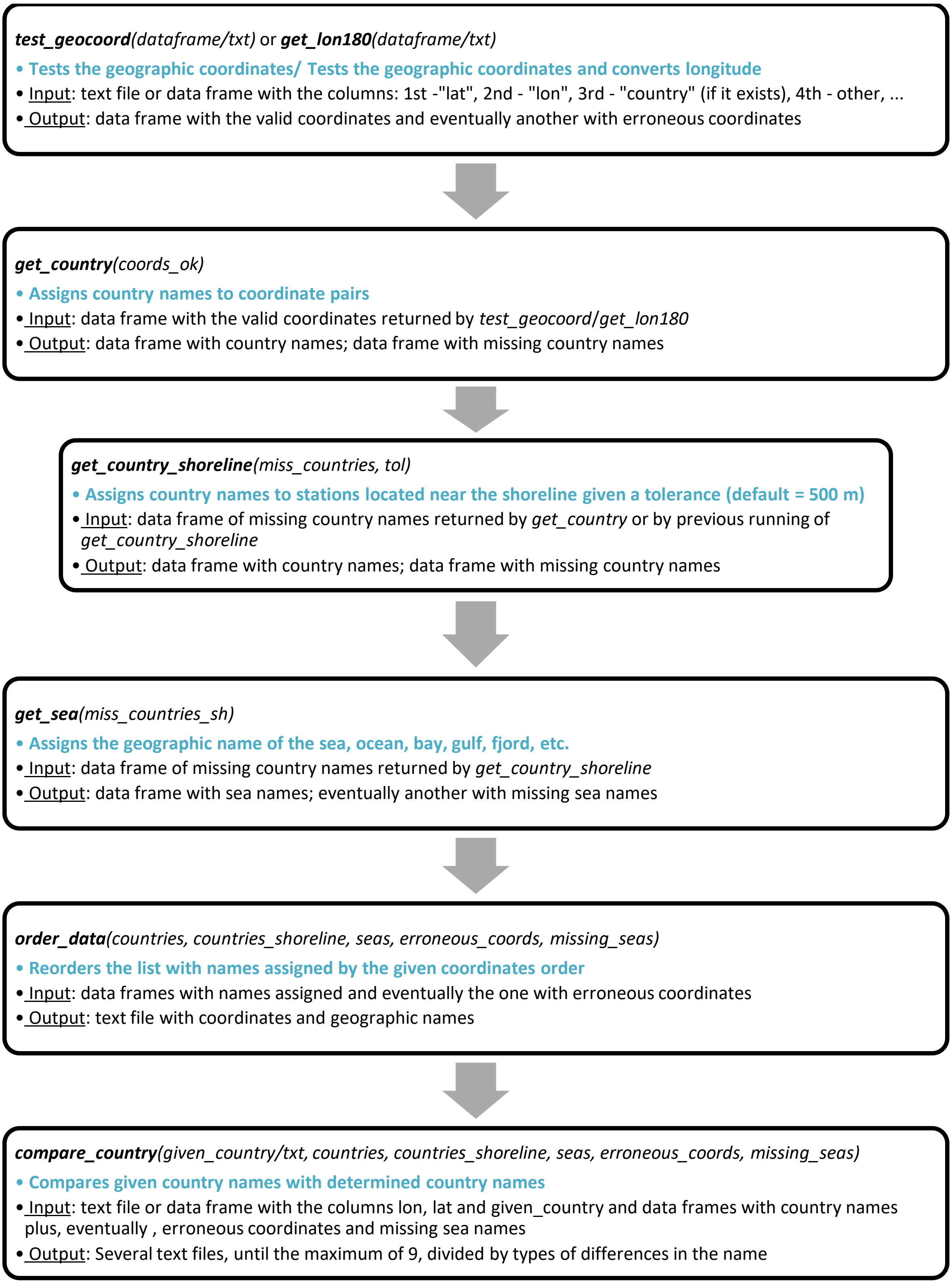
2.1.5.3. Submitting metadata to the C3S Data Rescue Service inventory

Submission of metadata inventories requires login after registration at the C3S Data Rescue Service portal. There are two ways to submit metadata to the Inventories in the Portal:

1. Insert each record (row) one by one by using the on-line form supplied by the metadata inventory
2. Download the Land Surface Observations Template available at the Metadata Exchange Facility (https://datarescue.climate.copernicus.eu/met), fill in the template, upload, validate and submit

The first process only allows the uploading of one record at a time and should be used for a small number of entries, or correction of already inserted metadata. The second method is ideal for submitting long inventories.

After completion of a Land Surface Inventory following Tables 1 and 2, also contained in the Guidelines for Metadata Inventories Standards and Formats (https://datarescue.climate.copernicus.eu/met), it can be validated by the Metadata Exchange Facility. Metadata quality control tests are applied to the inventory’s content, and if they finds errors or missing mandatory fields, these will be listed on the screen. The user can then proceed to correct each error and/or fill the missing mandatory fields.



test_geocoord*(dataframe/txt)* or ***get_lon180****(dataframe/txt)*

- **Tests the geographic coordinates/ Tests the geographic coordinates and converts longitude**
- Input: text file or data frame with the columns: 1st -"lat", 2nd - "lon", 3rd - "country" (if it exists), 4th - other, ...
- Output: data frame with the valid coordinates and eventually another with erroneous coordinates

get_country*(coords_ok)*

- **Assigns country names to coordinate pairs**
- Input: data frame with the valid coordinates returned by *test_geocoord/get_lon180*
- Output: data frame with country names; data frame with missing country names

get_country_shoreline*(miss_countries, tol)*

- **Assigns country names to stations located near the shoreline given a tolerance (default = 500 m)**
- Input: data frame of missing country names returned by *get_country* or by previous running of *get_country_shoreline*
- Output: data frame with country names; data frame with missing country names

get_sea*(miss_countries_sh)*

- **Assigns the geographic name of the sea, ocean, bay, gulf, fjord, etc.**
- Input: data frame of missing country names returned by *get_country_shoreline*
- Output: data frame with sea names; eventually another with missing sea names

order_data*(countries, countries_shoreline, seas, erroneous_coords, missing_seas)*

- **Reorders the list with names assigned by the given coordinates order**
- Input: data frames with names assigned and eventually the one with erroneous coordinates
- Output: text file with coordinates and geographic names

compare_country*(given_country/txt, countries, countries_shoreline, seas, erroneous_coords, missing_seas)*

- **Compares given country names with determined country names**
- Input: text file or data frame with the columns lon, lat and given_country and data frames with country names plus, eventually , erroneous coordinates and missing sea names
- Output: Several text files, until the maximum of 9, divided by types of differences in the name

Fig. 4. Workflow, concept and relation among functions of the stlocationqc software.

The different metadata Quality control tests include verification of consistency between the metadata and insertion of correct entries. There are limit tests for latitude, longitude and altitude, starting and ending dates of the series, variable names and units and cross-checks of the inserted country and the latitude and longitude (see Sect. 2.1.5.2), among other tests.

After all the errors found in the validation process have been corrected, the user can click on the "submit" button to add the metadata to the Land Surface Observations Inventory. The procedure is similar for other inventories (Upper Air Fixed and Moving Platforms and Marine Observations). The user needs to download the corresponding inventory template and follow the Guidelines for Metadata Inventories Standards and Formats both available at the C3S Data Rescue Service Portal (https://datarescue.climate.copernicus.eu/met).

2.2. Climate data formatting

2.2.1. Data format as interface between digitizers and database builders

Climate data rescue is the entire process of archiving, searching, finding, imaging, digitizing, and converting non-digital data to a machine-readable format. For the sake of simplicity, this can be roughly described by two main steps:

1. A digitization step finds observations archived on paper or any other media and produces digital versions of those observations - first as digital images and then as Excel spreadsheets or similar machine-readable format.
2. A database-building step converts the new digital observations into the format and schema used by an observation's database and adds the observations to the database.

Different persons usually undertake these two steps: the first by a large group of observations experts (we will call them transcribers for simplicity, although experts do not usually do physical transcription), each interested in a different set of to-be-digitized observations; the second by a small group of "synthesizers" trying to make the best possible database. The split between the steps causes problems: the output of step 1 (typically differently structured Excel spreadsheets) is poorly suited for the input to step 2. We cannot ask the transcribers to produce database-ready output, because this requires them to know too much about the precise and idiosyncratic details of each database, and we cannot expect the synthesizers to work with a too large amount of variably-structured Excel spreadsheets - partly because they would have to learn too much about the idiosyncrasies of each observation source, and partly because there are many fewer synthesizers than transcribers. The practical effect of this is that observations pile up in a transcribed-but-unusable state, and it takes too long to prepare them for use.

2.2.2. The C3S Data Rescue Service climate data format: The Station Exchange Format (SEF)

The Station Exchange Format (SEF) is a proposed new output for the transcription step. It will eliminate the bottleneck between the transcription and database-building steps by specifying a single data format that is suitable both as the output of the former and the input to the latter. This means the format must have two, somewhat contradictory properties:

1. It must be machine readable with NO human involvement – so it needs all the necessary metadata in an unambiguous arrangement.
2. It must be easy for non-experts to read, understand, and create.
3. The design of the SEF tries to be both simple enough to be obvious, and powerful enough to be useful, by having a core set of headers and columns which are obvious, and an arbitrarily extensible metadata section.

The latest SEF specifications can be found on the C3S Data Rescue Service website (https://datarescue.climate.copernicus.eu/st_formatting-land-stations), as well as tutorials and examples.

2.2.2.1 Who should use the SEF, and why

The SEF allows international data repositories to ingest rescued data more efficiently, reducing the average ingestion time from years to weeks. Data sets considered to be a low priority are often never ingested if they are in a non-standard format. Therefore, anyone who would like to see their newly rescued data being available to the international community in a timely fashion should adopt the SEF.

However, creating SEF files implies, in most cases (particularly for large data sets), a certain familiarity with at least one coding language (e.g., R, Python, Fortran…). Those who do not feel comfortable with coding and cannot allocate time for improving their skills are not recommended to try and produce SEF files, as they are likely to introduce errors in the data.

2.2.2.2 What is needed to produce the SEF

Before starting the conversion of raw digitized data into SEF, one should make sure that the following actions have been performed:

- All metadata have been collected for the records that are going to be formatted into SEF. This includes, in particular: station name, geographical coordinates, units, and observation times and their offset from UTC.
- If multiple data series are going to be converted: metadata are organized in a table or other format that is easy to read in an automated way. Ideally, they should follow the metadata structure described in Table 1.

- A unique identifier has been assigned to each station. The identifier should not contain special characters or blanks (for more details see the SEF documentation).
- Data are converted to modern units (preferably metric, see table of recommended units in the SEF documentation). However, note that the values in the original units should be provided in the SEF files too.
- A data license has been chosen, if possible one that allows commercial use of the data.
- If using R or Python: the latest version of the software tools provided by the C3S Data Rescue Service have been installed.
- The SEF documentation provided on the C3S Data Rescue Service website has been read thoroughly.

Before submitting SEF files to a data repository, it is important to validate the files with simple checks on ranges (e.g., month must be between 1 and 12) and consistency between fields. Software for this is also provided by the C3S Data Rescue Service. It is also recommended to try and use the SEF files (e.g., plot the data) before submitting them.

2.2.2.3 Time conversion to UTC

One of the main requirements of the SEF at the time of writing is that the observation times must be provided in UTC (also known as Greenwich Mean Time). This is fundamental for global use of the data and for many scientific applications but requires historical knowledge that only those who rescue the data can easily obtain (hence the requirement for SEF).

The software provided by the C3S Data Rescue Service has tools that facilitate this conversion. This usually involves a constant offset (e.g., 1 hour for Central Europe), but there have been numerous changes over history that must be considered. Daylight saving time is usually not adopted for weather observations, but this rule might not apply to all stations and must also be considered. For measurements taken before the introduction of standard time zones, mean local solar time is usually an adequate approximation and the conversion to UTC can be performed using the longitude of the station.

To ensure data traceability and facilitate quality control, it is also important to provide in the metadata fields of the SEF, the original time and date as digitised.

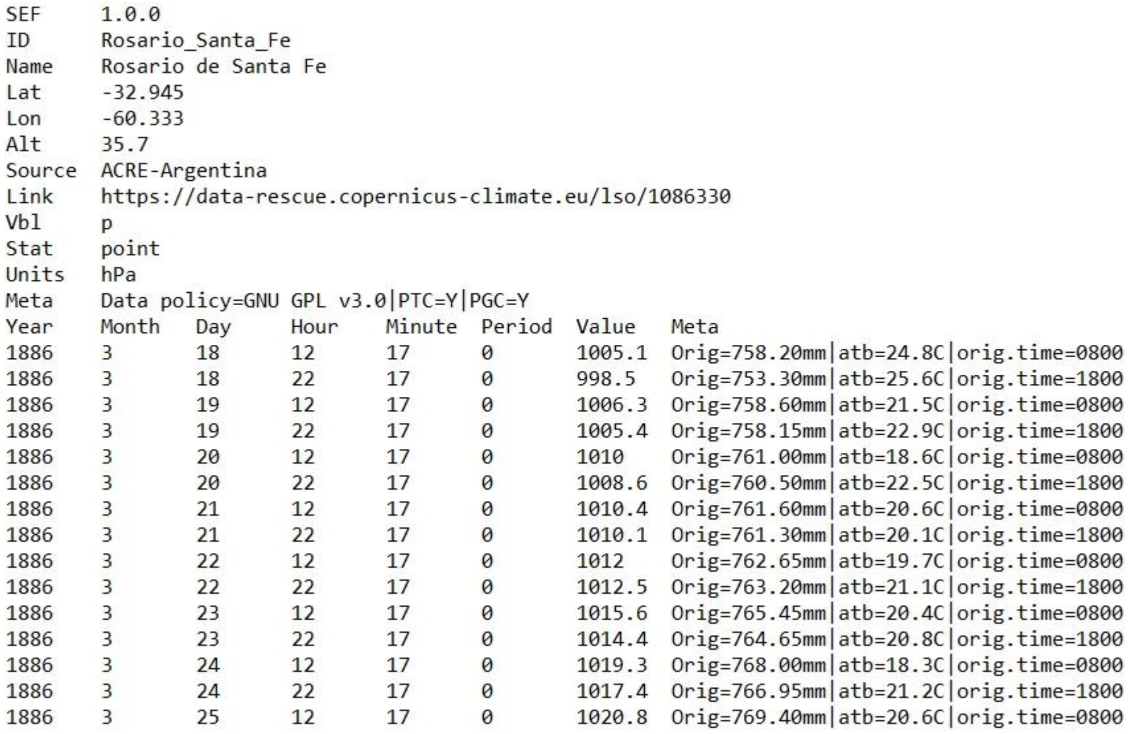
2.2.2.4 SEF examples

In this section examples of data converted into SEF (version 1.0.0) are shown. Future changes in the format are possible: always refer to the latest documentation available online.

The example in Fig. 5 contains an instantaneous pressure series. The “Meta” field in the header indicates the license under which the data can be used, and that the pressure in the “Value” column has already been corrected for temperature (PTC) and gravity (PGC). The “Meta” column gives the original value as written in the source (with the original unit), the temperature of the barometer (again as written in the source), and the original time.

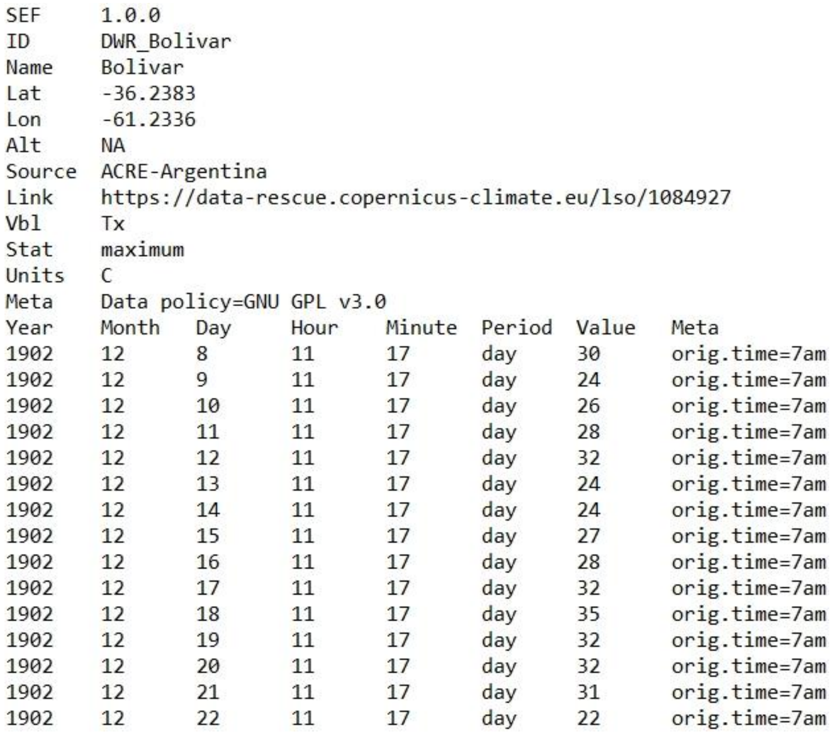
The example in Fig. 6 is for a daily maximum temperature series. The period is set to “day”, because the observation is always made at the same time (7am). If the observation time is not known, the columns “Hour” and “Minute” can be left empty. In this case, the original observation in the “Meta” column is not necessary because no conversion was performed (temperature was already expressed in degrees Celsius in the source).

The example in Fig. 7 shows the use of the “p” period code in a precipitation series. Here, for some days, precipitation is measured multiple times, but in others only once. The maximum interval between two of these observations is always 1 day (hence “p1day”).



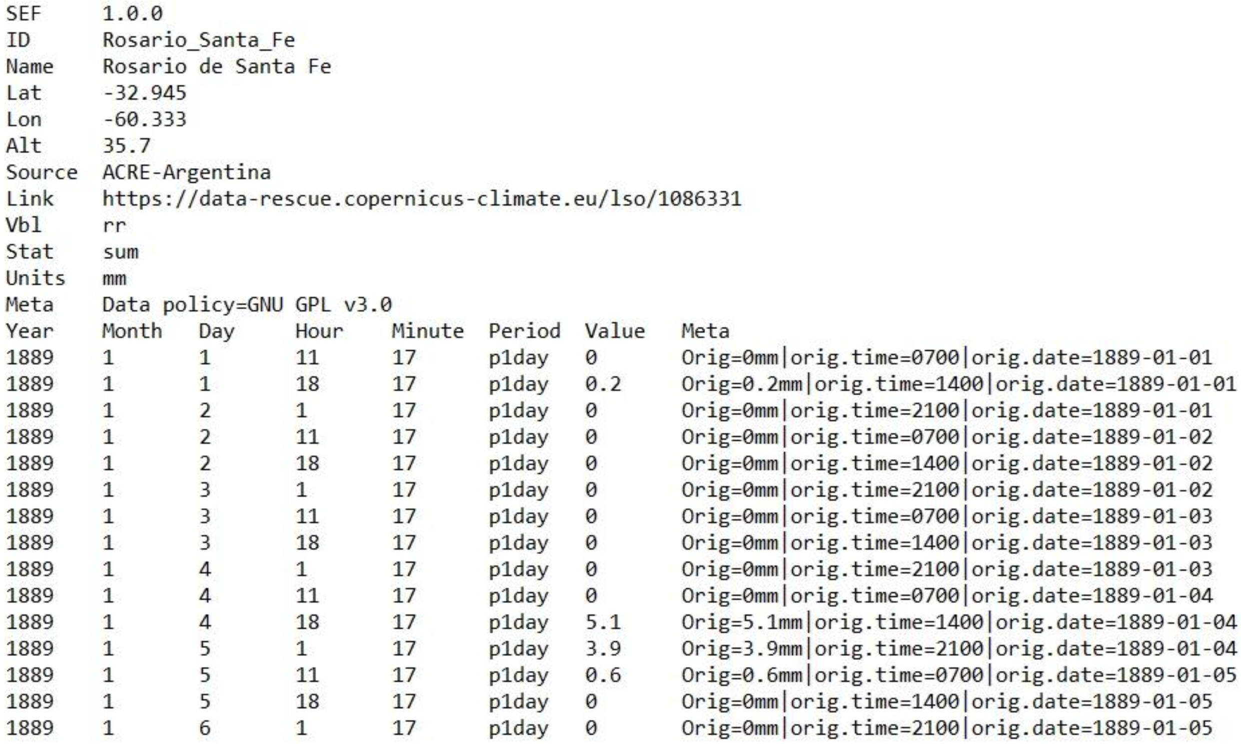
```
SEF     1.0.0
ID      Rosario_Santa_Fe
Name    Rosario de Santa Fe
Lat     -32.945
Lon     -60.333
Alt     35.7
Source  ACRE-Argentina
Link    https://data-rescue.copernicus-climate.eu/lso/1086330
Vbl     p
Stat    point
Units   hPa
Meta    Data policy=GNU GPL v3.0|PTC=Y|PGC=Y
Year    Month   Day     Hour    Minute  Period  Value   Meta
1886    3       18      12      17      0       1005.1  Orig=758.20mm|atb=24.8C|orig.time=0800
1886    3       18      22      17      0       998.5   Orig=753.30mm|atb=25.6C|orig.time=1800
1886    3       19      12      17      0       1006.3  Orig=758.60mm|atb=21.5C|orig.time=0800
1886    3       19      22      17      0       1005.4  Orig=758.15mm|atb=22.9C|orig.time=1800
1886    3       20      12      17      0       1010    Orig=761.00mm|atb=18.6C|orig.time=0800
1886    3       20      22      17      0       1008.6  Orig=760.50mm|atb=22.5C|orig.time=1800
1886    3       21      12      17      0       1010.4  Orig=761.60mm|atb=20.6C|orig.time=0800
1886    3       21      22      17      0       1010.1  Orig=761.30mm|atb=20.1C|orig.time=1800
1886    3       22      12      17      0       1012    Orig=762.65mm|atb=19.7C|orig.time=0800
1886    3       22      22      17      0       1012.5  Orig=763.20mm|atb=21.1C|orig.time=1800
1886    3       23      12      17      0       1015.6  Orig=765.45mm|atb=20.4C|orig.time=0800
1886    3       23      22      17      0       1014.4  Orig=764.65mm|atb=20.8C|orig.time=1800
1886    3       24      12      17      0       1019.3  Orig=768.00mm|atb=18.3C|orig.time=0800
1886    3       24      22      17      0       1017.4  Orig=766.95mm|atb=21.2C|orig.time=1800
1886    3       25      12      17      0       1020.8  Orig=769.40mm|atb=20.6C|orig.time=0800
```

Fig. 5. Example of an SEF file for the pressure series of Rosario de Santa Fe.



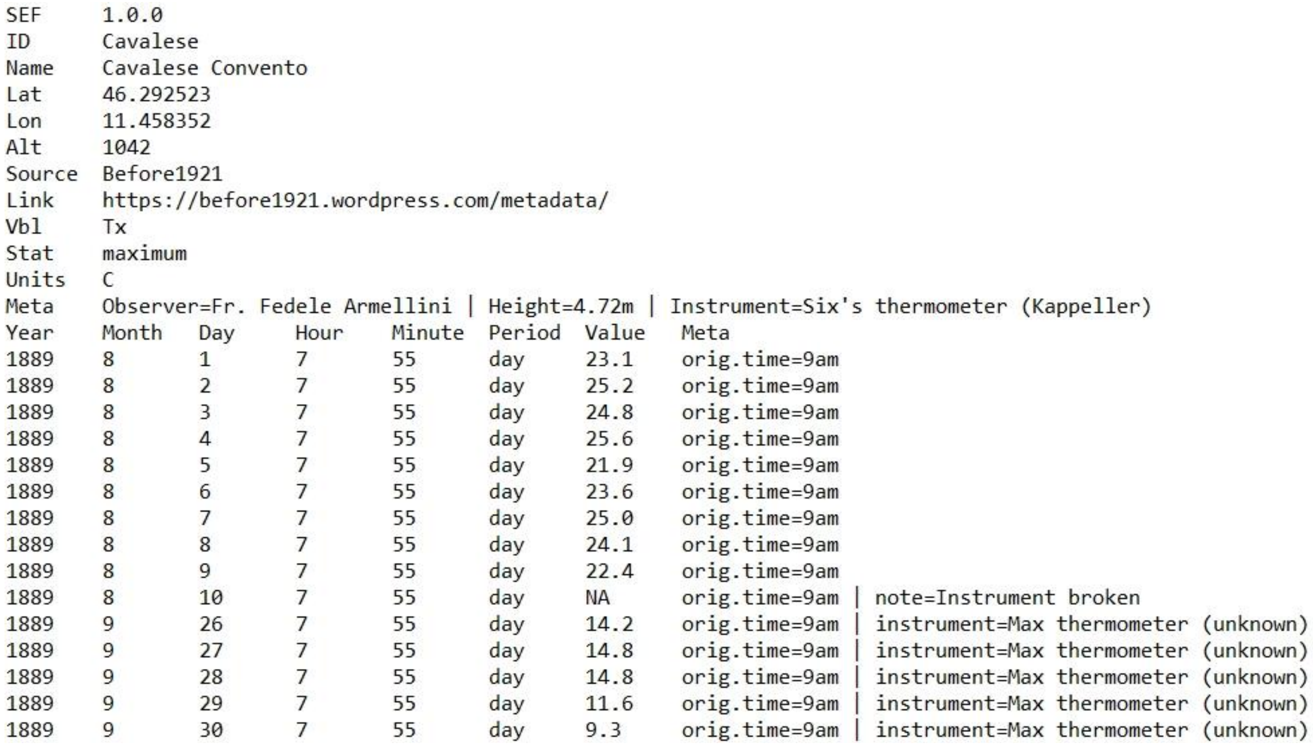
```
SEF     1.0.0
ID      DWR_Bolivar
Name    Bolivar
Lat     -36.2383
Lon     -61.2336
Alt     NA
Source  ACRE-Argentina
Link    https://data-rescue.copernicus-climate.eu/lso/1084927
Vbl     Tx
Stat    maximum
Units   C
Meta    Data policy=GNU GPL v3.0
Year    Month   Day     Hour    Minute  Period  Value   Meta
1902    12      8       11      17      day     30      orig.time=7am
1902    12      9       11      17      day     24      orig.time=7am
1902    12      10      11      17      day     26      orig.time=7am
1902    12      11      11      17      day     28      orig.time=7am
1902    12      12      11      17      day     32      orig.time=7am
1902    12      13      11      17      day     24      orig.time=7am
1902    12      14      11      17      day     24      orig.time=7am
1902    12      15      11      17      day     27      orig.time=7am
1902    12      16      11      17      day     28      orig.time=7am
1902    12      17      11      17      day     32      orig.time=7am
1902    12      18      11      17      day     35      orig.time=7am
1902    12      19      11      17      day     32      orig.time=7am
1902    12      20      11      17      day     32      orig.time=7am
1902    12      21      11      17      day     31      orig.time=7am
1902    12      22      11      17      day     22      orig.time=7am
```

Fig. 6. Example of an SEF file for the maximum temperature series of Bolivar.



```
SEF     1.0.0
ID      Rosario_Santa_Fe
Name    Rosario de Santa Fe
Lat     -32.945
Lon     -60.333
Alt     35.7
Source  ACRE-Argentina
Link    https://data-rescue.copernicus-climate.eu/lso/1086331
Vbl     rr
Stat    sum
Units   mm
Meta    Data policy=GNU GPL v3.0
Year    Month   Day     Hour    Minute  Period  Value   Meta
1889    1       1       11      17      p1day   0       Orig=0mm|orig.time=0700|orig.date=1889-01-01
1889    1       1       18      17      p1day   0.2     Orig=0.2mm|orig.time=1400|orig.date=1889-01-01
1889    1       2       1       17      p1day   0       Orig=0mm|orig.time=2100|orig.date=1889-01-01
1889    1       2       11      17      p1day   0       Orig=0mm|orig.time=0700|orig.date=1889-01-02
1889    1       2       18      17      p1day   0       Orig=0mm|orig.time=1400|orig.date=1889-01-02
1889    1       3       1       17      p1day   0       Orig=0mm|orig.time=2100|orig.date=1889-01-02
1889    1       3       11      17      p1day   0       Orig=0mm|orig.time=0700|orig.date=1889-01-03
1889    1       3       18      17      p1day   0       Orig=0mm|orig.time=1400|orig.date=1889-01-03
1889    1       4       1       17      p1day   0       Orig=0mm|orig.time=2100|orig.date=1889-01-03
1889    1       4       11      17      p1day   0       Orig=0mm|orig.time=0700|orig.date=1889-01-04
1889    1       4       18      17      p1day   5.1     Orig=5.1mm|orig.time=1400|orig.date=1889-01-04
1889    1       5       1       17      p1day   3.9     Orig=3.9mm|orig.time=2100|orig.date=1889-01-04
1889    1       5       11      17      p1day   0.6     Orig=0.6mm|orig.time=0700|orig.date=1889-01-05
1889    1       5       18      17      p1day   0       Orig=0mm|orig.time=1400|orig.date=1889-01-05
1889    1       6       1       17      p1day   0       Orig=0mm|orig.time=2100|orig.date=1889-01-05
```

Fig. 7. Example of an SEF file for the precipitation series of Rosario de Santa Fe.

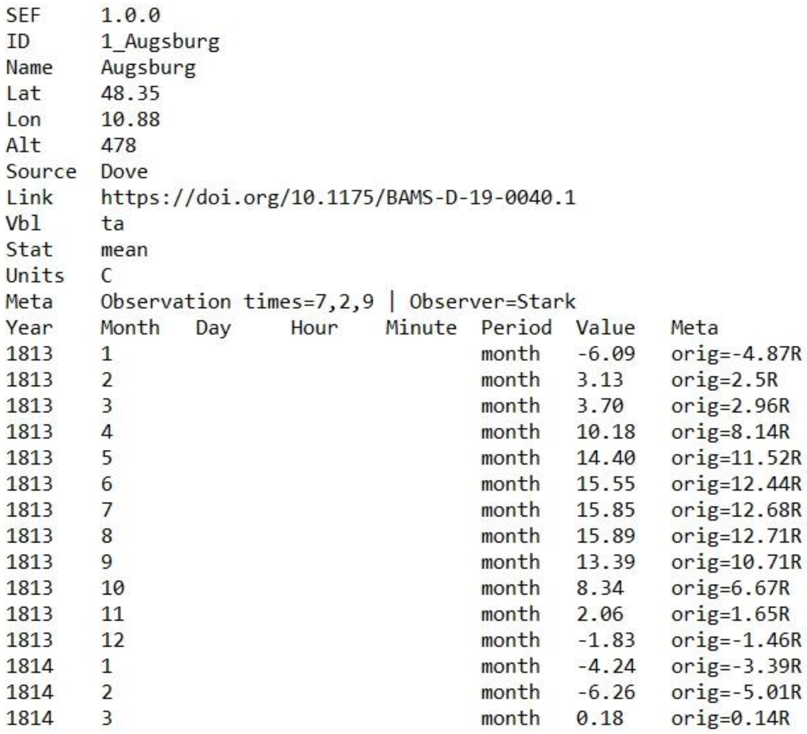


```
SEF     1.0.0
ID      Cavalese
Name    Cavalese Convento
Lat     46.292523
Lon     11.458352
Alt     1042
Source  Before1921
Link    https://before1921.wordpress.com/metadata/
Vbl     Tx
Stat    maximum
Units   C
Meta    Observer=Fr. Fedele Armellini | Height=4.72m | Instrument=Six's thermometer (Kappeller)
Year    Month   Day     Hour    Minute  Period  Value   Meta
1889    8       1       7       55      day     23.1    orig.time=9am
1889    8       2       7       55      day     25.2    orig.time=9am
1889    8       3       7       55      day     24.8    orig.time=9am
1889    8       4       7       55      day     25.6    orig.time=9am
1889    8       5       7       55      day     21.9    orig.time=9am
1889    8       6       7       55      day     23.6    orig.time=9am
1889    8       7       7       55      day     25.0    orig.time=9am
1889    8       8       7       55      day     24.1    orig.time=9am
1889    8       9       7       55      day     22.4    orig.time=9am
1889    8       10      7       55      day     NA      orig.time=9am | note=Instrument broken
1889    9       26      7       55      day     14.2    orig.time=9am | instrument=Max thermometer (unknown)
1889    9       27      7       55      day     14.8    orig.time=9am | instrument=Max thermometer (unknown)
1889    9       28      7       55      day     14.8    orig.time=9am | instrument=Max thermometer (unknown)
1889    9       29      7       55      day     11.6    orig.time=9am | instrument=Max thermometer (unknown)
1889    9       30      7       55      day     9.3     orig.time=9am | instrument=Max thermometer (unknown)
```

Fig. 8. Example of an SEF file with change of instrument.

The example in Fig. 8 shows how to represent a change of instrument. The same can be applied to any change that affects the entries of the “Meta” header (e.g., change of observer). Another way to represent a change in the metadata is to split a data series into multiple SEF files with different headers (but same ID). In case of a station relocation, a new SEF file with new coordinates is required.

The last example (Fig. 9) is for a monthly precipitation series. A practical example of the conversion of a digitization sheet to SEF is given in Sect. 3.



```
SEF     1.0.0
ID      1_Augsburg
Name    Augsburg
Lat     48.35
Lon     10.88
Alt     478
Source  Dove
Link    https://doi.org/10.1175/BAMS-D-19-0040.1
Vbl     ta
Stat    mean
Units   C
Meta    Observation times=7,2,9 | Observer=Stark
Year    Month   Day     Hour    Minute  Period  Value   Meta
1813    1                               month   -6.09   orig=-4.87R
1813    2                               month   3.13    orig=2.5R
1813    3                               month   3.70    orig=2.96R
1813    4                               month   10.18   orig=8.14R
1813    5                               month   14.40   orig=11.52R
1813    6                               month   15.55   orig=12.44R
1813    7                               month   15.85   orig=12.68R
1813    8                               month   15.89   orig=12.71R
1813    9                               month   13.39   orig=10.71R
1813    10                              month   8.34    orig=6.67R
1813    11                              month   2.06    orig=1.65R
1813    12                              month   -1.83   orig=-1.46R
1814    1                               month   -4.24   orig=-3.39R
1814    2                               month   -6.26   orig=-5.01R
1814    3                               month   0.18    orig=0.14R
```

Fig. 9. Example of an SEF file for monthly data.

2.3. Climate data QC

2.3.1. What is data QC and why should QC exercises be undertaken?

QC is the process to detect and label suspicious or potentially wrong values. This is necessary to avoid possible errors within datasets that could compromise the results of subsequent analysis.
We can distinguish between two fundamental types of errors that must be detected by the QC process:

1. Errors introduced during the digitization process (e.g., mistyped number) and also when transferring and managing data (e.g., misplacing a string of observations from one station into another). They can and should be corrected by going back to the original documents that have been scanned and digitized or by fixing bugs in the formatting algorithm.
2. Any error in the original documents, related to the observation procedure, or the transmission or publication of the observations. This can be for instance, a typographical error in a book, or a wrong reading of an instrument. Values affected by this kind of error must be flagged, not removed or corrected.

Typically, the observations affected by quality issues are of the order of 1% of the total, mostly related to digitization errors. This percentage can vary depending on the type of data, the quality of the source (e.g., larger for handwritten than for printed), and the digitization technique. Systematic errors (such as a radiation bias caused by poor sheltering) are not dealt with during the QC process.

2.3.2. QC stages: Detection, validation, flagging and summarizing QC results

Generally, climate data quality control procedures are composed of three stages, as shown in Fig. 10. The first one is the detection of the suspicious values, after subjecting the data series to a QC exercise. There can be three different QC approaches:

1. Manual/visual QC (cross-checking): Manual checks can be performed immediately on the raw output of the digitization process (e.g., spreadsheet) and are particularly effective in detecting digitization errors. Cross-checking consists of selecting a set of digitized values to be compared to the original source images. This can identify errors in the order of the data or columns/rows digitized that are difficult to detect with statistical procedures. In addition, manual checks can be to plot the data to visually identify aberrant values, calculate maxima and minima of each column, or any other statistics that also appears in the original source, and compare them to see whether they agree or not.
2. Semi-automatic QC: The detection of the suspicious values is done using statistical or logical tests that isolate suspicious values that exceed given thresholds (e.g., statistical outliers) or are physically inconsistent. It is usually necessary to convert the digitized data into a standard format (e.g., ASCII) to perform these tests. Several software packages already exist for the QC of climate data. In particular, the R package dataresqc provided by the C3S Data Rescue Service was specifically developed for newly digitized historical data. The output of the automatic tests is analyzed by a trained climatologist, who has the final word on which suspicious values to flag.
3. Automatic QC: The detection and the validation of the suspicious values is done automatically by the climate data QC tool. Unlike a semi-automatic QC, in this case, all suspicious values are flagged. Some expertise is still necessary to set the parameters of the QC so that an acceptable compromise between detection rate and false detections is achieved.

A combination of the first and second approaches gives the best results. But it requires the analysis of a trained climatologist and potentially a significant amount of time.

When checking the raw digitized data manually, one should check that the instructions given to the digitizer have been followed thoroughly, such as recommended in BPG1 on digitization. Then, large outliers or impossible values should be detected and, if confirmed to be errors of the digitizer, corrected. If the number of typographical errors is very large, it might be necessary to reassign the digitization.

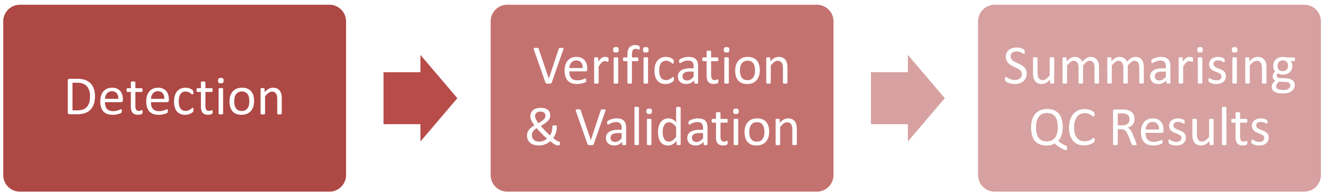


Fig. 10. Climate data quality control stages.

A useful approach, although normally rather expensive, is the multiple typing of the same document by different digitizers. In this case, typographical errors are easily recognized where the different versions disagree. In this regard, citizen science and crowd-sourcing initiatives (e.g., https://www.oldweather.org/) have shown their value compared to other digitization approaches in DARE exercises, since multiple typing is adopted but at the same time costs are reduced, making digitization more accurate and cheaper. For variables which provide enough redundancy, such as sub-daily temperature or pressure, multiple typing is often not necessary as typographical errors are efficiently detected during the QC. Multiple typing may however be necessary for variables such as precipitation, where otherwise error rates much above 1% will result.

In addition, if daily or monthly statistics (or similar) are digitized together with the raw observations, one can calculate the same statistics from the digitized data and compare them with the original ones. If there are inconsistencies, in principle, they should be caused by typographical error. One difficulty with this approach is that the error rate of the (handmade) calculations in old documents is often high, to the point that inconsistencies due to miscalculations (e.g., of monthly means and totals) can be more frequent than those caused by typographical errors.

Sometimes, it might be convenient to rearrange the structure of the spreadsheets in order to increase machine-readability. In any case, the original digitized spreadsheets as produced by the digitizer (before any QC) should be stored so that any change can be traceable and, if necessary, undone (errors can be introduced by the QC too!).

For the semi-automatic and automatic detection of suspicious values, different types of statistical QC tests can be applied. We can differentiate five different types of tests, following WMO Guidelines on Climate Metadata and Homogenization (Aguilar et al., 2003):

- Gross errors tests: to detect unrealistic values, data repetitions, date order, unrealistic dates, non-numeric value.
- Tolerance tests: to detect climatic outliers, unusual values considering the distribution.
- Temporal coherence tests: to detect values not consistent with the expected amount of change, for example, flat line tests, big jump tests.
- Inter-variable check tests: to detect inconsistencies between associated variables.
- Inter-station check tests: to detect inconsistencies between neighbouring stations.

The next stage is the tests' output verification and validation. This stage should be undertaken by an expert climatologist. Suspicious values should be compared with the original document from where they have been digitized, when possible, and by expert judgement when access to the original data source is not feasible and any correction cannot be applied with certainty. If the suspicious value is not a digitization error, it is possible to check value consistency between the previous and the next day, with other variables and with nearby stations. The procedure is illustrated in Fig. 11.

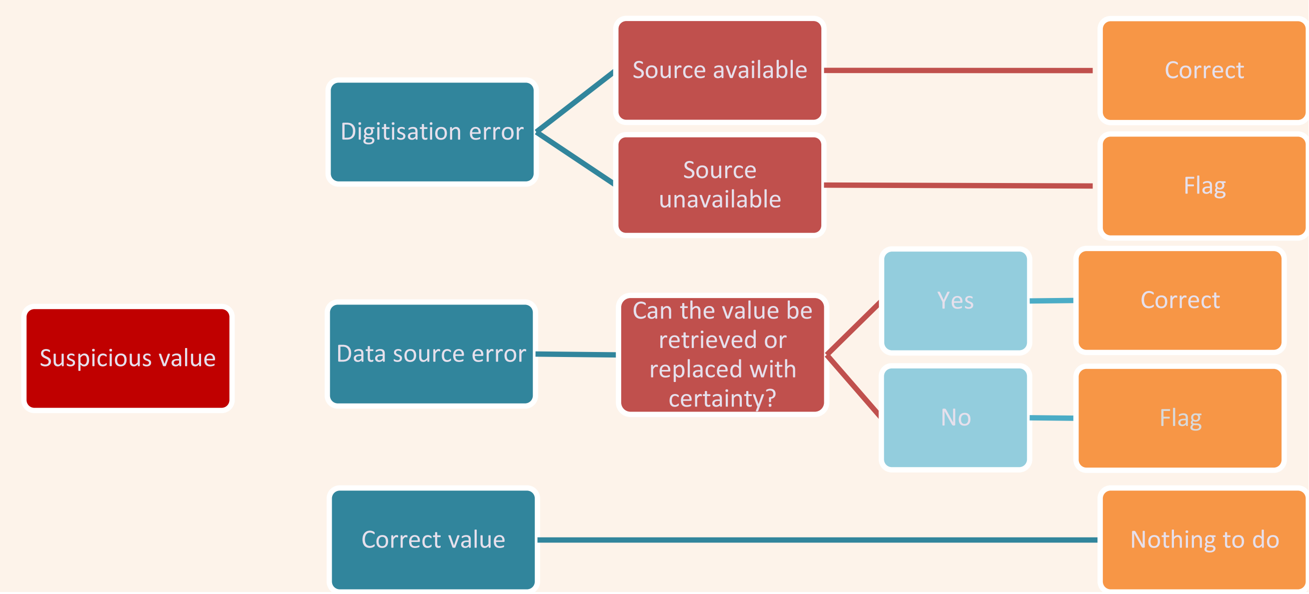


Fig. 11. Diagram showing the procedures to verify (validate or reject) suspicious values emerging from data QC.

Digitization errors can be recognized by comparing the digitized data with the document from where they have been copied, following the recommendations previously suggested in the manual/visual QC (cross-checking) approach. The original digitized data file (e.g., in a spreadsheet) has to be duplicated, leaving one of the files such as it was digitized (the original non-QC version) and another where the digitization and other obvious errors are corrected, to ensure they appear correctly in any successive data processing step and format.

Any suspicious value detected by the QC must be related to the statistical tests that have detected it and must be flagged. This means that the suspicious values are accompanied by additional information indicating that they should be considered unreliable for most uses. Ideally, each flag should also indicate the nature of the problem (e.g., which test was failed). In the SEF, the name and version of the software used for the QC should be added to the header, while the flags are indicated by the abbreviation "qc=" in the Meta column and can contain any text (more information on how to flag data in SEF files using the package dataresqc is given in the next section).

To ensure full traceability of the QC exercise undertaken and its results, it is important that the original (erroneous) version of the digitized data be kept, allowing others to assess the correctness of the applied QC, ensuring its reproducibility, and it will give hints to improve future QC statistical tests and their application.

2.3.3. The C3S Data Rescue Service land-surface climate data QC tools (package “dataresqc”)

2.3.3.1. Requirements and documentation

Dataresqc is a QC software package developed by the C3S Data Rescue Service. It is mainly intended to promote best practices in the quality control of newly rescued land surface observations.

It requires the R platform, which offers an open source environment for statistical computing (https://www.r-project.org/). R can be downloaded for free and is supported by all common operating systems. Although R can be used through a command line interface, most users prefer to use an integrated development environment (IDE) providing a graphical user interface, for example, RStudio (https://rstudio.com/).

Dataresqc is continuously updated. The latest version and its documentation is available at https://datarescue.climate.copernicus.eu/st_data-quality-control.

2.3.3.2 Input data

Dataresqc is optimized to work with SEF files and it includes functions that facilitate reading and writing the SEF. As an alternative, input data can also be in the form of R data frames. Typically, the data frame containing the observations must be supplemented by an additional data frame with metadata. The exact structure of the data frames is described in the built-in documentation of each function.

2.3.3.3. Workflow

The recommended workflow is summarized in Fig. 1. Here the most common case is described, in which data are typed manually into spreadsheets. A preliminary, manual QC should be performed on the spreadsheets. This step is particularly important if typing is performed by personnel with limited expertise (students, volunteers, etc.).

After the preliminary QC is completed, data should be converted into the SEF (see Sect. 2.2.2). This typically involves writing some code. If the code is written in R or Python, it is recommended to make use of the functions provided by the C3S Data Rescue Service (see example in Sect. 3). It is important to include the available metadata in the SEF files, as compiled in the metadata inventory (see Sect. 2.1).

The SEF files can then be analyzed with dataresqc. The software offers several automatic tests as well as plotting functions for visual inspection. These are described in detail in the next section. In most cases, the thresholds of the automatic tests can be changed by the user to suit the data better.

There should be at least two rounds of quality tests. After the first one, detected digitization errors that can be recoverable must be corrected. This must be conducted in the quality-controlled spreadsheets (or in the spreadsheet produced by the digitizers if no preliminary QC is carried out, taking care to rename the file so that the original version is not lost), which then must be converted again to SEF. The tests can now be repeated, and the values that did not pass any of the tests applied must be flagged using the dedicated function. The final output of the QC will then be SEF files containing data, metadata, and quality flags. Each change made to the original data must be tracked to ensure traceability, by retaining different versions of the data in the native or in an intermediate format.

To summarize, the workflow should be as follows:

1. Apply the quality tests;
2. Check the suspicious values in the original source;
3. Correct typographical errors in a duplicated spreadsheet (or equivalent; ideally one file should be created for each test applied), and/or fix bugs in the formatting procedure;
4. Convert the data again to the common format;
5. Repeat the previous 4 points until no suspicious values are related to the digitization process and only data source mistakes remain;
6. Remove false alarms from the output of the quality tests (i.e., validated values), or add new lines for suspicious values that were not detected by the tests;
7. Add the flags to the formatted data.

2.3.3.4. Overview of the C3S Data Rescue Service QC software “dataresqc”

The package dataresqc is a collection of functions developed in the framework of various international projects (e.g., ERA-CLIM[2], UERRA[3], DECADE[4]). The functions are run from the command line. Therefore, a basic knowledge of the programming language R is required.

Most functions produce either a table of suspicious values or a plot (or both). The table of suspicious values is a text file in which each row represents a suspicious observation, and the last column indicates which tests flagged that observation. One table is produced for each station and each

[2] https://www.ecmwf.int/en/research/projects/era-clim

[3] Uncertainties in Ensembles of Regional Reanalysis: http://www.uerra.eu/

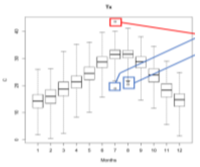
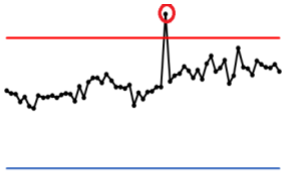
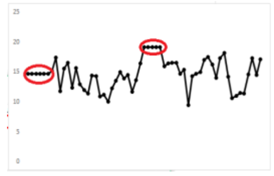
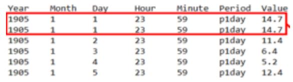
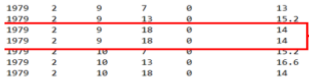
[4] Swiss-Bolivian-Peruvian DECADE project

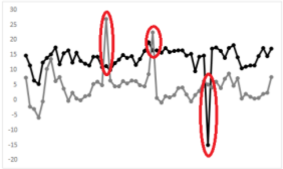
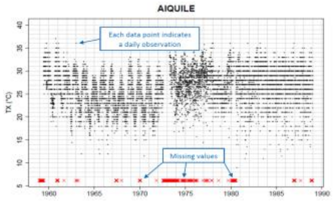
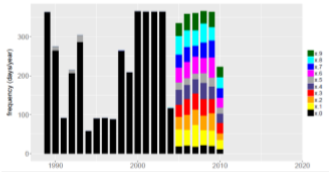
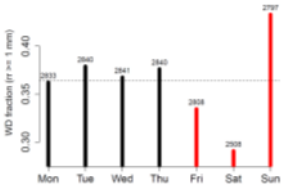
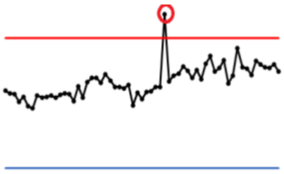
variable. The table of suspicious values can be easily translated into quality flags in SEF files by using the function write_flags.

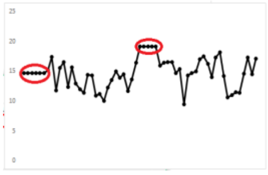
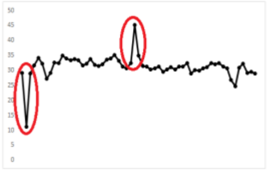
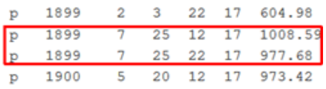
Table 3 gives an overview of the functions provided with the first release of the software. They all apply absolute tests to daily and/or sub-daily (i.e., instantaneous) observations. Additional functions might be added in future versions, including functions to analyse monthly means and to apply relative tests (e.g., spatial consistency). For more detailed information, the reader is referred to the online documentation available on the C3S Data Rescue Service website.

Instructions on how to use each function can be obtained from the R command line by typing the name of a function preceded by a question mark (e.g., '?plot_daily').

Table 3. Functions in the package dataresqc. For variable codes, see Table 1.

Function	Variables	Output	Description	Example
climatic_outliers	Tx, Tn, ta, rr, sc, sd, fs	plot, txt	Detects all values falling outside a range between *p25 - n interquartile ranges* (lower bound) and *p75 + n interquartile ranges* (upper bound). *n* depends on the variable and can be modified by the user.	
daily_out_of_range	Tx, Tn, rr, dd, w, sc, sd	txt	Detects daily values that exceed a determinate threshold set by the user.	
daily_repetition	any daily	txt	Detects equal consecutive values in daily data. The number of minimum equal consecutive values can be modified by the user.	
duplicate_dates	any daily	txt	Detects dates that appear more than once in daily data.	Year Month Day Hour Minute Period Value 1905 1 1 23 59 p1day 14.7 1905 1 1 23 59 p1day 14.7 1905 1 2 23 59 p1day 11.4 1905 1 3 23 59 p1day 6.4 1905 1 4 23 59 p1day 5.2 1905 1 5 23 59 p1day 12.4
duplicate_times	any sub-daily	txt	Detects times that appear more than one in sub-daily data.	1979 2 9 7 0 13 1979 2 9 13 0 15.2 1979 2 9 18 0 14 1979 2 9 18 0 14 1979 2 10 7 0 15.2 1979 2 10 13 0 16.6 1979 2 10 18 0 14
impossible_values	rh, n	txt	Detects values that are outside the physical range of bounded variables.	Negative relative humidity

Function	Variables	Output	Description	Example
internal_consistency	Tx, Tn, w, dd, sc, sd, fs	txt	Detects inconsistencies between pairs of variables (Tx – Tn, w – dd, sc – sd, fs – Tn, sd – Tn).	
plot_daily	any daily	plot	Produces a simple plot of daily data for any variable for direct visual inspection.	
plot_decimals	any	plot	Plots the distribution of decimals for each year. It is particularly suited for temperature data, but it can be used with any variable (some pre-processing might be necessary for bounded variables, e.g., removing the zeros in precipitation data).	
plot_subdaily	any sub-daily	plot	Produces a simple plot of sub-daily data for any variable for direct visual inspection.	January 1816
plot_weekly_cycle	rr	plot	Applies a binomial test to daily precipitation data in order to detect significant weekly cycles in the precipitation frequency. Two types of plots are produced: • one plot referred to each analyzed station showing the frequency of wet days for each weekday; • one plot giving an overview for the whole dataset on an annual scale.	
subdaily_out_of_range	ta, w, dd, sc, sd, fs	txt	Detects sub-daily values that exceed a determinate threshold set by the user.	

Function	Variables	Output	Description	Example
subdaily_repetition	any sub-daily	txt	Detects equal consecutive values in sub-daily data. The number of minimum equal consecutive values can be modified by the user.	
temporal_coherence	Tx, Tn, w, sc, sd, fs	txt	Detects too large differences between the values of two consecutive days. The thresholds can be modified by the user and are different for each variable.	
wmo_gross_errors	Tx, Tn, ta, w, td, p, mslp	txt	Detects observations that exceed the limits for suspect values recommended by the guidelines of the WMO. The limits depend on latitude and season.	Pressure > 1100 hPa
wmo_time_consistency	ta, td, p, mslp	txt	Detects values whose difference from the previous or next observation exceeds the limits for suspect values recommended by the guidelines of the WMO.	 p 1899 2 3 22 17 604.98 p 1899 7 25 12 17 1008.59 p 1899 7 25 22 17 977.68 p 1900 5 20 12 17 973.42
qc	any	txt	This is a wrapper that executes sequentially all the previous functions (except pure plotting functions). The functions are run with default parameters and no plots are produced.	

2.3.3.5. Guiding and advising the application of QC tests

Automatic QC is in most cases not very effective in detecting quality problems (i.e., biases due to poor equipment or resulting from documented or undocumented changes at the site or with the instruments). Each data series has its own peculiarities, and it is important to choose tests and thresholds wisely and to be able to interpret the results. Therefore, user expertise and extensive metadata are fundamentally important for a good QC. This also means that the QC process requires a significant amount of time and resources.

Each quality flag produced by the software has a description, by default the name of the test that caused it. The description can be changed manually in the text files produced by the functions (last column). Ideally, this should be a very short description of the problem (for example: “weekly cycle”).

Flagging an observation is not equivalent to deleting it. Depending on the problem, other data users may decide to ignore certain flags. For instance, the weekly cycle in precipitation (i.e., when manual rain gauges are not emptied at the weekend) can be ignored when calculating monthly totals; temperature in the sunlight could be useful to estimate cloud cover; etc. For this reason, data should never be deleted during the QC process, and the original digitized data must always be kept, since in the future new and more robust tests will be implemented, as well as better approaches to verify, validate, or correct the suspicious values, advances that will benefit from the comparison between the original digitized file and that being quality controlled.

It is recommended to start performing QC early in a data rescue project, when the digitization is still in progress. This can help with spotting systematic problems in the digitization process that would require additional resources if addressed later, or unexpected data issues that could affect future digitization priorities.

Specific recommendations on single tests are given in the software documentation. A practical example is also given in Sect. 3.

2.4. Data submission and consolidation

2.4.1. The C3S 311a Lot 2 Global Land and Marine Observations Database (GLAMOD)

Historical observational climate records are key in understanding climatic variability, extreme events and how climate change signals are manifested (or not) and allow us to make informed decisions to help society better adapt to climate change (e.g., Brunet et al., 2013; Kennedy et al., 2010; Murphy et al., 2017; Noone et al., 2017; Shapiro et al., 2010; Thorne et al., 2018; Wilby et al., 2006). Historical observations are also important for derived reanalysis products (Compo et al., 2011; Dee et al., 2011; Slivinski et al., 2019) and help evaluate and validate climate models (Flato et al., 2013).

Marine surface-based observations can be accessed from the International Comprehensive Ocean-Atmospheric Data Set (ICOADS), which provides a consolidated and integrated set of marine surface data (Freeman et al., 2017). The ICOADS dataset is currently meeting most data user needs. However, the data management situation with land-based surface observations differs considerably. Historically, most of the land-based data holdings that have been produced are either timescale or variable specific and are also either regionally or nationally specific. In addition, many of these holdings may be lacking in completeness or may differ in the specific data quality checks applied. These diverse data holdings mean there are many distinct data formats, gross duplication of stations with differing station identifiers, names and location inconsistencies. There are also issues with verifying data discovery, with many data holdings having a lack of traceability back to the original data source. These current issues with data management make it difficult for users not only in climate science but also in wider disciplines such as water management, ecology and engineering to obtain the full benefits of the available historical land meteorological holdings.

The GLAMOD (https://climate.copernicus.eu/global-land-and-marine-observations-database) aims to address the issues outlined above by producing a comprehensive set of global climate data holdings for both land- and marine-surface domains. These holdings will be integrated across essential climate variables (ECVs) and across time scales (sub-daily, daily and monthly). Initially, the first data releases will contain stations with temperature, pressure, water vapour, wind speed, wind direction and precipitation observations. However, it is planned to introduce other variables as the service develops in the future. Once compiled, the data will be provided via the C3S Climate Data Store (CDS) (https://cds.climate.copernicus.eu/cdsapp#!/home) in a common data model with all available supporting metadata and via the NOAA/NCEI data repository (https://www.ncdc.noaa.gov/data-access).

The GLAMOD team have produced a set of harmonized land and marine data for the first full data release which is expected to be available to users by the end of 2020 via the CDS. The second data release is expected to be available early 2021 with a plan to increase the temporal and spatial coverage in each data release.

GLAMOD team objectives for 2020/2021:

- Continue regular updates of land and marine data inventories including timely updates for selected data streams.
- Release updated global harmonized products for land and marine data.
- Develop and perform quality assurance and quality control checks on core data and maintain temporal consistency between sub-daily, daily and monthly data.
- Assess homogeneity of selected ECVs and provide information on breakpoints and dates in metadata.
- User and service documentation updated to reflect service status.
- Enrichment and harmonization of metadata.

2.4.2. Submitting climate observation data to GLAMOD

The C3S 311a Lot 2 service provides a "Data deposit service" to enable third parties to contribute data to its global databases of land and marine observations. Collections of observations that are successfully uploaded can be consolidated into the global database which is in turn accessible via the CDS.

Access to the Data deposit Server is managed as follows:

1. The Data Provider (DP) requests an account by providing a username and email address at: datadeposit.climate.copernicus.eu
2. The service manager will receive a notification and will either:
   a. Authorize the account, or
   b. Contact the DP to find out more information about their intended use of the service.

3. Upon authorization the DP will receive a notification that allows them to set a password and login to the service.
4. Before being allowed to upload data, the DP will be prompted to confirm that the user can provide a minimum required set of metadata and data.

At this stage, the DP is ready to add a "collection" to the service, and the DP can begin by adding detailed metadata to a form that includes information about:

- Domain
- Short/long name
- Source and usage rights
- Start and end year
- Funding source
- Citations

The DP is then directed to the main "upload" page where files and directories can be added, edited and deleted in preparation for submission. Some directories are pre-generated for purposes such as "data" and "documentation". There is also an option allowing the user to select either FTP or RSYNC as an alternative upload method, which is particularly appropriate for large volumes or numbers of files.

Once the DP has uploaded the necessary files to the server, the DP can then "submit" the collection to the service. It is important to note that the collection is no longer editable by the DP once it has been submitted.

Following submission, the service manager will be notified that the collection is ready to be considered for inclusion into the database. A message will be sent to the DP, explaining that the collection has been received (see Fig. 12 which shows the process workflow for a DP to submit their data collection).

If a DP wants to add to one of their existing collections that have already been submitted, they can do so by logging into their account and following the data submission steps. The DP should also include an updated description of the dataset update in the metadata form. The submitted, updated data will then be merged with the existing data collection and any new metadata will be used to update existing records.

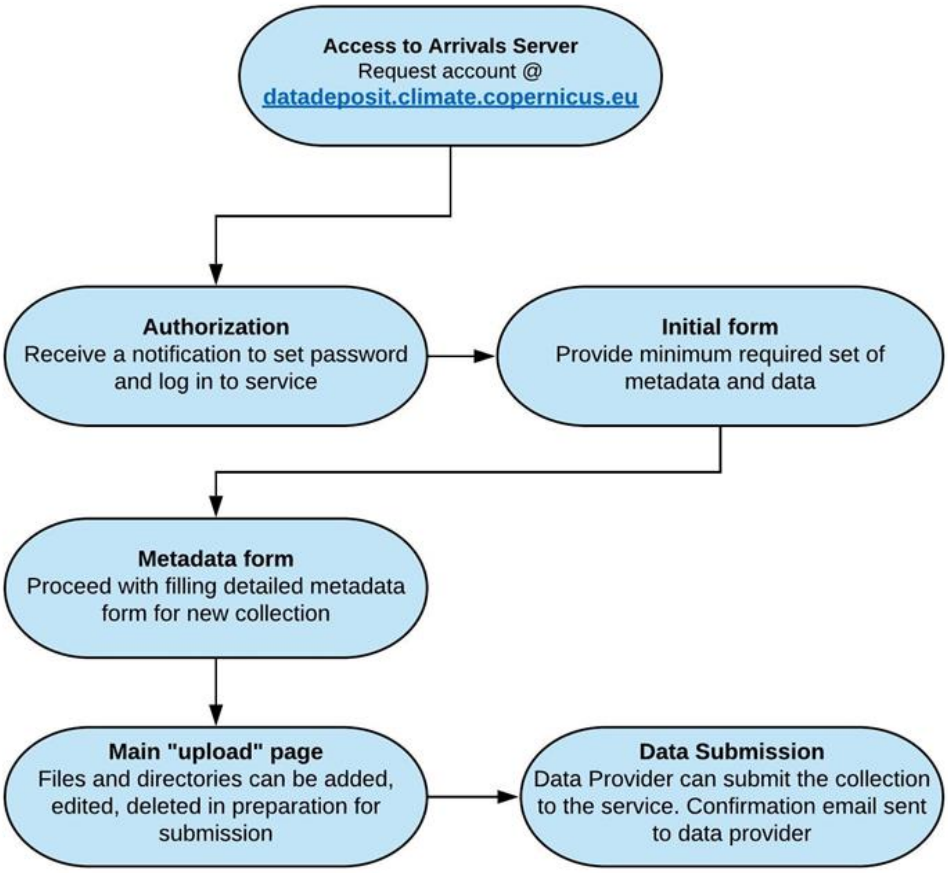

Access to Arrivals Server
Request account @
datadeposit.climate.copernicus.eu
Authorization
Receive a notification to set password and log in to service
Initial form
Provide minimum required set of metadata and data
Metadata form
Proceed with filling detailed metadata form for new collection
Main "upload" page
Files and directories can be added, edited, deleted in preparation for submission
Data Submission
Data Provider can submit the collection to the service. Confirmation email sent to data provider

Fig. 12. Data submission workflow schematic.

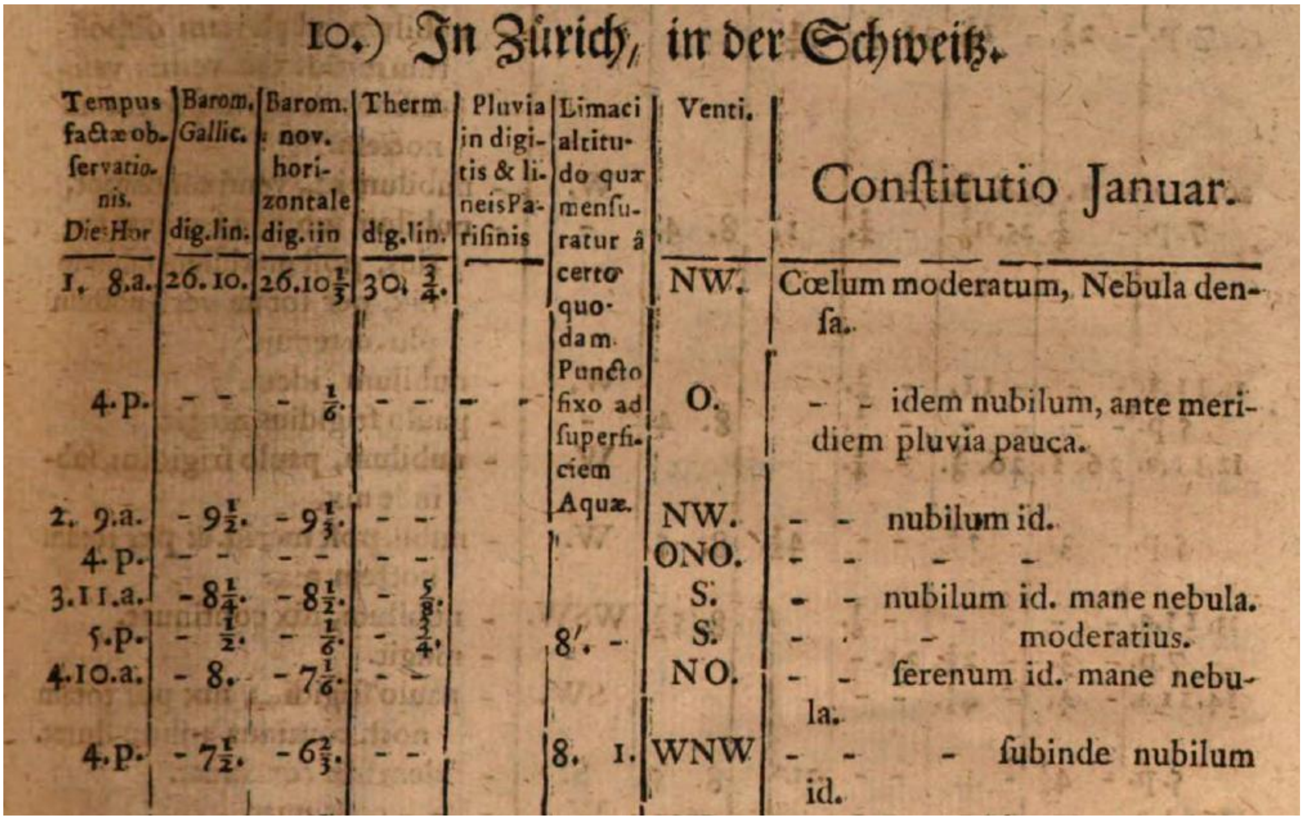
2.4.3. Climate data consolidation

Once data has been uploaded to the service via the Data deposit service portal it will be assessed and prioritized for inclusion in the GLAMOD merge/integration data process and subsequently served through the CDS. This will be more efficient if the data provider has provided the data and the fundamental metadata in SEF. If not, the dataset may not be assessed immediately. It would be ideal to acquire the data in the original raw format and SEF, with all available supporting metadata, including information on any QC checks, station moves etc. However, GLAMOD also accepts homogenized and adjusted data once all supporting documentation on methods and adjustments made are also provided. All source QC flags will be incorporated with internally generated QC flags and provided to the end data user of the CDS. The datasets will be prioritized based on data source provenance information, data access/usage policy, variables available, length of data years, location of stations and other supporting metadata. The process of data merge/integration from multiple sources to produce a set of truly integrated data holdings employs the methods described in Menne et al. (2012).

3. Step by step example of formatting and QC: The Zurich pressure record by J. J. Scheuchzer, 1718-1730

3.1. Formatting

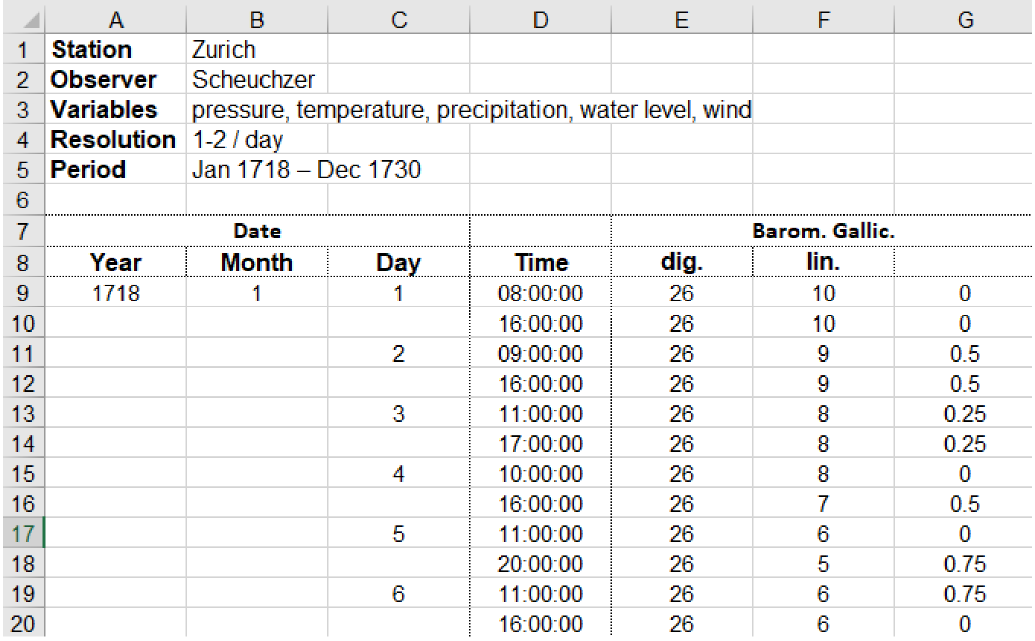
We will work with a pressure record for Zurich. Here is what the source looks like:



10.) In Zürich, in der Schweitz.

Tempus factæ obſervationis. Die Hor	Barom. Gallic. dig.lin.	Barom. nov. horizontale dig.iin	Therm dig.lin.	Pluvia in digitis & lineis Pariſinis	Limaci altitudo quæ menſuratur â certo quodam Puncto fixo ad ſuperficiem Aquæ.	Venti.	Conſtitutio Januar.
1. 8.a.	26.10.	$26.10\frac{1}{3}$	30. $\frac{3}{4}$.			NW.	Cœlum moderatum, Nebula denſa.
4.p.	- -	- $\frac{1}{6}$.	- -	- -		O.	- - idem nubilum, ante meridiem pluvia pauca.
2. 9.a.	- $9\frac{1}{2}$.	- $9\frac{1}{3}$.	- -			NW.	- - nubilum id.
4.p.	- -	- -	- -			ONO.	- - - -
3. 11.a.	- $8\frac{1}{4}$.	- $8\frac{1}{2}$.	- $\frac{5}{8}$.			S.	- - nubilum id. mane nebula.
5.p.	- $\frac{1}{2}$.	- $\frac{1}{6}$.	- $\frac{3}{4}$.		8′. -	S.	- - - moderatius.
4. 10.a.	- 8.	- $7\frac{1}{6}$.	- -			NO.	- - ſerenum id. mane nebula.
4.p.	- $7\frac{1}{2}$.	- $6\frac{2}{3}$.	- -		8. 1.	WNW	- - - ſubinde nubilum id.

The data were digitized in an Excel sheet that looks like this (for the sake of simplicity, times have been formatted to HH:MM:SS and repeated values have been written explicitly):



	A	B	C	D	E	F	G
1	**Station**	Zurich					
2	**Observer**	Scheuchzer					
3	**Variables**	pressure, temperature, precipitation, water level, wind					
4	**Resolution**	1-2 / day					
5	**Period**	Jan 1718 – Dec 1730					
6							
7	**Date**				**Barom. Gallic.**		
8	**Year**	**Month**	**Day**	**Time**	**dig.**	**lin.**	
9	1718	1	1	08:00:00	26	10	0
10				16:00:00	26	10	0
11			2	09:00:00	26	9	0.5
12				16:00:00	26	9	0.5
13			3	11:00:00	26	8	0.25
14				17:00:00	26	8	0.25
15			4	10:00:00	26	8	0
16				16:00:00	26	7	0.5
17			5	11:00:00	26	6	0
18				20:00:00	26	5	0.75
19			6	11:00:00	26	6	0.75
20				16:00:00	26	6	0

To convert the data to SEF 1.0.0 and perform the QC we will use R and the *dataresqc* package version 1.0.3 (this example might not work for more recent versions, please refer to the package documentation). First, we need to read the Excel sheet into a data frame. For this we use the R package *XLConnect*:

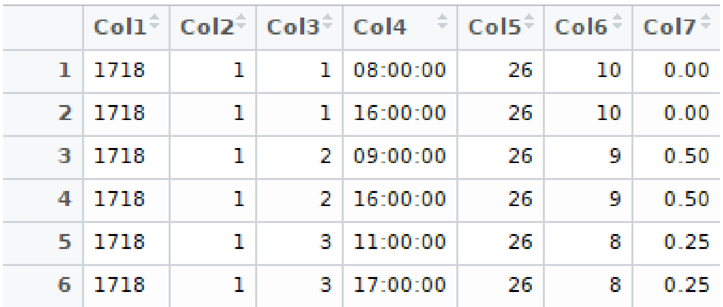
```r
library(XLConnect)
df <- readWorksheetFromFile("Zurich_Scheuchzer_1718-1730.xls", sheet = 1,
               startRow = 9, header = FALSE)
```

Next, we need to fill in the empty cells in the Date columns. For this we create a function "fill_variable":

```r
# Fill missing values with the value on the previous row
fill_variable <- function(timeseries) {
 for (i in 2:length(timeseries)) {
  if (timeseries[i] %in% c(NA, "-")) {
   if(!timeseries[i-1] %in% c(NA, "-")) {
    timeseries[i] <- timeseries[i-1]
   } else if (i > 2) {
    if (!timeseries[i-2] %in% c(NA, "-")) {
     timeseries[i] <- timeseries[i-2]
    }
   }
  }
 }
 return(timeseries)
}

for (j in 1:3) df[, j] <- fill_variable(df[, j])
```

So now the data frame will look like this:



	Col1	Col2	Col3	Col4	Col5	Col6	Col7
1	1718	1	1	08:00:00	26	10	0.00
2	1718	1	1	16:00:00	26	10	0.00
3	1718	1	2	09:00:00	26	9	0.50
4	1718	1	2	16:00:00	26	9	0.50
5	1718	1	3	11:00:00	26	8	0.25
6	1718	1	3	17:00:00	26	8	0.25

For the SEF file, we need the time to be divided into the hour and minute. We create two new variables (“hour” and “minute”) in the data frame and use the package *lubridate* to extract hour and minute from the observation times:

```
library(lubridate)

df$hour <- hour(df[, 4])
df$minute <- minute(df[, 4])
```

Pressure must be converted to hPa. We use the function “convert_pressure” from the package *dataresqc*, which will also correct for gravity. Note that we do not have the temperature of the barometer so we cannot correct for temperature. The converted pressure is stored in the variable “pressure” of the data frame.

```
library(dataresqc)
# First we combine the sub-units to obtain decimal values of Paris inches
# (one line is 1/12 of an inch)
# 1 Paris inch = 27.07 mm
df$pressure <- df[, 5] + (df[, 6] + df[,7]) / 12
df$pressure <- convert_pressure(df$pressure, f = 27.07, lat = 47.37162, alt = 418)
df$pressure <- round(df$pressure, 1)  # we round to 1 decimal place
```

We also need to arrange a column with the original observation and the original time for the meta column in the SEF file (new variable “meta” in the data frame):

```
df$meta <- paste(df[, 5], df[, 6], df[, 7], sep = ".")
df$meta <- paste0("orig=", df$meta, "Pin | orig.time=", df$hour, ":", df$minute, 0)
```

Now we just need to rearrange the data frame for the write_sef function; we create a new data frame “sef”:

```
sef <- data.frame(year = df[, 1],
         month = df[, 2],
         day = df[, 3],
         hour = df$hour,
         minute = df$minute,
         pressure = df$pressure)
```

and finally write the data into SEF:

```
write_sef(sef,
     outpath = getwd(),
     variable = "p",
     cod = "ZH01_Zurich_Scheuchzer",
     nam = "Zürich",
     lat = 47.37162,
     lon = 8.54398,
     alt = 418,
     sou = "CHIMES",
     link = "https://doi.org/10.5194/cp-15-1345-2019",
     units = "hPa",
     stat = "point",
     metaHead = "Observer=Johann Jakob Scheuchzer | PTC=N | PGC=Y",
     meta = df$meta,
     period = 0,
     time_offset = 8.54398 * 12 / 180)  # local solar time based on longitude
```

This is the result:

```
SEF     1.0.0
ID      ZH01_Zurich_Scheuchzer
Name    Zürich
Lat     47.37162
Lon     8.54398
Alt     418
Source  CHIMES
Link    https://doi.org/10.5194/cp-15-1345-2019
Vbl     p
Stat    point
Units   hPa
Meta    Observer=Johann Jakob Scheuchzer | PTC=N | PGC=Y
Year    Month   Day     Hour    Minute  Period  Value   Meta
1718    1       1       7       25      0       968.5   orig=26.10.0Pin | orig.time=8:00
1718    1       1       15      25      0       968.5   orig=26.10.0Pin | orig.time=16:00
1718    1       2       8       25      0       967     orig=26.9.0.5Pin | orig.time=9:00
1718    1       2       15      25      0       967     orig=26.9.0.5Pin | orig.time=16:00
1718    1       3       10      25      0       963.2   orig=26.8.0.25Pin | orig.time=11:00
1718    1       3       16      25      0       963.2   orig=26.8.0.25Pin | orig.time=17:00
1718    1       4       9       25      0       962.4   orig=26.8.0Pin | orig.time=10:00
1718    1       4       15      25      0       960.9   orig=26.7.0.5Pin | orig.time=16:00
1718    1       5       10      25      0       956.4   orig=26.6.0Pin | orig.time=11:00
1718    1       5       19      25      0       955.7   orig=26.5.0.75Pin | orig.time=20:00
1718    1       6       10      25      0       958.7   orig=26.6.0.75Pin | orig.time=11:00
1718    1       6       15      25      0       956.4   orig=26.6.0Pin | orig.time=16:00
```

A similar formatting exercise is provided on the website of the C3S Data Rescue Service for the latest version of SEF.

3.2. Quality control (QC)

First, we use the “qc” function to run all available tests at once. This will give us an idea on how problematic the series is and whether there have been systematic mistakes in the digitization or formatting:

```
qc("CHIMES_ZH01_Zurich_Scheuchzer_17180101-17300630_p.tsv", outpath = getwd())
```

The resulting qc file (qc_ZH01_Zurich_Scheuchzer_p_subdaily.txt) has 317 lines, that is 4% of the observations have been flagged as suspicious. This is a reasonable amount (a percentage larger than 10% would likely indicate a major systematic problem in the data).

Now we delete the qc file, and we run the tests one by one, adjusting the parameters according to our knowledge on the series. In the first round of tests, we will target digitization errors.

There are six tests available for sub-daily pressure in *dataresqc* version 1.0.3 (see also variable “Tests”):

- duplicate_columns
- duplicate_times
- subdaily_repetition
- climatic_outliers
- wmo_gross_errors
- wmo_time_consistency

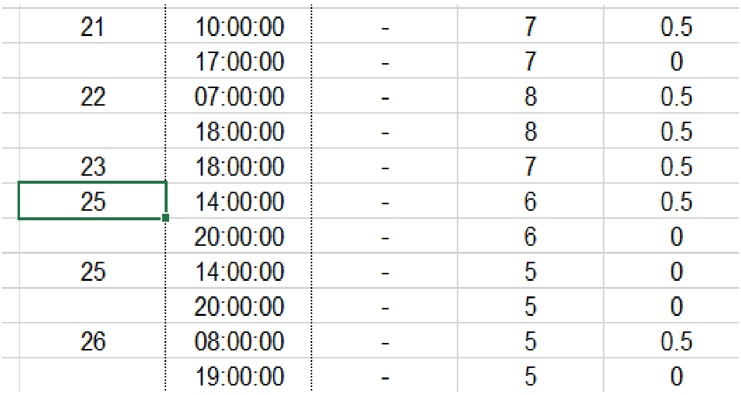
We start with **duplicate_times** (we skip **duplicate_columns** as it is only relevant where observation times are in different columns in the document):

```
duplicate_times("CHIMES_ZH01_Zurich_Scheuchzer_17180101-17300630_p.tsv", outpath=getwd()
)
```

This test finds errors in the dates and times, often arising from a wrong time conversion. Here is the result for our file:

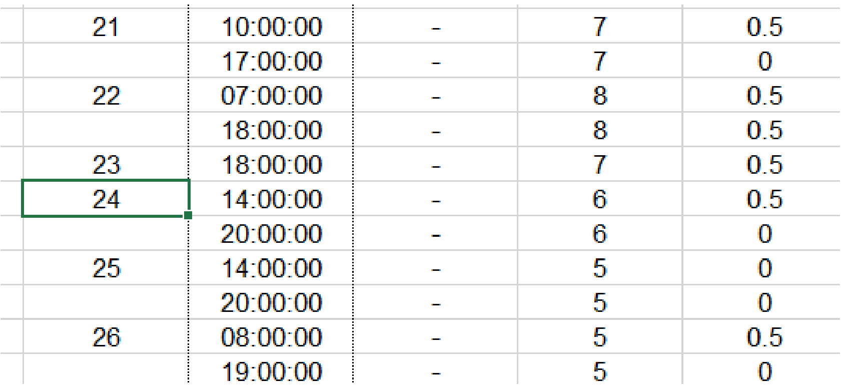
```
Var     Year    Month   Day     Hour    Minute  Value   Test
p       1723    8       25      13      25      957.9   duplicate_times
p       1723    8       25      13      25      953.4   duplicate_times
p       1723    8       25      19      25      956.4   duplicate_times
p       1723    8       25      19      25      953.4   duplicate_times
```

Observations for 25 August 1723 appear twice. This is because of a typographical error in the date:



21	10:00:00	-	7	0.5
	17:00:00	-	7	0
22	07:00:00	-	8	0.5
	18:00:00	-	8	0.5
23	18:00:00	-	7	0.5
25	14:00:00	-	6	0.5
	20:00:00	-	6	0
25	14:00:00	-	5	0
	20:00:00	-	5	0
26	08:00:00	-	5	0.5
	19:00:00	-	5	0

Correcting this mistake is straightforward:



21	10:00:00	-	7	0.5
	17:00:00	-	7	0
22	07:00:00	-	8	0.5
	18:00:00	-	8	0.5
23	18:00:00	-	7	0.5
24	14:00:00	-	6	0.5
	20:00:00	-	6	0
25	14:00:00	-	5	0
	20:00:00	-	5	0
26	08:00:00	-	5	0.5
	19:00:00	-	5	0

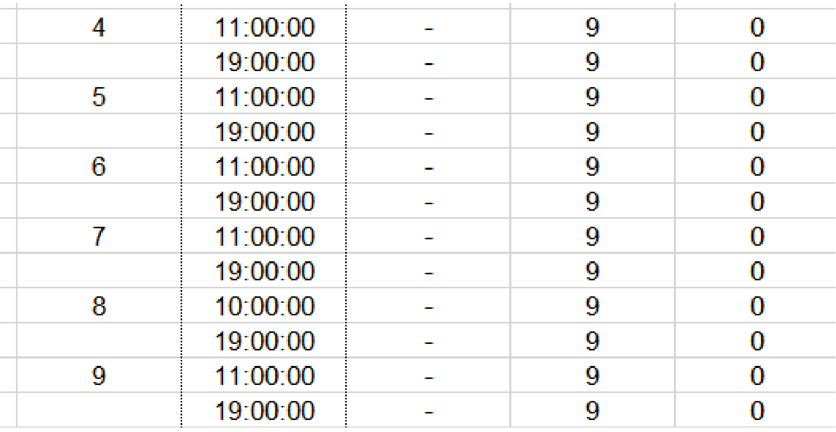
We now save the corrected spreadsheet with a new filename, for example 'Zurich_Scheuchzer_1718-1730_corrected_01_duplicate_times.xls', where '01' stands for 'first test applied' and keeps track of the correction sequence. For the next test that detects digitization errors, in this example wmo_time_consistency (see below), we will rename the file 'Zurich_Scheuchzer_1718-1730_corrected_02_wmo_time_consistency.xls', and so on.

The next test is **subdaily_repetition**. Here we look for values that are repeated for several observations in a row. Since the resolution of the pressure observations in this series is rather coarse (0.5 Paris lines = 1.5 hPa), we can expect consecutive identical observations to happen quite often. Therefore, we change the default setting of the test by increasing the minimum amount of consecutive repeated values required for a flag from the default (6) to 10:

```
subdaily_repetition("CHIMES_ZH01_Zurich_Scheuchzer_17180101-17300630_p.tsv",
         outpath = getwd(), n = 10)
```

We obtain four instances with at least 10 consecutive repetitions, the longest of which has 12 repetitions:

```
p       1721    9       4       10      25      965.5   subdaily_repetition
p       1721    9       4       18      25      965.5   subdaily_repetition
p       1721    9       5       10      25      965.5   subdaily_repetition
p       1721    9       5       18      25      965.5   subdaily_repetition
p       1721    9       6       10      25      965.5   subdaily_repetition
p       1721    9       6       18      25      965.5   subdaily_repetition
p       1721    9       7       10      25      965.5   subdaily_repetition
p       1721    9       7       18      25      965.5   subdaily_repetition
p       1721    9       8       9       25      965.5   subdaily_repetition
p       1721    9       8       18      25      965.5   subdaily_repetition
p       1721    9       9       10      25      965.5   subdaily_repetition
p       1721    9       9       18      25      965.5   subdaily_repetition
```



4	11:00:00	-	9	0
	19:00:00	-	9	0
5	11:00:00	-	9	0
	19:00:00	-	9	0
6	11:00:00	-	9	0
	19:00:00	-	9	0
7	11:00:00	-	9	0
	19:00:00	-	9	0
8	10:00:00	-	9	0
	19:00:00	-	9	0
9	11:00:00	-	9	0
	19:00:00	-	9	0

By consulting the original data source, we see that this is not a digitization error. The same applies to the other instances found by the test. Several days with nearly constant pressure are not uncommon in Zurich, particularly in the warm season, therefore none of the values constitute potential quality issues.

We then run **climatic_outliers**, leaving the parameters set to their default values and choosing to produce a plot:

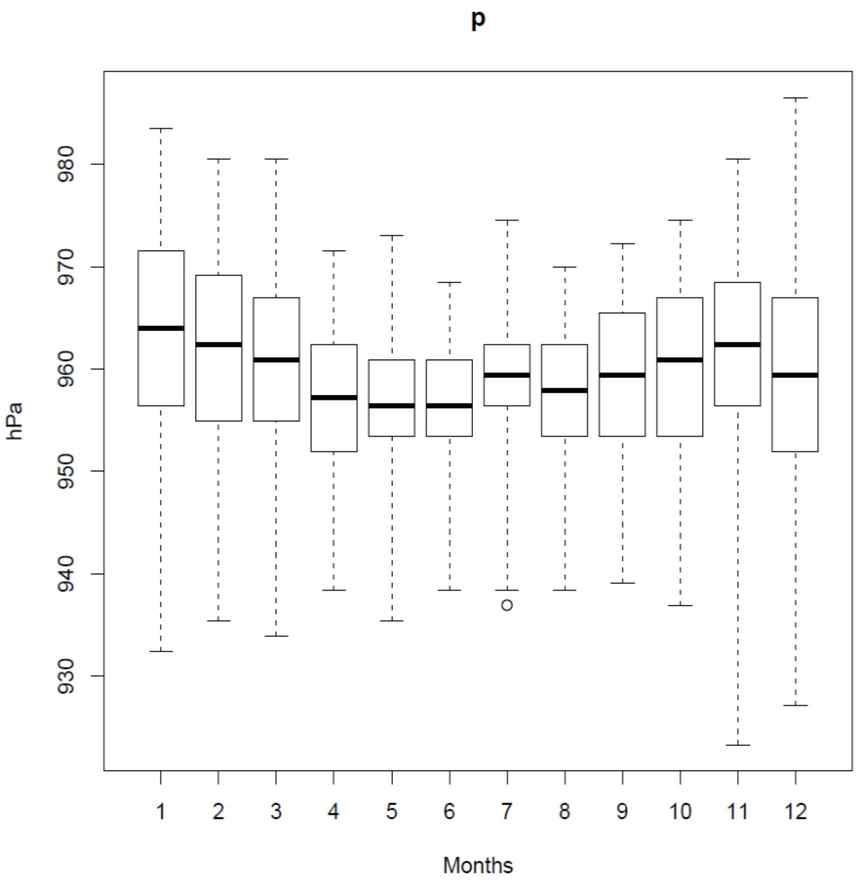
```
climatic_outliers("CHIMES_ZH01_Zurich_Scheuchzer_17180101-17300630_p.tsv",
        outpath = getwd(), bplot = TRUE)
```

No outliers are found. We then try with a lower threshold:

```
climatic_outliers("CHIMES_ZH01_Zurich_Scheuchzer_17180101-17300630_p.tsv",
        outpath = getwd(), IQR = 3, bplot = TRUE)
```

Now one outlier is found in July 1729:

```
p       1729    7       19      15      25      936.9   climatic_outliers
```



We verify that the outlier is not a digitization error. One could try even lower thresholds if the type of data and the underlying climate require it.

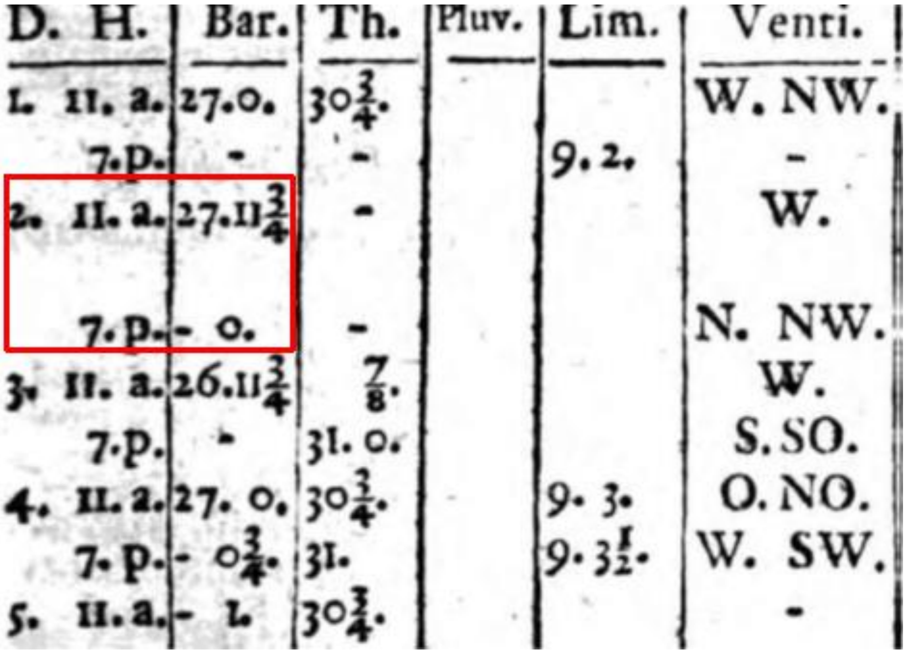
Running **wmo_gross_errors** is not necessary, as we can see in the previous plot that there are no unrealistic pressure values. Finally, we run **wmo_time_consistency**:

```
wmo_time_consistency("CHIMES_ZH01_Zurich_Scheuchzer_17180101-17300630_p.tsv",
           outpath = getwd())
```

This gives us 14 pairs of suspect values:

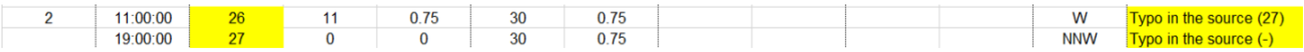
```
Var     Year    Month   Day     Hour    Minute  Value   Test
p       1718    2       12      10      25      968.5   wmo_time_consistency
p       1718    2       12      16      25      939.9   wmo_time_consistency
p       1721    2       2       10      25      973.7   wmo_time_consistency
p       1721    2       2       18      25      938.4   wmo_time_consistency
p       1722    2       26      10      25      971.5   wmo_time_consistency
p       1722    2       26      18      25      939.9   wmo_time_consistency
p       1722    5       21      10      25      965.5   wmo_time_consistency
p       1722    5       21      17      25      938.4   wmo_time_consistency
p       1723    8       25      13      25      953.4   duplicate_times;wmo_time_consistency
p       1723    8       25      13      25      957.9   duplicate_times;wmo_time_consistency
p       1723    8       25      19      25      953.4   duplicate_times;wmo_time_consistency
p       1723    8       25      19      25      956.4   duplicate_times;wmo_time_consistency
p       1723    11      19      20      25      974.5   wmo_time_consistency
p       1723    11      20      7       25      936.9   wmo_time_consistency
p       1725    10      11      10      25      942.1   wmo_time_consistency
p       1725    10      11      16      25      965.5   wmo_time_consistency
p       1726    1       2       16      25      950.4   wmo_time_consistency
p       1726    1       2       19      25      960.9   wmo_time_consistency
p       1728    4       17      9       25      944.4   wmo_time_consistency
p       1728    4       17      18      25      971.5   wmo_time_consistency
p       1728    8       11      18      25      953.4   wmo_time_consistency
p       1728    8       11      19      25      947.4   wmo_time_consistency
p       1729    11      20      17      25      945.9   wmo_time_consistency
p       1729    11      20      18      25      941.4   wmo_time_consistency
p       1730    1       21      10      25      977.5   wmo_time_consistency
p       1730    1       21      20      25      942.9   wmo_time_consistency
p       1730    3       10      10      25      935.4   wmo_time_consistency
p       1730    3       10      19      25      971.5   wmo_time_consistency
```

Two pairs are related to the time duplicate issue that we have already corrected. We need to check the remaining 12 pairs to see if there are digitization errors involved. In one case (2 February 1721), the error is caused by an obvious typographical error in the source:



D. H.	Bar.	Th.	Pluv.	Lim.	Venti.
1. 11. a.	27.0.	30¾.			W. NW.
7. p.	-	-		9. 2.	-
2. 11. a.	27.11¾	-			W.
7. p.	- 0.	-			N. NW.
3. 11. a.	26.11¾	⅞.			W.
7. p.	-	31. 0.			S. SO.
4. 11. a.	27. 0.	30¾.		9. 3.	O. NO.
7. p.	- 0¾.	31.		9. 3½.	W. SW.
5. 11. a.	- 1.	30¾.			-

Here a 27 was printed instead of a 26. We can be sure of this by knowing that the number of inches is only printed if it changes from the previous observation, which was 27.0 on the 1st of February. The 27 should have been printed one line below, in the afternoon observation. When correcting the value on the spreadsheet, we add a note about the typographical error:



2	11:00:00	26	11	0.75	30	0.75					W	Typo in the source (27)
	19:00:00	27	0	0	30	0.75					NNW	Typo in the source (-)

A similar, unequivocal typographical error is found in two other instances. Five of the flagged values are related to actual digitization errors. The remaining four pairs are dealt with in the next round.

Now that all digitization errors have been corrected (at least those that we could detect), we delete both the SEF and the qc file, and we repeat the formatting exercise (see above), with the only difference than we read from the file ‘Zurich_Scheuchzer_1718-1730_corrected.xls’ instead of ‘Zurich_Scheuchzer_1718-1730.xls’.

In a second round of QC, we repeat the tests one by one with the same parameters used before. However, we skip the subdaily_repetition test because we already found that the values detected with that test can be considered correct. We end up with these nine suspect values:

```
Var     Year    Month   Day     Hour    Minute  Value   Test
p       1718    2       12      10      25      968.5   wmo_time_consistency
p       1718    2       12      16      25      939.9   wmo_time_consistency
p       1723    11      19      20      25      974.5   wmo_time_consistency
p       1723    11      20      7       25      936.9   wmo_time_consistency
p       1725    10      11      10      25      942.1   wmo_time_consistency
p       1725    10      11      16      25      965.5   wmo_time_consistency
p       1729    7       19      15      25      936.9   climatic_outliers
p       1729    11      20      17      25      945.9   wmo_time_consistency
p       1729    11      20      18      25      941.4   wmo_time_consistency
```

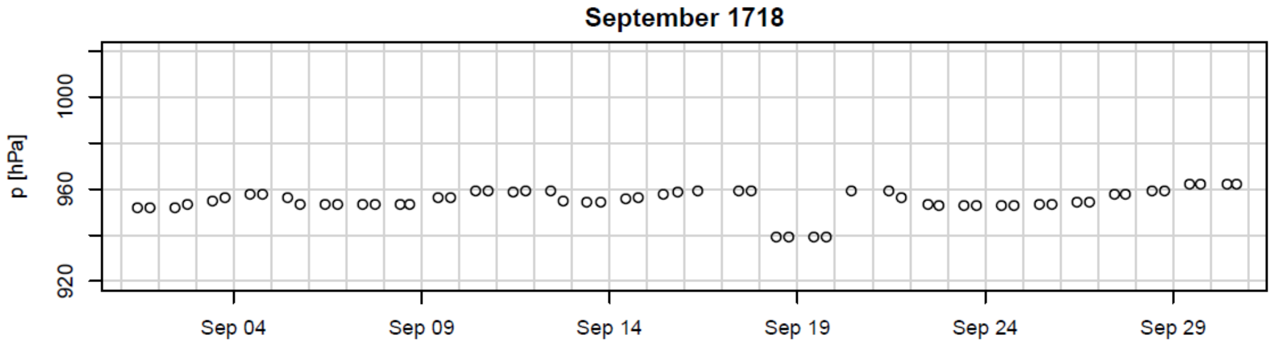
After further inspection, we consider the value detected by the climatic_outliers test to be a valid observation, as it is consistent with the previous and successive observations; therefore, we remove it from the qc file:

```
Var     Year    Month   Day     Hour    Minute  Value   Test
p       1718    2       12      10      25      968.5   wmo_time_consistency
p       1718    2       12      16      25      939.9   wmo_time_consistency
p       1723    11      19      20      25      974.5   wmo_time_consistency
p       1723    11      20      7       25      936.9   wmo_time_consistency
p       1725    10      11      10      25      942.1   wmo_time_consistency
p       1725    10      11      16      25      965.5   wmo_time_consistency
p       1729    11      20      17      25      945.9   wmo_time_consistency
p       1729    11      20      18      25      941.4   wmo_time_consistency
```

As a final check, we can plot the data with:

```
plot_subdaily("CHIMES_ZH01_Zurich_Scheuchzer_17180101-17300630_p.tsv",
       outfile = "Zurich_Scheuchzer")
```

This can help to spot errors that were overlooked by the automatic tests. We find something unusual in September 1718:



The sequence of low values between 18-19 September is probably a typographical mistake in the source (one number was not printed); however, we cannot be 100% sure (and we cannot know the correct value anyway), therefore we add those four observations manually to the qc file (by copying them from the SEF file):

```
Var     Year    Month   Day     Hour    Minute  Value   Test
p       1718    2       12      10      25      968.5   wmo_time_consistency
p       1718    2       12      16      25      939.9   wmo_time_consistency
p       1718    9       18      10      25      939.1   plot_subdaily
p       1718    9       18      18      25      939.1   plot_subdaily
p       1718    9       19      10      25      939.1   plot_subdaily
p       1718    9       19      18      25      939.1   plot_subdaily
p       1723    11      19      20      25      974.5   wmo_time_consistency
p       1723    11      20      7       25      936.9   wmo_time_consistency
p       1725    10      11      10      25      942.1   wmo_time_consistency
p       1725    10      11      16      25      965.5   wmo_time_consistency
p       1729    11      20      17      25      945.9   wmo_time_consistency
p       1729    11      20      18      25      941.4   wmo_time_consistency
```

No further problems are detected. The last step is to then flag in the SEF file the observations listed in the qc file. To do so, we use the function write_flags:

```
write_flags("CHIMES_ZH01_Zurich_Scheuchzer_17180101-17300630_p.tsv",
      "qc_ZH01_Zurich_Scheuchzer_p_subdaily.txt",
      outpath = getwd())
```

A similar exercise on quality control is provided on the website of the C3S Data Rescue Service for the latest version of dataresqc.

References

Aguilar, E., I. Auer, M. Brunet, T. C. Peterson, and J. Wieringa (2003) *Guidelines on climate metadata and homogenization.* WCDMP-No. 53, WMO-TD No. 1186. World Meteorological Organization, Geneva

Allan, R., P. Brohan, G. P. Compo, R. Stone, J. Luterbacher, and S. Brönnimann (2011a) The International Atmospheric Circulation Reconstructions over the Earth (ACRE) Initiative. *Bull. Amer. Meteorol. Soc.,* **92,** 1421-1425.

Allan, R., G. P. Compo, and J. Carton (2011b) Recovery of Global Surface Weather Observations for Historical Reanalyses and International Users. *Eos, Trans. AGU,* **92(18),** 154.

Allan, R., G. Endfield, V. Damodaran, G. Adamson, M. Hannaford, F. Carroll, N. Macdonald, N. Groom, J. Jones, F. Williamson, E.Hendy, P. Holper, P. Arroya, L. Hughes, R. Bickers, and A.-M. Bliuc (2016) Towards integrated historical climate research: the example of ACRE (Atmospheric Circulation Reconstructions over the Earth). *WIREs Climate Change,* **7**, 164–174.

Brunet, M. and P. D. Jones (2011) Emerging data rescue initiatives: bringing historical climate data into the 21st century. *Climate Research,* **47,** 29-40, DOI:10.3354/cr00960

Brunet, M., P. D. Jones, S. Jourdain, D. Efthymiadis, M. Kerrouche, and C. Boroneant (2013) Data sources for rescuing the rich heritage of Mediterranean historical surface climate data. *Geosci. Data J.,* **1,** 61–73.

Compo, G. P., J. S. Whitaker, P. D. Sardeshmukh, N. Matsui, R. J. Allan, X. Yin, B. E. Gleason, R. S. Vose, G. Rutledge, P. Bessemoulin, S. Brönnimann, M. Brunet, R. I. Crouthamel, A. N. Grant, P. Y. Groisman, P. D. Jones, M. Kruk, A. C. Kruger, G. J. Marshall, M. Maugeri, H. Y. Mok, Ø. Nordli, T. F. Ross, R. M. Trigo, X. Wang, S. D. Woodruff, S. J. Worley (2011) The Twentieth Century Reanalysis Project. *Q. J. R. Meteorol. Soc.,* **137,** 1-28.

Dee, D. P., S. M. Uppala, A. J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A. C. M. Beljaars, L. van den Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geer, L. Haimberger, S. B. Healy, H. Hersbach, E. V. Hólm. L. Isaksen, P. Kållberg, M. Köhler, M. Matricardi, A. P. McNally, B. M. Monge-Sanz, J. J. Morcrette, B. K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J. N. Thépaut, and F. Vitart (2011) The ERA-Interim reanalysis:configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.,* **137,** 553–597.

Flato, G., J. Marotzke, B. Abiodun, P. Braconnot, S. C. Chou, W. Collins, P. Cox, F. Driouech, S. Emori, V. Eyring, C. Forest, P. Gleckler, E. Guilyardi, C. Jakob, V. Kattsov, C. Reason, and M. Rummukainen (2013) *Evaluation of Climate Models.* In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T. F., D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P. M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Freeman, E., S. D. Woodruff, S. J. Worley, S. J. Lubker, E. C. Kent, W. E. Angel, D. I. Berry, P. Brohan, R. Eastman, L. Gates, W. Gloeden, Z. Ji, J. Lawrimore, N. A. Rayner, G. Rosenhagen, and S. R. Smith (2017) ICOADS Release 3.0: A Major Update to the Historical Marine Climate Record. *Int. J. Climatol.,* **37**, 2211–2232.

Hyland, R. W. and A. Wexler (1983) Formulations for the Thermodynamic Properties of the saturated Phases of H2O from 173.15K to 473.15K. *ASHRAE Trans.,* **89(2A),** 500-519.

Kennedy, J. J., P. W. Thorne, T. C. Peterson, R. Ruedy, P. A. Stott, D. E. Parker, S. A. Good, H. Titchner, and K. Willett (2010) How do we know the world has warmed? *Bull. Amer. Meteorol. Soc.,* **91(7),** S26–S27.

Menne, M. J., I. Durre, R. S. Vose, B. E. Gleason, and T. G. Houston (2012) An Overview of the Global Historical Climatology Network-Daily Database. *J. Atmos. Oceanic Technol.,* **29,** 897–910.

Murphy, C., C. Broderick, T. P. Burt, M. Curley, C. Duffy, J. Hall, S. Harrigan, T. K. R. Matthews, N. Macdonald, G. McCarthy, M. P. McCarthy, D. Mullan, S. Noone, T. J. Osborn, C. Ryan, J. Sweeney, P. W. Thorne, S. Walsh, and R. L. Wilby (2018) A 305-year continuous monthly rainfall series for the island of Ireland (1711–2016). *Clim. Past,* **14,** 413–440.

Noone, S., C. Broderick, C. Duffy, T. Matthews, R. Wilby, and C. Murphy (2017) A 250-year drought catalogue for the island of Ireland (1765–2015). *Int. J. Climatol,* **37,** 239-254.

Shapiro, M., J. Shukla, G. Brunet, C. Nobre, M. Béland, R. Dole, K. Trenberth, R. Anthes, G. Asrar, L. Barrie, P. Bougeault, G. Brasseur, D. Burridge, A. Busalacchi, J. Caughey, D. Chen, J. Church, T. Enomoto, B. Hoskins, Ø. Hov, A. Laing, H. Le Treut, J. Marotzke, G. McBean, G. Meehl, M. Miller, B. Mills, J. Mitchell, M. Moncrieff, T. Nakazawa, H. Olafsson, T. Palmer, D. Parsons, D. Rogers, A. Simmons, A. Troccoli, Z. Toth, L. Uccellini, C. Velden, and J. M. Wallace (2010) An Earth-System Prediction Initiative for the Twenty-First Century. *Bull. Amer. Meteor. Soc.,* **91,** 1377–1388.

Slivinski, L. C., G. P. Compo, J. S. Whitaker, P. D. Sardeshmukh, B. Giese, C. McColl, P. Brohan, R. Allan, X. Yin, R. Vose, H. Titchner, J. Kennedy, N. Rayner, L. J. Spencer, L. Ashcroft, S. Brönnimann, M. Brunet, D. Camuffo, R. Cornes, T. A. Cram, R. Crouthamel, F. Domínguez-Castro, J. E. Freeman, J. Gergis, E. Hawkins, P. D. Jones, S. Jourdain, A. Kaplan, H. Kubota, F. Le Blancq, T. C. Lee, A. Lorrey, J. Luterbacher, M. Maugeri, C. J. Mock, G. W. K. Moore, R. Przybylak, C. Pudmenzky, C. Reason, V. C. Slonosky, C. Smith, B. Tinz, B. Trewin, M. A. Valente, X. L. Wang, C. Wilkinson, K. Wood, and P. Wyszynski (2019) Towards a more reliable historical reanalysis: Improvements to the Twentieth Century Reanalysis system. *Q. J. Roy. Meteorol. Soc.,* **145,** 2876-2908.

Thorne, P. (2017) *Defining a Common Data Model, C3S_311a_Lot 2: Global Land and Marine Observations Database,* Deliverable C3S_D311a_Lot2.2.1.1. ECMWF Copernicus Report. Copernicus Climate Change Service.

Thorne P. W., R. J. Allan, L. Ashcroft, P. Brohan, R.J.H Dunn, M. J. Menne, P. Pearce, J. Picas, K. M. Willett, M. Benoy, S. Brönnimann, P. O. Canziani, J. Coll, R. Crouthamel, G. P. Compo, D. Cuppett, M. Curley, C. Duffy, I. Gillespie, J. Guijarro, S. Jourdain, E. C. Kent, H. Kubota, T. P. Legg, Q. Li, J. Matsumoto, C. Murphy, N. A. Rayner, J. J. Rennie, E. Rustemeier, L. Slivinski, V. Slonosky, A. Squintu, B. Tinz, M. A. Valente, S. Walsh, X. L. Wang, N. Westcott, K. Wood, S. D. Woodruff, and S. J. Worley (2017) Towards an integrated set of surface meteorological observations for climate science and applications. *Bull. Amer. Meteorol. Soc.,* **98,** 2689–2702.

Valente, M. A. (2019) *Guidelines for inventory metadata standards and formats, Deliverable C3S_D311a_Lot1.2.1_201812.* ECMWF Copernicus Report. Copernicus Climate Change Service.

Venema V, B. Trewin, X. Wang, T. Szentimrey, M. Lakatos, E. Aguilar, I. Auer, J. A. Guijarro, M. Menne, C. Oria, W. S. R. Likeba Louamba, and G. Rasul (2020). *Guidance on the homogenisation of climate station data.* WMO no. 1245, World Meteorological Organization Geneva.

Wilby, R. L. (2006) When and where might climate change be detectable in UK river flows? *Geophys. Res. Lett.,* **33(19),** L19407.

Wilkinson C., S. Brönnimann, S. Jourdain, E. Roucaute, R. Crouthamel & IEDRO Team, P. Brohan, A. Valente, Y. Brugnara, M. Brunet, and G. P. Compo (2019) *Best Practice Guidelines for Climate Data Rescue.* Copernicus Climate Change Service.

WMO/CIMO (2014) *Guide to Instruments and Methods of Observation. Part I. Measurement of meteorological variables.* WMO no. 8, World Meteorological Organization, Geneva.

WMO/WIGOS (2015) *The WIGOS metadata standard.* World Meteorological Organization, Geneva, http://www.wmo.int/pages/prog/sat/meetings/documents/IPET-SUP-1_INF_05-02_WIGOS-Metadata-Standard-V0.2.pdf

WMO/WIGOS (2017) *WIGOS Metadata Standard*, WMO no. 1192, World Meteorological Organization, Geneva.

Acronyms

ACRE	Atmospheric Circulation Reconstructions over the Earth
C3S	Copernicus Climate Change Service
CDM	C3S Common Data Model
CDS	C3S Climate Data Store
CSV	Comma-separated values
DARE	Data Rescue
DECADE	Data on climate and Extreme weather for the Central Andes, Swiss National Science Foundation project
ECV	Essential Climate Variable
ERA-CLIM	European Reanalysis of Global Climate Observations EU-project
GCOS	Global Climate Observing System
GLAMOD	Global Land and Marine Observations Database
I-DARE	International Data Rescue
IEDRO	International Environmental Data Rescue Organization
MEDARE	WMO Mediterranean Data Rescue Initiative
NMHS	National Meteorological and Hydrological Services
QC	Quality Control
SEF	Station Exchange Format
UERRA	Uncertainties in Ensembles of Regional Reanalysis
WMO	World Meteorological Organization
WMO/CIMO	WMO Commission for Instruments and Methods of Observation
WMO/WIGOS	WMO Integrated Global Observing System