ECMWF Copernicus Procurement

Invitation to Tender

Copernicus Climate Change Service

Volume II

Rescue, Collection, and Processing of In Situ Observations

ITT Ref: C3S2_311_bis
ISSUED BY: ECMWF Administration Department Procurement Section
Date: 24 March 2025
Version: Final

Funded by the European Union

Implemented by ECMWF

Table of Contents

1 Introduction

ECMWF as the Entrusted Entity for the Copernicus Climate Change Service (C3S) invites tenders for services related to the rescue, collection, and processing of in situ observations in support of climate services. The term in situ is to be understood in the sense that it takes within the World Meteorological Organization (WMO) Integrated Global Observing System (WIGOS), i.e., any observation acquired by surface-based observing stations and platforms, and excluding observations collected by satellites. The overall objective for these services is to improve access to available in situ instrumental data records and to data streams from observing networks, as needed for climate change monitoring and climate science. The technical work described in this document shall provide enhanced continuation of prior activities carried out in earlier phases of the Copernicus programme, as described at https://climate.copernicus.eu/observations. All relevant results to date are available to Tenderers, including documentation, on request via ECMWF. Additional technical clarifications can be provided on request, and questions should be raised via the Messaging Board on the ECMWF eProcurement Portal. To avoid duplication of work, synergies with other Copernicus services shall be exploited.

Goals to be addressed by the services described in this document include: improvement of the historical instrumental record by means of support to data rescue activities, quality control and homogenisation of in situ observations; harmonisation of access to data from major archives of climate observations as well as specialised networks of observing sites; development of merged collections of in situ observations suitable for climate reanalysis; development of high-resolution gridded datasets from in situ observations over Europe that are suitable for climate monitoring.

Central requirement is the open and free provision of in situ observations via the unified C3S Climate Data Store[1] (CDS; Raoult *et al.*, 2017). Scientific requirements on observations and derived information products for the CDS are based on the framework provided by the Global Climate Observing System (GCOS). The CDS provides information about the past, present and future climate in terms of Essential Climate Variables (ECVs; Bojinski *et al.*, 2014) and derived climate indicators. The CDS also provides a comprehensive set of software (earthkit[2]) which enables users to develop custom-made applications. These applications make use of the content of the CDS to analyse, monitor and predict the evolution of both climate drivers and impacts.

Following Thorne *et al. (*2017), the collection of in situ climate records can be grouped into three categories: (1) from reference networks that provide metrologically traceable observations and are, though sparse, suitable for the absolute calibration of Climate Data Records (CDR) for ECVs, (2) from baseline-observing networks, which lack full traceability but do provide long-term records and (3) comprehensive; the much larger set of any observations that are not contained in the first two categories. The present service shall cover in situ observations from all these categories, as each has its own climate user base.

Since the previous contracts on in-situ activities lead by C3S since its inception, several important developments have taken place. Following substantial activities to digitally image records at risk from the African Centre of Meteorological Applications for Development (ACMAD), the need for an open image repository has emerged, to preserve Image Data Records (IDR) in a form that is suitable for downstream transcription to CDR and further iterations/feedback. Also, it is now possible to benefit from additional observations to increase the spatio-temporal coverage and richness (diversity) of observables with near-real-time (NRT) updates from the ECMWF WMO Information System (WIS) Global Telecommunication System (GTS) node following adoption of the WMO Unified Data Policy[3] and emerging open-access data. Technical

[1] https://cds.climate.copernicus.eu/

[2] https://cds.climate.copernicus.eu/earthkit

[3] https://wmo.int/wmo-unified-data-policy-resolution-res1

solutions such as RODEO[4] are emerging as a result of the Directive (EU) 2019/1024[5] of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information and comparable national or local regulations, prompted by wide international recognition of the value of such opening (e.g., G8 Open Data Charter of 2013). Also, to accommodate station timeseries, the CDS includes since 2024 an Observation Repository (see section 5.2.1) that implements a Common Data Model for observations[6]. From the station timeseries, gridded products shall be obtained, after suitable spatio-temporal aggregation. Both types of products are covered by this tender. An important element of this new contract is the integration of the observations as a single value chain, aiming to extend the service closer to present times to empower the climate monitoring or climate attribution activities that require 'recent' data, with proper labelling of differences in quality ('reprocessed' versus 'initial').

The contract shall provide enhanced continuity for the services already available to users on the Climate Data Store: a data rescue portal and several in-situ observation datasets. The contract shall target observation datasets with open, unrestricted access, in accordance as much as possible with the open and free principles of the Copernicus Licence. This requires provision of the data in well-organized online repositories that can be queried directly by users via a simple CDS interface. Specific requirements are provided in section 3.3 and Tenderers should read those carefully since they will have to interface to the technical solution described therein. Services shall include a guarantee on uptime, provision of up-to-date documentation, expert user support and contribution to the C3S evaluation and quality control function (EQC) and Climate Intelligence (CI). The Successful Tenderer shall work with data providers as needed, liaise between network providers and ECMWF. On the other hand, the Successful Tenderer shall ensure that Intellectual Property Rights (IPRs) are respected for all public-facing data on the CDS. In addition, the Successful Tenderer shall engage in outreach and training activities to maximise the impact of this activity.

Although the in-situ observation datasets will serve a wide range of users and applications, a particular emphasis shall be on the support for climate reanalysis. This entails the selection of datasets in regions and periods where most value to reanalysis is added, emphasis on particular geophysical quantities and the provision of adequate metadata.

Specific objectives and technical requirements are described in section 2 of this document. General requirements are presented in section 3. Information about the tender format and content is given in section 4, and section 5 contains a list of reference documents, appendices and acronyms.

2 Technical Requirements

ECMWF intends to award a single multi-annual service contract (for a duration indicated in section 3.1) for services in the areas of in situ data rescue, the provision of station timeseries data from comprehensive, baseline, and reference networks of high value for climate activities, the provision of high-resolution gridded datasets, covering regions and variables of interest to European citizens to tackle climate change issues, and more generally to enhance the value of services delivered by C3S. The activities shall also support the development of the GCOS with a specific action on data obtained from Global Navigation Satellite Systems (GNSS).

[4] EU Digital Europe Programme project on the Provision of Open Access to Public Meteorological Data and Development of Shared Federated Data Infrastructure for the Development of Information Products and Services, a joint effort by eleven European National Meteorological and Hydrological Services (NMHS), ECMWF, and the network of 31 European NMHS (EUMETNET)

[5] https://eur-lex.europa.eu/eli/dir/2019/1024/oj

[6] https://github.com/ecmwf-projects/cdm-obs

2.1 Scope of Service

The Successful Tenderer shall:

1. organise and facilitate efforts to digitise documented instrumental historical data records,
2. support selected data rescue projects with academia, public, and private sectors,
3. provide access to digital images of data rescue material at risk,
4. forward data rescue transcribed results to internationally-recognised data archives,
5. develop and implement solutions to expose progress in data rescue activities,
6. make rescued station data visible at WIGOS highest level,
7. service in-situ-observation-based station climate timeseries on the CDS, with regular exchanges with comparable international archives as relevant,
8. service in-situ-observation-based gridded climate products on the CDS,
9. maintain a global repository of GNSS fundamental records acquired by ground stations,
10. engage in relevant international fora to ensure uptake of the overall activity and efficient contributions to the GCOS,
11. contribute to cross-cutting C3S activities: Evaluation and Quality Control, Climate Intelligence, user training, and communications and outreach.

2.2 Specification of Work, by Work-Package, including Schedule of Deliverables

The tasks to be carried out may be divided in four themes: climate data rescue, station climate timeseries, gridded climate products, and additional contributions to the C3S value chain, as well as management. Each theme is mapped into a contract work-package (WP). The tasks are numbered within each WP.

In each task, deliverables can be required to be "Annual", in which case this amounts to 2 outputs over the entire duration of the contract, given the maximum foreseen contract duration of 34 months; the extra months shall be used to distribute the load and stage deliverables in a way that avoids that annual deliverables are all due within the same month. Other deliverables, apart from annual, may either be half-yearly (every 6 months), or single deliverables. In the latter case, the requirements thereafter may include an indicative timing, "H1" refers to the first half of the contract duration, and "H2" refers to the second half of the contract duration.

2.2.1 WP0: Management and Coordination

All management and coordination activities for the services shall be included in a single work package. The following activities shall be listed and described in the Tenderer's proposal:

- Planning, coordination, and monitoring of all technical Work Packages activities and corresponding resources.
- Contractual obligations as described in the Volume V Framework Agreement, especially in its Clause 2.3 "Reporting and Planning" and Annex 5 "Report content".
- Organising and attending meetings:
    - ECMWF and the Successful Tenderer will organise a Kick-Off Meeting during the first month of the contract's implementation.
    - ECMWF and the Successful Tenderer's Service Manager and Technical Lead (two different roles, must be held by two different persons) will organise Progress Review Meetings (teleconferences), linked to Payment Milestones, on a quarterly basis unless otherwise agreed.
    - ECMWF will organise annual C3S General Assemblies. The Successful Tenderer is required to participate in these meetings, with the sole requirement being the presence of the Service Manager or an appointed representative of the full service. The Successful Tenderer will organise annual in-person Service Readiness Review meetings, where at least a

representative per dataset proposed will be required. The main purpose is to conduct a thorough review of the status of every dataset. The meeting also aims to identify any potential issues and make decisions about moving forward based on the findings, lessons learnt, and recommendations presented.

- As needed, ECMWF will host additional teleconference meetings to discuss service-specific or technical topics. The Service Manager and/or Technical Lead and/or a representative of the WP under discussion (as relevant) will represent the Successful Tenderer in such meetings.
- The Tenderer can propose additional project meetings, whose added value must be precisely substantiated, as part of their Tender.

- Quality assurance and control as well as risk management, including in what concerns the Sub-contractors' activities if any. The final quality check of all deliverables (contents, use of ECMWF's templates for deliverables and reports, format, deliverables/milestones numbering and naming, typing errors, etc.) shall be made by the Successful Tenderer.
  - The Tenderer shall outline and justify the proposed management methodology for this contract in its technical proposal. The Tenderer shall provide a list of its quality assurance processes and management systems and if applicable, any quality related accreditations or certifications it holds. The Tenderer shall also describe in its Tender how the Volume V Framework Agreement, in particular Clause 2.9 "Sub-contracting", has been flowed down to all their Sub-contractors.
- Proactive and dynamic communication towards and between all parties involved in the contract.
- Management of personal data and how this meets the requirements of Clause 2.8 and Annex 6 "Personal Data Protection" of the Volume V Framework Agreement.
- Sub-contractor management in accordance with the Volume V Framework Agreement, including dispute resolution (the Successful Tenderer will be responsible for settling disagreements, although advice/approval from ECMWF may be sought on the subject).

The table hereinafter provides the Tenderers with the complete list of deliverables and milestones, as well as the corresponding schedule and due dates, for WP0:

Deliverable / Milestone ID	***Resp.***	***Nature***	***Deliverable / Milestone title***	***Due date***
List of deliverables				
WP0-QIR-YYYYQQ	Tenderer	Report	Quarterly Implementation Report YYYYQQ *YYYYQQ being here the previous quarter (e.g. 2024Q3)*	Quarterly on 15/04, 15/07 and 15/10
WP0-AIR1-YYYY	Tenderer	Report / Other	Annual Implementation Report for year YYYY – Part 1 *including both:* · *the Quarterly Implementation Report YYYYQ4 and* · *the requested preliminary financial information for year YYYY* *YYYY being here the Year n-1*	Annually on 15/01
WP0-AIR2-YYYY	Tenderer	Report	Annual Implementation Report for year YYYY – Part 2	Annually on 28/02

			YYYY being here the Year n-1	
WP0-FIR	Tenderer	Report	Final Implementation Report	Not later than 60 days after the end of contract and once all other activities duly performed
WP0-AIP-YYYY	Tenderer	Report	Annual Implementation Plan for year YYYY *YYYY being here the Year n+1*	Annually on 30/09
WP0-FIN-YYYY	Tenderer	Other	Copy of Prime Contractor's general financial statements and audit report for year YYYY *YYYY being the Year n-1*	Annually by 30/06 (no associated cost)
WP0-KOM	Tenderer	Presentation and MoM	Kick-Off Meeting	Not later than 30 days after the start of contract
WP0-PRMxx	Tenderer	Presentation and MoM	Progress Review Meeting #xx *xx being the iteration number of the PRM*	Circa every 3 months(1)
List of milestones				
WP0-C3SGA-YYYY	Tenderer	Attendance	C3S General Assembly YYYY *YYYY being here the concerned year*	Annually, not later than on 15/12(1)

Table 1: WP0 Deliverables and Milestones

(1) These due dates are indicated to frame the corresponding deliverables and milestones schedule only, consequently the following shall be considered by the Tenderer:

- the general financial statements shall be sent by the Successful Tenderer as soon as available,
- the schedule of the Progress Review Meetings shall be aligned with the different Payment Milestones during the contract negotiation (i.e. each Payment Milestone shall have at least one corresponding Progress Review Meeting),
- depending on the year, the C3S General Assembly may take place at a different period of the year.

2.2.2 WP1: Climate Observations Data Rescue

The objective of this work-package is to recover climate observations, considering physical or digital media for which transcribed records are not yet available in digital archives. To achieve this objective, several elements need to be improved along the data rescue chain. Several tasks are specified as follows, along with deliverables and their estimated number in parentheses. Several terms indicated in bold (the first time they are mentioned) are defined in section 2.5 of the present ITT Volume II.

Task 1.1: Data rescue coordination

This task requires to continue the coordination role taken by C3S in association with WMO International Data Rescue (I-DARE), operating the existing Joint WMO-C3S Data Rescue Service Portal (DRSP). Data rescue activities covered in this task are geared towards digitisation of documented instrumental data records extending back a century or more. The portal must be maintained and kept operational, and new users shall continue to be able to register or else send updates for keeping the projects up-to-date.

A review on the status of all projects shall be made, at a frequency to be agreed with the other partners of the portal. The presentation on the website should make it possible to easily display only projects updated within a time period that is on par with this frequency. This should help make more visible projects that progress, and conversely assist in spotting projects that are stalled and in possible need of assistance. The task must include provision of technical and user support.

This task also calls for animation of the data rescue community, in a shape and format that is left at the initiative of the Successful Tenderer: e.g., web forum, social media posts, special events at conferences. Prime partners of the portal as ACRE and IEDRO shall be associated to these actions.

Deliverables required (2): **Annual Report (AR)** on Joint WMO-C3S Data Rescue Service Portal operations and updates and data rescue community animation actions.

Task 1.2: Safeguarding climate image data records

Expanding on concepts defined in earlier work (see pre-existing input underneath), a repository of digital images obtained from climate records at risk of disappearance or decay shall be defined and developed, and a prototype shall be put in function. The capacity shall cover at least the equivalent of twice the size of the ACMAD collection(i.e., 2 x 64 TB = 128 TB). The repository may be populated with this collection, and/or other ones, depending on images availability.

The repository shall be designed to handle Image Data Records (IDR). For example, geolocation tags and temporal tags shall be enabled. The repository shall also aim to support storing of annotations, for example for images whose contents are only partially transcribed. This layer of information would allow later a feedback step, e.g., to verify correspondence between the transcribed numbers and the original hand-/type-writing or printing, and to train/improve advanced transcription algorithms.

For technical implementation, a partnership with the private sector, e.g. internet cloud providers, may be sought.

Deliverables required (2): H1: **System Requirements Document (SRD)** for the C3S repository of image data records; H2: Prototype instance of the C3S repository of image data records, with documentation.

Task 1.3: Data deposition service

This task requires to continue operating the existing C3S Data Deposition Service (DDS) and to improve its readiness to become part of an automated workflow in the long-term. This facility presents a web interface through which data providers can be initially authenticated and then connect to upload datasets for potential inclusion into larger collections. Data can be pushed by a range of transfer protocols. They must then be manually assessed/translated or else input into a traceable workflow so the key quality control/translation steps can be traced with a reproducible data ingestion tool (e.g., Kent *et al.*, 2025). This service should allow for the harvesting of new sources of historical observations from the land surface and the marine surface, before they can be included into future versions of international data collections (see Tasks 2.3 and 2.4 of the present ITT Volume II). The Successful Tenderer shall operate this service, document the current requirements that it should be serving, conduct initial quality assessments on data received, forward to international data collections, and report annually on the progress. Efforts to feed-back information to the initial providers, with, e.g., unique identifiers and a deposit certificate, shall be pursued to enhance transparency and enhance user engagement. To prepare for this in the medium-term, this task is responsible in this activity to evaluate the outcomes of Tasks 1.5 to 1.7, by providing a **Data Receipt Acceptance Report (DRAR)** in each case.

Deliverables required (6): System Requirements Document for the C3S Data Deposition Service; Annual report on records harvested by the C3S Data Deposition Service; Data Receipt Acceptance Report for data submitted by Task 1.5; Data Receipt Acceptance Report for data submitted by Task 1.6; Data Receipt Acceptance Report for data submitted by Task 1.7.

Task 1.4: Inventorying climate data rescue station timeseries

Station timeseries inventories, initially for the land surface component, shall be made available online, constituting a new facility. This facility shall offer visual means (e.g., interactive map) of patching the information gap regarding exposing the available data and the need for data rescue. Such inventories exist already for large international data collections such as the C3S comprehensive land surface dataset. The facility shall enable import and export of such tabulated information. The information should be displayed on a web page in a way that it should address the most recurrent question whether or not data from a particular location have already been rescued, or may be being rescued (e.g., if there is a registered project relevant to the location of interest, if there are relevant images in the image repository, or if records were submitted to the deposition service). The inventories will need to consider, at the minimum, the station locations, with various possible names and identifiers, the variables observed, the reporting times (schedule), and the dates of availability. This task will need to make multiple links, (a) starting with existing archives such as Task 2.3 to establish availability (and possibly the ability to form a query allowing to obtain the data, e.g. for more detailed inspection of the exact reporting dates/times sequence), (b) with the data deposition service in Task 1.3 (for records that were submitted there), (c) with the image repository in Task 1.2 (if images are known to exist in place of yet-transcribed records), and (d) with the projects inventory in Task 1.1 (possibly via the current country name, or a spatiotemporal search function). Given that nearly all elements of this ecosystem are 'work in progress', the Contractor will need adopt a pragmatic selection in the number of metadata that this inventory covers and that the links that are developed, keeping in mind that WMO Observing Systems Capability Analysis and Review Tool (OSCAR) Surface[7] is the natural endpoint for all station metadata. The station inventories requested here are not intended to replicate this capacity, but only to guide the data rescue, and allow to better allocate resources to this end. The result may be rudimentary to begin with, but most importantly it should allow interested users to find where further answers to their questions may be found. The resulting inventories may be linked from the data rescue portal. This facility shall cover the land surface component, and, optionally (by decreasing order of priority), the surface marine component, and the upper-air balloon ascents component.

Deliverables required (2): H1: System Requirements Document for the C3S station timeseries inventory facility; H2: Prototype instance of the C3S station timeseries inventory facility, with documentation.

Task 1.5: Data rescue actions with the public sector

At least one project registered on the I-DARE portal or known to the Successful Tenderer shall be selected for providing financial assistance, and the work must be carried out by a public sector entity. Both the inputs (when in digital form, data policy permitting, and if not already available online) and the outputs (transcribed records) shall be made easily accessible for further processing and analysis, to (respectively) the image repository and the data deposition service. The WMO guidelines on data rescue must be applied. The timeline of completion of this task must leave sufficient time to Task 1.3 to issue a corresponding Data receipt acceptance report before the contract ends.

Deliverables required (3): H1: Mid-term report on the data rescue action; H2: **Manifest Of Data Submitted (MODS)** to the Data Deposition Service; H2: Report on the methods used for data rescue.

Task 1.6: Data rescue actions with academia

Similarly to task 1.5, at least one project registered on the I-DARE portal or known to the Successful Tenderer shall be selected for providing financial assistance, except the work must be carried out by an academic entity. Another difference is the last deliverable, the task shall improve the maturity of data rescue educational or pedagogical material, so it can be used as part of higher education curriculums. The other requirements laid out for Task 1.5 apply.

[7] https://oscar.wmo.int/surface/

Deliverables required (3): H1: Mid-term report on the data rescue action; H2: Manifest Of Data Submitted to the Data Deposition Service; H2: Compendium of material or references suitable for data rescue in an educational context.

Task 1.7: Data rescue actions with the private sector

Similarly to task 1.5, at least one project registered on the I-DARE portal or known to the Successful Tenderer shall be selected for providing financial assistance, except the work must be carried out by a private sector entity. Another difference is the specific requirement for this task to employ AI-based methods for part of the work, which could include transcription, but not only. Other use of AI include, for example, automatic identification of the forms to be transcribed or verification of the results (outliers/consistency). The other requirements laid out for Task 1.5 apply.

Deliverables required (3): H1: Mid-term report on the data rescue action; H2: Manifest Of Data Submitted to the Data Deposition Service; H2: Report on the methods used for data rescue.

Task 1.8: Registration of stations in WIGOS

This task registers on behalf of C3S the stations with rescued or published digital records and that are no longer operating (noting the term 'station' is used loosely here, to designate also moving platforms, including over oceans) into WMO OSCAR/Surface. This task shall involve in association the Institutions mandated by WMO to be responsible in this area. For example, ad hoc partnerships or shall be sought to cover special areas such as marine platforms, and land stations over mountains. The task should result in C3S-minted WIGOS station identifiers (WSIs). This work should be undertaken at least for land stations, and optionally (by decreasing order of priority) for marine platforms and for upper-air radiosondes, assembling metadata as required by WIGOS.

Deliverables required (2): Annual report on stations registered by C3S in WMO OSCAR/Surface.

2.2.3 WP2: Station Climate Observation Timeseries

The objective of this work-package is to increase the quality and quantity of climate observations readily available to support climate research. To meet this objective, records from multiple sources must be acquired, reconciled in case of duplicates, to form multivariate station timeseries. The data must be subject to state-of-the-art quality controls and the results must be made available on the CDS. Several deliverables mentioned subsequently in this WP are defined in the present ITT Volume II section 3.3.1. This work-package includes two 'service' tasks and six 'dataset' tasks. The expected deliverables are listed for each task (with the number of deliverables in parentheses).

Task 2.1: Evolution of the common data model for observations

The common data model for observations (CDM-Obs), as well as its "core" extract, developed in previous contracts, shall continue to be maintained in GitHub. Tables shall be reformatted so they can be viewed in a user-friendly way on the web-based code repository (without need for a rendering step to transform this code into a documentation). Requests logged by datasets providers via the GitHub platform to update the data model shall be answered within one calendar month, either by providing a development branch that includes changes to the CDM-Obs, or by providing alternate solutions that do not require changes to the CDM-Obs. Changes that have consequences on existing dataset entries served by the CDS shall be propagated to the relevant dataset entries for testing before implementation. It is recommended to adopt a versioning system to tag 'frozen' versions used for production. In addition, the task requires that the Successful Tenderer takes part in the relevant WMO instance (presently the Task Team on Climate Data Models, TT-CDM) to ensure continued inter-operability with evolving standards. The documentation of the CDM-Obs shall be entirely contained at the same location on GitHub. Technical solutions shall be put in place to make the CDM-Obs machine-readable and usable by APIs.

Deliverables (2): Annual report on updates made to the CDM-Obs.

Task 2.2: Observation timeseries repository staging service

This service task shall upload datasets holdings, already CDM-Obs(-Core) compliant, into the C3S Observation Repository (COR). The upload format must be agreed format such as CSV or NetCDF files. The upload service may need to make minor changes to the COR code when the type of data is not readily supported, and to feed these changes back to the GitHub repository where the code is maintained. Major changes to the code are outside the scope of this contract, so the dataset tasks thereafter shall interface early with this service task to verify that their requirements can be supported. Excluded are changes that would require significant development time, or disabling functionality in datasets already in production and COR-powered. The staging service task assists in the processing of the input data to the point they are accessible in the C3S object store, and that a series of corresponding dataset metadata are generated. These metadata include, for example, spatio-temporal coverage. This task requires to acquire a thorough understanding of the workings of the observation repository (see section 5.2.1) and its supporting python packages (list available on demand to ECMWF). As needed, automatic monthly updates to datasets shall involve this task.

Deliverables (2): Annual report on updates made to the C3S Observation Repository.

Task 2.3: Serving a comprehensive dataset of land surface observations

Climate observations from land surface stations (since the 18th century) until the recent past (one month behind real-time) shall be collected from multiple sources. The sources shall include:

A. output of Task 1.3 (**Data deposition service**),
B. large international collection(s) of land surface observations, such as European Climate Assessment and Dataset (ECA&D), U.S. National Oceanic and Atmospheric Administration National Centers for Environmental Information[8] (NOAA-NCEI) Global Historical Climatology Network (GHCN; Menne *et al.*, 2012, 2018, 2023) as well as (campaign/field data in particular) archives of the National Center for Atmospheric Research Research Data Archive[9] (NCAR-RDA) and the Earth Observing Laboratory[10] (NCAR-EOL),
C. national archives and NMHS that have implemented APIs to serve open-access data holdings (possibly *via* RODEO),
D. large public data repositories (e.g., zenodo, pangea),
E. observations from the GTS accessed *via* ECMWF (see section 3.6). Regarding this data source, two-way exchanges shall be pursued with source(s) above fed in NRT from GTS node(s) outside Europe susceptible to receive substantially different data feeds than ECMWF's. This measure shall help provide safe redundancy between the largest international collections, and help catch transmission issues that are unavoidable in NRT operations, thus reducing future data rescue needs.

The input data obtained from the sources shall be ingested, inventoried, and metadata shall be kept or acquired when/as needed. The provenance of the data, along with data policy and citation requirements, shall be retained. Other requirements include geo-location and timing, instrument characteristics as much as possible, and any information that may inform about data quality or other attributes that facilitate reprocessing and exploitation in climate applications. The merging of records from multiple sources shall be done considering state-of-the-art, automated quality controls. To facilitate downstream applications, the 'data reprocessed flag' envisioned in the CDM-Obs shall be exploited as relevant so users can quickly differentiate between NRT sources and delayed-mode acquisitions, supposedly of higher maturity than the former. The temporal aggregation of the data shall be consistent with the original collection method and climate applications, i.e., sub-daily, daily, and monthly. Use of the WSI shall be undertaken when the station

[8] https://www.ncei.noaa.gov/

[9] https://rda.ucar.edu/

[10] https://www.eol.ucar.edu/all-field-programs

already exists in WMO OSCAR/Surface, or else the station metadata elements shall be provided to Task 1.8 (**Registration of stations in WIGOS**) to obtain a WSI (existence of a WSI being an objective, and not a requirement for data publication). The data products shall be CDM-Obs(-Core) compliant and submitted to Task 2.2 (**Observation timeseries repository staging service**).

A dataset entry will need to be prepared in the CDS development instance. After a review procedure by ECMWF, which may require changes to the dataset entry, possible re-ingestion in the observation repository, and further review(s), the dataset should appear on the public-facing CDS. Key deliverables must be accompanied by documentation so that users can fully exploit the deliverable.

Towards the end of the contract efforts shall move towards advancing the automatic procedures for monthly updates using NRT data. For this last part of the activity, reports shall serve as deliverables, where a gradual improvement in the maturity of the service to generate automatic updates is expected to be demonstrated, trying to exercise the mechanisms all the way to the point of Testing. These two reports shall indicate the outcome of the trials, the difficulties encountered, the measures that were implemented to address these difficulties, and remaining challenges, if any, before follow-on activities may move to an automated workflow to provide full monthly updates.

Deliverables required (12): One release of the comprehensive land surface observations dataset, consisting of the following deliverables: **Software, with Documentation (SD)**; **Dataset Licence (DL)**; **Product User Guide and Specifications (PUGS)**; **Algorithm Theoretical Baseline Document (ATBD)**; **Source and Station Inventory (SSI)**; **Data Qualification Report (DQR)**; **Data Rendering Report (DRR)**; **Data Staging Report (DSR)**; **Dataset Testing Report (DTR)**; **Dataset pUblication Report (DUR)**; 2 monthly update trials (in the second half of the contract) with a report for each trial.

Task 2.4: Serving a comprehensive dataset of marine surface observations

This task is similar to Task 2.3, with the difference that it covers the marine surface, primarily ships (since the 19th century) and data buoys (since the 20th century). The data sources shall also consider the European-based capacity in the area of the Marine Climate Data System (MCDS), notably with several Global Data Assembly Centers (GDAC) hosted in countries participating in Copernicus. The largest international collection relevant to this task is NOAA-NCEI International Comprehensive Ocean-Atmosphere Data Set (ICOADS; Freeman *et al.*, 2017, 2019) including NRT updates (Liu *et al.*, 2022).

Particular attention shall be given to the issue of reconstructing tracks and reconciling multi-variate sources, revisiting procedures for quality control, and quantifying biases, as appropriate (e.g., Kent *et al.*, 2017).

Other elements described in Task 2.3 shall apply as appropriate, with a view to ensure the dataset produced is as inclusive as possible of rescued surface marine observations.

Deliverables required (12): One release of the comprehensive marine surface observations dataset, consisting of the following deliverables: Software, with Documentation; Dataset Licence; Product User Guide and Specifications; Algorithm Theoretical Baseline Document; Source and Station Inventory; Data Qualification Report; Data Rendering Report; Data Staging Report; Dataset Testing Report; Dataset pUblication Report; 2 monthly update trials (in the second half of the contract) with a report for each trial.

Task 2.5: Serving a comprehensive dataset of upper-air observations

This task is similar to Task 2.3, with the difference that it covers upper-air measurements. The platforms to be covered include at the minimum balloon ascents. The largest international collection relevant to this task is NOAA-NCEI International Global Radiosonde Archive (IGRA; Durre *et al.*, 2018).

Homogenisation corrections to the observations for instrumental artefacts (such as radiative or instruments biases, changing instruments, or incorrect North reference to measure the wind) shall be provided as additional quantities, so users can optionally apply these corrections (e.g., Haimberger *et al.*, 2012). Balloon

trajectories shall be reconstructed when feasible to locate the observations as accurately as possible (e.g., Voggenberger *et al.*, 2024). Estimates of uncertainties shall be included.

Additional types of balloon-borne instruments may be included in this dataset, such as, for example, ozone-sondes, or descending balloons, dropsondes, rocketsondes, or long-range altitude balloons.

Most other elements (but not all) described in Task 2.3 shall also apply (as appropriate to the observations).

Deliverables required (12): One release of the comprehensive upper-air observations dataset, consisting of the following deliverables: Software, with Documentation; Dataset Licence; Product User Guide and Specifications; Algorithm Theoretical Baseline Document; Source and Station Inventory; Data Qualification Report; Data Rendering Report; Data Staging Report; Dataset Testing Report; Dataset pUblication Report; 2 monthly update trials (in the second half of the contract) with a report for each trial.

Task 2.6: Serving baseline observation networks datasets

This task is similar to Task 2.5, except it shall cover baseline network(s). At least one dataset shall be produced in this task. The choice is left at the discretion of the Tenderer. Possible networks are the Baseline Surface Radiation Network (BSRN, since the 1990s), or other baseline observation networks that monitor variables relevant to climate change, aggregating the products in a way that the offer is distinctive and specific to climate monitoring. Of particular interest are observations collected with high instrument calibration standards by atmospheric composition monitoring networks (e.g., greenhouse gases concentrations) and by renewable energy sectors (e.g., wind, solar radiation).

Deliverables required (10 at the minimum): For each baseline dataset (at least one) proposed to be serviced: one dataset release, consisting of the following deliverables: Software, with Documentation; Dataset Licence; Product User Guide and Specifications; Algorithm Theoretical Baseline Document; Source and Station Inventory; Data Qualification Report; Data Rendering Report; Data Staging Report; Dataset Testing Report; Dataset pUblication Report.

Task 2.7: Serving reference observation networks datasets

This task is similar to Task 2.6, except it covers reference networks, such as (at the minimum) the GCOS Reference Upper Air Network (GRUAN, since 2006), the U.S. Climate Reference Network (USCRN, since 2006), and GNSS ground-based stations (since the 1990s). Additional networks may be proposed in the offer, such as stations from the Doppler Orbitography by Radiopositioning Integrated on Satellite network (DORIS, since the 1990s), or more generally reference observation networks that monitor variables (including atmospheric composition) relevant to climate change. The datasets served must include estimates of uncertainties at the observation datum level.

Deliverables required (10x3 at the minimum): For each reference dataset (at least three) proposed to be serviced: one dataset release, consisting of the following deliverables: Software, with Documentation; Dataset Licence; Product User Guide and Specifications; Algorithm Theoretical Baseline Document; Source and Station Inventory; Data Qualification Report; Data Rendering Report; Data Staging Report; Dataset Testing Report; Dataset pUblication Report.

Task 2.8: GNSS global repository contribution to GCOS

This task fills a capacity gap identified by the GCOS, and that C3S has pledged to help address. Raw observation data (RINEX files) collected by ground-based GNSS stations shall be collected, with a target horizontal resolution of 100 km. The provenance of the data, along with data policy and citation requirements, shall be retained. Other requirements include geo-location and timing, instrument characteristics as much as possible, and any information that may inform about data quality or other attributes that facilitate reprocessing and exploitation in climate applications. The Successful Tenderer shall develop partnerships with the key agencies and the geodesy community, and develop synergies within the

European landscape such as the ESA GNSS Science Support Centre (GSSC). The repository is to be implemented on an ECMWF-provided Virtual Machine (VM).

Deliverables required (6): Software, with Documentation; Dataset Licence; Product User Guide and Specifications; Algorithm Theoretical Baseline Document; Source and Station Inventory; Data Rendering Report.

2.2.4 WP3: Gridded climate observation products and ensuring impact of the in-situ programme

Recognising that gridded observation products are the most popular observation-based products on the CDS in terms of numbers of users, the objective of this work-package is to make available a greater number of such products for climate services, and to showcase the use of in-situ observations in public-understandable languages. The need for an offer that caters to commercial applications is also a primary objective. The impact of the activity will also be felt over the long-term, and for this reason a special task of deep archive is foreseen. The tasks are specified as follows. Modalities of deliveries of gridded datasets are specified in section 3.3.2 of this ITT Volume II.

Task 3.1: Serving high-resolution gridded observation climate products over Europe

At least one high-resolution dataset of land surface observations over Europe shall be serviced, with half-yearly updates, using a regular latitude, longitude grid (target: 0.1^{o}x0.1^{o}), for the following variables: temperature, humidity, wind speed, radiation, and pressure, from daily observations. As a target priority, the geographical area shall include all Copernicus participating countries (at the time of writing: EU 27 countries, plus Iceland, Norway, and the United Kingdom) and the data policy shall meet the Copernicus license (free and open, without commercial restriction).

The products shall include a 'best estimate' of the gridded quantity, plus some form of estimate of the uncertainty due to the gridding process and observation sampling following well-established methodologies (e.g., Cornes *et al.*, 2018).

Additional datasets can be proposed to be serviced by the Tenderer. Each additional dataset may cover a wider domain (e.g., WMO RA VI) or a smaller region (e.g., European mountain range, European sea basin, European river basin). For these additional datasets the Tenderer is free to propose a different grid, and possibly not regular but opt to grid quantities at an administrative resolution (e.g., EEA Nomenclature of territorial units for statistics of small regions (so-called NUTS-3; Eurostat, 2024) as used by EUMETNET-supported systems, including MeteoAlarm and RODEO). These additional datasets may present varying types of data policies and may include also marine observations as appropriate, and the choice of variable and temporal resolution is also left at the discretion of the Tenderer. The offer must indicate the particular interest that each proposed dataset (i.e., choice of region, grid/horizontal resolution, variables, and temporal resolution) represents for European citizens, regarding climate change.

In order to ensure synergy and availability of the input data for all C3S activities and users, minimal efforts shall be placed on station timeseries data acquisition or observation network access, which should be taken as much as possible from WP2 (in particular Task 2.3, **Serving a comprehensive dataset of land surface observations**). Exceptions requiring to duplicate this work will have to be clearly articulated in the proposal.

Deliverables required: For each high-resolution gridded climate dataset (at least one) proposed to be serviced: Dataset licence; Product User Guide and Specifications; Algorithm Theoretical Baseline Document; Data Qualification Report; Data Rendering Report (half-yearly, i.e. 5 instances); Dataset Testing Report; Dataset pUblication Report.

Task 3.2: Contribution to C3S Climate Intelligence

This task requires the Successful Tenderer to assist in the preparation of Climate Intelligence (CI) products, such as (for example), the climate bulletins, seasonal look backs, and the European State of the Climate (ESOTC).

As part of the CI activities, C3S produces information and knowledge products based on the data stored on CDS. Should the inclusion in any of the reports mentioned above, for timeless or scope reasons, not be feasible, the Tenderer shall propose an alternative form of contribution that showcases the relevance in the context of climate variability, climate change or climate impacts of the variable in question, subject to agreement with ECMWF.

Deliverables required (2): Annual Report on contribution to C3S Climate Intelligence.

Task 3.3: Contribution to Evaluation and Quality Control, User Engagement and Training, and Communications and Outreach

This task requires to assist in reviewing the concepts proposed and developed by the EQC component for each dataset published on the CDS. Similarly, Jupyter notebooks must be prepared to illustrate how to read the data (or to support more advanced use cases), for each dataset serviced by the contract, for use in training activities. Finally, this task also requires occasional contributions to prepare social media posts, news items, or web pages, and to provide support with user queries.

Communication actions undertaken under section 3.5 are part of this task.

User support: ECMWF has a well-established centralised User Support to provide multi-tiered technical support to all users of C3S data, products, tools and services. A service desk system is used for ticketing user requests and distributing these requests to specialists as needed. Dedicated staff at ECMWF promote and maintain self-help facilities (C3S Knowledge Base (CKB), user forum, FAQs and tutorials etc.) and also provide individualised support on technical queries related to the CDS, data formats, data access etc. In addition, ECMWF staff members provide specialised scientific support to address questions related to its industrial contributions to C3S, e.g., in the areas of global reanalysis and seasonal forecasting.

All C3S contractors are expected to contribute to the delivery of multi-tiered technical support for the data and/or services they provide. The Successful Tenderer shall provide expert (Level-2) support through a) the Jira ticketing system with agreed KPIs (for example, 85% of Level-2 tickets should be resolved within 15-working days) and/or b) the user forum[11] by monitoring topics and providing responses.

Deliverables required (2+8 at the minimum): Annual Report on contribution to C3S cross-cutting activities; Jupyter notebooks (at least one per dataset, so a minimum of 8 is foreseen).

Task 3.4: Long-term data stewardship supervision

This task requires to consider holistically all the input and output of this service and isolate the data elements that are unique and not presently sufficiently backed-up. For example, a unique institution serving a particular dataset, not available anywhere else, represents a case of single point-of-failure (SPOF). This task shall list of the input/output of the service and identify which data are presently stored as SPOF. Each entry in the plan shall contain: the name of the data(set), the type (e.g. network, variable), the time range (decadal level sufficient), the order of magnitude of the total size (MB/GB/TB), and where the data are held (primary copy), along with a second or third copy. Entries with only a primary copy will be considered SPOF. Two SPOF situations may arise.

(A) The data may be held uniquely at C3S/ECMWF. For each data element in this category, the Tenderer shall propose where to host a back-up.
(B) The data may be held uniquely at a 3rd-party institution. In this case, ECMWF will provide access to archiving on its Data Handling System (DHS) tape system to address the need thus identified. For storage on ECMWF File Storage System[12] (ECFS), the Tenderer will need to prepare/arrange the files in large bundles (1-2 GB size), using tar and compression, avoiding however proprietary compression

[11] https://forum.ecmwf.int/

[12] https://confluence.ecmwf.int/display/UDOC/ECFS+user+documentation

algorithms to ensure the data can still be decompressed in the medium term. The data specifically backed-up to ECMWF in the framework of this task shall be listed in dedicated annual reports.

Deliverables required (3): Data Management Plan; Annual Report on data backed-up at ECMWF.

2.3 Interfaces with Key Stakeholders and Pre-existing Elements

Key stakeholders and pre-existing components listed thereafter must be considered when drafting the offer. When marked with the letter Ⓜ (Ⓞ), partnership or use of the corresponding component is mandatory (optional, respectively) in the offer – noting that partnership may take the form of a signed support letter for in-kind partners, or funded participation in the work to be carried out (given Copernicus funding eligibility rules).

#	*Element*	*Desired interaction type*	*Ⓜandatory Ⓞptional*
1.	Largest international data and metadata collections	Partnership & Data exchange	Ⓜ
2.	Joint C3S/WMO Data Rescue Portal[13]	Part of this ITT	Ⓜ
3.	Data Deposition Service[14]	Part of this ITT	Ⓜ
4.	WIGOS Oscar/Surface	Data exchange	Ⓜ
5.	Draft C3S procedure to mint WSIs in accordance with WMO rules	Part of this ITT	Ⓜ
6.	Common Data Model for Observations	Part of this ITT	Ⓜ
7.	WMO Task Team on Climate Data Models[15]	Information exchange	Ⓜ
8.	Observation repository	Part of this ITT	Ⓜ
9.	Imaged ACMAD collection	Part of this ITT	Ⓜ
10.	C3S station timeseries and high-resolution gridded datasets	Part of this ITT	Ⓞ
11.	OceanOPS	Partnership	Ⓞ
12.	GEO Mountains (observing stations inventory[16])	Partnership	Ⓞ
13.	Educational material on data rescue	Part of this ITT	Ⓞ

Table 2: Key stakeholders and components to interface with in this activity

[13] https://datarescue.climate.copernicus.eu/

[14] https://datadeposit.climate.copernicus.eu/

[15] https://community.wmo.int/en/governance/commission-membership/commission-observation-infrastructure-and-information-systems-infcom/standing-committee-information-management-and-technology-sc-imt/expert-team-metadata-standards-et-metadata/task-team-climate-data-model-tt-cdm

[16] https://geomountains.org/resources/resources-surveys/inventory-of-in-situ-observational-infrastructure

Ⓜ NOAA-NCEI maintains the world largest collection of climate observations (e.g., GHCN, ICOADS, and IGRA), along with a comprehensive set of station metadata.

Ⓜ The Joint C3S/WMO Data Rescue Portal provides an overview of existing data rescue activities worldwide and continues data rescue coordination under the umbrella of WMO I-DARE. This portal also provides rescue projects metadata supportive information, tutorials, guidelines, and tools for data rescue activities.

Ⓜ The aforementioned portal links to a Data Deposition Service. This service allows providers, once authenticated, to contribute rescued observations from anywhere in the world to the C3S in situ databases. Behind the public eye, this part of the service includes the ingestion of such submitted records into future versions of consolidated archives and the improvement and harmonisation of such versions. This activity also includes exchange of rescued data collections with comparable international initiatives such as the 'Send2NCEI'[17].

Ⓜ The WIGOS Oscar/Surface portal can display locations of stations once WSIs have been assigned.

Ⓜ A draft procedure for C3S to mint WSIs in accordance with WMO rules has already been developed. This procedure ensures that WMO Members are first contacted to avoid double allocation.

Ⓜ A Common Data Model for Observations was developed by C3S. The source code is stored on GitHub[18]. The Core extract of the CDM-Obs is published online[19].

Ⓜ The WMO Task Team on Climate Data Models, TT-CDM, oversees the evolution of climate data models.

Ⓜ An observation repository was developed by C3S to ingest station timeseries in an object store.

Ⓜ A previous C3S-funded activity has imaged approximately 4 million pages of weather and climate records from 44 countries in Africa[20]. Some progress has been made to start transcription, but a considerable amount of pages still remain to be transcribed.

Ⓞ Several in-situ observation datasets have already been published on the CDS, either station timeseries or high-resolution gridded datasets (e.g., E-OBS, NGCD, LaPrec).

Ⓞ For marine platforms, OceanOPS has technical solutions in place to assign WSIs. Using this capacity for a large number of platforms may incur costs that may be covered by the present service provided they are part of the offer.

Ⓞ For land stations in mountain regions, the Mountain Research Initiative has developed an inventory of observational infrastructure along with connections to local network operators, under the umbrella of GEO Mountains.

Ⓞ Educational material was produced by several projects in the past (e.g., Ryan *et al.*, 2018; Noone *et al.*, 2024). This material may be used as a basis to further advance on the topic. Besides delivering potentially new records valuable to climate research, such an approach delivers foremost valuable teaching and learning experiences, and helps communicate at large with the next generation of citizens on the issue of climate change. Of particular interest is the budding interest of training datasets for AI & ML.

[17] https://www.ncei.noaa.gov/archive/send2ncei/

[18] https://github.com/ecmwf-projects/cdm-obs

[19] https://dast.copernicus-climate.eu/documents/in-situ/CDM-Obs-Core.pdf

[20] https://datarescue.climate.copernicus.eu/en/project-highlight-acmad-data-rescue-initative

2.4 Effort Distribution between the Tasks

The following table presents guidelines for the distribution of the effort. To support planning and resourcing, we have included an estimated share of effort for each task. These figures are provided as guidelines only and are intended to help Tenderers understand the relative workload and likely distribution of work across the various activities. Tenderers are free to propose their own allocation of effort, provided all requirements are adequately addressed.

WP / Task		***Guidelines***	
WP0: Management and coordination			
		Sub-total	**7%**
WP1: Climate observations data rescue			
1.1	Data rescue coordination	**4%**	
1.2	Safeguarding climate image data records	**5%**	
1.3	Data deposition service	**5%**	
1.4	Inventorying climate data rescue station timeseries	**3%**	
1.5	Data rescue actions with the public sector	**3%**	
1.6	Data rescue actions with academia	**3%**	
1.7	Data rescue actions with the private sector	**3%**	
1.8	Registration of stations in WIGOS	**3%**	
		Sub-total	**29%**
WP2: Station climate observation timeseries			
2.1	Evolution of the common data model for observations	**2%**	
2.2	Observation timeseries repository staging service	**6%**	
2.3	Serving a comprehensive dataset of land surface observations	**10%**	
2.4	Serving a comprehensive dataset of marine surface observations	**8%**	
2.5	Serving a comprehensive dataset of upper-air observations	**8%**	
2.6	Serving baseline observation networks datasets	**2%**	
2.7	Serving reference observation networks datasets	**6%**	
2.8	GNSS global repository contribution to GCOS	**2%**	
		Sub-total	**44%**
WP3: Gridded climate observation products and ensuring impact of the in-situ programme			
3.1	Serving high-resolution gridded observation climate products over Europe	**9%**	
3.2	Contribution to C3S Climate Intelligence	**5%**	
3.3	Contribution to Evaluation and Quality Control, User Engagement and Training, and Communications and Outreach	**3%**	
3.4	Long-term data stewardship supervision	**3%**	
		Sub-total	**20%**
		TOTAL	**100%**

Table 3: Guideline for the distribution of effort to be devoted to the tasks

To accelerate the integration of the data delivered into the C3S value chain, computations shall be carried out in ECMWF computing environment as much as possible. This will also help ensure that further activities remain eligible to open tender.

In all cases, the activities may entail close engagement with network operators or data providers, with a selection of sources that shall consider in priority those that allow alignment to the open and free principles of the Copernicus Licence.

2.5 Deliverables and Documentation

This section describes the deliverables required to be produced in section 2.2 of the present ITT Volume II.

- **Annual Report (AR), for a given task**: Describes the work undertaken to implement the activities prescribed by the task description. For each such a report, a draft outline shall be sent to ECMWF a few months prior to submission of the deliverable, to ensure no major topic has been left out.
- **System Requirements Document (SRD)**: Describes the business agreement to develop a particular component of the service. It contains three major pieces of information: about the context of the activity (how it fits, why it is needed), about the context of the document (norms and standards, common practices), and the list of requirements that the component is supposed to address.
- **Data Receipt Acceptance Report (DRAR)**: Describes to what extent a given data submission received via the Data deposition service could be ingested, given the SRD of the service. It is not necessary to describe all the steps that were needed to process the data submission but this report should focus on obstacles or issues that prevented processing (in part of fully), or that required additional effort from the part of the Tenderer. A score shall be given (on a scale from 0 to 100) using criteria that the Tenderer will deem as relevant: a high score shall indicate that all data could be processed, that no major issue was found, and that the Data deposition service may not need further improvements; on the contrary, a low score shall indicate, e.g., when a data submission was of insufficient quality, could not be processed with reasonable efforts with the present state of the Data deposition service, and that further developments of the Data Deposition Service are needed.
- **Manifest Of Data Submitted (MODS)**: Lists the number of files, their sizes (including the total size), with a summary of their contents in the form of a statistical profile. The statistical profile shall indicate, for each element of the dataset (e.g., latitude, longitude, report_timestamp, observation_value, ...): a count of all values, a count of missing values, a count of non-missing values, and for these non-missing values, if they are numeric values, minimum, maximum, average, and standard deviation, or, for non-empty strings, the 100 most common strings (ordered by descending ranking of occurrence). This statistical profile shall be refined, e.g., by observed variable or source/network, in case the data delivered covers several of them.
- **Algorithm Theoretical Basis Document (ATBD)**: Describes the physical and mathematical basis of algorithms and systems used to generate data products, e.g., data dependencies; use and source of auxiliary data; all aspects of data processing and quality control; calibration and bias adjustment; filtering, interpolation, transformation; uncertainty estimation, etc. It shall contain sufficient detail to be able to serve as a reference document for implementing the production systems, and ensure full traceability to the source. To avoid un-necessary duplication of work, the ATBD, or parts of it, may incorporate or refer to **existing open-access peer-reviewed papers**, or existing consolidated documents on the condition that it comprehensively addresses all sections of the template that will be provided by ECMWF.
- **Product User Guide and Specification (PUGS):** This document shall be designed and written for non-specialist C3S users with varying backgrounds. It shall contain descriptions of all data products and the observing instruments used to produce them, and any information needed for traceability (e.g.

algorithm name and version, processing level, etc.). It shall include the data specification, and any specific information and aspects to consider when using the data, including data format and file names, product content and attributes, quality indicator and flags, data masks and filtering (including information on gap filling strategy). **This document shall also include any known issues and limitations, data disclaimers and/or suitability for specific sectors/applications, as well as best practices to use the dataset for climate variability and trend analysis**. The document shall include a section on quality assessment. The approach to product quality assurance and methods used for product validation, as well as the assessments performed on the provided datasets, including any application-specific assessments if available, shall be included. Dataset validation should follow best practices protocols where available. In all cases, timeseries covering the entire domain of the dataset shall be featured. In case of dataset updates (e.g. temporal extension), these timeseries will need to be updated.

- **Dataset license (DL):** Contains the applicable data policy (or policies). This document may contain several entries, possibly with link(s) to other documents, e.g., in case several data sources with multiple data policies are integrated. This information is needed for exposure to the CDS users. The formatting of the core part of this document should be minimal, featuring Unicode text and hyperlinks, to allow speedy integration in the CDS environment.
- **Software, with Documentation (SD)**: The software developed for the activity shall be made available on a public software repository (e.g., Github), when requested as a deliverable. User documentation should be included in the software repository. This documentation shall include instructions on how to install and deploy the software, and selected examples on how to use it.
- **Source and Station Inventory (SSI)**: A single CSV-type file, with the header containing the name of the dataset and its version number, shall contain the following entries on each line: station name(s) (e.g., primary/secondary), location coordinates (latitude, longitude, elevation), identifier(s), start and end dates, observable (one line per observable), data license, source or provenance. The header shall describe the columns (and the referential of the information, when relevant). For completeness, further tables may be provided as supplement (e.g., table of sources, with an identifier number and a long name, possibly with DOI or URI). Additional details may also be included in the inventory as needed, such as record number in the case of multi-sourced mingled timeseries or the total number of non-missing observation entries. In this fashion, further refinements may be carried out depending on the need (e.g., one inventory per month).
- **Data Qualification Report (DQR)**: Demonstrates that the data provided are compatible with expectations. See section 3.3.1.1 of the present ITT Volume II for WP2 deliverables, and section 3.3.2.1 of the present ITT Volume II for WP3 deliverables.
- **Data Rendering Report (DRR)**: This report consists of two parts. The first part includes date/time of ingestion, the name of the ECMWF VM onto which the data were uploaded. The second part is a MODS (see above). See section 3.3.1.2 of the present ITT Volume II for WP2 deliverables, and section 3.3.2.2 of the present ITT Volume II for WP3 deliverables.
- **Data Staging Report (DSR)**: Demonstrates that the data provided could be ingested into the observation repository. See section 3.3.1.3 of the present ITT Volume II.
- **Dataset Testing Report (DTR)**: Demonstrates that a dataset entry could be deployed in the development instance of the CDS. See section 3.3.1.4 of the present ITT Volume II.
- **Dataset pUblication Report (DUR)**: Demonstrates that a dataset entry could be deployed in the production instance of the CDS, and is open for public service. See section 3.3.1.5 of the present ITT Volume II.

Use of Confluence: Selected documents **SRD**, **ATBD**, **PUGS**, **SRD**, as well as their potential updates, shall be exposed in the Confluence-based documentation workspace dedicated to this project, prior to publication in the C3S Knowledge Base (CKB). ECMWF will provide guidelines to port documentation to this reserved space.

Re-use of existing documentation: Where feasible, **the Successful Tenderer is encouraged to maximise and provide updates to existing documentation**, as long as they are available in the Confluence-based workspace.

Templates: ECMWF will provide templates for each of the documents described in this section to the Successful Tenderer at the start of the contract.

Datasets: The following table reminds the minimum list of datasets to be served by the activity. This information is also found in the relevant task descriptions.

Task number	***Minimum number of datasets to be served***
2.3	1
2.4	1
2.5	1
2.6	1
2.7	3
3.1	1
Total	8

Table 4: Minimum numbers of datasets to be serviced by the activity

Assuming these minima, the total number of technical deliverables expected is **133**, distributed as follows:

Deliverable ID	***Deliverable title***
WP1: Climate observations data rescue (22 deliverables)	
Task 1.1: Data rescue coordination	
D311_bis_1.1.1_DRSP_AR.1 D311_bis_1.1.2_DRSP_AR.2	First (and second) Annual Report on Joint WMO-C3S Data Rescue Service Portal operations and updates and data rescue community animation actions
Task 1.2: Safeguarding climate image data records	
D311_bis_1.2.1_IREP_SRD	System Requirements Document for the C3S repository of images from climate-records-at-risk
D311_bis_1.2.2_IREP_Proto	Prototype instance of the C3S repository of images from climate-records-at-risk, with documentation
Task 1.3: Data deposition service	
D311_bis_1.3.1_DDS_SRD	System Requirements Document for the C3S Data deposition service
D311_bis_1.3.2_DDS_AR.1 D311_bis_1.3.2_DDS_AR.2	First (and second) Annual Report on records harvested by the C3S data deposition service
D311_bis_1.3.4_DDS_DRAR.1.5 D311_bis_1.3.5_DDS_DRAR.1.6 D311_bis_1.3.5_DDS_DRAR.1.7	Data receipt acceptance report for data submissions by data rescue action in Task 1.5 (respectively: Task 1.6, Task 1.7)
Task 1.4: Inventorying climate data rescue station timeseries	

D311_bis_1.4.1_STIF_SRD	System Requirements Document for the C3S station timeseries inventory facility
D311_bis_1.4.2_STIF_Proto	Prototype instance of the C3S station timeseries inventory facility, with documentation
Tasks 1.5 / 1.6 / 1.7: Data rescue actions with the public sector / Data rescue actions with academia /Data rescue actions with the private sector	
D311_bis_1.5.1_DRA_RepMid D311_bis_1.6.1_DRA_RepMid D311_bis_1.7.1_DRA_RepMid	Mid-term report on the data rescue action in Task 1.5 (respectively: Task 1.6, Task 1.7)
D311_bis_1.5.2_DRA_MODS D311_bis_1.6.2_DRA_MODS D311_bis_1.7.2_DRA_MODS	Manifest of data submitted by Task 1.5 (respectively: Task 1.6, Task 1.7) to the data deposition service
D311_bis_1.5.3_DRA_RepFinal D311_bis_1.6.3_DRA_RepFinal D311_bis_1.7.3_DRA_RepFinal	Report on the methods used by Task 1.5 (respectively: Task 1.6, Task 1.7) for data rescue
Task 1.8: Registration of stations in WIGOS	
D311_bis_1.8.1_WSI_AR.1 D311_bis_1.8.1_WSI_AR.2	First (and second) Annual Report on stations registered by C3S in WMO OSCAR/Surface
WP2: Station climate observation timeseries (86 deliverables)	
Task 2.1: Evolution of the common data model for observations	
D311_bis_2.1.1_CDM_AR.1 D311_bis_2.1.1_CDM_AR.2	First (and second) Annual Report on updates made to the CDM-Obs
Task 2.2: Observation timeseries repository staging service	
D311_bis_2.2.1_COR_AR.1 D311_bis_2.2.1_COR_AR.2	First (and second) Annual Report on updates made to the C3S Observation Repository
Task 2.3: Comprehensive Land surface observations Dataset (CLD)	
D311_bis_2.3.1_CLD_SD	CLD Software, with documentation
D311_bis_2.3.2_CLD_DL	CLD Dataset licence
D311_bis_2.3.3_CLD_PUGS	CLD Product User Guide and Specifications
D311_bis_2.3.4_CLD_ATBD	CLD Algorithm Theoretical Baseline Document
D311_bis_2.3.5_CLD_SSI	CLD Source and Station Inventory
D311_bis_2.3.6_CLD_DQR	CLD Data Qualification Report
D311_bis_2.3.7_CLD_DRR	CLD Data Rendering Report
D311_bis_2.3.8_CLD_DSR	CLD Data Staging Report
D311_bis_2.3.9_CLD_DTR	CLD Dataset Testing Report
D311_bis_2.3.10_CLD_DUR	CLD Dataset pUblication Report
D311_bis_2.3.11_CLD_NRT.1	CLD First monthly update trial report
D311_bis_2.3.12_CLD_NRT.2	CLD Second monthly update trial report
Task 2.4: Comprehensive Marine surface observations Dataset (CMD)	
D311_bis_2.4.1_CMD_SD	CMD Software, with documentation
…	...

D311_bis_2.4.12_CMD_NRT.2	CMD Second monthly update trial report
Task 2.5: Comprehensive Upper-air observations Dataset (CUD)	
D311_bis_2.5.1_CUD_SD	CUD Software, with documentation
...	...
D311_bis_2.5.12_CMD_NRT.2	CUD Second monthly update trial report
Task 2.6: Baseline X dataset (BXD)	
D311_bis_2.6.1_BXD_SD	BXD Software, with documentation
…	...
D311_bis_2.6.10_BXD_DUR	BXD Dataset pUblication Report
Task 2.7: Reference X dataset (RXD), Reference Y dataset (RYD), Reference Z dataset (RZD)	
D311_bis_2.7.1_RXD_SD	RXD Software, with documentation
…	...
D311_bis_2.7.10_RXD_DUR	RXD Dataset pUblication Report
D311_bis_2.7.11_RYD_SD	RYD Software, with documentation
...	...
D311_bis_2.7.20_RYD_DUR	RYD Dataset pUblication Report
D311_bis_2.7.21_RZD_SD	RZD Software, with documentation
…	...
D311_bis_2.7.30_RZD_DUR	RZD Dataset pUblication Report
Task 2.8: GNSS Global Repository (GREP)	
D311_bis_2.8.1_GREP_SD	GREP Software, with documentation
D311_bis_2.8.2_GREP_DL	GREP Dataset licence
D311_bis_2.8.3_GREP_PUGS	GREP Product User Guide and Specifications
D311_bis_2.8.4_GREP_ATBD	GREP Algorithm Theoretical Baseline Document
D311_bis_2.8.5_GREP_SSI	GREP Source and station inventory
D311_bis_2.8.6_GREP_DRR	GREP Data Rendering Report
WP3: Gridded climate observation products and ensuring impact of the in-situ programme (25 deliverables)	
Task 3.1: High-Resolution gridded climate dataset #1 (HRD1)	
D311_bis_3.1.1_HRD1_DL	HRD1 Dataset licence
D311_bis_3.1.2_HRD1_PUGS	HRD1 Product User Guide and Specifications
D311_bis_3.1.3_HRD1_ATBD	HRD1 Algorithm Theoretical Baseline Document
D311_bis_3.1.4_HRD1_DQR	HRD1 Data Qualification Report
D311_bis_3.1.5_HRD1_DRR.2026S1	HRD1 Data Rendering Report 2026 First semester
D311_bis_3.1.6_HRD1_DRR.2026S2	HRD1 Data Rendering Report 2026 Second semester
D311_bis_3.1.7_HRD1_DRR.2027S1	HRD1 Data Rendering Report 2027 First semester
D311_bis_3.1.8_HRD1_DRR.2027S2	HRD1 Data Rendering Report 2027 Second semester
D311_bis_3.1.9_HRD1_DRR.2028S1	HRD1 Data Rendering Report 2028 First semester

D311_bis_3.1.10_HRD1_DTR	HRD1 Dataset Testing Report
D311_bis_3.1.11_HRD1_DUR	HRD1 Dataset pUblication Report
Task 3.2: Contribution to C3S Climate Intelligence	
D311_bis_3.2.1_CI_AR.1 D311_bis_3.2.1_CI_AR.2	First (and second) Annual Report on contribution to C3S Climate Intelligence
Task 3.3: Contribution to Evaluation and Quality Control, User Engagement and Training, and Communications and Outreach	
D311_bis_3.3.1_CC_AR.1 D311_bis_3.3.1_CC_AR.2	First (and second) Annual Report on contribution to C3S cross-cutting activities
D311_bis_3.3.3_JN_CLD_1	Jupyter Notebook #1 showcasing CLD
D311_bis_3.3.4_JN_CMD_1	Jupyter Notebook #1 showcasing CMD
D311_bis_3.3.5_JN_CUD_1	Jupyter Notebook #1 showcasing CUD
D311_bis_3.3.6_JN_BXD_1	Jupyter Notebook #1 showcasing BXD
D311_bis_3.3.7_JN_RXD_1	Jupyter Notebook #1 showcasing RXD
D311_bis_3.3.8_JN_RYD_1	Jupyter Notebook #1 showcasing RYD
D311_bis_3.3.9_JN_RZD_1	Jupyter Notebook #1 showcasing RZD
D311_bis_3.3.10_JN_HRD1_1	Jupyter Notebook #1 showcasing HRD1
Task 3.4: Long-term data stewardship supervision	
D311_bis_3.4.1_DMP	Data Management Plan
D311_bis_3.4.2_EBUP_AR.1	First (and second) Annual Report on data backed-up at ECMWF

Table 5: Comprehensive list of WP1, WP2, and WP3 deliverables expected assuming 8 datasets

2.6 Coordination and Synergies with other Copernicus Services

The Successful Tenderer shall engage with the Copernicus In Situ Component (currently entrusted to the EEA) on request. This involves enhancement of data sharing, contribution to Copernicus-wide in situ communication activities and the provision of over-arching information to the EC.

To avoid duplication of work, the Successful Tenderer shall identify and leverage synergies with other Copernicus services conducting similar observation-related activities (e.g., in different physical domains). If Tenderers or their Sub-contractors are engaged in related contracts under other Copernicus services, this must be disclosed in the Tenderer's response to this ITT. The response should include a detailed explanation of the measures in place to monitor, avoid, and mitigate potential duplication of activities, ensuring efficiency and complementarity across services.

3 General Requirements

3.1 Implementation

ECMWF intends to award a single Framework Agreement for a period of maximum 34 months which shall be implemented via a single Service Contract expected to commence in September 2025.

The Tenderer shall provide a detailed implementation plan of proposed activities for the full period of the contract.

The processing of datasets shall preferentially be conducted by the Tenderer on its own (or Sub-contractor-provided) computing infrastructure, with the exception of the processing of datasets that are to be delivered or for the components of the service that can enable to increase the delivery efficiency. Examples include, e.g., when input/output data shall both be from/to ECMWF, in which case the ECMWF computing infrastructure shall be used, indicating clearly in the proposal the resources that would be required to do so and the reasons that lead to support that need. For key components of the dataset preparation, such as data rescue workflows, preference shall be given to computing solutions that adopt the European Weather Cloud (Abellan *et al.*, 2024).

3.2 Deliverables and Milestones

The Tenderer shall provide the list of deliverables and milestones (cf. ITT Volume IIIA "Pricing and deliverables", Excel spreadsheet "Deliverables List") for each Work Package. All deliverables and milestones must be consistent with the activities and objectives described in Section 2 of this ITT Volume II:

- A deliverable is a substantial, tangible or intangible good or service produced as a result of a project (see also the deliverable definition in this ITT Volume V Clause 1.2 and Clause 3.2). In other words, a deliverable is a verifiable outcome produced in response to the specific objectives of the contract and is subject to approval by both ECMWF's Technical Officer (TO) and Contract Management Officer (CMO) before being considered as contractually approved. When requested as such in section 2 of this ITT Volume II, document deliverables shall be periodically updated and versioned accordingly.
- Milestones should be designed as markers of demonstrable progress in service development and/or quality of service delivery during the contract implementation (see also the milestone definition in this ITT Volume V Clause 1.2). They should not duplicate deliverables.

The following shall apply to the deliverables and milestones:

- The deliverables and milestones should be consistent with and meet the technical requirements specified in section 2 of this ITT Volume II;
- All contract deliverables shall be produced in English;
- The quality of reports shall be equivalent to the standard of peer-reviewed publications and practice;
- Regarding the format of the deliverables, unless otherwise specified in the contract, or requested by ECMWF during the contract implementation:
    - For the WP0, the deliverables shall be made available to ECMWF in electronic format (PDF/Microsoft Word/Microsoft Excel/HTML or compatible, while all other formats - if any - must be agreed during the contract negotiation), via the Copernicus Deliverables Repository portal.
    - For the other WPs, the deliverables shall be made available to ECMWF in electronic format (PDF/Microsoft Word/Microsoft Excel/HTML or compatible, while all other formats - if any - must be agreed during the contract negotiation) via the platform agreed between parties for the other WPs deliverables and milestones. Furthermore, when necessary, the Successful Tenderer shall make the outputs of their work available on a server accessible by ECMWF using standard protocols such as FTP or https. The data formats to be used shall be agreed during the contract negotiation. ECMWF will only accept data in formats that follow internationally recognised standards. Such standards must be open (i.e. non-proprietary), managed by a recognised international standardisation body (e.g. ISO, WMO, OGC, etc.), or any de-facto standard. Open-source software should also exist that can read and write files of these standards. Serialisation formats (e.g. NetCDF) should be supported by standard schemas and conventions.
    - In addition, dataset documentation shall be published on Confluence.

The following shall apply in ITT Volume IIIA "Pricing and deliverables" (cf. Excel spreadsheet "Deliverables List"):

- Deliverables and milestones shall respectively follow the referencing system used in section 2 of this ITT Volume II. Any additional deliverables and milestones, to implement the work requested in sections 2.2.2 to 2.2.4 of this ITT Volume II, shall follow similar same referencing system.
- Each deliverable shall have an associated resource allocation and price (cf. column I "Nb of PM allocated" and column J "Estimated price"), while the only resource type to be considered is "payroll" (the total of these allocated resources and prices shall therefore amount to the total price associated with payroll in Volume IIIA spreadsheet "Costs and Prices").
- Milestones shall not attract the budget under Volume IIIA in the Excel spreadsheet "Deliverables list".

The Tenderer shall provide a due date for each proposed deliverable and milestone (in accordance with those indicated in section 2 for each Task):

- Where monthly NRT dataset extensions are requested (Tasks 2.3 and 2.4), the aim is to reach operational status, for which a timely delivery of services is essential. **The Tenderer shall therefore ensure that the proposed due dates for all deliverables and milestones are realistic and achievable**, i.e. the Tenderer shall also consider dependencies*, the source of original data and assess the risk accordingly.
- It is advised to schedule the submission/completion of the last deliverables and/or milestones associated to a Payment Milestone not later than 15 days before the expected date of completion of the said Payment Milestone (i.e. when all deliverables have been submitted by the contractor and all milestones have been completed by the concerned parties).

(*) Please note that any dependencies on input data, whose origin must be specified, shall be detailed by the Tenderer, and also accounted for in the risk register (cf. ITT Volume IIIB Section 5.6).

3.3 Delivery of In-Situ Observation Datasets to the CDS

Several datasets are to be delivered by the activity covered by the present Tender (see ITT Volume II, section 2.2.3 for WP2, and section 2.2.4 for WP3). The datasets will be made accessible to users via the CDS, under responsibility of ECMWF. The Tenderer is however expected to participate in this publication process. This section presents the steps that are to be included by the Tenderer in their workplan.

The CDS has been designed as a distributed system that provides access to datasets and tools through a unified web interface. A general description of the design and functionality is given by Raoult *et al.* (2017). The Successful Tenderer shall provide the data in a way that is compatible with the working practice of the CDS, but this is not limited to the data format and standard but also covers metadata and documentation. The main result of the data integration and data publication processes is a Data Catalogue Entry in the CDS. Data suppliers shall contribute to those processes. A Data Catalogue Entry is a hypertext document providing access to a collection of data or datasets. Typically, the entry has its own Digital Object Identifier (DOI) and citation (which can differ from the DOIs and citations associated with the underlying data or datasets).

The DOI attribution to each dataset published facilitates its citation. However, this does not guarantee in itself a high level of scientific maturity and community adoption. Consequently, the Tenderer is encouraged to submit one manuscript on each dataset to peer-review, *e.g.* to geoscience or scientific data journals. This point is left at the Tenderer's discretion but it is noted that such effort will help in particular to ensure that CDS datasets can be cited in support of the work of the Intergovernmental Panel on Climate Change (IPCC).

General information on the **Data Integration** and **Data Publication Processes**, as well as on the **Deprecation and Replacement of datasets in** the CDS is available in the Appendix 2 document referenced in section 5.2.2.

The following sub-sections present the details that are pertinent for the integration and publication of in-situ observation datasets, which will come either as station timeseries (from WP2) or as gridded products (from WP3).

3.3.1 Station Climate Observation Timeseries produced in WP2

The observation station timeseries shall be ingested into the observation repository and the results shall be made available to the public as a CDS entry in the production environment. Technical documentation of key steps referenced hereafter is provided in the Observation Provider Guide (OPG) chapter, which is part of the Appendix 1 document referenced in section 5.2.1. The deliverables **DQR**, **DRR**, **DSR**, **DTR**, and **DUR** mentioned in various tasks in WP2 correspond respectively to the completion of steps described hereafter: (1) **Q**ualification, (2) **R**endering, (3) **S**taging, (4) **T**esting, and (5) p**U**blication.

3.3.1.1 Qualification

This step qualifies the data as being compatible with the observation repository infrastructure. This entails reaching the technical stage where OPG step 'make-cdm' is passed. This may require modifications to the CDM-Obs (e.g., declaration of new variables, addition of metadata) or to the dataset (e.g., adaptation of the dataset, updating of units) or, exceptionally, to the observation repository itself (this is not expected but cannot be ruled out entirely). As indicated in the OPG, this step requires preparation of a draft service definition and configuration (see Observations Service Definition chapter, which is part of the Appendix 1 document referenced in section 5.2.1). Proof of passing the qualification stage shall be shown with a printout (screenshot) of the 'make_cdm' task, to be included in the **DQR**. The data sample for this stage does not need to cover the entire a dataset, but shall include at least a day of data. This step can be initiated without any access to ECMWF VM.

3.3.1.2 Rendering

This step verifies that data are delivered to an ECMWF VM. This requires the Tenderer to access such a resource. Request for such access shall be made via the ECMWF Contractors Portal[21] interface (registration required), under 'Cloud resources (VM)', once the approximate size of the dataset to be uploaded (in the right format, see previous step) is known, so that sufficient disc space can be made available on the VM. In this step, the data are to be uploaded to the VM (secured SSH copy; no remote access). Proof of passing this stage shall be shown by providing the same information foreseen in the **DRR**. All the CDM-Obs elements featured in the dataset shall be surveyed in the statistical profiling.

3.3.1.3 Staging

This step entails passing the OPG step 'make-production'. It guarantees that the data are uploaded onto the development instance of the observation repository's object store. Proof of passing this stage shall be provided with a snapshot of running the OPG instruction 'catalogue-dataset-info' in the **DSR**. This step requires to improve the service definition and the configuration, so the object store entities are of reasonable size and numbers, for performance reasons.

3.3.1.4 Testing

This step entails deploying a dataset entry on the CDS development instance. This step shall be initiated via the ECMWF Contractors Portal interface, under 'Dataset ingestion (CDSI)'. More details on this step will be provided by the ECMWF Technical team after the contract starts. Briefly, the Tenderer will have to provide the following elements for each dataset: the Dataset Licence (DL), an abstract, and an overview image (see Information document template[22]; practical examples will also be provided by ECMWF). Based on this information and tools made available by ECMWF, the Tenderer will be instructed how to create a CDS catalogue entry. At this point the 'Publication process' to be followed is described in the Appendix 2 document referenced in section 5.2.2. As soon as a first instance of the dataset appears on the CDS development website, the **DTR** can be prepared, featuring snapshots of the dataset entry webpage.

[21] https://jira.ecmwf.int/plugins/servlet/desk/site/copco

[22] https://confluence.ecmwf.int/display/COPSRV/Information+document+template

3.3.1.5 Publication

This step deploys the CDS dataset entry into the production environment. This entails undergoing review and making necessary changes to the test environment. This process ends when the dataset is published and visible on the main CDS website, at which point the **DUR** can be prepared (snapshots of the dataset entry public-facing webpage).

3.3.2 Gridded climate observation products

The following four steps shall be included in the proposal, for each delivered dataset: (1) qualification, (2) rendering, (3) testing, and (4) publication.

3.3.2.1 Qualification

This step entails preparing a subset of the data in a way that the data organization and access can be verified. The gridded observation products can be implemented in the CDS distributed infrastructure either by:

(a) **Push mode:** uploading datasets to a designated ECMWF VM.

(b) **Pull mode:** providing datasets via web services. (This does not require ECMWF VM resources)

ECMWF has a strong preference for push mode. In case pull mode is implemented, the Successful Tenderer shall ensure that the data products are stored in one or more EU members countries. The Successful Tenderer is responsible for storing the data for at least 6 years after the contract has come to an end.

The Successful Tenderer is encouraged to prepare their data in non-proprietary file format (NetCDF, csv, shape files, etc.). The "ECMWF metadata recommendations for NetCDF[23]" document provides recommendations for encoding the datasets in NetCDF. In case of pull mode, a web service must be accessible by simple commands like wget. Requests for access to those web services will originate from the CDS, as part of a workflow run on behalf of an end-user. ECMWF will therefore need to have the necessary credentials to invoke these services. ECMWF will not provide information on the end user's identity when invoking the web services, nevertheless it will collect usage statistics for all aspects of C3S.

The data shall be prepared in a way it can be accessed by unique filenames constructs. Considering the 'facets' of the queries that users may build, each single query must map to a single data filename. The deliverable for this step is a **DQR**. It shall consist of a printout (screenshot) of a file listing that considers at least two different values for each 'facet' (e.g. two dates, two variables), along with a dump of a single file header. In each file, only main one variable must be present, possibly with additional uncertainties or ways of characterizing the data as appropriate. The data sample for this stage does not need to cover the entire dataset, but shall include at least a day of data. This step can be started without any access to ECMWF VM.

3.3.2.2 Rendering

In case of using 'push' mode, this step is identical to section 3.3.1.2 above. In case of 'pull' mode, the only difference is that no ECMWF VM access is required, and the statistical profiling in the **MODS** (part of the **DRR**) shall be prepared outside the ECMWF environment.

3.3.2.3 Testing

This step is identical to section 3.3.1.4 above.

3.3.2.4 Publication

This step is identical to section 3.3.1.5 above.

[23] https://confluence.ecmwf.int/x/9IsjDQ

3.3.3 Web services

Any web services and/or portals developed under contract with C3S shall be fully integrated in the C3S web portal following the guidance provided in the table below.

Activity	***Guidance***
Design	The existing templates and styles for the main service website (http://climate.copernicus.eu) must be used. The ECMWF Copernicus web officer will provide these on request.
Domain	Off-platform sites must be registered as a sub sub-domain of the main C3S sites (http://project.climate.copernicus.eu). The name will be agreed with the Copernicus web officer and registered by the European Commission once approved.
User journey	The user journey must start on the main C3S website via a dedicated landing page for the project. The sub sub-domain URL shall point to this page.
Content	All corporate and 'About us' project content will be published on the main service website and not duplicated on the microsite.
Navigation	A home button shall take users to the main websites' homepage.
Logos	Supplier logos shall not appear on the microsites. There will be a page on the service main website that reflects the contribution of suppliers.
Reporting	We require monthly Google Analytics reports for the microsites. These shall include at minimum: • Visits • Unique visits • Bounce rate • Traffic source • Document downloads There shall be an accompanying short explanation of the trends shown by the data.

Table 6: Web services guidance

3.4 Evaluation and Quality Control Framework

C3S has established an evaluation and quality control (EQC) framework for all its products and services to ensure that users are served well and that this will continue to be the case as their needs evolve. A clear distinction is made between quality assurance and quality assessment to address these two very different categories of user requirements.

Quality assurance serves to inform users that data, metadata and documentation comply with a well-defined set of verifiable technical requirements. It provides evidence that this compliance has been checked independently from the producers. Each requirement is formulated as a verifiable statement about data records, metadata, documentation, or all three combined. Evaluators verify each statement and enter the result with commentary attached in a database. The outcomes are being published along with the dataset on the CDS. The Successful Tenderer shall be aware of the EQC quality assurance requirements which will be made available as guideline.

The purpose of quality assessments is to provide science-based information about accuracy, uncertainties, strengths and weaknesses of a dataset in the context of real use cases. Taken together, the outcomes of these activities provide the key information needed to determine fitness for purpose. The Successful Tenderer will have access to EQC content prior to public release.

3.5 Communication

The Successful Tenderer shall support ECMWF in its communication activities for the C3S services, where they are related to the activities described in this ITT, especially in what regards the contributions to the European State of the Climate report. Additional activities such as C3S website news items, C3S brochures and flyers, may be discussed on a case-by-case basis during the contract implementation.

All communication activity must be agreed with the ECMWF Copernicus Communication team in advance. This includes, but not exhaustively, communication planning, branding and visual style, media outreach, website and social media activity, externally facing written and graphic content and events. Such agreed communication activity would also need to be evaluated and reported on, once complete, so that success measures and KPIs can be provided to the European Commission.

3.6 Data and IPR

It is a condition of EU funding for C3S that ownership of any datasets developed with C3S funding passes from the suppliers to the European Union via ECMWF. Ownership will pass from the date of creation of the datasets. Suppliers will be granted a non-exclusive licence to use the datasets which they have provided to C3S for any purpose.

All software and products used by the Successful Tenderer to produce the C3S datasets will remain the property of the Successful Tenderer, except for those components which are acquired or created specifically for C3S purposes, with C3S funding, and which are separable and useable in isolation from the rest of the Successful Tenderer's production system. The identity and ownership of such exceptional components will be passed to the European Union via ECMWF annually. The Successful Tenderer will be granted a non-exclusive licence to use them for any purpose.

A distinction ought to be made between:

- those datasets (or relating documentation) specifically created as a result of this ITT, which, as Deliverables, will be fully owned by the EU, and
- pre-existing datasets (or documentation), which are simply brokered / made accessible as part of the services.

Such brokered datasets (or documentation) will continue to be owned by their original owner. The Successful Tenderer will licence the relevant brokered data/documentation to ECMWF/EU or will procure on behalf of ECMWF a licence directly from the owner. Such licence will ensure the best available terms of accessibility and redistribution, bearing in mind the purpose of the Copernicus Programme and the free and open terms of accessibility and redistribution, established for Copernicus products in the Copernicus Data Regulation (see respective definition in Volume V Clause 1.2). At a minimum, the Successful Tenderer shall grant, or procure on behalf of ECMWF, the right for the brokered datasets (or documentation) to be made available via the Climate Data Store (CDS) on terms consistent with any applicable specifications of ECMWF and the Copernicus Data Regulation. The Successful Tenderer will be responsible to provide the license terms to ECMWF in a suitable format in order for ECMWF to make the brokered datasets (or documentation) available via the CDS. The Successful Tenderer will inform ECMWF of any updates to such terms. In this case, ECMWF is procuring a service, rather than the datasets (or the documentation) themselves.

The Successful Tenderer selected for this activity will have access to datasets published on the Climate Data Store (no particular/special access will be granted, the general access conditions of the CDS apply). Other national or international data collections of observations will need to be collected by the Successful Tenderer. To expand data access to observation networks over the European Union, the Successful Tenderer can solicit assistance from ECMWF so that solutions managed by Copernicus In Situ may be implemented to facilitate

access when possible, without any guarantee however that such solicitation is completed satisfactorily before the contract ends.

In order to implement this tendered activity, which is essential to support GCOS, a programme co-sponsored by WMO, the Successful Tenderer selected for this activity will be granted access to observation data received by ECMWF under WMO arrangements as follows:

- From July 1995 to October 2021, pursuant to 12th WMO Congress (Cg-XII, 30 May - 21 June 1995) resolution number 40[24], excerpt thereafter:
  "
  (1) Six-hourly surface synoptic data from RBSNs, e.g. data in SYNOP, BUFR or other general purpose WMO Code;
  (2) All available in situ *observations from the marine environment, e.g. data in SHIP, BUOY, BATHY, TESAC codes, etc.;*
  (3) All available aircraft reports, e.g. data in AMDAR, AIREP codes, etc.;
  (4) All available data from upper air sounding networks, e.g. data in TEMP, PILOT, TEMP SHIP, PILOT SHIP codes etc.;
  (5) All reports from the network of stations recom- mended by the regional associations as necessary to provide a good representation of climate, e.g. data in CLIMAT/CLIMAT TEMP and CLIMAT SHIP/CLIMAT TEMP SHIP codes, etc.;
  "
- From November 2021 onwards, pursuant to the WMO Unified Data Policy adopted 18 October 2021 by the Extraordinary Session of the World Meteorological Congress (Cg-Ext(2021)[25].

In both cases above, the access will be granted only for the sake of implementing the activity as described in the present tender, and for no other purpose. Accordingly, the scope of observation data to be accessed must conform to the tasks undertaken, staying within the limits of the aforementioned WMO resolutions.

The proposal shall thus provide a clear distinction between both cases by setting the nature of:

- the datasets (or relating documentation) specifically created as a result of this ITT to "Dataset" or "Report",
- the brokered datasets (or relating documentation) made accessible as part of the service to "Brokerage Dataset" or "Brokered Report".

Please note that, in both cases, the Tenderer shall warrant that it has all necessary rights to either pass on ownership to the ECMWF/EU or, alternatively, that it has all necessary rights to grant the required license to ECMWF and the EU in respect of brokered datasets (or documentation), as described above. Please refer to the ITT Volume V (Framework Agreement) for further details of the license required.

Foreseen Assets, Background IPR, Improvements and Brokerage Datasets (as defined in the ITT Volume V Clause 3) shall also be described in the proposal.

Detailed contractual terms, including terms to give effect to the arrangements described above, are set out in the ITT Volume V.

[24] https://library.wmo.int/viewer/58735/download?file=827_en.pdf&type=pdf&navigator=1

[25] https://library.wmo.int/viewer/58009/download?file=WMO_Unified_Data_Policy_brochure_en.pdf&type=pdf&navigator=1

3.7 Key Performance Indicators (KPIs)

The Successful Tenderer shall report to ECMWF on a set of Key Performance Indicators (KPIs) suitable for monitoring the following aspects of the service performance:

- Data and service quality (accuracy, stability, coverage, maturity).
- User support.
- Contract management.

The tables below provide examples of KPIs that may be used by the Tenderer, along with examples of performance targets, frequency of delivery and explanations, to build their Tender. Note that KPI.D1, KPI.D2, KPI.U1 and KPI.C1 in the tables below must be part of the contract. The Tenderer may propose additional KPIs suitable for their datasets but shall limit them to the sole KPIs whose reporting may help to optimize the performance of the contract in case of deviation per comparison with the performance targets.

- List of **data and service quality KPIs**:

KPI #	KPI Title	Performance Target and Unit of Measure	Frequency of Delivery	Explanations / Comments
KPI.D1	Self-assessment of the operational system's maturity		Once per year	Include a maturity matrix, as in the supplementary material of Bates, J. J. and Privette, J. L., (2012), to enable self-assessment of the service's operational capability.
KPI.D4.1	….	….	….	….
KPI.D4.2	….	….	….	….
KPI.D4	….	….	….	….

- List of **user support KPIs**:

KPI #	KPI Title	Performance Target and Unit of Measure	Frequency of Delivery	Explanations / Comments
KPI.U1	User Support ticket response during last quarter	85% within 15 working days	Quarterly	Resolve user issue
KPI.Ux	….	….	….	….

- List of **contract management KPIs**:

KPI #	KPI Title	Performance Target and Unit of Measure	Frequency of Delivery	Explanations / Comments

KPI.C1	Deliverables submitted on time for review during last quarter	100%	Quarterly	Due dates are the deadlines (inclusive) for the deliverables to be submitted for review by ECMWF
KPI.Cx	….	….	….	….

All KPIs shall be labelled and numbered as indicated in the tables above.

During the contract implementation, all KPIs shall be duly reported by the Successful Tenderer in the Quarterly Implementation Reports (QIR) in accordance with their frequency of delivery.

For the sake of clarity, the Tenderer shall provide preliminary versions of the completed tables as part of their Tender.

3.8 Payment Plan

The Tenderer can propose a Payment Plan in ITT Volume IIIA "Pricing and deliverables" (cf. Excel spreadsheet "Payment Plan preparation"):

- The Payment Milestones should relate to the deliverables and milestones delivered during the corresponding Payment Milestone period (e.g. the payment covering the period January-June would only relate to the deliverables and milestones whose due dates are part of the same period).
- It is recommended to have Payment Milestones, and therefore payments, with an anticipated date of completion every 6 months.
- The frequency of Progress Review Meetings might be adapted to synchronise with the anticipated date of completion of each Payment Milestone.
- In case of request for a payment at contract signature, please note that this should be duly substantiated (e.g. in terms of necessary investment prior to implementation or during first weeks/months for ensuring the initial set up of the project). It is necessary to relate this payment to activities subject to other Payment Milestones.

4 Tender Format and Content

General guidelines for the Tender are described in Volume IIIB of this ITT. This section describes specific requirements to prepare the proposal for this particular Tender, along with guidelines for minimum content expected to be included in the proposal, additional to the content described in the general guidelines of Volume IIIB. This is not an exhaustive description and additional information may be necessary depending on the Tenderer's response.

4.1 Page Limits

As a guideline, it is expected that individual sections of the Tenderer's response do not exceed the page limits listed below. These are advisory limits and should be followed wherever possible, to avoid excessive or wordy responses.

Section	***Page Limit***
Executive Summary	2
Track Record	2 (for general) and 2 (per entity)

Quality of Resources to be Deployed	2 (excluding Table 1 in Volume IIIB and CVs with a maximum length of 2 pages each)
Technical Solution Proposed	20 (Table 2 in Volume IIIB, the section on references, publications, patents and any background IP is excluded from the page limit and has no page limit)
Management and Implementation	6 (excluding Table 4 and Table 5 in Volume IIIB) + 2 per each Work Package description (Table 3 in Volume IIIB)
Pricing Table	No limitation

Table 7: Page limits

4.2 Specific additional instructions for the Tenderer's response

The following is a guide to the minimum content expected to be included in each section, additional to the content described in the general guidelines of Volume IIIB. This is not an exhaustive description and additional information may be necessary depending on the Tenderer's response.

4.2.1 Executive Summary

The Tenderer shall provide an executive summary of the proposal, describing the objectives, team and service level.

4.2.2 Track Record

The Tenderer shall demonstrate for itself and for any proposed Sub-contractors that they have experience with relevant projects in the public or private sector at national or international level. ECMWF may ask for evidence of performance in the form of certificates issued or countersigned by the competent authority.

4.2.3 Quality of Resources to be deployed

The Tenderer shall propose a team possessing the skills required for providing operational services that meet the technical requirements set out in Section 2. The team shall include a Service Manager with at least 5 years of experience in management of large-scale projects. The Tenderer shall describe the experience of the Service Manager and the technical project team in performing activities related to the various aspects of this Tender.

4.2.4 Technical Solution Proposed

The Tenderer shall give a short background to the proposed solution to demonstrate understanding of the state-of-the-art in the C3S context and hence justify their proposed solution. This section shall also include information on any other third party suppliers that are used as part of the technical solution, with a table listing them all for clarity.

For each additional dataset proposed beyond those indicated to be provided 'at a minimum' in the present Tender, the Tenderer shall present an analysis of fitness for purpose (in terms of quality, uniqueness, etc.). Where brokered in-situ data products beyond those already in place at present are proposed as part of the technical solution, their inclusion shall be justified (e.g., in terms of originality and uniqueness), and the agreement of the third-party supplier shall be provided as part of the proposal.

This section shall also provide a statement of compliance for each requirement formulated throughout this document, describing how the proposed solution maps to the requirements. Additionally, **where equivalent data products are also available through other Copernicus services or major research programmes, the Tenderer should detail the differences that justifies the production in C3S, keeping in mind that the case must be sufficiently solid in the proposal and that the negotiation stage is too late to provide further justification.**

4.2.5 Management and Implementation

As part of the general project management description, and in addition to the guidance provided in Volume IIIB, Tenderers shall consider the elements described in section 2.2.1 above. Note that costs associated with fulfilling WP0 requirements shall not exceed 10% of the total price of the Tender.

Furthermore, should any Sub-contractors be proposed in the Tender, in order to ensure a comprehensive and realistic proposal, it is a mandatory requirement for the Tenderer to actively involve all such Sub-contractors in the development of the proposal. This involvement should include, but is not limited to, collaborative planning, clear communication of project timelines, and agreement on deliverables and deadlines. The Tenderer must provide documented evidence of this collaboration, demonstrating that each Sub-contractor has been consulted and has agreed to their respective roles, responsibilities, and deadlines as outlined in the proposal. This requirement is instituted to promote a cohesive and feasible project plan, reflecting a true and committed partnership among all participating entities.

5 Additional information

5.1 References

Abellan, X., V. Baousis, R. Correa, R. Cuccu, C. Duma, C. Kominos, S. Langlois, and U. Modigliani, 2024: The European Weather Cloud is now operational. *ECMWF Newsletter*, **180** (Summer 2024), 37–43. https://doi.org/10.21957/xawe573ul9

Bojinski, S., M. Verstraete, T.C. Peterson, C. Richter, A. Simmons, and M. Zemp, 2014: The Concept of Essential Climate Variables in Support of Climate Research, Applications, and Policy. *Bull. Amer. Meteor. Soc.*, **95**, 1431–1443. http://journals.ametsoc.org/doi/pdf/10.1175/BAMS-D-13-00047.1

Cornes, R. C., G. van der Schrier, E.J., van den Besselar, and P.D. Jones, 2018: An ensemble version of the E-OBS temperature and precipitation data sets. *J. Geophys. Res. Atmos.*, **123**(17), 9391-9409. https://doi.org/10.1029/2017JD028200

Durre, I., X. Yin, R. S. Vose, S. Applequist, and J. Arnfield, 2018: Enhancing the Data Coverage in the Integrated Global Radiosonde Archive. *J. Atmos. Oceanic Technol.*, **35**, 1753–1770. https://doi.org/10.1175/JTECH-D-17-0223.1

Eurostat, 2024: Regions in the European Union Nomenclature of territorial units for statistics (NUTS), 2024 edition. Publications Office of the European Union. ISBN 978-92-68-05481-9, ISSN 2315-0815, https://doi.org/10.2785/714519

GCOS, 2015: Status of the Global Observing System for Climate. GCOS-195. http://www.wmo.int/pages/prog/gcos/Publications/GCOS-195_en.pdf

Freeman, E., S.D. Woodruff, S.J. Worley, S.J. Lubker, E.C. Kent, W.E. Angel, D.I. Berry, and Co-authors, 2017: ICOADS Release 3.0: a major update to the historical marine climate record. *Int. J. Clim.*, **37**(5), 2211-2232. https://dx.doi.org/doi:10.1002/joc.4775

Freeman, E., E.C. Kent, P. Brohan, T. Cram, L. Gates, B. Huang, C. Liu, and Co-authors, 2019: The International Comprehensive Ocean-Atmosphere Data Set—Meeting users needs and future priorities. *Front. Mar. Sci.*, **6**, 435. https://doi.org/10.3389/fmars.2019.00435

Haimberger, L., C. Tavolato, and S. Sperka, 2012: Homogenization of the global radiosonde temperature dataset through combined comparison with reanalysis background series and neighboring stations. *J. Climate,* **25**, 8108-8131. https://doi.org/10.1175/JCLI-D-11-00668.1

Kent, E.C., J.J., Kennedy, T.M. Smith, S. Hirahara, B. Huang, A. Kaplan, D.E. Parker, and Co-authors, 2017: A Call for New Approaches to Quantifying Biases in Observations of Sea Surface Temperature. *Bull. Amer. Meteor. Soc.*, **98**, 1601–1616. https://doi.org/10.1175/BAMS-D-15-00251.1

Kent, E., R. Cornes, J. Siddons, and A. Ward, 2025: A marine data pipeline: data rescue to archive. Figshare. Presentation. https://doi.org/10.6084/m9.figshare.28342709.v1

Liu, C., E. Freeman, E.C. Kent, D.I. Berry, S.J. Worley, S.R. Smith, B. Huang, and Co-authors, 2022: Blending TAC and BUFR Marine In Situ Data for ICOADS Near-Real-Time Release 3.0.2. *J. Atmos. Oceanic Technol.*, **39**, 1943–1959. https://doi.org/10.1175/JTECH-D-21-0182.1

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. doi.10.1175/JTECH-D-11-00103.1

Menne, M.J., C. N. Williams, B. E. Gleason, J. J. Rennie, and J. H. Lawrimore, 2018: The Global Historical Climatology Network Monthly Temperature Dataset, Version 4. *J. Climate*, **31**, 9835–9854. https://doi.org/10.1175/JCLI-D-18-0094.1

Menne, M.J., S. Noone, N.W. Casey, R.H. Dunn, S. McNeill, D. Kantor, P.W. Thorne, and Co-authors, 2023: Global Historical Climatology Network-Hourly (GHCNh). NOAA National Centers for Environmental Information. https://doi.org/10.25921/jp3d-3v19

Noone, S., C. D'Arcy, S. Donegan, W. Durkan, B. Essel, K. Healion, H. Hersbach, and Co-authors, 2024: Investigating the potential for students to contribute to climate data rescue: Introducing the Climate Data Rescue Africa project (CliDaR-Africa). *Geosci. Data J.*, **11**, 758–774. https://doi.org/10.1002/gdj3.248

Raoult, B., C. Bergeron, A. L. Alós, J-N. Thépaut, D. Dee, 2017: Climate service develops user-friendly data store. *ECMWF Newsletter*, 151, 22-27. https://doi.org/10.21957/p3c285

Ryan, C., C. Duffy, C. Broderick, P.W. Thorne, M. Curley, S. Walsh, C. Daly, and Coauthors, 2018: Integrating Data Rescue into the Classroom. *Bull. Amer. Meteor. Soc.*, **99**, 1757–1764, https://doi.org/10.1175/BAMS-D-17-0147.1

Thorne, P.W., F. Madonna, J. Schulz, T. Oakley, B. Ingleby, M. Rosoldi, E. Tramutola, A. Arola, M. Buschmann, A.C. Mikalsen, and R. Davy, 2017: Making better sense of the mosaic of environmental measurement networks: a system-of-systems approach and quantitative assessment. *Geosci. Instrum. Methods Data Syst*. Nov 6;6(2):453-72. https://doi.org/10.5194/gi-6-453-2017

Voggenberger, U., L. Haimberger, F. Ambrogi, and P. Poli, 2024: Balloon drift estimation and improved position estimates for radiosondes. *Geosci. Model Dev.*, **17**, 3783–3799. https://doi.org/10.5194/gmd-17-3783-2024

5.2 Appendices

5.2.1 Appendix 1 Guidelines for preparing a Observation Repository based dataset for the CADS

Tenderers should refer to the separate document attached. Note that the document may include links to other ECMWF and/or web resources, some of which may not be publicly available. This document is provided to facilitate Tenderers' understanding of what is expected in order to prepare an observation dataset for the Climate Data Store observation repository, as well as to facilitate the assessment and costing of resources which Tenderers should allocate in their response for any such activities. Full access to ECMWF internal resources shall be provided to the Successful Tenderer at the start of the contract.

5.2.2 Appendix 2 Dataset Integration in the Catalogue

Tenderers should refer to the separate document attached. Note that the document may include links to other ECMWF and/or web resources, some of which may not be publicly available. This document is provided to facilitate Tenderers' understanding of what is expected in order to support the publication of a dataset entry in the Climate Data Store, as well as to facilitate the assessment and costing of resources which Tenderers should allocate in their response for any such activities. Full access to ECMWF internal resources shall be provided to the Successful Tenderer at the start of the contract.

5.3 Acronyms

ACMAD	African Centre of Meteorological Application for Development
ACRE	Atmospheric Circulation Reconstructions over the Earth
AIP	Annual Implementation Plan
AIR	Annual Implementation Report
AR	Annual Report
ASCII	American Standard Code for Information Interchange
ATBD	Algorithm Theoretical Basis Document
BSRN	Baseline Surface Radiation Network
C3S	Copernicus Climate Change Service
CDM-Obs	C3S Common Data Model for Observations
CDR	Climate Data Record
CDS	Climate Data Store
CI	Climate Intelligence
CKB	C3S Knowledge Base
COR	C3S Observation Repository
CSV	Comma-separated values ASCII format
DHS	Data Handling System
DL	Dataset Licence
DORIS	Doppler Orbitography by Radiopositioning Integrated on Satellite
DQR	Data Qualification Report
DRAR	Data Receipt Acceptance Report
DRR	Data Rendering Report
DRSP	Data Rescue Service Portal
DSR	Data Staging Report
DTR	Dataset Testing Report
DUR	Dataset pUblication Report
EC	European Commission
ECA&D	European Climate Assessment and Dataset
ECFS	ECMWF's File Storage system
ECMWF	European Centre for Medium-Range Weather Forecasts
ECV	Essential Climate Variable
EEA	European Environment Agency
E-OBS	European gridded dataset based on ECA&D information
EOL	Earth Observing Laboratory
EQC	Evaluation and Quality Control function
ESA	European Space Agency
ESOTC	European State of the Climate
EU	European Union
EUMETNET	European NMHS network

FAQ	Frequently Asked Question
FIR	Final Implementation Report
GCOS	Global Climate Observing System
GDAC	Global Data Assembly Center
GHCN	Global Historical Climatology Network
GNSS	Global Navigation Satellite System
GRUAN	GCOS Reference Upper-Air Network
GSSC	GNSS Science Support Centre
GTS	Global Telecommunication System
HTML	Hyper Text Markup Language
ICOADS	International Comprehensive Ocean-Atmosphere Data Set
I-DARE	International Data Rescue Portal
IDR	Image Data Record
IEDRO	International Environmental Data Rescue Organization
IGRA	Integrated Global Radiosonde Archive
IPR	Intellectual Property Right
ISO	International Organization for Standardization
ITT	Invitation to tender
KOM	Kick-Off Meeting
KPI	Key Performance Indicator
LAPrec	Long-term Alpine precipitation reconstruction
MCDS	Marine Climate Data System
MODS	Manifest Of Data Submitted
MOOC	Massive Open Online Course
NCEI	National Centers for Environmental Information
NDACC	Network for the Detection of Atmospheric Composition Change
NetCDF	Network Common Data Form
NGCD	Nordic Gridded Climate Dataset
NMHS	National Meteorological and Hydrological Service
NOAA	U.S. National Oceanic and Atmospheric Administration
NRT	Near-Real Time
OGC	Open Geospatial Consortium
OPG	Observation Provider Guide
OSCAR	WMO Observing Systems Capability Analysis and Review Tool
PDF	Portable Document Format
PM	Person Month
PUGS	Product User Guide and Specifications
QIR	Quarterly Implementation Report
RDA	Research Data Archive
RINEX	Receiver INdependent Exchange
SD	Software, with Documentation
SPOF	Single Point Of Failure
SRD	System Requirements Document
SSI	Source and Station Inventory
URI	Unique Resource Identifier
USCRN	United States Climate Reference Network
VM	Virtual Machine
WIGOS	WMO Integrated Global Observing System
WIS	WMO Information System
WMO	World Meteorological Organisation

WP	Contract Work Package
WSI	WIGOS Station Identifier

C3S2_311_bis Volume II
Appendix 1

Guidelines for preparing a Observation Repository based dataset for the CADS

Copernicus Contractors

Exported on 03/20/2025

Table of Contents

- Observation Provider guide (see page 5)
- Observations Service Definition (see page 29)

In development power point presentaiton of process

https://docs.google.com/presentation/d/1HZXLhIfVBIaCscyv1AwkmZkNiEOuAn6C/edit#slide=id.p1

1 Observation Provider guide

1.1 Introduction

The CopDS Observation Catalogue Manager is distributed as a python package.Its goal is to process the data submitted to the CDS by data providers and to store it in the CoPDS Observation repository. It also does check that the units and the variable names are consistent with the observations common data model. For this is provides a Command Line Interface, and also a python API that can be used for more flexibility.

It does interact with three components:

- Input data: The input data can be in principle in any format, as long as there is a reader that can read it and transform it to a pandas.DataFrame. This dataframe will be partitioned in time and space and stored in the observation repository. This repository has two parts: A catalogue and the object storage. Each partition will be persisted as a netCDF file in the object storage, and an entry will be created in the catalogue with metadata like (time coverage, latitude coverage, longitude coverage, size of the data, variables in the partition, etc)
- Catalogue database: A postgres database where the entries will be stored.
- Object storage: An S3 compatible object storage.

See the configuration section to see how to configure the observation catalogue manager so these services are accesible.

An example of a complete workflow followd to add a new dataset would be as it follows:

- Install[1] and configure[2] the package in case it is not installed.
- Ensure that the input data are in a format understood by one of the available readers[3]. If not, create a new one[4].
- Create a service definition file[5] for the new dataset. validate-service-definition (see page 17) tool can be used to validate it.
- Call the make-cdm[6] tool to check that the catalogue manager is able to successfully read the data and transform it to the observation repository format. Check the warnings of the CDM checker and fix them if possible.
- Run the make-production[7]tool. Several processes can be spawn in case an extra speedup is needed.
- Run the retrieve[8] tool to check that the data can be retrieved successfully and that the performance is good.

1 https://confluence.ecmwf.int/display/CDSM/Observation+Provider+guide#ObservationProviderguide-Install
2 https://confluence.ecmwf.int/display/CDSM/Observation+Provider+guide#ObservationProviderguide-Configuration
3 https://confluence.ecmwf.int/display/CDSM/Observation+Provider+guide#ObservationProviderguide-Availablereaders
4 https://confluence.ecmwf.int/display/CDSM/Observation+Provider+guide#ObservationProviderguide-Addinganewreader
5 https://confluence.ecmwf.int/display/CDSM/Observation+Provider+guide#ObservationProviderguide-Servicedefinitionfiles
6 https://confluence.ecmwf.int/display/CDSM/Observation+Provider+guide#ObservationProviderguide-make-cdm
7 https://confluence.ecmwf.int/display/CDSM/Observation+Provider+guide#ObservationProviderguide-make-production
8 https://confluence.ecmwf.int/display/CDSM/Observation+Provider+guide#ObservationProviderguide-retrieve

- Run the get-dataset-jsons[9] tool to produce the JSONS needed by C3S to define the dataset in the CDS and send them to the CDS staff.

1.2 Install

The CopDS Observation Repository Catalogue Manager must currently be installed from the github repository, using a conda environment in linux. It has not been tested in windows or Mac, although all the libraries are compatible in principle.

conda needs a reasonably up to date linux version, see here[10] for details. It is recommended to use the mamba command instead of conda for increased speed. See here[11] to install conda/mamba.

To install the Observation Repository Catalogue Manager, run the following commands:

```
git clone https://github.com/ecmwf-projects/cads-obs-catalogue-manager.git
cd cads-obs-catalogue-manager
mamba create -n cads-obs -c conda-forge python=3.10
mamba activate cads-obs
mamba env update --file environment.yml
pip install .
```

1.3 Configuration

1.3.1 Main configuration file

The configuration of the catalogue manager is set up in a single YAML file. An example is available in cdsobs/data/cdsobs_config_template.yml However, each dataset must be configured in a service definition[12] file.

```
# Configuration of the ingestion database where much of the input data to fill the
# repository lives on. More than one can be configured. main is the default.
ingestion_databases:
  main:
    db_user: someuser
    pwd: somepassword
    host: ingestiondb.domain.com
    port: 5432
    db_name: somename
# Configuration of the database of the observation repository catalogue.
catalogue_db:
  db_user: your_catalogue_user
```

9 https://confluence.ecmwf.int/display/CDSM/Observation+Provider+guide#ObservationProviderguide-get-forms-jsons
10 https://docs.anaconda.com/anaconda-repository/admin-guide/install/requirements/#software-requirements
11 https://github.com/conda-forge/miniforge
12 https://confluence.ecmwf.int/display/CDSM/Observation+Provider+guide#ObservationProviderguide-Servicedefinitionfiles

```
  pwd: your_catalogue_pass
  host: catalogue.domain.com
  port: 5433
  db_name: cataloguedbtest
# Configuration of the S3 storage of the observation repository.
# namespace will be prepended to the bucket names, to prevent collisions.
s3config:
  access_key: some_access_key
  secret_key: some_secret_key
  host: s3.cds.ecmwf.int
  port: 443
  secure: true
  namespace: cds2-obs-dev
# The following are configuration variables specific for each dataset.
# Supported keywords are:
# name: dataset_name following CoPDS conventions
# lon_tile_size: size of the partitions in the longitude axis, in degrees.
# lat_tile_size: size of the partitions in the latitude axis, in degrees.
# time_tile_size: Optional, can be month and year, default is month.
# available_cdm_tables: Optional, CDM tables to be used for parsing the data.
 Default is
# to use observations_table, header_table and station_configuration.
# reader: Optional, default is
"cdsobs.ingestion.readers.sql.read_header_and_data_tables".
# Reader function to use to read the data into the observation repository. Can be a
mapping
# to specify a different reader for each dataset_source.
# reader_extra_args: mapping to pass to the reader as extra arguments, for example
# an input dir containing data files.
datasets:
- name: insitu-observations-woudc-ozone-total-column-and-profiles
  lon_tile_size: 180
  lat_tile_size: 90
  time_tile_size: year
- name: insitu-observations-igra-baseline-network
  lon_tile_size: 45
  lat_tile_size: 45
  available_cdm_tables:
  - observations_table
  - header_table
  - station_configuration
  - sensor_configuration
  - uncertainty_table
  reader:
    IGRA: "cdsobs.ingestion.readers.sql.read_header_and_data_tables"
    IGRA_H: "cdsobs.ingestion.readers.sql.read_singletable_data"
- name: insitu-observations-gruan-reference-network
  lon_tile_size: 20
  lat_tile_size: 20
- name: insitu-observations-near-surface-temperature-us-climate-reference-network
  lon_tile_size: 20
  lat_tile_size: 20
  reader: "cdsobs.ingestion.readers.sql.read_singletable_data"
```

```
- name: insitu-comprehensive-upper-air-observation-network
  lon_tile_size: 20
  lat_tile_size: 20
  available_cdm_tables:
  - observations_table
  - header_table
  - station_configuration
  - sensor_configuration
  reader: "cdsobs.ingestion.readers.cuon.read_cuon_netcdfs"
  reader_extra_args:
    input_dir: "test"
```

The configuration is validated by pydantic using the class in cdsobs.config.CDSObsConfig.

1.3.2 Environment variables

- CDSOBS_CONFIG: Can be used to set the path of the configuration file. If not set, by default is $HOME/.cdsobs/cds_config.yml. Note that it can also be directly passed to the command line tools with the "-c" flag.
- CADSOBS_LOGGING_LEVEL: Logging level, by default is INFO.
- CADSOBS_LOGGING_FORMAT: Logging format. By default is CONSOLE. It can also be JSON, in order to get parseable JSONl files (each line is a JSON).
- CADSOBS_AVOID_MULTIPROCESS: Only matters for the CUON reader. It can be used to switch off multiprocessing in the readers that use it.
- CLI_DEBUG: It can be used to get very detailed error messages when using the command line tools.

1.3.3 Advanced configuration of tile sizes

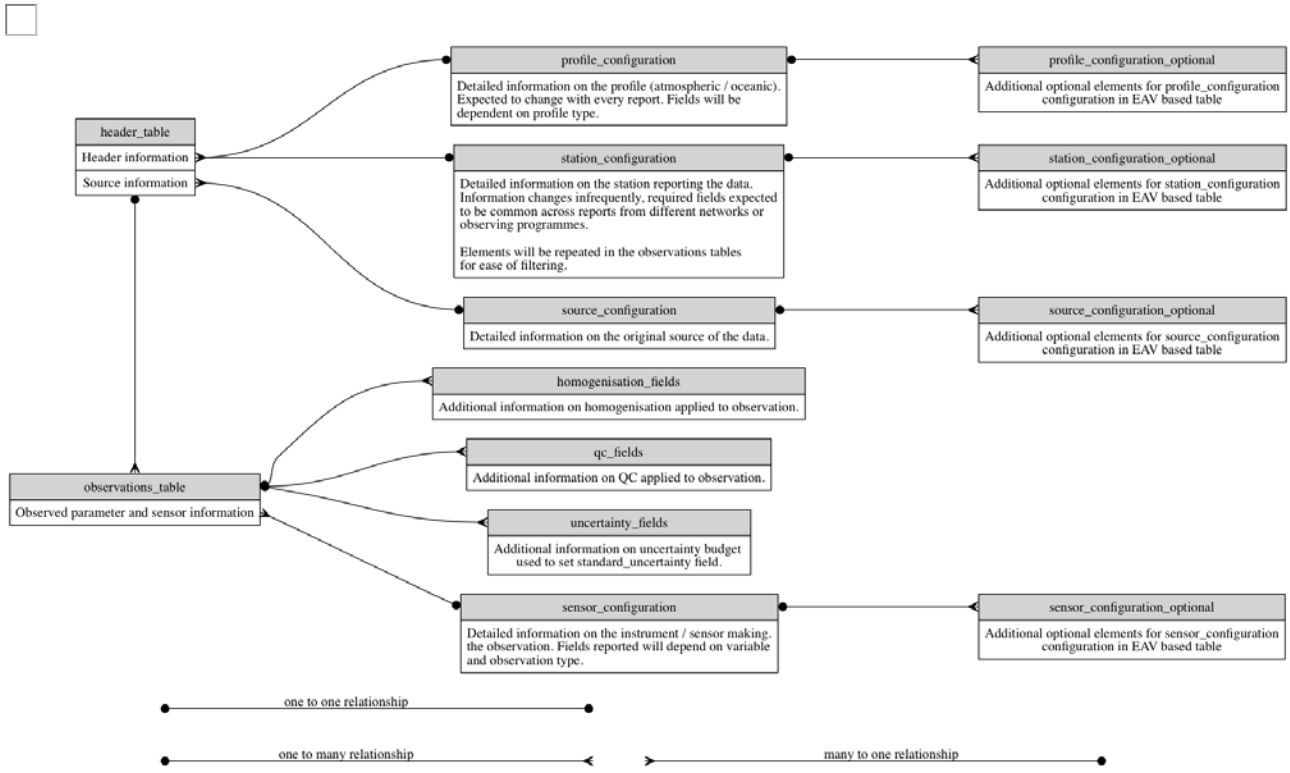
For big or complex datasets, longitude and latitude tile sizes can be optimized by making then dependent on the period (year interval) and the dataset source. In the following example, we setup global tiles for monthly data, which is lightweight, and smaller tiles for daily data. Subdaily data is made smaller for the last decade. Keep in mind that the computer that does run the ingestion (the make-production tool) needs to be able to at least hold one of these tiles in memory.

```
- name: insitu-observations-near-surface-temperature-us-climate-reference-network
  lon_tile_size:
    USCRN_MONTHLY: 90
    USCRN_DAILY: 90
    USCRN_HOURLY: 90
    USCRN_SUBHOURLY:
      2006-2010: 30
      2010-2030: 20
  lat_tile_size:
    USCRN_MONTHLY: 90
    USCRN_DAILY: 90
```

```
    USCRN_HOURLY: 90
    USCRN_SUBHOURLY:
      2006-2010: 30
      2010-2030: 20
  available_cdm_tables:
  - observations_table
  - header_table
  - station_configuration
  - sensor_configuration
  - uncertainty_table
  reader: "cdsobs.ingestion.readers.sql.read_singletable_data"
```

1.3.4 Common Data Model for observations

A detailed Common Data Model (CDM-OBS) has been defined by the C3S observation contracts C3S 311a Lot 2 and C3S 311a Lot 2. In is defined in the github repository https://github.com/ecmwf-projects/cdm-obs. A pdf document is available (cdm_latest.pdf) with all the details. This CDM is a **relational model** designed with a relational database (Postgresql) in mind. It defines several tables which are linked by relations that can be one-to-one, one-to-many, many-to-one or many-to-many.



These tables are defined in the tables_definition and tables folders. The CopDS catalogue manager does have a module that parses these tables into python objects, so the CDM can be explored from python (in the cdsobs.cdm) namespace.

1.3.5 How CDM compliance is checked

The CopDS catalogue manager does check for CDM compliance in the input data, although it does not stop if a violation is detected, it will just issue a warning and continue. Some warnings will be relatively harmless, but other will require attention from the data providers. Ideally, there should be no warnings. This is implemented in the cdsobs.cdm.api.check_cdm_compliance, which checks:

- If a field name is ambiguous, this is, if it is a field that can be mapped to one or more CDM fields. As the CDM has different tables, some different variables belonging to different tables do share the same name.
- It will raise a warning if the primary key of a table is not found, for example, primary_id for station configuration. In case NaNs are found in the primary if, it will raise an error.
- It will check the data types agains those defined in the CDM tables. Note that the types defined in the CDM tables as postgres types, while the Catalogue manager maps them to numpy/pandas types. In case a mismatch is found, a warning will be raised. Inc ase of timestamp variables, a warning will be raise if to timezone information is available, and UTC will be assumed.
- If a field is not defined in the CDM, a warning will be raised. In case a variable is not defined in the variable tables of the CDM, an error will be raised.

The CDM tables used by a given dataset can be configured in the main configuration file. Other tables will be ignored in the checks.

1.3.6 Output data format (CDM-OBS-core)

The format of the files delivered by the Climate Data Store is a simplified, single table, version of the CDM that is called CDM-lite. The documentation is not public yet but it will be added here when available. It defines a minimum set of variables that must exist for compliance. Additional fields may exist depending on the specific dataset.

1.3.7 Service definition files

Each dataset added to the observations repository needs a service definition file, that will contain information needed by the make-production[13] tool to generate the dataset. Previously, these files were JSON files. Now a new format in YML is available. It is defined here Observations Service Definition (see page 29). A set of service definition files is available in cdsobs/data/${dataset}/service_definition.yml Legacy JSON files are also available so they can be compared.

1.3.8 Available readers

For each dataset, a reader must be set in the configuration file entry for that dataset, unless we want to use the default reader "cdsobs.ingestion.readers.sql.read_header_and_data_tables". The reader is a python function responsible of reading the input data into a single pandas.DataFrame, which is the main format used internally by the catalogue manager.The following readers are available:

13 https://confluence.ecmwf.int/display/CDSM/Observation+Provider+guide#ObservationProviderguide-Makeproduction

- cdsobs.ingestion.readers.sql.read_header_and_data_tables: Reads data from a header/data table binomial from the ingestion database. The ingestion database must ve setup in the configuration file. This is the default reader.
- cdsobs.ingestion.readers.sql.read_singletable_data: Also reads from the ingestion database, but only from a single table, not a header/data table format.
- cdsobs.ingestion.readers.cuon.read_cuon_netcdfs: This reader is specific to the "insitu-comprehensive-upper-air-observation-network dataset".
- cdsobs.ingestion.readers.netcdf.read_flat_netcdfs: This reader is to be used if input data is in a simple, single table, netCDF format.
- cdsobs.ingestion.readers.netcdf.read_flat_csvs This reader is to be used if input data is in a simple, single table, CSV format. Extra arguments to be provided are the input_path (where files must be located) and optionally the filename_pattern (e.g. {dataset_name}_{source}_{year}_{month}.nc) and the separator (by default is ",").
- cdsobs.ingestion.readers.netcdf.read_flat_parquet: This reader is to be used if input data is in a simple, single table, parquer format. Extra arguments to be provided are the same of the read_flat_csvs, except the separator, which is not needed.

1.3.9 Adding a new reader

A new reader must comply with the cdsobs.ingestion.core.DatasetReaderFunctionCallable protocol

```
class DatasetReaderFunctionCallable(Protocol):
    """Signature of the dataset readers."""

    def __call__(
        self,
        dataset_name: str,
        config: CDSObsConfig,
        service_definition: ServiceDefinition,
        source: str,
        time_batch: Optional[TimeSpaceBatch],
        **kwargs,
    ) -> pandas.DataFrame:
        ...
```

Dataset name is simply the name of the dataset, the config is the object containing the configuration of the observations manager. Service definition is an object that contains the service definition file, source is the sub-dataset as defined in the service_definition file, time_batch is an object specifying which data to read in space and time (currently only supported by cdsobs.ingestion.readers.cuon.read_cuon_netcdfs), and finally kwargs are extra arguments accepted by the reader. Let's see an example, the cdsobs.ingestion.readers.netcdf.read_flat_netcdfs reader used in the previous section.

```
import xarray

from cdsobs.config import CDSObsConfig
from cdsobs.ingestion.api import EmptyBatchException
from cdsobs.ingestion.core import TimeSpaceBatch
```

```
from cdsobs.service_definition.service_definition_models import ServiceDefinition
from cdsobs.utils.logutils import get_logger

logger = get_logger(__name__)


def read_flat_netcdfs(
    dataset_name: str,
    config: CDSObsConfig,
    service_definition: ServiceDefinition,
    source: str,
    time_space_batch: TimeSpaceBatch,
    input_dir: str,
) -> pandas.DataFrame:
    if time_space_batch.space_batch != "global":
        logger.warning("This reader does not support subsetting in space.")
    time_batch = time_space_batch.time_batch
    netcdf_path = Path(
        input_dir,
        f"{dataset_name}_{source}_{time_batch.year}_{time_batch.month:02d}.nc",
    )
    if netcdf_path.exists():
        data = xarray.open_dataset(netcdf_path).to_pandas()
        data_types = data.dtypes
    else:
        raise EmptyBatchException
    return data
```

Once it has been defined, a new reader needs to be added to cdsobs.config.AvailableReaders so the configuration validation by pydantic will not fail.

1.3.10 Adding data from netCDF files

Currently, the observation catalogue manager supports reading data both from a SQL database and from a set of files. Actually it supports reading it from anywhere, by writing custom reader functions, as long as a data table with columns consistent with the Common Data Model is retrieved. See Write a custom adaptor

In order to read data in netCDF format, the “cdsobs.ingestion.readers.netcdf.read_flat_netcdfs” has to be used. An example of a dataset using it is available in the configuration file used for the tests (tests/data/cdsobs_config_template):

```
- name: insitu-observations-woudc-netcdfs
  lon_tile_size: 180
  lat_tile_size: 90
  reader: "cdsobs.ingestion.readers.netcdf.read_flat_netcdfs"
  reader_extra_args:
    input_dir: "test"
```

This reader will parse the files in the “input_dir” path following the pattern “{dataset}_{source}_YYYY_mm.nc (see page 5)”. For example “insitu-observations-woudc-netcdfs_OzoneSonde_1969_01.nc[14]”

These files must contain containing all variables and stations, as a simple 2D table. From a pandas.DataFrame, this can be directly achieved by chaining the to_xarray() and to_netcdf() methods. Column names do not need to follow the Common Data Model names, but they must be mapped to them by the service_definition.yml file to be provided with the data. Data types and units should follow the Common Data Model conventions. This data format does repeat a lot of values, compared to a normalized relational model. However, note netCDF compression can be used to achieve large compression rates and save space. Aside from netCDF, it is possible to use other formats, such as parquet. CSV text files are however discouraged because they can’t properly encode metadata such as data types, text encodings, etc. Test netCDFs can be downloaded from the following link. The following is a subset of the ncdump output of one of these files:

```
netcdf insitu-observations-woudc-netcdfs_OzoneSonde_1969_01 {
dimensions:
    index = 404 ;
    string1 = 1 ;
    string4 = 4 ;
    string5 = 5 ;
    string3 = 3 ;
variables:
    int64 index(index) ;
        index:_Storage = "chunked" ;
        index:_ChunkSizes = 404 ;
        index:_DeflateLevel = 3 ;
        index:_Shuffle = "true" ;
        index:_Endianness = "little" ;
    char primary_station_id(index, string1) ;
        primary_station_id:_Storage = "chunked" ;
        primary_station_id:_ChunkSizes = 404, 1 ;
        primary_station_id:_DeflateLevel = 3 ;
        primary_station_id:_Shuffle = "true" ;
    float height_of_station_above_sea_level(index) ;
        height_of_station_above_sea_level:_FillValue = NaNf ;
        height_of_station_above_sea_level:_Storage = "chunked" ;
        height_of_station_above_sea_level:_ChunkSizes = 404 ;
        height_of_station_above_sea_level:_DeflateLevel = 3 ;
        height_of_station_above_sea_level:_Shuffle = "true" ;
        height_of_station_above_sea_level:_Endianness = "little" ;
    int64 report_id(index) ;
        report_id:_Storage = "chunked" ;
        report_id:_ChunkSizes = 404 ;
        report_id:_DeflateLevel = 3 ;
        report_id:_Shuffle = "true" ;
        report_id:_Endianness = "little" ;
    char reference_model(index, string4) ;
        reference_model:_Storage = "chunked" ;
```

14 http://insitu-observations-woudc-netcdfs_OzoneSonde_1969_01.nc

```
        reference_model:_ChunkSizes = 404, 4 ;
        reference_model:_DeflateLevel = 3 ;
        reference_model:_Shuffle = "true" ;
    char sensor_model(index, string4) ;
        sensor_model:_Storage = "chunked" ;
        sensor_model:_ChunkSizes = 404, 4 ;
        sensor_model:_DeflateLevel = 3 ;
        sensor_model:_Shuffle = "true" ;
    float ozone_reference_time_mean(index) ;
        ozone_reference_time_mean:_FillValue = NaNf ;
        ozone_reference_time_mean:_Storage = "chunked" ;
        ozone_reference_time_mean:_ChunkSizes = 404 ;
        ozone_reference_time_mean:_DeflateLevel = 3 ;
        ozone_reference_time_mean:_Shuffle = "true" ;
        ozone_reference_time_mean:_Endianness = "little" ;
    int64 report_timestamp(index) ;
        report_timestamp:units = "days since 1969-01-30 05:45:00" ;
        report_timestamp:calendar = "proleptic_gregorian" ;
        report_timestamp:_Storage = "chunked" ;
        report_timestamp:_ChunkSizes = 404 ;
        report_timestamp:_DeflateLevel = 3 ;
        report_timestamp:_Shuffle = "true" ;
        report_timestamp:_Endianness = "little" ;
    char sensor_id(index, string5) ;
        sensor_id:_Storage = "chunked" ;
        sensor_id:_ChunkSizes = 404, 5 ;
        sensor_id:_DeflateLevel = 3 ;
        sensor_id:_Shuffle = "true" ;

    (...)
```

Note that the index variable is a dummy variable. Also note that char data types are used for strings instead of the string data type available in netCDF4, which is very slow.

The service_defitinion.json required to parse these files is available in the repo in tests/data/insitu-observations-woudc-netcdfs/service_definition.json.

See also this example for GRUAN dataset.

1.4 User guide

1.4.1 Basic usage

In order to upload a new insitu dataset to the CDS, the basic workflow is to run make-production (see page 17), then get_cds_forms (see page 26) and finally gecko.

1.4.2 Extending the data in time

- To **extend the data in time,** adding a new month or year of data, we just have to repeat the previous steps. make-production, get_cds_forms and gecko.
- Just be sure that the new data is contained in the time period defined by "--start-year" and "--end-year".
- Months with existing data will not be modified. **Update must be done in monthly steps**.
- There is an "--update" that will attempt to merge new and old data, for adding new stations and variables. This is advanced and not recommended to be used right now.

1.4.3 Command Line Interface

The command line interface is available under the cadsobs command:

```
 Usage: cadsobs [OPTIONS] COMMAND [ARGS]...

 Copernicus Climate & Atmoshpere Data Store Observation Manager Command Line
Interface
```

```
╭─ Options ─────────────────────────────────────────────────────────────────────╮
│ --install-completion          Install completion for the current shell.       │
│ --show-completion             Show completion for the current shell, to copy it or
customize the installation.     │
│ --help                        Show this message and exit.                     │
╰───────────────────────────────────────────────────────────────────────────────╯
╭─ Commands ────────────────────────────────────────────────────────────────────╮
│ catalogue-dataset-info          Get catalogue info for certain dataset.       │
│ check-consistency               Check if catalogue db and object storage are
consistent.                     │
│ copy-dataset                    Copy all catalogue datasets entries and its S3
assets.                         │
│ delete-dataset                  Permanently delete the given dataset from the
catalogue and the storage.      │
│ get_forms_jsons                 Save the geco output json files in a folder,
optionally upload it.           │
│ list-catalogue                  List entries in the catalogue. Accepts arguments
to filter the output.           │
│ list-datasets                   List all datasets and versions.               │
│ make-cdm                        Will prepare the data to be uploaded to the
observation repository, without │
│                                 actually uploading it. It can be used for     │
 testing.
│ make-production                 Upload datasets to the CADS observation
repository.                     │
│ retrieve                        Retrieve datasets from the CADS observation
repository.                     │
│ validate-service-definition     Validate a service definition YAML file.      │
╰───────────────────────────────────────────────────────────────────────────────╯
```

1.4.3.1 validate-service-definition

Simply validates the correctness of the service definition YAML file. All the errors encountered will appear on the log.

```
Usage: cadsobs validate-service-definition [OPTIONS]
                                           SERVICE_DEFINITION

Validate a service definition YAML file.

╭─ Arguments ──────────────────────────────────────────────────────────────────
│ *    service_definition      TEXT  Path to YAML file [default: None] [required]
╰──────────────────────────────────────────────────────────────────────────────
╭─ Options ────────────────────────────────────────────────────────────────────
│ --help          Show this message and exit.
╰──────────────────────────────────────────────────────────────────────────────
```

An example of a valid service defintion is available at tests/data/insitu-observations-woudc-ozone-total-column-and-profiles/service_definition.yml

See this confluence page for the specification of the service definition file.

1.4.3.2 make-production

Runs the ingestion_pipeline. This pipeline passess the input data trhough a series of steps in order to save it into the observations catalogue and the storage components.

The cdsobs_config file must contain database credentials, tile size and log level. A config template is available at tests/data/cdsobs_config_template.yml. Note that filled values on the template are for the test instance and not valid for production.

```
Usage: cadsobs make-production [OPTIONS]
```

```
 Upload datasets to the CADS observation repository.
 Read input data for a CADS observations dataset, homogenises it, partitions it and
uploads
 it to the observation catalogue and storage.

╭─ Options ──────────────────────────────────────────────────────────────────────────
│ *  --dataset                -d      TEXT     Dataset name [required]
│ *  --service-definition     -s      PATH     Path to the service_definition YAML
[required]    │
│ *  --start-year                    INTEGER  Year to start processing the data
[required]     │
│ *  --end-year                      INTEGER  Year to stop processing the data
[required]      │
│    --config                 -c      PATH     Path to the cdsobs_config yml. If not
provided,   │
│                                              the function will search for the file
          │
│                                              $HOME/.cdsobs/cdsobs_config.yml
          │
│                                              [env var: CDSOBS_CONFIG]
        │
│    --source                        TEXT     Process only a given source, by default
 it       │
│                                              processes all
          │
│                                              [default: all]
        │
│    --verbose                -v              Sets log level to debug
          │
│    --update                 -u              If set, data overlapping in time (year
and       │
│                                              month) with existing partitions will be
read in   │
│                                              order to check if it changed these need
to be     │
│                                              updated. By default, these time
intervals will    │
│                                              be skipped.
          │
│    --help                                    Show this message and exit.
          │
╰────────────────────────────────────────────────────────────────────────────────────
```

1.4.3.3 make-cdm

A sort of "dry run" of make production, it will run the same process, but without uploading any data to the catalogue or the storage. Inc an be used to check that the data is successfully processed before uploading anything.

```
 Usage: cadsobs make-cdm [OPTIONS]

 Prepare the data to be uploaded without actually uploading it.

╭─ Options ──────────────────────────────────────────────────────────────────────────────────────────────────────────╮
│ *  --dataset                -d      TEXT     Dataset name [required]                                              │
│ *  --service-definition     -s      PATH     Path to the service_definition.yml [required]                        │
│ *  --start-year                     INTEGER  Year to start processing the data [required]                         │
│ *  --end-year                       INTEGER  Year to stop processing the data [required]                          │
│    --config                 -c      PATH     Path to the cdsobs_config yml. If not provided, the function        │
│                                              will search for the file $HOME/.cdsobs/cdsobs_config.yml             │
│                                              [env var: CDSOBS_CONFIG]                                             │
│    --source                         TEXT     Process only a given source, by default it processes all             │
│                                              [default: all]                                                       │
│    --output-dir             -o      PATH     Directory where to write the output if --save-data is enabled.       │
│                                              [default: /tmp]                                                      │
│    --save-data              -s               If set, will save netCDF files in --output-dir                       │
│    --help                            Show this message and exit.                                                  │
╰────────────────────────────────────────────────────────────────────────────────────────────────────────────────────╯
```

1.4.3.4 retrieve

Retrieve CLI tool launches the retrieve_pipeline. It will download and filter the data in the storage according

to the parameters passed in the JSON file passed to –retrieve-params argument (see below for an example JSON). It will save this data as a netCDF file located in the directory defined by –output-dir argument.

```
 Usage: cadsobs retrieve [OPTIONS]

 Retrieve datasets from the CADS observation repository.

╭─ Options ──────────────────────────────────────────────────────────────────────────────────────────────────────────╮
│    --config           -c       PATH     Path to the cdsobs_config yml. If not provided, the function │
│                                         will search for the file $HOME/.cdsobs/cdsobs_config.yml      │
│                                         [env var: CDSOBS_CONFIG]                                      │
│ *  --retrieve-params  -p       PATH     Path to a JSON file with the retrieve params. [required]      │
│ *  --output-dir       -o       PATH     Directory where to write the output file. [required]          │
│    --size-limit       -sl      INTEGER  Specify a size limit for the data size retrieved in bytes.    │
│    --np                        INTEGER  Number of processes to be used by dask, default is 8.if the   │
│                                         partitions are large, this may use too much memory.           │
│                                         [default: 8]                                                  │
│    --help                               Show this message and exit.                                   │
╰────────────────────────────────────────────────────────────────────────────────────────────────────────────────────╯
```

The following is an example of the JSON file to be passed to –retrieve-params.

```
[
  "insitu-observations-woudc-ozone-total-column-and-profiles",
  {
    "dataset_source": "OzoneSonde",
    "time_coverage": [
      "1961-01-01 00:00:00",
      "2022-12-31 00:00:00"
    ],
    "variables": [
      "air_temperature"
    ],
    "stations": [
      "69",
```

```
      "111",
      "76"
    ],
    "format": "netCDF"
  }
]
```

1.4.3.5 list-datasets

This CLI tool simply lists the datasets available in an observation repository.

```
Usage: cadsobs list-datasets [OPTIONS]

List all datasets and versions.

╭─ Options ──────────────────────────────────────────────────────────────────
│ --config        -c      PATH     Path to the cdsobs_config yml. If not provided,
the function will search for the │
│                                  file $HOME/.cdsobs/cdsobs_config.yml
                                 │
│                                  [env var: CDSOBS_CONFIG]
                                 │
│ --page                  INTEGER  Results are paginated by 50 entries, choose page
[default: 0]                     │
│ --print-format          TEXT     Format to display results, either table or json [d
efault: table]                  │
│ --help                           Show this message and exit.
                                │
╰──────────────────────────────────────────────────────────────────────────
───────────────────────────────────────────────────────────────────────────
─────────────────────────╯
```

1.4.3.6 catalogue-dataset-info

This tool shows a summary of the data loaded in the observation repository for a given dataset.

```
Usage: cadsobs catalogue-dataset-info [OPTIONS] DATASET [SOURCE]
```

```
 Get catalogue info for certain dataset.

╭─ Arguments ──────────────────────────────────────────────────────────────────
│ *    dataset      TEXT      dataset name [default: None] [required]
│      source       [SOURCE]  dataset source, if not provided all dataset sources
will be displayed
╰──────────────────────────────────────────────────────────────────────────────
╭─ Options ────────────────────────────────────────────────────────────────────
│ --config  -c      PATH  Path to the cdsobs_config yml. If not provided, the
function will search for the file
│                         $HOME/.cdsobs/cdsobs_config.yml
│                         [env var: CDSOBS_CONFIG]
│ --help                  Show this message and exit.
╰──────────────────────────────────────────────────────────────────────────────
```

1.4.3.7 list-catalogue

Another tool to explore the data loaded into the repository. list-catalogue will list all the entries of the catalogue. Accepts several arguments to filter the output by variables, stations, etc.

```
 Usage: cadsobs list-catalogue [OPTIONS]

 List entries in the catalogue. Accepts arguments to filter the output.

╭─ Options ────────────────────────────────────────────────────────────────────
│ --config        -c      PATH     Path to the cdsobs_config yml. If not provided,
the function will search for the
│                                  file $HOME/.cdsobs/cdsobs_config.yml
```

```
|                                  [env var: CDSOBS_CONFIG]
|
| --page                  INTEGER  Results are paginated by 50 entries, choose a
page number [default: 1]           |
| --dataset               TEXT     filter by dataset name
|
| --source                TEXT     filter by source name
|
| --time                  TEXT     Filter by an exact date or by an interval of two
dates. For example: to retrieve |
|                                  all partitions of year 1970: 1970-1-1,1970-12-31
|
| --latitudes             TEXT     Filter by an exact latitude or by an interval of
two latitudes                      |
| --longitudes            TEXT     Filter by an exact longitude or by an interval of
two longitudes                     |
| --variables             TEXT     Filter by a variable or a list of variables. For
example:to retrieve all            |
|                                  partitions that contain variables air_pressure
and/or air_temperature:            |
|                                  air_pressure,air_temperature
|
| --stations              TEXT     Filter by a station or a list of stations
|
| --print-format          TEXT     Format to display results, either table or json [d
efault: table]                     |
| --help                           Show this message and exit.
|
╰──────────────────────────────────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────────────
──────────────────────╯
```

1.4.3.8 check-consistency

This tool runs a sanity check over the repository. It checks every asset in the storage has a catalogue entry and vice versa.

```
 Usage: cadsobs check-consistency [OPTIONS] DATASET

 Check if catalogue db and object storage are consistent.
 That means that every asset has a catalogue entry and vice versa.

╭─ Arguments
──────────────────────────────────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────────────
──╮
```

```
 * dataset      TEXT  dataset name. If provided will only check entries for that
dataset. [default: None]
                        [required]
```

```
 Options
```

```
 --config -c      PATH  Path to the cdsobs_config yml. If not provided, the
function will search for the file
                        $HOME/.cdsobs/cdsobs_config.yml
                        [env var: CDSOBS_CONFIG]
 --help                Show this message and exit.
```

1.4.3.9 copy-dataset

The copy dataset command is intended to be used to move a dataset from a test environment to a production environment.

```
 Usage: cadsobs copy-dataset [OPTIONS]

 Copy all catalogue datasets entries and its S3 assets.
 Choose to copy inside the original databases with a new name (dest_dataset) or
provide a different
 Catalogue DB and S3 credentials to insert copies.
 Parameters ---------- cdsobs_config_yml dataset dest_config_yml dest_dataset -------
```

```
 Options
```

```
 --config          -c      PATH  Path to the cdsobs_config yml. If not
provided, the function
                                 will search for the file $HOME/.cdsobs/
cdsobs_config.yml
                                 [env var: CDSOBS_CONFIG]
```

```
| *  --dataset                  TEXT  dataset to copy [default: None] [required]
|    --dest-config-yml          PATH  Path to the cdsobs_config.yml with destination
database           |
|                                     credentials. Must contain both s3 and
catalogue db credentials.   |
|                                     If not provided, the dataset will be copied
inside the initial    |
|                                     databases
|
|                                     [default: None]
|
|    --dest-dataset             TEXT  Destination name for the dataset. If not
provided, the original   |
|                                     name will be used.
|
|                                     [default: None]
|
|    --help                           Show this message and exit.
|
╰──────────────────────────────────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────────────
─╯
```

1.4.3.10 delete-dataset

This command is to be used to completely remove a dataset. Sometimes updating it will not be possible and a fresh “make_production” command will have to be run from zero. In this case, note that the dataset version will be reset to 1.0 in the catalogue.

```
 Usage: cadsobs delete-dataset [OPTIONS]

 Permanently delete the given dataset from the catalogue and the storage.

╭─ Options ──────────────────────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────────────╮
|    --config   -c      PATH  Path to the cdsobs_config yml. If not provided, the
function will search  |
|                             for the file $HOME/.cdsobs/cdsobs_config.yml
|
|                             [env var: CDSOBS_CONFIG]
|
| *  --dataset         TEXT  dataset to delete [default: None] [required]
|
```

```
│    --help                    Show this message and exit.
                │
╰──────────────────────────────────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────────────
──╯
```

1.4.3.11 get-forms-jsons

```
 Usage: cadsobs get_forms_jsons [OPTIONS]

 Save the geco output json files in a folder, optionally upload it.

 ╭─ Options
──────────────────────────────────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────────────╮
│ *  --dataset     -d      TEXT  Dataset name [required]
                │
│    --config      -c      PATH  Path to the cdsobs_config yml. If not provided, the
function will     │
│                                search for the file $HOME/.cdsobs/cdsobs_config.yml
                │
│                                [env var: CDSOBS_CONFIG]
                │
│    --output-dir  -o      PATH  Directory where to write the output if --save-data
is enabled.       │
│                                [default: /tmp]
                │
│    --upload      -u
                │
│    --help                      Show this message and exit.
                │
╰──────────────────────────────────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────────────
──╯
```

1.4.4 Python API

It is recommended to use the Command Line Interface, as is it simpler to use than the python API. However, here we document the main entrypoints.

- The main function is cdsobs.api.run_ingestion_pipeline() which runs the ingestion pipeline that populates the observations repository with data.
- The second key function is cdsobs.retrieve.api.retrieve_observations() which is the way of retrieving data from the repository in netCDF format.

1.5 Development

Anyone can contribute to the CopDS Observation Repository Catalogue Manager in github, by forking the repository (https://github.com/ecmwf-projects/cads-obs-catalogue-manager) and creating a pull request.

For best experience create a new conda environment (e.g. DEVELOP) with Python 3.10:

```
mamba create -n DEVELOP -c conda-forge python=3.10
mamba activate DEVELOP
```

Before pushing to GitHub, run the following commands:

1. Update conda environment: `make conda-env-update`
2. Install this package: `pip install -e .`
3. Sync with the latest template[15] (optional): `make template-update`
4. Run quality assurance checks: `make qa`
5. Run tests: `make unit-tests`
6. Run the static type checker: `make type-check`
7. Build the documentation (see Sphinx tutorial[16]): `make docs-build`

In order to run the tests locally, some test services are required, which are deployed with docker.

The CI pipeline will run the tests in github so deploying the tests services locally is not necessary. In case you are interested on doing it, here are the instructions.

- First, download the dump: test_ingestiondb.sql[17] file and store it at tests/docker.
- Download also cuon_data.tar.gz[18] and extract it in tests/data/cuon_data. This contains a couple of netCDFs from the CUON soundings dataset for the tests.
- Add a .env file at tests/docker containing the pass credentials for the 3 db:

```
TEST_INGESTION_DB_PASS=xxxx
CATALOGUE_PASSWORD=xxxx
STORAGE_PASSWORD=xxxx
```

15 https://github.com/ecmwf-projects/cookiecutter-conda-package
16 https://www.sphinx-doc.org/en/master/tutorial/
17 https://cloud.predictia.es/s/R9a6z8fBZQcPrAQ
18 https://cloud.predictia.es/s/dTb87RQXfgJ6S6S

- Note that credentials must coincide with the ones at the cdsobs_config template.
- Finally start the docker containers by running:

```
cd tests/docker
docker-compose up -d
```

These are the test ingestion db tables available as of 1 April? 2024:

- woudc_ozonesonde_header
- woudc_ozonesonde_data
- woudc_totalozone_header
- woudc_totalozone_data
- guan_data_header
- guan_data_value
- header
- harmonized_data
- gruan_data_header
- gruan_data_value
- uscrn.unc_subhourly
- uscrn.unc_hourly
- uscrn.unc_daily
- uscrn.unc_monthly

2 Observations Service Definition

2.1 First version

2.1.1 Changes from original

- The format has changed from JSON to YAML to improve readability and to be able to include comments
- Introduction of ***cdm_mapping*** field per source. It will add the following sub-fields:
    - *rename* : includes all translations from original names to CDM names. This means that the other fields that included those translations will not be needed anymore. For example the *descriptions* field or the *header_columns* field will now only reference CDM names
    - *unit_changes*: per variable, the old and new units will be declared, including the corresponding unit conversion (scale, offset). The ingestor will handle those unit changes if they are CDM consistent.
    - *melt_columns*: if defined, the data format will "unpivot" to be displayed into two different columns: *observed_variable* and *observation_value*. Most importantly, here we also define the some fields that need to be aligned with these melted values. Namely the uncertainty, the quality flag and the process level.
- Introduction of a "main_variables" field that lists the variables that are going to be shown in the form of the dataset. The rest of the data is considered metadata. This field replaces the "variables" group in the "products" section of the legacy service definition files.
- Removal of the "products" section of the legacy service definition files
- Entries in the description fields now must only have the following fields: "description", "dtype" and "units". The keys of the description must be already the final name for this field, the is no "name_for_output". In order to rename fields, the rename section of the cdm_mapping section must be used.

2.1.2 Examples

Example for GRUAN CDM compliant netCDF files.

2.1.3 Field by field description

Field name	Location	Required	Description
global_attributes	Top level	No	free dictionary, it will be added to the final NetCDF global attributes
space_columns	Top level or source level	Yes	Define which columns are the spatial coordinate name using the keys "x" and "y". In the case of soundings, the vertical coordinate can also be defined using "z".
sources	Top level	Yes	Dictionary with source name and associated source information fields
cdm_mapping	Source level	No	Contains the information needed to map the input data to a CDM compliant data table.
rename	cdm_mapping level	No	Dictionary of all column translations (original name to CDM name)
unit_changes	cdm_mapping level	No	Dictionary containing all variable unit conversions

Field name	Location	Required	Description

melt_columns	cdm_mapping level	No	If defined, the main variables will be "melted" to be arranged into two columns, as defined in the CDM: *observed_variable* and *observation_value*. Most importantly, here we also define the some fields that need to be aligned with these melted values. Namely the uncertainty, the quality flag and the process level. These need to be defined in a mapping with the following structure:
			• uncertainty: must be a mapping with uncertainty types as keys. Each value of the mapping will be a list of mappings with the following items:
			  • main_variable: The name of a main variable that has an estimate for this uncertainty type in the input data. For example "air_temperature".
			  • name: Name of the field in the input data containing the values for this uncertainty type and variable. For example "air_temperature_negative_total_uncertainty"
			  • units: Units for this uncertainty type. Usually they will be the units of the variable, but not always. For example "K".
			• quality_flag: Must contain a mapping with a single key also named "quality_flag". This key must contain a list of mappings with the following items:
			  • main_variable: The name of a main variable that has a quality flag. For example "solar_irradiance"
			  • name: Name of the field in the input data containing the values for this quality flag.

Field name	Location	Required	Description

			• processing_level: Must contain a mapping with a single key also named "processing_level". This key must contain a list of mappings with the following items:
			  • main_variable: The name of a main variable that has a processing level. For example "soil_temperature".
			  • name: Name of the field in the input data containing the values for this processing level. For example "soil_temperature_processing_level".

remap_categories	cdm_mapping level	No	Codes for enumerated columns (columns with categorical data such as quality flags). Will contain the CDM name for the column and a dictionary of original name values and their associated code
header_columns	Source level	No	List of header column names (CDM names). This is not needed if there is only a data table.
header_table	Source level	No	Name of the header table. This is not needed if there is only a data table.
data_table	Source level	Yes	Name of the data table
join_ids	Source level	No	In case there is a header table, specify connections between header and data columns.
descriptions	Source level	Yes	Dictionary containing the column CDM name as key, and the following attributes as values: • Description: column description. • dtype: column dtype • long_name: column name • units: column units
mandatory_columns	Source level	Yes	List of mandatory columns. CDM names.

Field name	Location	Required	Description
order_by	Source level	No	List specifying order of columns. CDM names.

These files are validated by using pydantic. And afterwards, also using the CDM tables. The pydantic models are defined in https://github.com/ecmwf-projects/cads-obs-catalogue-manager/blob/main/cdsobs/service_definition/service_definition_models.py

C3S2_311_bis Volume II
Appendix 2

Dataset integration in the Catalogue: pre-publication, publication and post-publication

Copernicus Contractors

Exported on 03/19/2025

Table of Contents

About this page

Scope	This page describes the main steps needed to integrate data in the Catalogue. It **does not describe** what is needed to integrate documentation, the role of the technical officers and other aspects that are very important for a successful integration but that are the scope of other wiki pages.
Intended audience	*Copernicus data providers.*
Outline	*The focus of this page is on what the data provider needs to supply to the CADS team and how to do it.*
Disclaimer	*The information in this page is not guaranteed to describe exactly the actual processes which are subject to change from time to time.* *But, the CADS team intends to keep the information in this page as close as possible of the actual practices.*

1 Provider's role: summary

List of the expected contributions

The provider is expected to work closely with CDS team and the technical officer in order to resolve any issues that come up during the various stages of the publication process and afterwards. Communication is preferred through Jira ticket.
The steps below are laid out in chronological order and contain the main contributions.

1. Registers the Dataset: Applies only to ECV datasets - Dataset registration[1] (*Integration process*)
2. Supplies the Information document: Information document template[2] *(Integration process*)
3. Supplies manifest file (See more about manifests below at *Manifest and pseudo-manifest files*. *Integration process.*)
4. Help the CDS team member to reply to the reviewer's and Editorial Board's comments (*Review process*)
5. Provide previous existent DOIs, licences and citations associated with each part of the data (see below: *DOI, citation and licence. Review process.)*
6. After entry published in the CDS Catalogue, the data provider should keep the manifest's filename and path exactly the same for the whole duration of the contract, even when the contents of the manifest is changed. (post-publication *processes: automatic updates*)
7. The data provider is expected to help the CDS team on keeping the entry working as expected when the data provider has the knowledge and the resources to do it (post-publication *processes: Maintenance)*
8. Follows CDS procedures for deprecating data (see below: *Versions, deprecation of entries, replacement of data. Post-publication processes: Maintenance*)

1 https://confluence.ecmwf.int/display/COPCO/Dataset+registration
2 https://confluence.ecmwf.int/display/COPCO/Dataset+Information+Document+Templates

2 How to start the integration of your data in the Catalogue

JIRA ticket, Information document and manifest file

Provider's role	Description
JIRA When asked by the CDS management the provider is expected to register the dataset[3] and create a JIRA ticket on the ECMWF Contractors Portal (COPCO[4]) - Data Ingestion	All information concerning the creation, modification, merging, updating, deprecation, additions of of data or documentation, DOIs, Citations, etc, is supposed to be managed through the JIRA ticket.
Manifest Have a pseudo-manifest file (or a manifest file) prepared.	This is the central piece of information needed by the CDS. So important that we have a whole section about it at the bottom of this page.
Information document Attach to the JIRA ticket an "Information document" filled in with the information associated to the data that you are delivering for publication in the CDS Catalogue. The template for the information document can be found here: Information document template[5].	The information document is the starting point for the integration process. In order to arrive at an agreed draft entry to submit for review, additional inputs may be required. The document contains fields and tables that should be completed with the information relevant for your data. Guidelines are provide along those fields and tables intending to help you to understand exactly which information is required and in which format.

3 https://confluence.ecmwf.int/display/COPSRV/Dataset+registration
4 https://jira.ecmwf.int/plugins/servlet/desk/category/cds-data-ingestion
5 https://confluence.ecmwf.int/display/COPCO/Dataset+Information+Document+Templates

3 Main processes in which you are expected to participate

Pre-publication process

The aim of this step is to check and agree on the main inputs for the subsequent publication process. In the case of URL based datasets, this covers the following aspects:

- path and filename conventions,
- the size of the files,
- the number of variables per file,
- where the data will be stored,
- manifest file
- for more details on these points please refer to Guidelines for preparing a URL based dataset for the CADS[6]

The CDS team expects to have access to the information in the list above as soon as possible through a JIRA ticket and at least 2 month before the actual delivery of the data.
Please consult How to start the integration process above in this page.

For data for which the contents and the container is still modifiable, the CDS team expects to interact with the provider in order to influence the way the data is stored making it more suitable for the needs of the Catalogue and the needs of the Toolbox.

Publication process

The process that goes from the initial trigger of the integration of your data, to the publication of the Catalogue entry in the public Catalogue, is referred as the "publication process".
The publication process has two processes in sequence: the integration process and the review process.
Your role in these two processes is summarised below.

Provider's role	Process	Description

6 https://confluence.ecmwf.int/display/COPCO/Guidelines+for+preparing+a+URL+based+dataset+for+the+CADS

<table>
<tr>
<td>Creates JIRA ticket
Supplies the Information document
Supplies manifest file
Replies to CDS team queries</td>
<td>Integration process (analogous to creating a draft of a paper to be submitted to a scientific journal)</td>
<td>
<table>
<tr><td>Inputs</td><td>JIRA ticket and Information document</td></tr>
<tr><td>Outputs</td><td>Draft Catalogue entry judged to be good enough to be submitted to review by the CDS team, the technical officer and the data provider</td></tr>
<tr><td>Work</td><td>Based on the manifest file and the Information Document, a CDS team member (or associated) creates one or more possible drafts for the future entry in the Catalogue. When agreed that the draft is good enough to be submitted for publication this process ends.</td></tr>
</table>
</td>
</tr>
<tr>
<td>Help the CDS team member to reply to the reviewer's and Editorial Board's comments.</td>
<td>Review process (analogous to the review process of a paper submitted to a scientific journal)</td>
<td>
<table>
<tr><td>Inputs</td><td>Draft entry</td></tr>
<tr><td>Outputs</td><td>Modified entry reviewed and approved by the Editorial Board published in the public CDS Catalogue</td></tr>
<tr><td>Work</td><td>A CDS team member (or associated) runs an internal review process to guarantee that the entry respects the CDS expectations.</td></tr>
</table>
</td>
</tr>
</table>

Post-publication processes

These are the two main processes where you may be asked to participate:

Provider's role	Process	Description

<table>
<tr>
<td>Keep the manifest's filename and path exactly the same. Contents of the manifest is expected to change. But new additions to the contents, other than time extensions, should be discussed with the CDS team. See more about manifests below at Manifest and pseudo-manifest files.</td>
<td>Automatic updates (Updates date and time related widgets in the download form. This allows the automatic release of time extensions of data. Does not work for other widget's updates like new variables. new versions etc.)</td>
<td>
<table>
<tr><td>Inputs</td><td>Entry already published in the Catalogue
Manifest file or equivalent
Update frequency agreed
EC-Flow suite implemented</td></tr>
<tr><td>Outputs</td><td>Entry updated with new dates</td></tr>
<tr><td>Work</td><td>EC-Flow suite will read the manifest file and run CDS scripts able to recreate the download form.</td></tr>
</table>
</td>
</tr>
<tr>
<td>Notify the CDS team via the ECMWF Contractors Portal (COPCO[7]) when:
download form not providing the expected data
documentation tab not providing the expected documentation
mismatch between data and documentation
Provide a Known issues table: See below</td>
<td>Maintenance (new programmed versions, new documentation, deprecating data, unexpected issues with the data and the documentation, licences, etc)</td>
<td>
<table>
<tr><td>Inputs</td><td>Published entry
Request for modification of the published entry</td></tr>
<tr><td>Outputs</td><td>Modified entry</td></tr>
<tr><td>Work</td><td>A CDS team member (or associated) modifies the entry as requested. The CDS team evaluates when the required modification needs agreement from the Editorial Board.</td></tr>
</table>
</td>
</tr>
</table>

Known-Issues guidance

7 https://jira.ecmwf.int/plugins/servlet/desk/site/copco

3.1 Known-Issues guidance

Issue detection date	Status	Which data are impacted	Description	
2022-02-15	Closed	Version: 21.03 Period: whole period Variable: all variables	The filenames have been changed to include the version in them. Example: name before the change: "NGCD_TN_type1_20191130.nc[8]" name after the change: "NGCD_TN_type1_version_21.03_20191130.nc[9]".	No action required

Issue detection date	Status	Which data are impacted	Description	Guidance to users
2022-04-01	Ongoing	For some of the years there were a few days impacted by the sparse availability of satellite data.	In most of the impacted days the problematic regions have missing values but, in other of those days, the problematic regions have the value 0. GPCP data has been corrected for these data gaps but the correction has been made by rewriting the files without modifying the version. The previous version of the files can be made available to users on request only.	Redownload the data.

8 http://NGCD_TN_type1_20191130.nc
9 http://NGCD_TN_type1_version_21.03_20191130.nc

4 Preparing your data and delivering manifests

Preparing and structuring your data

The CDS can serve data that is stored in various ways, therefore the prepartion guidelines differ depending on your use-case. The currently operational options are:

1. URL datasets
    a. Datasets served as a series of files each with their own unique URLs
    b. This is the most common option for project data providers
    c. Guidelines for preparing a URL based dataset for the CADS[10]
    d. You MUST provide a manifest as described in the following section
2. MARS datasets
    a. Datasets stored on a MARS archive
    b. This is generally only used for internal ECMWF datasets (e.g. ERA5, SEAS5, EFAS and GLOFAS)
    c. The MARS and GRIB standards mean that preparation requirements are enforced by the infrastructure
3. ESGF WPS server
    a. Dataset stored on a ESGF WPS server which are accessible via Rooki
    b. Such datasets require an intake inventory which fully describes the dataset to be exposed
    c. Preparation requirements are deteremined by rooki requirements
    d. Please contact the CDS team if you believe your dataset is compatible
4. Observation databases
    a. For small in size, but high in complexity datasets
    b. This is currently a bespoke set up for people producing insitu observations under a C3S contract
        i. If you are working on such a contract and require further details on preperation guidelines, please contact the CDS Team or your techinical officer.

Manifest and pseudo-manifest files for URL datasets

4.1 Content of the Manifest

The manifest should contain the path and the file name for every file that the CDS catalogue is supposed to provide to the users. **Nothing more nothing else.** No empty lines, no comments.
For instance:

First ten lines of a manifest file for a dataset accessible through URL addresses

10 https://confluence.ecmwf.int/display/COPCO/Guidelines+for+preparing+a+URL+based+dataset+for+the+CADS

head ./Integration_of_satellite-earth-radiation-budget/manifest_c3s_312b_lot1_erb_c3s_icdr_latest.txt
http://gws-access.ceda.ac.uk/public/cds_c3s_cloud/c3s_312b_lot1/data/erb/c3s/icdr/r01/monthly/2017/01/C3S-312bL1-L3C-MONTHLY-ERB-SLSTR_ORAC_Sentinel-3a_201701_fv3.1.nc
http://gws-access.ceda.ac.uk/public/cds_c3s_cloud/c3s_312b_lot1/data/erb/c3s/icdr/r01/monthly/2017/02/C3S-312bL1-L3C-MONTHLY-ERB-SLSTR_ORAC_Sentinel-3a_201702_fv3.1.nc
http://gws-access.ceda.ac.uk/public/cds_c3s_cloud/c3s_312b_lot1/data/erb/c3s/icdr/r01/monthly/2017/03/C3S-312bL1-L3C-MONTHLY-ERB-SLSTR_ORAC_Sentinel-3a_201703_fv3.1.nc
http://gws-access.ceda.ac.uk/public/cds_c3s_cloud/c3s_312b_lot1/data/erb/c3s/icdr/r01/monthly/2017/04/C3S-312bL1-L3C-MONTHLY-ERB-SLSTR_ORAC_Sentinel-3a_201704_fv3.1.nc
http://gws-access.ceda.ac.uk/public/cds_c3s_cloud/c3s_312b_lot1/data/erb/c3s/icdr/r01/monthly/2017/05/C3S-312bL1-L3C-MONTHLY-ERB-SLSTR_ORAC_Sentinel-3a_201705_fv3.1.nc
http://gws-access.ceda.ac.uk/public/cds_c3s_cloud/c3s_312b_lot1/data/erb/c3s/icdr/r01/monthly/2017/06/C3S-312bL1-L3C-MONTHLY-ERB-SLSTR_ORAC_Sentinel-3a_201706_fv3.1.nc
http://gws-access.ceda.ac.uk/public/cds_c3s_cloud/c3s_312b_lot1/data/erb/c3s/icdr/r01/monthly/2017/07/C3S-312bL1-L3C-MONTHLY-ERB-SLSTR_ORAC_Sentinel-3a_201707_fv3.1.nc
http://gws-access.ceda.ac.uk/public/cds_c3s_cloud/c3s_312b_lot1/data/erb/c3s/icdr/r01/monthly/2017/08/C3S-312bL1-L3C-MONTHLY-ERB-SLSTR_ORAC_Sentinel-3a_201708_fv3.1.nc
http://gws-access.ceda.ac.uk/public/cds_c3s_cloud/c3s_312b_lot1/data/erb/c3s/icdr/r01/monthly/2017/09/C3S-312bL1-L3C-MONTHLY-ERB-SLSTR_ORAC_Sentinel-3a_201709_fv3.1.nc
http://gws-access.ceda.ac.uk/public/cds_c3s_cloud/c3s_312b_lot1/data/erb/c3s/icdr/r01/monthly/2017/10/C3S-312bL1-L3C-MONTHLY-ERB-SLSTR_ORAC_Sentinel-3a_201710_fv3.1.nc

4.2 Pseudo-manifest

Dataset suppliers to the CDS shall provide a comprehensive description of their data at least two months prior to delivery, using a data registration process established by ECMWF.
For the CDS team this means the delivery of a pseudo-manifest file.

A pseudo-manifest is a manifest file with expected path and filenames for the expected data to be created. Note that the pseudo-manifest should be as close as possible of the final delivery but the CDS team understands that modifications may be needed.
If a pseudo-manifest is provided, then a Catalogue entry can be created and its design agreed and tested. Filenames and paths can be checked to see if they allow a good building of the download form.

4.3 Name of the manifest and updates of the contents of the manifest

The manifest should be named "manifest_<Contract tag>_<ECV_name_tag or SIS_name_tag>_<optional_tag>_yyyymmdd.txt" where yyyymmdd is the date where this manifest was created.
It is expected that the providers replace the strings <...> in the manifest filename with the actual names for the dataset they are providing.

When a new manifest file is added to the the providers site, that manifest should also be copied to "manifest_<Contract tag>_<ECV_name|SIS_name>_<optional_tag>_**latest**.txt".
Remove the date and leave just the string "latest".

This convention is central for the CDS computers to find and access the correct manifest.

Manifests for datasets stored in MARS (mars.list files)

To construct catalogue entries based on MARS datasets the CDS need a mars.list file which describes the contents to be exposed. If this needs to be regularly updated then please contact the CDS team to ensure the regular updates are in place.

Example mars.list (from UERRA)

```
class=ur,expver=prod,levtype=sol,origin=eswi,stream=oper,type=an,param=260199/260360,
levelist=1/2/3,time=00:00:00/06:00:00/12:00:00/18:00:00,date=1961-01-01/1961-01-02/19
61-01-03/1961-01-04/1961-01-05/1961-01-06/1961-01-07/1961-01-08/1961-01-09/1961-01-10
/1961-01-11/1961-01-12/1961-01-13/1961-01-14/1961-01-15/1961-01-16/1961-01-17/1961-01
-18/1961-01-19/1961-01-20/1961-01-21/1961-01-22/1961-01-23/1961-01-24/1961-01-25/1961
-01-26/1961-01-27/1961-01-28/1961-01-29/1961-01-30/1961-01-31
class=ur,expver=prod,levtype=sol,origin=eswi,stream=oper,type=an,param=260199/260360,
levelist=1/2/3,time=00:00:00/06:00:00/12:00:00/18:00:00,date=1961-02-01/1961-02-02/19
61-02-03/1961-02-04/1961-02-05/1961-02-06/1961-02-07/1961-02-08/1961-02-09/1961-02-10
/1961-02-11/1961-02-12/1961-02-13/1961-02-14/1961-02-15/1961-02-16/1961-02-17/1961-02
-18/1961-02-19/1961-02-20/1961-02-21/1961-02-22/1961-02-23/1961-02-24/1961-02-25/1961
-02-26/1961-02-27/1961-02-28
class=ur,expver=prod,levtype=sol,origin=eswi,stream=oper,type=an,param=260199/260360,
levelist=1/2/3,time=00:00:00/06:00:00/12:00:00/18:00:00,date=1961-03-01/1961-03-02/19
61-03-03/1961-03-04/1961-03-05/1961-03-06/1961-03-07/1961-03-08/1961-03-09/1961-03-10
/1961-03-11/1961-03-12/1961-03-13/1961-03-14/1961-03-15/1961-03-16/1961-03-17/1961-03
-18/1961-03-19/1961-03-20/1961-03-21/1961-03-22/1961-03-23/1961-03-24/1961-03-25/1961
-03-26/1961-03-27/1961-03-28/1961-03-29/1961-03-30/1961-03-31
```

Manifests for ESGF-WPS datasets (intake.yaml inventories)

To construct catalogue entries based on ESGF-WPS datasets the CDS need an intake.yaml file which describes the contents to be exposed, and any additional post-processing funcitonality. If this needs to be regularly updated then please contact the CDS team to ensure the regular updates are in place.

Example intake.yaml (from CMIP6)

```
- path: CMIP/NUIST/NESM3/historical/r1i1p1f1/Amon/evspsbl/gn/v20190705
  ds_id: c3s-cmip6.CMIP.NUIST.NESM3.historical.r1i1p1f1.Amon.evspsbl.gn.v20190705
  var_id: evspsbl
  array_dims: time lat lon
  array_shape: 1980 96 192
  time: 1850-01-16T12:00:00 2014-12-16T12:00:00
  latitude: -88.57 88.57
  longitude: 0.00 358.12
- path: ScenarioMIP/CNRM-CERFACS/CNRM-CM6-1-HR/ssp245/r1i1p1f2/Amon/pr/gr/v20191202
  ds_id: c3s-cmip6.ScenarioMIP.CNRM-CERFACS.CNRM-CM6-1-
HR.ssp245.r1i1p1f2.Amon.pr.gr.v20191202
  var_id: pr
  array_dims: time lat lon
  array_shape: 1032 360 720
```

```
  time: 2015-01-16T12:00:00 2100-12-16T12:00:00
  latitude: -89.62 89.62
  longitude: 0.00 359.50
- path: CMIP/CNRM-CERFACS/CNRM-CM6-1/historical/r1i1p1f2/Amon/tas/gr/v20180917
  ds_id: c3s-cmip6.CMIP.CNRM-CERFACS.CNRM-CM6-1.historical.r1i1p1f2.Amon.tas.gr.v2018
0917
  var_id: tas
  array_dims: time lat lon
  array_shape: 1980 128 256
  time: 1850-01-16T12:00:00 2014-12-16T12:00:00
  level: 2.00 2.00
  latitude: -88.93 88.93
  longitude: 0.00 358.59
```

5 Deprecation of data, versions, DOI, citation, acknowledgement and licence

Versions, deprecation of entries, replacement of data

Amount of data to deprecate	Provider's role	CDS team
Large amount of data	• Provide old and new data in the same updated manifest file • Keep old and new data • Remove deprecated data and corresponding lines from the manifest at the end of the deprecation period	• Deprecate the whole entry and create a new one. The deprecated entry will not be searchable in the CDS, but API request will continue to work. (This prevents new users to find and download deprecated data, allowing at the same time scientific traceability and reproducibility), Example: Deprecated SST[11]. New entry with corrected data: Corrected SST[12] • Remove the deprecated data after 1 to 3 year deprecation period
Small amount of data	• Create new files with a different version tag for the corrected data • Include those files in the manifest • Manifest should contain both old and new versions • Remove deprecated data and corresponding entries in the manifest at the end of the deprecation period	• Deprecate the version of the data corresponding to the wrong data • Modify overview to explain the deprecation or use a new widget called "Known issues" under the Documentation tab • Modify the download form making clear the deprecated version of the data. (When the CDS will have the tools to do it: the deprecated data will only be accessible through the API). • Remove the deprecated data after 1 to 3 year deprecation period
No data will be replaced	• Provide information to fill in the Known-issues table following the guidance provided below	• Add the Know-Issues table to the documentation tab • Decide if further action are needed concerning user support

11 https://cds-test.climate.copernicus.eu/cdsapp#!/dataset/satellite-sst-esa-cci?tab=overview
12 https://cds.climate.copernicus.eu/cdsapp#!/dataset/satellite-sea-surface-temperature?tab=overview

DOI, citation, licence and acknowledgement

DOI

Type of data in the CDS Catalogue entry	Provider's role	CDS team
Data **without** DOI issued before the publication in the CDS Catalogue	• No active role	• Provides a DOI to the Catalogue entry (which can be see as a DOI for the data themselves) Example: https://cds.climate.copernicus.eu/cdsapp#!/dataset/cems-glofas-reforecast?tab=overview

Data **with** DOIs issued before publication in the CDS Catalogue	• Provides a mapping between the data and the previous DOIs	• Create a DOI box allowing for multiple DOIs.
		• DOI's box will show
		  • all the DOIs supplied by the data provider (with a clear association to which data they refer to)
		  • the DOI of the Catalogue entry itself
		Example:
		https://cds-test.climate.copernicus.eu/cdsapp#!/dataset/satellite-total-column-water-vapour?tab=overview
		https://cds-dev.copernicus-climate.eu/cdsapp#!/dataset/satellite-cloud-properties?tab=overview

Mixing of data **with and without** DOIs issued before the publication in the CDS Catalogue	• Provides a mapping between the data and the DOIs	• Create a DOI box allowing for multiple DOIs.
		• DOI's box will show
		  • all the DOIs supplied by the data provider (with a clear association to which data they refer to)
		  • the DOI of the Catalogue entry itself
		  • data with no DOI attribute will be associated with the string: "no specific DOI"
		Example:
		https://cds-dev.copernicus-climate.eu/cdsapp#!/dataset/satellite-surface-radiation-budget?tab=overview

Citations

Citations are like file formats, there are a few available, no one better than all the others in all situations.
The "Citation" link in the Catalogue entry **does not say how** people should cite the data, that depends on the journal, site and publisher where the data will be cited.
The "Citation" link in the Catalogue entry **is the Catalogue citing** the contents that it is exposing.
In this way it also shows how to cite the data, but that is just an example of how to cite the data and contents from where **people can extract all information** required to cite the data using other formats in other places.

Type of data in the CDS Catalogue entry	**Provider's role**	**CDS team**
Data **without** citation issued before the publication in the CDS Catalogue	• Interact with the CDS team on this. Most probably you will be asked for the names of the authors of the data.	• Interact with the provider and create a Citation following the Catalogue citation format

Data **with** citation issued before publication in the CDS Catalogue	• Provides those citations to the CDS team	• Create a Citation box allowing for multiple citations.
		• Citation's box will show
		  • all the Citations supplied by the data provider (with a clear association to which data they refer to)
		  • the Citation of the Catalogue entry itself
		
		Example:
		https://cds-test.climate.copernicus.eu/cdsapp#!/dataset/satellite-total-column-water-vapour?tab=overview
		https://cds-dev.copernicus-climate.eu/cdsapp#!/dataset/satellite-cloud-properties?tab=overview

Mixing of data **with and without** citations issued before the publication in the CDS Catalogue	• Provides those citations to the CDS team	• Create a Citation box allowing for multiple citations.
		• Citation's box will show
		  • all the Citations supplied by the data provider (with a clear association to which data they refer to)
		  • the Citation of the Catalogue entry itself
		Example:
		https://cds-dev.copernicus-climate.eu/cdsapp#!/dataset/satellite-surface-radiation-budget?tab=overview

Licence

Provider's role	CDS team
Provide all licences related to the data and a mapping between the licences and the parts of the data they are related to	For datasets with multiple licences use a "Origin" button in the download form making related to the name of the licence Example: https://cds-dev.copernicus-climate.eu/cdsapp#!/dataset/satellite-surface-radiation-budget?tab=form

Acknowledgement

Provider's role	CDS team
No active role but may want to have a look at: How to acknowledge and cite a Climate Data Store (CDS) catalogue entry and the data published as part of it[13] https://cds-test.climate.copernicus.eu/cdsapp#!/dataset/satellite-total-column-water-vapour?tab=overview	For datasets with multiple licences use a "Origin" button in the download form making related to the name of the licence Example: https://cds-dev.copernicus-climate.eu/cdsapp#!/dataset/satellite-surface-radiation-budget?tab=form

13 https://confluence.ecmwf.int/display/CKB/How+to+acknowledge+and+cite+a+Climate+Data+Store+%28CDS%29+catalogue+entry+and+the+data+published+as+part+of+it

6 **See Publishing under FAIR principles[14] for general principles.**

14 https://confluence.ecmwf.int/display/PS/Publishing+under+FAIR+principles

Publishing under FAIR principles

Production Section

Exported on 03/19/2025

Table of Contents

Search

1 Core interpretation of the FAIR principles

The core principles of the FAIR guidelines have not changed since they were first published in 2016 (*), and have since been widely adopted by the scientific community as a way to improve the quality and usability of research data. However, the principles are intended to be flexible and adaptable, and different organizations and communities may have different interpretations and implementations of the principles. It's also possible that the principles may be updated or refined over time as the field of data science and technology evolves.

FAIR principles are a set of guidelines for making data more Findable, Accessible, Interoperable, and Reusable:

1. Is the data Findable?
   Can the data be easily discovered by those who need it, using relevant keywords and metadata?
2. Is the data Accessible?
   Can the data be accessed, read, and understood by a machine or a human? Is it available in a widely used, open format?
3. Is the data Interoperable?
   Can the data be easily integrated with other data sources, using common standards and formats?
4. Is the data Reusable?
   Can the data be used and reused for multiple purposes, without significant effort or additional licensing restrictions?

If the data meets all of these criteria, it can be considered "FAIR." It's important to note, however, that the FAIR principles are guidelines rather than strict rules, and different organizations and communities may have different interpretations and implementations of the principles.

(*) Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Source : Australian National Data Service (ANDS)
Entire FAIR resources graphic is licensed under a Creative Commons Attribution 4.0 International License

2 ECMWF data: Extending the FAIR principles to all our data

At the ECMWF, we are committed to making our data as useful and accessible as possible. That's why we aim to publish all our data in accordance with the FAIR principles.

Our data is carefully curated and described using relevant metadata, that provides detailed information about the variables and parameters included in the data. For each variable, we provide a clear definition, specify the units, and include any relevant notes or caveats that users should be aware of, ensuring that the data can be used accurately and reliably.

We use DOIs (Digital Object Identifiers) to provide persistent, stable links to our data, allowing users to easily find and access the data they need. We also use open, standardized formats for our data and provide API (Application Programming Interface) access, allowing users to easily integrate our data with other systems and applications.

And we provide clear licensing information, enabling users to freely reuse and repurpose the data for their own purposes.

By following the FAIR principles, we are helping to make our data more valuable and useful for a wide range of users, from meteorologists and researchers to policymakers and the general public. We are proud to be part of the growing community of organizations that are working to make data more FAIR (*).

(*) OGC FAIR Climate Services: ECMWF is co-chair for the OGC Climate Resilience Domain Working Group[1]

- Decommissioning plan of ECMWF public datasets service[2]

1 https://www.ogc.org/blog/4460
2 https://confluence.ecmwf.int/display/PS/Decommissioning+plan+of+ECMWF+public+datasets+service