Are our climate data fit for your purpose?

Dick Dee, a André Obregon, b Carlo Buontempo b

a Planet-A consulting, Hudson, NY
b European Centre for Medium-Range Weather Forecasts, Bonn, Germany

Corresponding author: Dick Dee, dick.dee@planet-a.consulting
ABSTRACT

This paper describes the vision and objectives for evaluation and quality control of datasets available on the Climate Data Store (CDS) of the Copernicus Climate Change Service (C3S). The CDS is a cloud service providing reliable, open, and free access to climate datasets from many different sources, including observations and derived climate data records, global and regional reanalysis products, and output from global and regional climate models. CDS data and information are widely used to monitor climate change, for policy development, environmental impact studies, business plans, public awareness, and many other purposes.

CDS users include policy makers, technical consultants, data scientists and others who are climate-literate but not necessarily specialised in climate science. A key objective for C3S is to simplify data discovery and access, and to ensure that all information provided is fit for purpose. The role of evaluation and quality control is to generate useful statements about the technical and scientific quality of the datasets available on the CDS, presented in a form that can be mapped directly to user requirements associated with specific application contexts. The ambition is to enable users to select, locate, access, and process climate data that are fit for purpose, simply by expressing their requirements in their own words.

SIGNIFICANCE STATEMENT

This article is about making it easier for non-specialists to use climate data effectively for purposes that extend beyond scientific analysis, such as policy development, environmental impact studies, business plans, public awareness, etc. In addition to simplifying access to the wealth of climate data available today, we need to provide more information about data quality that users can understand and use to determine fitness for purpose.

CAPSULE (BAMS ONLY)

Users of climate data need relevant information about data quality.

Introduction

Imagine a dialogue between an engineer working in the wind energy sector and an AI chatbot designed to provide guidance on the use of climate data, as follows:

>> Hello ChatC3S: I need reliable information on future trends for wind-energy capacity factors in NW Europe. Can you help me with this?
Hello C3S user: Trends in capacity factors for wind energy are mainly driven by climate variability and change. However, CMIP (global) and CORDEX (regional) climate model projections do not provide definitive trend estimates for average wind patterns within the specified region. Seasonal forecasts may offer useful information about prevailing broad scale circulations, but they have limited skill in predicting wind speed for this region.

Estimates of capacity factors for recent decades can be obtained from reanalysis products such as ERA5 (global coverage) and CERRA (European coverage). Reanalysis data can be used to estimate trends for well-observed meteorological parameters, but it is advisable to compare those estimates with direct observations, such as provided by E-OBS.

Local estimates of capacity factors in coastal areas are available in the IPCC Global Wind Atlas, derived from reanalysis data and bias adjustments based on machine learning techniques. Leveraging these tools could provide more accurate descriptions and predictions for your site of interest.

It is worth noting that capacity factors can also be affected by changes in land use. I recommend reviewing any recent shifts in land characteristics in the area of interest, for example by using the Corine Land Cover dataset derived from satellite observations.

Would you like me to point you to additional information about these topics? I can also provide the code needed to access the datasets I mentioned.

The remarkable ability of large language models to process and summarize massive amounts of information in response to free-form user prompts has been demonstrated in many application areas, including development of climate services (Koldunov and Jung, 2024). The strengths and limitations of these models essentially depend on having an abundance of high-quality data available for training. A generalized AI assistant as suggested above cannot be realized without access to a great deal of reliable information concerning technical and scientific quality aspects of climate data. Today, such information is not readily available, or not easily accessible, for most of the datasets used in climate science and climate services.

The purpose of this paper is to describe the approach adopted by the Copernicus Climate Change Service (C3S, Buontempo et al. 2022) for evaluation and quality control of the
climate datasets available on its Climate Data Store (CDS, Raoult et al. 2017). The approach is focussed on generating helpful information for users who need to determine whether a dataset is fit for purpose, given their use case and any specific concerns about data quality—with or without help from an AI assistant.

**Background**

The Copernicus Climate Change Service (C3S, Buontempo et al. 2022) was established by the European Commission in 2014 to support adaptation and mitigation policies of the European Union and increase public awareness of climate change and its impact on society. The mission of C3S is to provide reliable access to consistent and authoritative information about climate change on global and regional scales, and to enable and support climate services in the EU member states. All C3S data products and most of its services are free and open to unrestricted use worldwide.

A core objective for C3S is to simplify access to quality-controlled data about the climate and to make it much easier to use those data effectively in a variety of applications. The focus is on serving users who need climate information for policy development, environmental impact studies, business plans, public awareness, and many other purposes that extend beyond scientific analysis. Here are just a few examples from the C3S webpages (https://climate.copernicus.eu/):

- Assessing health impacts of climate change in Europe due to increased seasonal activity of tiger mosquitos.
- Mapping changes in rainfall erosivity and soil loss in Italy under future climate scenarios.
- Exploring and visualizing global glacier elevation and mass changes since 1850.
- Assessing the impacts of climate change on hydrological variables for European river basins.
- Estimating the impacts of past and future climate change on energy supply and demand in Europe.

C3S users involved in developing these and many other applications include domain experts, data scientists, technical consultants and others who are climate-literate but not necessarily specialised in climate science.

Simplifying access to scientific data about the climate is challenging, given the multiplicity of data sources and the various ways they are managed. To illustrate, the latest
IPCC Assessment Report refers to 283 different observational datasets (IPCC 2021, Annex I), including in situ reference measurements, data records derived from satellite observations, and several different reanalysis products. The means of access to those data, their file formats, data structures, available metadata and documentation vary greatly. Output from climate model simulations (IPCC 2021, Annex II) is somewhat better organised, although the variety of available model versions, their defining properties and simulation types can be daunting even to specialists.

The IPCC Assessment Reports represent many years of work by experienced climate scientists. Some have helped to create the datasets; most are trained to use them. C3S users have different skillsets and probably less time available for investigating and deciphering data sources that are unfamiliar to them. To serve those users well, the challenge for C3S is twofold. First, build a technical infrastructure that makes it much easier to locate, access, process and combine different types of climate data from disparate sources. Second, find a way to provide users with the information needed to select the right data for their applications and to use them effectively.

The first part of this challenge is well addressed by the C3S Climate Data Store (CDS; http://cds.climate.copernicus.eu). The CDS is a cloud service that provides open and free access to a wide variety of climate datasets from different sources. All datasets can be accessed via a consistent web interface with dynamic forms for selecting and downloading data, or by using an application programming interface (API) provided by C3S. The CDS contains extensive documentation for all datasets and offers tools for working with the data online to create simple dynamic web applications. There are currently more than 300,000 registered users of this service, with approximately 2,500 active users daily and delivery of 1 petabyte of data per week. Real-time usage information for the CDS is available at https://cds.climate.copernicus.eu/live/.

The second part of the challenge, which is to develop useful information about data quality and fitness for purpose, is much more complicated. C3S already provides good documentation, technical user support and training services along with the datasets on offer. However, this does not fully address the needs of all users. Some questions are not easily answered. Which dataset meets my specific requirements? If there are alternatives, how do I select the most suitable? How do uncertainties in the data affect the information that I need? Are there other use cases similar to mine, and what can I learn from them?
Clearly, such questions depend on context and are best addressed in dialogue with users. Besides user support and training, C3S conducts targeted outreach events and collaborates directly with users as part of its user engagement strategy. Those activities generate valuable information about the users: who they are, how they work and what they need. C3S uses this information to improve its products and services and thus to increase user uptake. User outreach activities also create opportunities to engage with users on the topic of data quality.

Most repeat users of the CDS obtain data via the API rather than the web interface. First-time users are more likely to visit the CDS web pages to explore datasets and select the data they need, based on the information available there. The vision for the future is that users can locate and select data simply by expressing requirements in their own language. This can be realised with machine learning tools based on large language models, assuming that suitable training data can be supplied. The training dataset must include accurate information about technical attributes, scientific quality aspects, and fitness for purpose associated with a wide range of applications, for all datasets available on the CDS. The aim of this paper is to explain how C3S will generate this information.

_evaluation and quality control of CDS datasets_

C3S invests substantial resources in evaluation and quality control for all its products and services. The goal is to ensure that users are served well and that this will continue to be the case as their needs evolve. Achieving this goal requires an effective strategy for obtaining clear and realistic requirements from users, combined with an operational framework for conducting independent checks and quality assessments on all service elements.

C3S has implemented a transparent and traceable process for collecting and analyzing user requirements. These can take many forms, e.g. a climate data record for permafrost, faster access to reanalysis data, better documentation of a specific data set. Such requirements are collected by various means, e.g. proactively in workshops and training events, via user surveys, questionnaires, and other interactions with actual and potential users as opportunities arise. C3S maintains a comprehensive database of user requirements with technical details, context and source of each requirement. Requirements are continuously analyzed, clustered and then prioritized for implementation based on feasibility, benefit to users and cost effectiveness.
To date, most of the evaluation and quality control activities have focused on the CDS. Several tools have been implemented for monitoring quality of service provided by the CDS infrastructure as a first step. A prototype dashboard with real-time information about key performance indicators, such as availability, response time, and data transfer rates, is available on the CDS web pages (https://eqc.climate.copernicus.eu/monitoring). The main challenge, however, is to provide information about the quality of CDS data in a form that can benefit users.

Evaluation and quality control of the datasets available on the CDS is—and will remain—a work in progress. It takes time to learn what non-specialist users mean by data quality, which kind of information is useful to them, and how is it best communicated. In addition, the CDS catalogue, which today includes more than 150 datasets from different sources, continues to evolve. Data include observations and derived estimates of climate variables, global and regional reanalyses, climate predictions for different timescales, scenario-based climate projections, and other types of numerical data related to climate change. Some examples are:

- A subset of CMIP6 climate projections from multiple models.
- Hourly reanalysis data for atmospheric variables from ERA5.
- Seasonal forecasts of ocean variables from multiple models.
- In situ observations from the GRUAN reference network.
- Multi-model seasonal forecasts of river discharge for Europe.
- Lake water levels estimated from satellite observations.
- Glaciers elevation and mass change data from 1850 to present.
- Sea surface temperature data estimated from satellite observations.

Each dataset includes a coherent set of data records together with associated metadata and documentation. C3S works closely with data producers and domain experts to prepare and organise each dataset and to ensure reliable access and a consistent user interface. Once integrated in the CDS infrastructure, the dataset can be accessed by anyone with an internet connection, either from a web browser or by using software that connects directly to the CDS via the API. A web page associated with the dataset shows a brief overview of data content and its provenance (Figure 1, panel b). The tabs along the top link to separate sections for selecting and downloading data, information about data quality, and available documentation.
The data quality section of the web page contains information generated by evaluation and quality control (e.g. panels c and d in Figure 1). Most users today access the CDS via the API, which enables automated retrieval and remote processing of data. The API will be extended to enable access to information about data quality. This will open the door to development of AI assistants that can help users with questions about data selection and fitness for purpose.

Figure 1: Web access to the CDS (panel a). A search for “Seasonal forecasts” leads to a monthly forecast dataset generated by the C3S multi-system seasonal forecast service (panel b). From this page, users can select and download data, locate all relevant documentation, and access a variety of information about data quality generated as described in this paper (panels b, c, d).

Previous work by C3S on evaluation and quality control of CDS datasets (Nightingale et al. 2019; Lacagnina et al. 2022; Yang et al. 2022) has focused on creating comprehensive quality assurance reports, which contain detailed technical information about data, metadata and documentation, assessments of data maturity, results of routine checks on data integrity and completeness, output from various diagnostic tools, and descriptions of relevant use cases. Much of this content is useful, but there is too much of it, and the presentation of the
information to users has not been effective. As a result, the approach was revised to focus more on user requirements and the type of questions that users may have about data quality and fitness for purpose.

Information about data quality is especially valuable to users if it helps them find the right data for their work and make good use of it. “Good use” can mean many things: following established best practices; accounting for uncertainties in the data; finding new applications, etc. The central question is about fitness for purpose: “How, and how well, can I use these data for my application?” As mentioned, this question and its possible answers can take many forms depending on context. It may lead to further questions and consideration of alternative choices, and ultimately result in new and innovative ways of using data.

Fitness for purpose can be defined as the degree to which a given dataset meets user requirements for a specific application. Accordingly, the main goal for the evaluation and quality control of CDS datasets is to develop precise statements about data quality that pertain to well-identified use cases. Those statements, in combination with other documented information about the datasets, constitute a knowledge base that can help users to assess fitness for purpose, given their needs and requirements.

The vision is that a user can employ a range of tools for locating, selecting, and accessing climate data, simply by expressing requirements in their own language. To make this possible, it will be necessary to map user requirements to the relevant properties of a dataset. Some of those properties correspond to technical requirements that are quite straightforward and easily verified. For example: requirements on spatial coverage and representation; temporal extent and resolution; availability of estimates for specific variables; access to near-real time data updates; standard file formats; documentation content; etc. Such requirements are typically imposed by the original data producers or by C3S as the provider of the dataset.

1 A slightly different definition of data quality is provided by the ISO 2502 Data Quality model: “the quality of a data product is the degree to which data satisfy the requirements defined by the product-owner organization” (ISO, 2008).
Scientific requirements related to uncertainties and consistency, on the other hand, are more complicated. For example: representation of regional trends in precipitation; accurate off-shore wind speed estimates relative to in-situ measurements; physically consistent land surface parameters; accurate representation of the hydrological cycle. More precise statements can be formulated for specific applications, but they may not always be easy to verify.

Accordingly, we make a clear distinction 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. The purpose of quality assessments, on the other hand, is to provide science-based information about accuracy, uncertainties, strengths, and weaknesses of a dataset in the context of realistic use cases. Taken together, the outcomes of these activities provide the key information needed to determine fitness for purpose (Figure 2).
Figure 2: Outline of the C3S framework for evaluation and quality control of CDS datasets.

To illustrate, an important use case for C3S is the preparation of the annual European State of the Climate report (https://climate.copernicus.eu/ESOTC), which has a European
focus but also includes the global context. Various CDS datasets are used for this publication, including climate data records derived from satellite observations, \textit{in-situ} observations, and reanalysis products. The data sources used meet various technical requirements related to spatial coverage, timeliness, length of record, documentation, and traceability. They also meet scientific requirements essentially related to representation of climate change and variability, with physical consistency across multiple variables. Fitness for purpose in this case can be well ascertained based on outcomes of quality assurance and quality assessments of the datasets involved.

\textbf{Implementation}

Evaluation and quality control of C3S datasets involves a dedicated team of evaluators, including experts on observations, climate data records, reanalysis, climate predictions and climate models. The evaluation team is responsible for quality assurance of all CDS datasets and for conducting a variety of assessments to expose the scientific quality of the datasets. Based on the outcomes of these activities, the team will also generate summary reports on fitness for purpose of each dataset. Evaluators work closely with data producers and the CDS data management team to ensure that the information they produce is accurate and up to date.

\textbf{Quality assurance on technical requirements}

Quality assurance for each CDS dataset is implemented by verifying a set of well-defined technical requirements associated with the dataset. Many of those requirements correspond to specifications used for the production and/or procurement of the dataset. Others apply to the systems and processes used by C3S as the provider of the data. In any case, the normal expectation for a user of a CDS dataset is that all technical requirements pertinent to the dataset are always met. The role of the evaluators is to check that this is in fact the case, and if not, to explain why.

A useful analogy is that of a modern car, which has a dashboard feature that displays the status of various technical systems that are needed to drive safely. When starting the engine, the dashboard will indicate normal operation for all systems (good to go), flags for any anomalies (attention needed), and urgent information in exceptional cases (not safe to drive). The systems being checked are technically complex, but the driver does not need to be confronted with the details unless there is a problem. The main value is knowing that those systems are routinely verified.
To support the quality assurance process for CDS datasets, a database has been created that contains all technical requirements for all datasets available on the CDS. There are typically 20 to 30 requirements for each dataset. 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 the database. The outcomes can then be published along with the dataset on the CDS. Verification is repeated if needed, for example when a new version of the dataset becomes available, documentation is updated, an error is corrected, or any other event causing a significant change in the dataset.

Figure 3 shows a high-level organization (or dashboard) of all technical requirements for CDS datasets, divided in four categories with multiple sub-groups. The first category contains basic requirements on data management and stewardship:

- “All information provided with the dataset must be accurate and consistent.”
- “Access to all information must be open, free and reliable.”
- “Adequate version control and archiving policies must be in place.”

These requirements apply equally to all datasets published on the CDS. They are compatible with the FAIR principles for scientific data management and stewardship (Wilkinson et al. 2016), which are incorporated in the design and operation of the CDS infrastructure. Nevertheless, for each dataset published on the CDS, expert evaluators conduct a thorough inspection of data records, metadata and documentation to verify these requirements.
1. **Data Management**
   1. **Accuracy and Consistency**: All information provided with the dataset must be accurate and consistent.
   2. **Reliable Access**: Access to all information must be open, free and reliable.
   3. **Versioning and Archiving**: Adequate version control and archiving policies must be in place.

2. **Data Records**
   1. **Consistency**: The data must be complete and internally consistent.
   2. **Uncertainty**: The data must include information about uncertainties.
   3. **Updates**: Updates of the data records must be provided on a predictable schedule.

3. **Metadata**
   1. **Discovery and Use**: Metadata must include all necessary information for discovery and proper use of the data records.
   2. **Interoperability**: Metadata must comply with relevant international standards.

4. **Documentation**
   1. **Content**: The content of data records must be fully documented.
   2. **Scientific Basis**: The scientific methodology used to produce the data must be fully documented.
   3. **Quality Control**: Quality control and validation activities must be fully documented.
   4. **User Guidance**: Detailed information must be available to assist users in using the data.

Figure 3: High-level technical requirements for all CDS datasets.

All other technical requirements for CDS datasets can be associated to one of the three other categories and sub-groups shown in Figure 3. The requirements in those subgroups tend to be different for different types of datasets.

For example, one of the consistency requirements (2.1 in Figure 3) for the CMIP6 climate projections available on the CDS is: “Data records include climate simulations for several shared socio-economic pathways (SSPs) generated by a variety of global climate models.” This statement is compatible with the CMIP6 guidance for experiment design (Eyring et al. 2016) and can be verified by evaluators by inspecting the subset of CMIP6 data available on the CDS. A different formulation of the consistency requirement applies to climate data records derived from satellite observations: “Climate data records are periodically reprocessed in their entirety based on improved algorithms and/or newly available observations,” as recommended by GCOS (Dowell et al. 2013). An evaluator may need to consult with the producer of the dataset to verify this requirement.
An example of a requirement on documentation for a dataset with observations from the GRUAN reference network is: “Documentation includes detailed information about observation data coverage in space and time”. Two requirements for a seasonal forecast dataset: “The technical implementation of each seasonal forecast system used to produce forecast and re-forecast data must be documented” and “A description of known issues with the data must be available”. For metadata: “A digital object identifier (DOI) is associated with the dataset.” Most of these statements are easy to verify, and all of them contain information that may help users assess fitness for purpose.

The outcomes of the quality assurance process for each dataset are displayed on the CDS web pages in the form of a summary checklist like Figure 3. The user can click on the checklist to see the actual statements of technical requirements for the dataset, how they are verified, and additional information about non-compliance if needed. It may be the case, for example, that one of the requirements is only partly met, or not at all. In that case the evaluator will have provided an explanation with recommendations for the user.

**Quality assessments on scientific requirements**

Many salient aspects of data quality cannot be captured by technical requirements that are easily verifiable. Examples are bounds on accuracy and precision, sources of uncertainty, temporal consistency, strengths and weaknesses, and other information about a dataset that may be essential for determining fitness for purpose. Providing such information requires scientific assessments carried out by experts. In many instances, especially for datasets that are widely used, relevant information is available in the published literature. In either case, it requires a dedicated effort to generate or extract this type of information and present it in a form that is helpful to CDS users.

Clearly, the methods that can be used for quality assessments depend on the use case as well as type of data, e.g., observations versus predictions. For example, a use case for reanalysis data is the estimation of wind energy potential in different locations around the world. In this context, a user may have specific questions about the accuracy and temporal consistency of reanalysis data on wind speed and direction at 100m altitude. This can be addressed by an assessment that explores uncertainty information included with the reanalysis dataset, combined with in situ reference measurements where available. It is likely that relevant information for this use case can be found in the published literature as well.
The results of the assessment could include maps of average biases and standard errors, as well as timeseries or other means to indicate temporal consistency.

Other representative use cases for CDS datasets relevant to different industrial sectors have been published on the CDS (https://cds.climate.copernicus.eu/cdsapp#!/search?type=application). One of those is a tool for mapping changes in soil loss and rainfall erosivity in Italy under future climate scenarios. A user of this application might question the accuracy of those parameters in the reference period, which are estimated from in-situ observations combined with information from the ERA5-Land reanalysis. Further questions could be asked about the main sources of uncertainty in the EURO-CORDEX projections used in the tool, and how those uncertainties affect the results. A thorough assessment of those questions, based on published literature and additional data analysis if needed, could provide valuable information for users of this application as well as related applications based on similar methods.

C3S evaluators are tasked with developing quality assessments that are designed to generate useful statements about fitness for purpose of CDS datasets. The assessments must address concrete questions about data quality associated with real use cases. Assessments must build on relevant scientific literature, including published documents developed by producers and users of the datasets. If needed, the information extracted from the literature can be supplemented with additional analysis and visualizations implemented in Jupyter notebooks that can be shared, re-used, and modified by users.

Assessments may involve multiple CDS datasets and other sources of reference data. For example, a comparison of climate data records derived from satellite observations with reanalysis data can provide useful information about uncertainties, strengths, and weaknesses of the individual datasets. Using comparable in-situ reference observations in the assessment, if available, may sharpen the conclusions that can be drawn.

The quality section of the web pages for a CDS dataset (Figure 1) includes a searchable list of all quality assessments available for that dataset, with links to the literature, Jupyter notebooks and other relevant information.

**Conclusion**
C3S was established to support and stimulate the development of effective climate services based on consistent and authoritative information about climate change. Evaluation and quality control is essential to success; it is the only way to know whether user needs are met and how to improve products and services. Furthermore, reliable information about data quality based on scientific expertise and framed in terms of real use cases can be highly valuable to C3S users.

The approach to evaluation and quality control adopted by C3S is focussed on developing concrete statements about data quality that can be directly mapped to user requirements. Implementation of the approach is underway, but it will take time and a great deal of effort to assess most of the CDS datasets in ways that are helpful to most users. Moreover, as long as the CDS and its content continue to evolve along with user needs and requirements, evaluation and quality control will continue as well.

ECMWF will launch version 2.0 of the Climate Data Store by mid 2024, with improved performance, better interfaces, and new content and tools for working with the data. Information on data quality, following the approach described in this paper, will be a prominent feature of this new implementation of the CDS. In the meantime, development of AI solutions for helping users to access the right data for the job will continue.

Acknowledgments.

The Copernicus Climate Change Service is funded by the European Union and implemented by ECMWF in collaboration with many European institutions. The ideas and methods described in this paper represent the collective expertise and experience gained from this collaboration. In particular, the authors owe their gratitude to the many capable scientists and technical specialists who have contributed to the development of the framework for evaluation and quality control and its implementation.

Data Availability Statement.

The datasets discussed in this paper, together with documentation and information about data quality are openly available via the Copernicus Climate Data Store at http://cds.climate.copernicus.eu.
REFERENCES


