ECMWF Feature article

METEOROLOGY

Climate service develops user-friendly data store
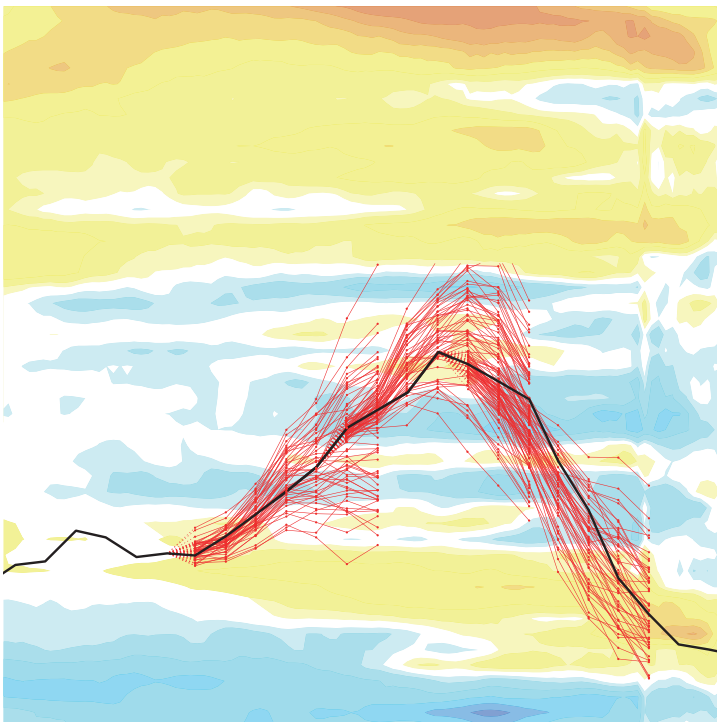
Climate service develops user-friendly data store

Baudouin Raoult, Cédric Bergeron, Angel López Alós, Jean-Noël Thépaut, Dick Dee

The European Commission has entrusted ECMWF with the implementation of the Copernicus Climate Change Service (C3S). The mission of C3S is to provide authoritative, quality-assured information to support adaptation and mitigation policies in a changing climate. At the heart of the C3S infrastructure is the Climate Data Store (CDS), which provides information about the past, present and future climate in terms of Essential Climate Variables (ECVs) and derived climate indicators.

The CDS is a distributed system. It provides improved access to existing but dispersed datasets through a unified web interface. Among other things, the CDS contains observations, historical climate data records, Earth-observation-based ECV datasets, global and regional climate reanalyses, global and regional climate projections and seasonal forecasts.

The CDS also provides a comprehensive set of software (the CDS toolbox) which enables users to develop custom-made applications. The applications will make use of the content of the CDS to analyse, monitor and predict the evolution of both climate drivers and impacts. To this end, the CDS includes a set of climate indicators tailored to sectoral applications, such as energy, water management, tourism, etc. – the so-called Sectoral Information System (SIS) component of C3S. The aim of the service is to accommodate the needs of a highly diverse set of users, including policy-makers, businesses and scientists.

The CDS is expected to be progressively extended to also serve the users of the Copernicus Atmosphere Monitoring Service (CAMS), which is also operated by ECMWF on behalf of the European Commission.

Requirements

The purpose of the C3S Climate Data Store is to:

- provide a central and holistic view of all information available to C3S

- provide consistent and seamless access to existing data repositories that are distributed over multiple data suppliers

- provide a catalogue of all available data and products

- provide quality information on all data and products

- provide access to a software toolbox, allowing users to perform computations on the data and products

- provide an operational system that is continuously monitored in terms of usage, system availability and response time.

A Copernicus Climate Data Store Workshop was held at ECMWF from 3 to 6 March 2015 to discuss more specific user requirements. The main messages from workshop participants were that:

- User engagement in the creation of the CDS is paramount.

- The CDS must provide different views to different users, based on their level of expertise.

- Login should only be required when downloading data or accessing the toolbox.

- There should be a user support desk, a user forum and training material.

- A 'find an expert' facility should be provided, so that users can get help from a knowledgeable source.

- All products and graphs in the CDS should share a common look and feel, with a special focus on representing uncertainty information.

- Data and product suppliers will have to follow agreed data management principles, in particular the use of digital object identifiers (DOIs).

- All content must be freely available.

- Although data will primarily be in binary form, the ability to display them in graphical form is required.

- It should be possible to sub-set, re-grid or re-project any product, as well as performing format conversions.

- The toolbox should provide advanced data analysis functions that operate on the content of the CDS and that can be combined into workflows, which can be shared between users.

- Workflows are invoked using the CDS web portal and can be parameterised by the end user.

- Tools should be based on existing software, such as Climate Data Operators or Metview.

- Results of workflows should be visualised on the portal.

- Workflows should record which data and tools have been used in order to provide traceability and preserve quality information (provenance).

- The CDS should not stop working under heavy usage; a quality of service module should be implemented.

- All functionalities of the CDS must be available via an application programming interface (API).

- The CDS must be continually monitored and assessed; key performance indicators (KPIs) must be defined and measured.

- The CDS must be based on standards and must be interoperable with other information systems (WIS, INSPIRE, GEOSS, GFCS, WCRP, etc.).

To this list should be added the requirement from the European Commission that the CDS must rely on existing infrastructure and that the data and products must remain with their producers. The CDS therefore had to be developed as a distributed system.

The CDS has been designed to meet these requirements and it will continue to evolve in line with user needs. In particular, the CDS is able to support users with very different needs, from researchers interested in very large volumes of raw data, such as outputs of climate projections, to decision-makers looking for a simple graph showing the result of a statistical analysis based on several sources of heterogeneous data. Users with different levels of expertise can interact with the CDS in different ways to transform data into information, as indicated schematically in Figure 1.
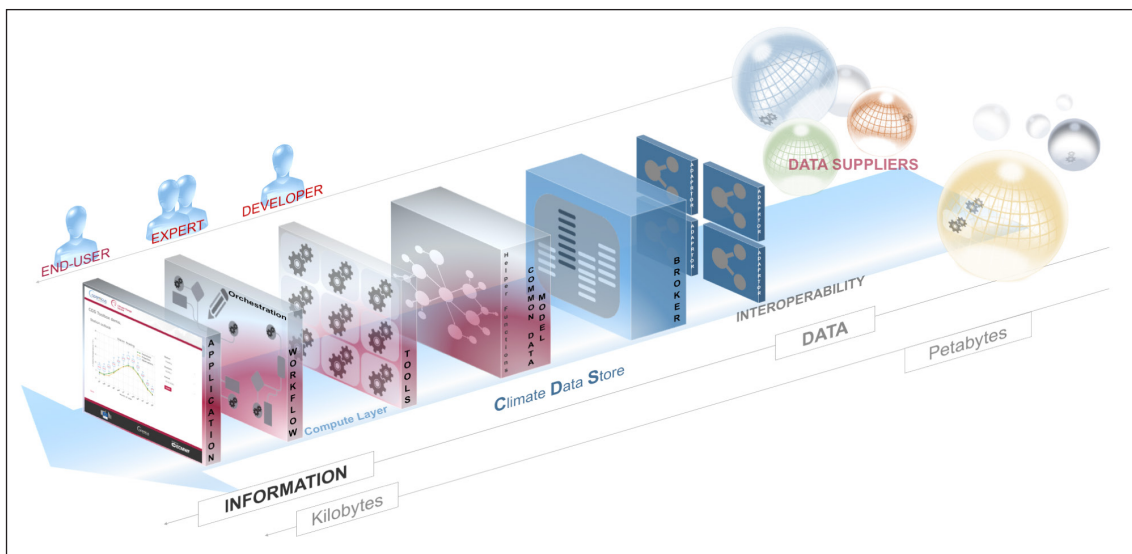
Figure 1 The CDS is designed to meet the needs of different users, including developers, expert users and a wide range of other users, who can interact with the CDS at different levels.

The CDS also supports a variety of data types (observations, fields…) as well as several data formats (NetCDF-CF, WMO-GRIB…). These data are stored using various technologies, such as relational databases, file systems or tape archives, at different sites, and they are accessible via different protocols, such as web services or APIs. The CDS aims to provide all available distributed data using a seamless and consistent user interface.

Although all data and products available through the CDS are to be freely available, they are subject to various licences.
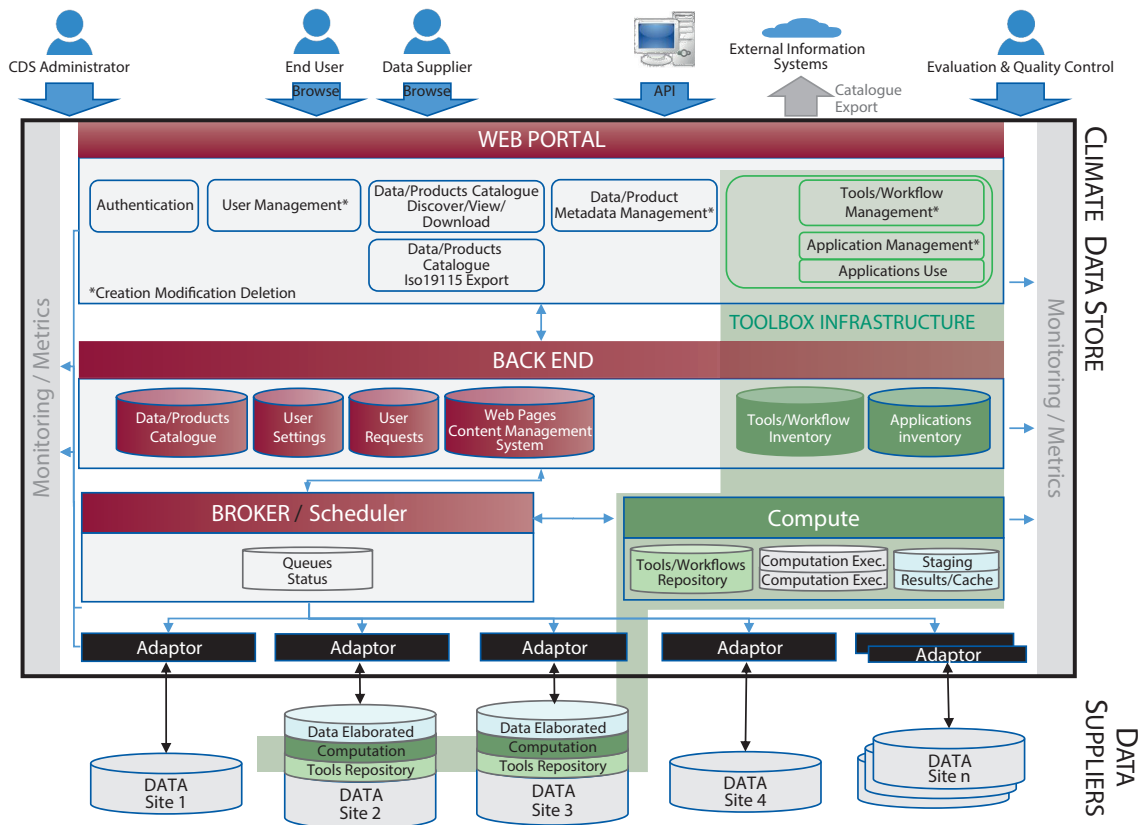
CDS infrastructure

The CDS infrastructure (Figure 2) aims to provide a 'one stop shop' for users to discover and process the data and products that are provided through the distributed data repositories. It also provides the environment to manage users, product catalogues and the toolbox.

Web portal

The web portal is the single point of entry for the discovery and manipulation of data and products available in the CDS. It allows users to browse and search for data and products in the CDS, perform data retrievals, invoke tools from the toolbox and visualise or download results. The web portal also displays the latest information on events and other news regarding the CDS as well as providing a help desk facility, FAQs, and a user forum.

A strong focus on the usability and user experience of the web portal and seamless integration between the different components is paramount. For example, the portal's search facility can be used to search across news articles, the product catalogue and the toolbox catalogue in a single, unified fashion.

Figure 2 Overview of the CDS infrastructure. The role of each component is described in the text. ECMWF is one of many data sites supplying data to the CDS.

Broker

This component schedules and forwards data and compute requests to the appropriate data repository (or the compute layer) via a set of adaptors. It also implements a quality of service module that queues requests and schedules their execution according to a series of rules, taking into account various parameters such as the user profile, the type of request, and the expected request cost (volume of data, CPU usage, etc.). The quality of service module is necessary to protect the CDS from denial-of-service attacks (DoS). Very large requests are given very low priority, and the number of simultaneous requests from a single user is limited. This guarantees that the system always remains responsive.

Adaptors

The role of these components is to translate data and computation requests issued by the broker on behalf of the user into requests that are understood by the infrastructure of each of the data providers. Each supplier is, as far as possible, expected to provide access to its data, products and computation facilities according to some agreed standards, such as OpenDAP, ESGF, OGC Web Coverage Service or a web-based REST API. Adaptors for each of the relevant standards are being developed.

Backend

The backend contains various databases, in particular a web content management system for news articles; a centralised catalogue that describes the data and products in the CDS; a toolbox catalogue; quality control information; and user-related information, such as the list of data licences users have agreed to and the status of their requests. The content of all databases is browsable and searchable in a seamless fashion since it is indexed in a single search engine. The databases also contain information that is useful for the operation of the CDS:

- information on how to sub-set large datasets

- terms and conditions and licences attached to datasets

- the location, within the CDS, of data and products, access methods and protocols

- information about what tools are applicable to given datasets.

All products and data will be assigned a digital object identifier (DOI). The data and products are described using the ISO19115 metadata record standard and are made available through the OAI-PMH and OGC-CSW protocols for interoperability with the World Meteorological Organization Information System (WIS) and the EU's INSPIRE initiative, respectively.

Compute

This component hosts computations that need to be performed on a combination of data retrieved from possibly several remote data repositories. Computations are limited to those provided by the tools from the toolbox. They can be performed at the data repositories where possible or in the cloud environment hosting the CDS.

Toolbox

The toolbox is a catalogue of software tools that can be classified as:

- *tools* that perform basic operations on data, such as computation of statistics, sub-setting, averaging, value at points, etc.

- *workflows* that combine the output of tools and feed this as input into other tools to produce derived results

- *applications*, which are interactive web pages that allow users to interrogate the CDS through parameterisation; applications make use of workflows and selected data and products in the CDS.

Applications are made available to users of the CDS via the web portal, allowing them to perform computations on the data and products in the CDS and to visualise and/or download their results. The toolbox is described in more detail below.

Application programming interface

A web-based, programming-language-agnostic API allows users to automate their interactions with the CDS. It provides batch access to all CDS functions, such as data and product retrievals and invocations of the toolbox.

Monitoring and metrics

All aspects of CDS operations will be measured and monitored. Any issues will be notified to 24/7 operators, who can refer them to on-call analysts. The information collected will be used to establish KPIs. Such indicators will measure the system in terms of responsiveness, speed, capacity, search effectiveness, ease of use, etc. These KPIs will be used when reporting to the European Commission and to enable continuous improvements in the service (capacity planning).

Toolbox

The toolbox is used by developers to create web-based applications that draw on the datasets available in the CDS. These applications can then be made available to users. Users are given some control over the applications by interacting with web form elements. This could involve the selection of a range of dates or a geographical area of interest, which is then used to parameterise the application.

All computations are executed within the CDS infrastructure, in a distributed, service-oriented architecture. The data used by the applications do not leave the CDS, and only the results are made available to the users. These results typically take the form of tables, maps and graphs on the CDS data portal. Users may be offered the ability to download these results.

The variety of data types, formats and structures makes their combined use highly challenging. The aim of the toolbox is to provide a set of high-level utilities that allow developers to implement applications without the need to know about these differences.

The toolbox hides the physical location of the datasets, access methods, formats, units, etc. from those who are developing the applications. Developers are presented with an abstract view of all the data available in the CDS. It is anticipated that all the datasets share some common attributes. Most of them represent variables which are either measured or forecast (e.g. temperature, amount of precipitation, etc.), and which are located in time and space. A common data model that is able to represent all datasets in the CDS has been adopted.

The toolbox also provides a series of tools that perform some basic operations on the datasets, such as averaging, calculating differences, sub-setting, etc.

All tools are registered in the toolbox database and documented within the CDS. Application developers implement workflows that invoke the various tools available to them on the datasets in the CDS. These workflows will also present the output from some of the tools as the input into other tools, in order to implement more complex algorithms. As is the case for tools, workflows are also registered and documented so that they can be reused.

Finally, application developers create web pages that present users with a series of widgets (check boxes, drop down menus, etc.) as well as various tables, maps or graphs, and associate the page with a selected workflow. The tables, maps and graphs represent different views on the result of executing the workflow parameterised with the values selected in the widgets. These views are dynamically updated as end users interact with the widgets.

After a validation process, applications may be published on the CDS web portal and made available to all users.

Common data model

The purpose of the common data model (CDM) is to provide a uniform description (conventions, structures, formats etc.) of all data and products in the CDS, so that they can be combined and processed by the toolbox in a consistent fashion. This entails the choice of a common naming convention, the use of agreed units, and a common representation of spatiotemporal dimensions.

The common data model is based on the CF (Climate and Forecast) convention, which is the convention used by the climate research community, in particular for the storage and exchange of climate projections.

Tools

The tools are simple Linux programs that transform some input data into some output data and can be parameterised. Good examples of such tools are the Climate Data Operators from the Max Planck Institute for Meteorology, which are widely used in the climate community to perform basic operations on NetCDF files. Such tools are invoked remotely through the broker as remote procedure calls when located in the compute layer, or via the adaptors when located at the data suppliers' location.

The tools have the ability to perform a wide range of operations over gridded data (point-wise operations) and/or time series data, such as basic mathematical operations (+, - , / , * , cos, sin, sqrt, …), statistical operations (minimum, maximum, average, standard deviation, …), field manipulations (sub-setting, re-gridding in time and space, interpolation, format conversion) and plotting (maps and graphs).

All tools are described in the toolbox catalogue, which will be available on the CDS website. The description provides information about input and output data types and formats, possible parameterisation and what algorithms are being used.

Workflows

A workflow can be defined as an orchestrated and repeatable pattern of processing steps that produce an outcome. In the context of the CDS, workflows can be viewed as Python scripts that invoke the tools remotely, through remote procedure calls, dispatched by the broker to either the compute layer or the data suppliers. The orchestrator is the process that executes the workflows. Workflow processing is carried out in isolation (sandboxing) with the use of containerisation software such as Docker. Like tools, workflows can be parameterised, can make use of input data, and produce output data. They are also described in the toolbox catalogue.

Applications

In the context of the CDS, applications can be seen as web pages that are associated with a workflow, some data sources, and some parameterisations. Each time the user interacts with the widgets on the web page, the workflow is executed, parameterised with the values of the widgets, and any resulting plots or graphs are updated. Figure 3 is a mock-up of what an application might look like. When a user selects a new period or a new area using the drop-down menus, the map and graph are automatically updated accordingly. There are several actions associated with applications:

- *Authoring*: a user can create/update an application by selecting a workflow, providing some fixed parameterisation, laying out some text, maps and graphs and adding some widgets.

- *Publishing*: once authored, an application can be published and catalogued so that other users can use it. Users can clone published applications and modify them for their own use or publish a modified version.

- *Using*: users are limited to the interaction provided by the widgets.

A hypothetical toolbox use case is presented in Box A. It demonstrates how tools, workflows and applications fit together and interact with the orchestrator and the broker.
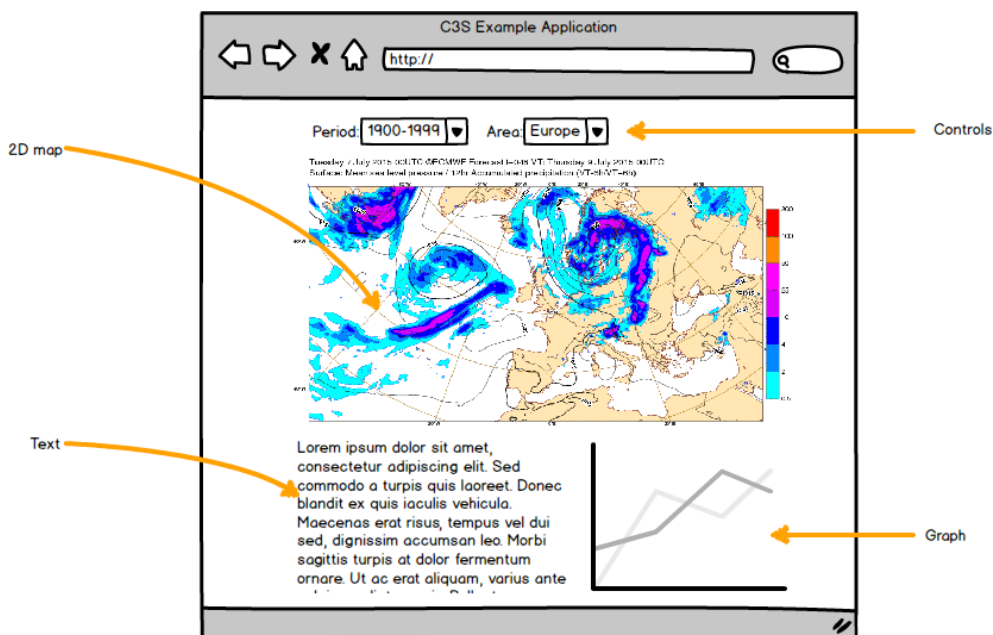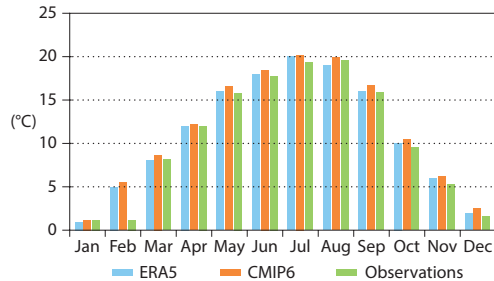

Figure 3 Mock-up of what an application might look like.

A

Toolbox use case

This toolbox use case describes a hypothetical application (i.e. an interactive web page on the data portal) in which users can select a city from a drop-down menu and are then presented with the following chart:


Example output of a Toolbox application that combines data obtained from three different data suppliers. The vertical axis is the average monthly temperature in °C.

Each of the bars represents a monthly average of the surface temperature at the selected location, where:

- *ERA5* is a global reanalysis located at ECMWF, covering 1979 to today. The dataset is composed of global fields encoded in WMO-GRIB and the temperature is expressed in Kelvin.

- *CMIP6* is a 200-year climate projection scenario located in an ESGF (Earth System Grid Federation) node accessible via the CDS. The dataset is composed of global fields encoded in NetCDF-CF and the temperature is expressed in Kelvin.

- *Observation*s is a time series of observed temperatures located in an SQL database at a hypothetical CDS data supplier called ClimDBase. The temperature is expressed in Celsius.

For the purposes of the example, we assume that there is a service that can map a city name (e.g. Reading, UK) to a latitude/longitude (e.g. 51°N, 1°W). The following additional assumptions are made:
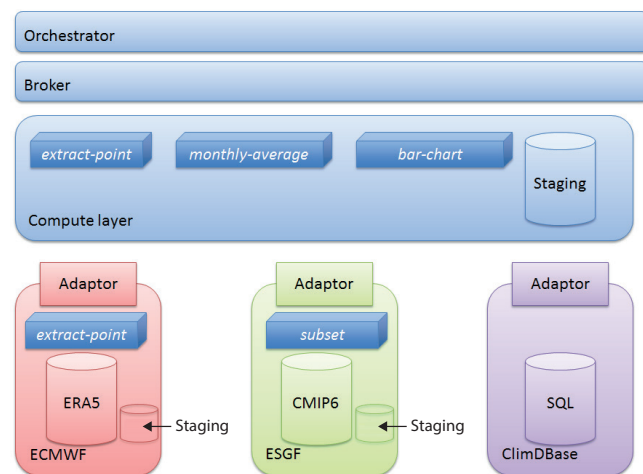
- The ECMWF data repository holds an extract-point tool that can be run directly on the data to interpolate the value of the data at the requested location from neighbouring points.

- The ESGF data repository holds a subset tool than can be run directly on the data to return data over a small area (four points) around a requested point.

- The ClimDBase data repository holds long time series of observations at various measuring stations and can be queried by station name.

- The compute layer of the CDS is equipped with many more tools including extract-point to interpolate values from neighbouring points at a given location; monthly-average to compute the monthly average from a time-series; and bar-chart to plot a bar chart from several time series.

The diagram below illustrates this setup.

Assuming the end user selects 'Reading' in the application, the workflow associated with the application is passed to the orchestrator with city=Reading as a parameterisation. The workflow is then executed.

Each tool is invoked in the right order via the broker, which schedules the execution of the tool at the given location, either at a data supplier's location or in the compute layer. Each tool writes its output to a file in a staging area and returns a URL.

This use case illustrates the following points:

- Data and products in the CDS are available in different formats, with different units etc. For data and products to be processed and/or combined seamlessly by the tools, a common data model must be designed and implemented.

- Tools produce intermediate results: there is a need for staging areas to temporarily hold these results, and for housekeeping jobs to clean up the staging areas regularly.

- The orchestration of workflows relies on knowing what is where, i.e. where are the datasets located in the CDS (which data suppliers), where are the tools located (i.e. at the data supplier or in the compute layer), where are intermediate results located (i.e. in which staging area). The workflows must then make use of this knowledge to call the most appropriate service when invoking a tool. If a tool is available at several locations, the workflow selects the one that will make the best use of available resources (e.g. minimising data transfers).

- Calls to services, as well as their results, only consist of references to the actual data and rely on URLs to point to them.

- Intermediate results are cached to improve the performance of the system.

The toolbox makes it possible to combine tools in the CDS with tools available at the data suppliers to generate the desired output.

Status and plans

In order to develop the CDS, ECMWF issued two tenders: one for the infrastructure and one for the toolbox. The tenders were awarded to different consortia made up of organisations ranging from software development and web design to meteorology and climate, bringing a wide spectrum of expertise to the project. One of the main messages voiced by the 2015 workshop participants was that user engagement is the key to building a successful CDS. As a result, the principles of 'Agile' software development were adopted. With Agile, software is delivered in small increments, every few weeks. This makes it possible to provide continuous feedback to developers and to cater for additional user requirements as the project progresses. Outsourcing software development is a new activity for ECMWF. The use of Agile principles has enabled us to work very closely with the selected consortia. At the time of writing, most of the building blocks forming the infrastructure and the toolbox have been developed. There is nevertheless still a lot of work to do before a beta version of the CDS can be offered to selected users, which is expected to happen by July 2017.

It was decided at the outset that the CDS will be deployed into a Cloud infrastructure. A commercial Cloud was selected for the duration of the development phase. This has enabled ECMWF to get acquainted with such technologies. A tender will be issued to select a Cloud solution that will host the operational CDS.

It was also decided that the CDS should be based on open source software where possible, so that other instances could be deployed if necessary. This is particularly important for the development of the toolbox: there is a vibrant community developing scientific libraries in Python, such as Numpy, Scipy, Pandas, xarray, dask, matplotlib etc. These libraries provide many of the algorithms required, and users from the weather and climate communities are already familiar with them. Making use of those libraries will therefore make it easier for users to contribute new additions to the toolbox.