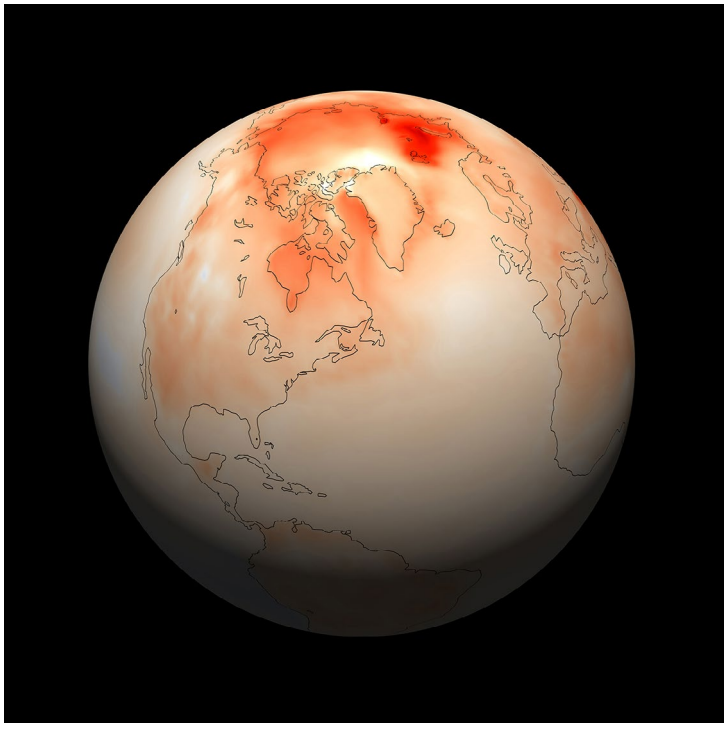
ECMWF Feature article

from Newsletter Number 159 – Spring 2019

COMPUTING

The ECMWF Production Data Store



Cover image: Point-wise linear trend in surface temperature between 1979 and 2018 from ERA5

www.ecmwf.int/en/about/media-centre/media-resources

doi: 10.21957/83deq5lgc0

This article appeared in the Computing section of ECMWF Newsletter No. 159 – Spring 2019, pp. 35-40.

The ECMWF Production Data Store

Laurent Gougeon

To ensure the timely delivery of ECMWF's forecasts to Member and Co-operating States and other users, observations from around the globe have to be collected fast and forecasts have to be delivered to users reliably and to a tight schedule. These processes of collection and dissemination used to be taken care of by separate software applications. However, the two activities have similar requirements for data services such as storage, transmission, scheduling, security and monitoring. After careful consideration, in 2016 ECMWF decided to combine the two applications into a single one, the ECMWF Production Data Store (ECPDS).

Since then, the ECPDS software has been developed in-house to support the goals of ECMWF's Strategy with the following objectives in mind:

- secure the evolving needs of the forecasting system for larger volumes (higher resolution) and a greater variety of observations
- support the increasing number of parameters for our forecast users
- speed up the delivery of our forecast products
- enable the ECMWF Data Services function to deliver products to national meteorological and hydrological services around the world within the framework of the World Meteorological Organization (WMO) and to a steadily growing number of commercial customers
- make the transition to cloud computing infrastructure for increased scalability and reliability.

This article presents the solutions that enable ECPDS to meet these objectives and explains its main features.

Basic features

ECPDS has been designed as a multi-purpose repository, hereafter referred to as the Data Store, delivering three strategic data-related services (Figure 1):

- Data Acquisition: the automatic discovery and retrieval of observational data from data providers
- Data Dissemination: the automatic distribution of meteorological products to our Member and Co-operating States and other forecast users
- Data Portal: the pulling of meteorological products and pushing of observational data initiated by remote sites.

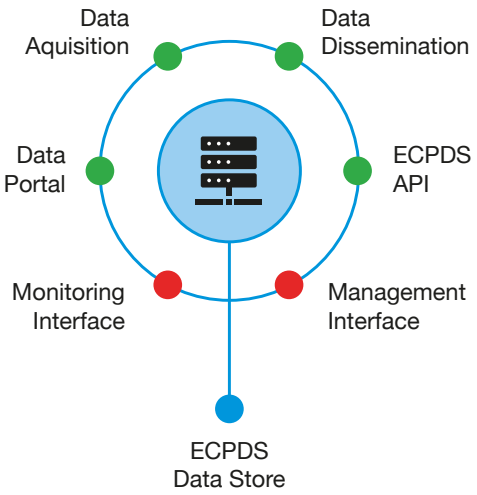


Figure 1 Schematic representation of the main components of ECPDS.

doi: 10.21957/83deq5lgc0

Data Acquisition and Data Dissemination are active services which are initiated by ECPDS, whereas the Data Portal is a passive service triggered by incoming requests from remote sites. In other words, the Data Portal service provides interactive access to the Data Dissemination and Data Acquisition services.

Unlike a conventional data store, ECPDS does not necessarily physically store the data in its persistent repository but rather works like a search engine, by crawling and indexing metadata from data providers. However, ECPDS can cache data in its Data Store. This is useful to ensure data availability in ECPDS without relying on instant access to the data providers. Feeding data into the Data Store can be achieved via different mechanisms:

- the Data Acquisition service discovering and fetching data from data providers
- data providers actively pushing data through the Data Portal
- data providers using the ECPDS API to register metadata: the ECPDS retrieval scheduler will asynchronously fetch data from the data providers or the data will be retrieved on the fly when needed.

Data products can be searched for online in ECPDS by name or metadata and their content can be either pushed by the Data Dissemination service or pulled from the Data Portal by users to their place of choice. Whenever data is required, ECPDS will either stream it on the fly from the data provider or send it straight from its Data Store, providing it was previously fetched by the ECPDS retrieval scheduler.

One key aspect of ECPDS is that it can easily interact with all sorts of environments without putting the burden on data providers or users to support one protocol or another:

- outgoing connections initiated by the Data Acquisition and Dissemination services use the most common protocols, such as ftp, sftp, ftps, GlobusFtp, http/s and AmazonS3
- incoming connections to the Data Portal service can be made using various protocols, such as ftp, sftp, scp and https.

Subsequent options vary depending on the selected protocol and authentication method (e.g. password-based vs key-based authentication, plaintext vs cyphertext connection, connection pooling vs straight connection, parallel connection vs serial connection, and more).

The ECPDS software is modular and every protocol is deployed as an extension to the system. All extensions make use of a common API which hides the complexity from the user and gives a unified view of the Data Store. On a regular basis, new protocols will be added when they become available or are required by some of our users. With the emergence of cloud computing, the AmazonS3 protocol was added recently and support for Microsoft Azure is currently in the pipeline. These new protocols are being introduced in ECPDS to make it possible to disseminate meteorological products directly to cloud object storage.

Another key aspect of ECPDS is its ability to reach a wide range of data providers (since forecast skill depends on the availability of observations from a range of sources around the world), users in Member and Co-operating States and others by operating through multiple networks:

- the Internet with a high-speed connection
- the Regional Meteorological Data Communication Network (RMDCN)
- the ECaccess Network with many ECaccess gateways deployed worldwide
- superfast bandwidth networks with dedicated leased lines.

Object storage

ECPDS stores data as ‘objects’, in other words as data with associated metadata and a globally unique identifier. The object storage system has been built on top of a file-system-based solution, with an efficient replication mechanism that allows continuous data availability across ECMWF and cloud platforms. To this end, ECPDS duplicates data in geographically separated storage locations, such as the US or Canada. This also brings ECMWF products closer to forecast users, enhancing data transfer performance to provide users with more immediate access to forecast products at scheduled release times. The ECPDS data persistence layer, along with its underlying raw storage, is also capable of interfacing with other object storage systems: in the future, ECPDS could expand its data storage capacity across object stores in the cloud, which would be useful in terms of scalability and reliability.

doi: 10.21957/83deq5lgc0

Like any object storage system, the one used in ECPDS is hierarchy-free and has no nested tree structure to store its data. However, ECPDS is able to emulate a directory structure when required. For instance, data providers can create directories when they transmit data to the Data Portal, and ECPDS will store this information as metadata along with the data. Later on, when the data is requested, ECPDS will be able to reuse this information to build a virtual file system with directory trees. Using the metadata enables ECPDS to present different views of the same data, depending on the configuration. For example, one user might see the data organised by date and another might see the data organised by type, depending on the users' preferences, managed via their ECPDS profile.

Other specificities of the ECPDS Data Store are:

- Data compression: this enables the Data Dissemination service to compress data. Compression can be performed either on the fly while transferring the data to the remote site, or in advance while the data is sitting in the queue before transmission. Most commonly used compression algorithms are supported, including zip, gzip, bzip or lzma. This is important because compression reduces dissemination time and therefore enables faster access to forecast products.
- Data checksumming: this makes it possible to generate, in advance or on the fly, MD5, CRC32, SHA-1 or SHA-256 one-way hashes to preserve the integrity of data against corruption. The Data Dissemination and Data Portal services can provide the cryptographic hashes along with the data, for verification. This feature is used routinely to identify corrupted files at remote sites.
- Garbage collection: every piece of data recorded in ECPDS is given an expiry date. The 'garbage collector' makes it possible to automatically remove data that no longer need to be stored. When required, the garbage collector can also take care of cleaning up on the data provider side. It is worth mentioning that there is no limit on the expiry date, which means that, if required, data can stay in the Data Store for ever.
- Data backup: this makes it possible to map entire sets of data in ECPDS to a wide variety of archiving systems (e.g. ECFS).

Useful links

Access to the ECPDS monitoring and managing interface:

- *https://ecpds-monitor.ecmwf.int*

Access to the Data Portal through the Internet:

- *https://dissemination.ecmwf.int*
- *https://acquisition.ecmwf.int*

Access to the Data Portal through the RMDCN:

- *https://acquisition-rmdcn.ecmwf.int*

Monitoring and management

ECPDS offers an uninterrupted 24/7 service and support for the acquisition and dissemination activities. Its monitoring interface is essential for the smooth operation of ECPDS. The software is interfaced with a Nagios server so that ECMWF operators can monitor its internal operations, but the system also has its own monitoring interface with dedicated tools to trace and debug issues specific to the ECPDS services.

An administration interface is also available to manage the various components of ECPDS:

- data storage: managing metadata, data content, transfer groups and Data Movers (activating or deactivating servers)
- transmission: managing destinations, transfer hosts (remote site settings) and transfer requests
- access control: managing users who perform monitoring tasks and users who have access to the Data Portal

doi: 10.21957/83deq5lgc0

- monitoring: a monitoring display for the Dissemination and Acquisition services.

Theses functionalities are available either through a website or a REST API for easy integration with other systems.

ECPDS concepts for users

Understanding some key ECPDS concepts will help users to benefit fully from the tool's capabilities:

- data files
- data transfers
- destinations
- dissemination and acquisition hosts.

A user connecting to the ECPDS web interface will come across each of these entities, which are related to each other through the different services.

A **data file** is a record of a product stored in the ECPDS Data Store with a one-to-one mapping between the data file and the product. The data file contains information on the physical specifications of the product, such as its size, type, compression and entity tag (ETag) in the Data Store, as well as the metadata associated with it by the data provider (e.g. meteorological parameters, name or comments concerning the product).

A **data transfer** is linked to a unique data file and represents a transfer request for its content, together with any related information (e.g. schedule, priority, progress, status, rate, errors, history). A single data file can be linked to several data transfers as many remote sites might be interested in obtaining the same products from the Data Store.

A **destination** should be understood as a place where data transfers are queued and processed in order to deliver data to a unique remote place, hence the name 'destination'. It specifies the information the Data Dissemination service needs to disseminate the content of a data file to a particular remote site (Figure 2).

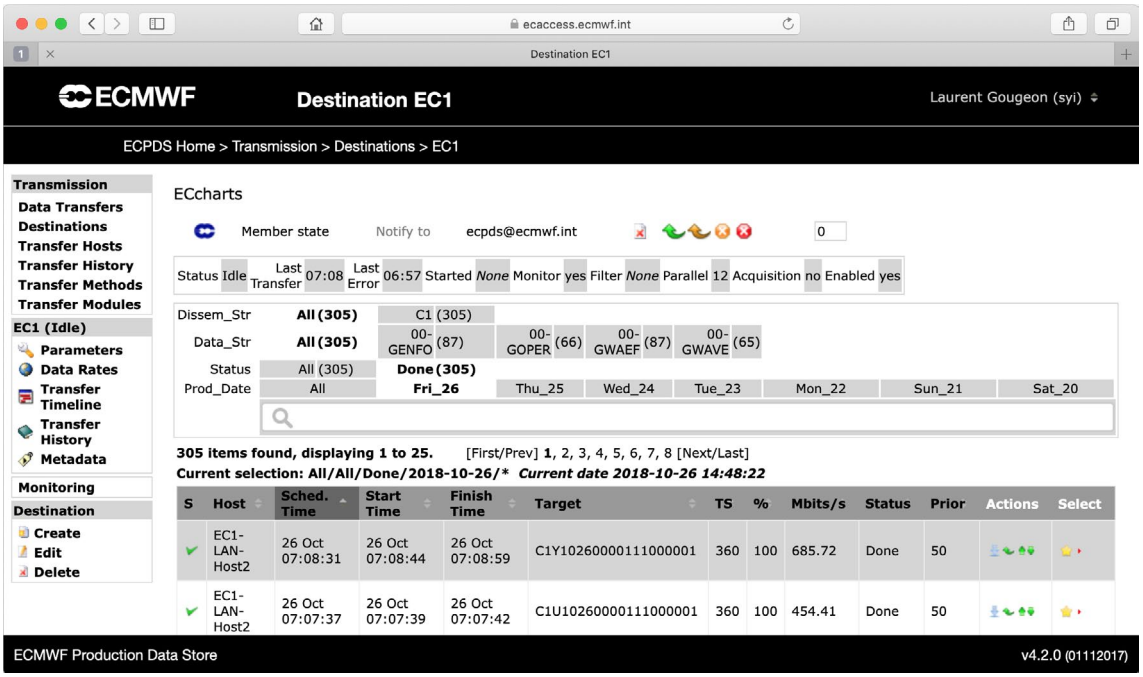


Figure 2 The ECPDS interface for internal and external users. A breadcrumb trail at the top shows where a user currently is in the tool. In this case, a user has created a destination called EC1. Users with the right credentials can see the status of this destination and can review the progress of data transmission. They can manage the destination by, for example, requesting data transfers, changing priorities and stopping or starting data transmissions.

Each destination implements a transfer scheduler with its own configuration parameters, which can be fine-tuned to meet the remote site's needs. These settings make it possible to control various things, such as how to organise the data transmission by using data transfer priorities and parallel transmissions, or how to deal with transmission errors with a fully customisable retry mechanism.

In addition, a destination can be associated with a list of **dissemination hosts**, with a primary host indicating the main target system where to deliver the data, and a list of fall-back hosts to switch to if for some reason the primary host is unavailable.

A dissemination host is used to connect and transmit the content of a data file to a target system. It enables users to configure various aspects of the data transmission, including which network and transfer protocol to use, in which target directory to place the data, which passwords, keys or certificates to use to connect to the remote system, and more.

If the data transfers within a destination are retrieved by remote users through the Data Portal service, then there will be no dissemination hosts attached to the destination. In this particular case, the destination can be seen as a 'bucket' (in Amazon S3 terms) or a 'blob container' (in Microsoft Azure terms). The transfer scheduler will be deactivated, and the data transfers will stay idle in the queue, waiting to be picked up through the Data Portal.

A destination can also be associated with a list of **acquisition hosts**, indicating the source systems where to discover and retrieve files from remote sites. Like their dissemination counterparts, the acquisition hosts contain all the information required to connect to the remote site, including which network, transfer protocol, source directory and credentials to use for the connection. In addition, the acquisition host also contains the information required to select the files at the source. Complex rules can be defined for each source directory, type, name, timestamp and protocol, to name just a few options.

A destination can be a dissemination destination, an acquisition destination or both. It will be a dissemination destination as long as at least one dissemination host is defined, and it will be an acquisition destination as long as at least one acquisition host is defined. When both are defined, then the destination can be used to automatically discover and retrieve data from one place and transmit it to another, with or without storing the data in the Data Store, depending on the destination configuration. This is a popular way of using ECPDS. For example, this mechanism is used for the delivery of some regional near-real-time ensemble air quality forecasts produced at ECMWF for the EU-funded Copernicus Atmospheric Monitoring Service implemented by the Centre.

There is also the concept of destination aliases, which makes it possible to link two or more destinations together, so that whatever data transfer is queued to one destination is also queued to the others. This mechanism enables processing the same set of data transfers to different sites with different schedules and/or transfer mechanisms defined on a destination basis. Conditional aliasing is also possible in order to alias only a subset of data transfers.

Developing all these features has made it possible for ECMWF to meet the requirements of our main user groups and stakeholders:

- Member and Co-operating States
- the WMO community
- commercial customers.

It also means that ECMWF can flexibly interface with all the observation providers around the world the Centre depends on.

 doi: 10.21957/83deq5lgc0

Physical architecture

ECPDS is a distributed application with three main software components (Figure 3):

- the ECPDS Master implements the business logic of the application
- the MySQL Database enables persistent storage for configurations, metadata and history
- Data Movers make it possible to store objects and to perform incoming and outgoing data transfers.

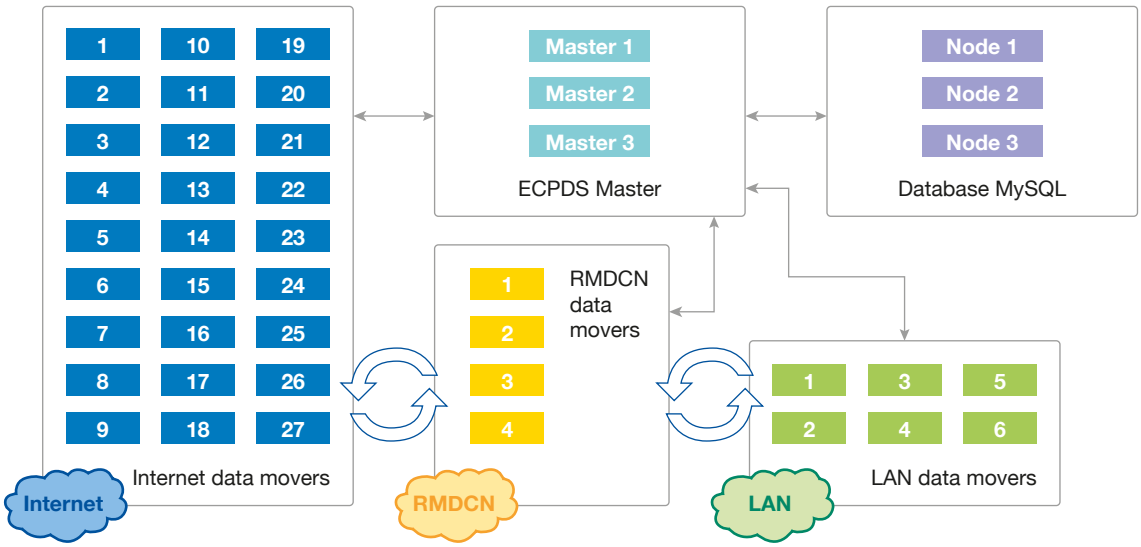


Figure 3 ECPDS physical architecture. The key components are the ECPDS Master, the MySQL database and the Data Movers.

The ECPDS Master and MySQL Database both run in a highly resilient environment. The Internet, RMDCN and LAN Data Movers currently run on physical servers deployed in their own secure environment. There are currently 37 Data Movers available for a total capacity of 1.2 petabytes.

In order to prevent downtime and data loss, which is a critical requirement of ECMWF's Member and Co-operating States and is needed for the real-time running of the forecasting system at ECMWF, the following mechanisms have been implemented:

- The ECPDS Master and MySQL Database are each replicated on three servers in an active/passive mode: one active instance handles requests and two passive instances are on standby. When the active instance fails or requires maintenance, one of the passive instances takes over and the service resumes as normal.
- Each Data Mover group includes multiple servers. In order to guarantee the availability of the files in a group, a replication process is run to copy the files across the Data Movers (each file is replicated three times). If relevant, replication is also performed between ECMWF Data Movers and Data Movers located in the cloud on other continents.

A load balancer is configured to distribute incoming requests to the ECPDS Data Portal amongst the Data Movers. The use of multiple Data Movers and load balancing increases availability through redundancy.

ECMWF upgrades its forecasting system on a regular basis, and such upgrades are usually accompanied by a large increase in the volume of data to distribute. The way to scale up ECPDS in this situation is by adding more Data Movers. Each Data Mover adds CPU and I/O capability and disk space resources. More Data Movers can handle a bigger workload and more data.

doi: 10.21957/83deq5lgc0

Future challenges

In order to support their delivery of forecast services to society, ECMWF's growing number of forecast users require quick and reliable access to ECMWF forecasts. As a result, the average data transmission volume handled by ECPDS is approaching 1 petabyte per month with an exponential increase in Internet traffic. ECMWF forecast products are disseminated to 547 places in 78 countries, and observational data are retrieved from 557 places in 34 countries. Figure 4 shows a map of locations from where data are retrieved in Europe. The huge rise in traffic observed in recent years has successfully demonstrated the reliability, availability and scalability of ECPDS. However, there are other challenges that will need to be addressed in the near future:

- ECMWF's new data centre in Bologna: there will be two independent data halls, a new network topology and a new high-performance computing facility. ECPDS will have to adapt simultaneously to a changing infrastructure and changing technologies on top of increasing traffic.
- ECPDS is turning to cloud computing to expand its potential: ECPDS is already running Data Movers in the cloud, but this is just a beginning and our engagement with the wider community as part of the European Weather Cloud and the HiDALGO and LEXIS EU-funded projects should help us to consolidate ECPDS's position in this field.

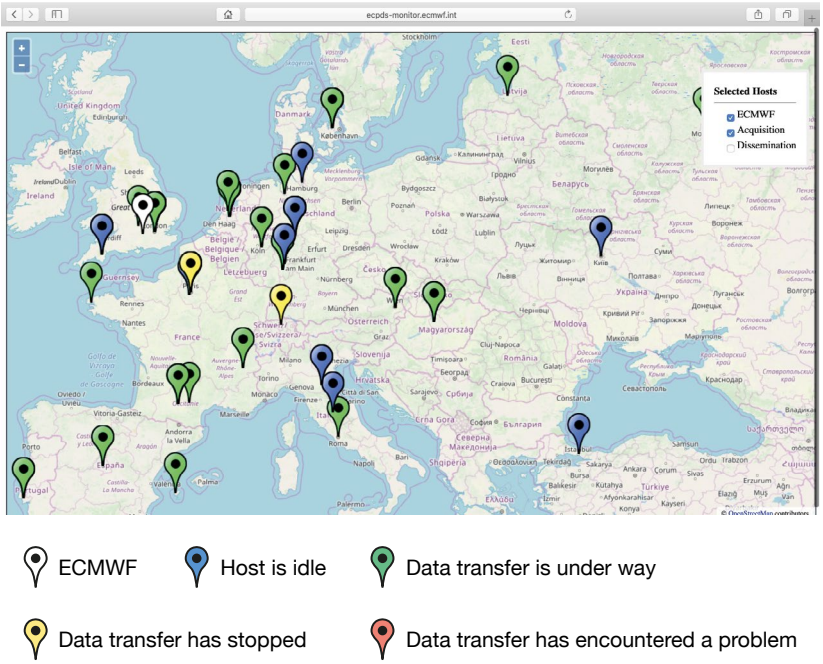


Figure 4 The interactive map in the ECPDS user interface shows the locations of hosts across the world, in this case over Europe for acquisition hosts only. A similar map is available for dissemination hosts. Different colours indicate the status of data transfer requests. A concentration of red hosts might indicate a network issue in the region. (Copyright OpenStreetMap *www.openstreetmap.org/copyright*)

Overall, ECPDS is now a mature solution which has helped to significantly improve efficiency and productivity of our data services by using proven and innovative technologies. With ECPDS, ECMWF delivers a portable and adaptable application which can fit diverse environments as well as providing a user-friendly tool to run data-related services. Stay tuned!

 doi: 10.21957/83deq5lgc0