Centre for Environmental Data Analysis (CEDA)

Centre for Environmental Data Analysis

SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

Annual Report 2020

(April 2019 to March 2020)

Victoria Bennett,
Poppy Townsend (Editors)

UKRI
Science and Technology Facilities Council
Natural Environment Research Council

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

CONTENTS

1. INTRODUCTION

The Centre for Environmental Data Analysis (CEDA) is based in the Science and Technology Facilities Council (STFC)'s RAL Space department. CEDA operates data centres and delivers data infrastructure, primarily for the Natural Environment Research Council (NERC), and undertakes project work for a range of national and international funders. CEDA's mission is to provide data and information services for environmental science: this includes curation of scientifically important environmental data for the long term, and facilitation of the use of data by the environmental science community.

CEDA was established in 2005, as a merged entity incorporating two NERC designated data centres: the British Atmospheric Data Centre, and the NERC Earth Observation Data Centre. The data centre function of CEDA (and its predecessor organisations) celebrated 25 years of existence in October 2019: the CEDA birthday cake is shown in Figure 1!

Since April 2018, CEDA has been a component part of the NERC Environmental Data Service, which brings together the five NERC data centres into a single service commissioned by NERC as National Capability.

Figure 1: CEDA Staff with 25th Birthday Cake, Oct 2019

JASMIN is the data intensive supercomputer which provides the infrastructure upon which the CEDA archives and services are delivered. Increasingly, JASMIN provides flexible data analysis capabilities to a growing community, who benefit from high performance compute and a private cloud, co-located with petascale data storage. The role of CEDA staff continues to evolve to include services and support for users of increasingly large and complex datasets. Last year saw further investment in, and development of, JASMIN, enabling us to continue to grow and improve the capabilities offered to users.

In addition, as in previous years, CEDA staff are involved in nearly all the major atmospheric science programmes underway in the UK, in many earth observation programmes, and in a wide range of informatics activities.

This annual report presents key statistics for the past year (2019- 2020) as well as a series of highlights reports showcasing a cross section of our activities. Topics covered include:

- A selection of funded projects we lead, or contribute to (funded by the European Commission and the European Space Agency)
- Examples of sharing our expertise with others (e.g. on data standards for climate data, and technical committees for data services)
- Behind the scenes development work on our infrastructure and services
- New and growing data holdings, and
- Engaging with our users

Key metrics are also provided. I hope it provides an interesting insight into another successful year for CEDA.

Victoria Bennett, Head of CEDA, RAL Space

PART 1: HIGHLIGHTS AND MAJOR ACTIVITIES

In this section we provide a selection of descriptions of key activities from the year. We have included highlights selected to showcase some CEDA activities supported through different funding streams, and a range of key areas of focus for CEDA staff this year.

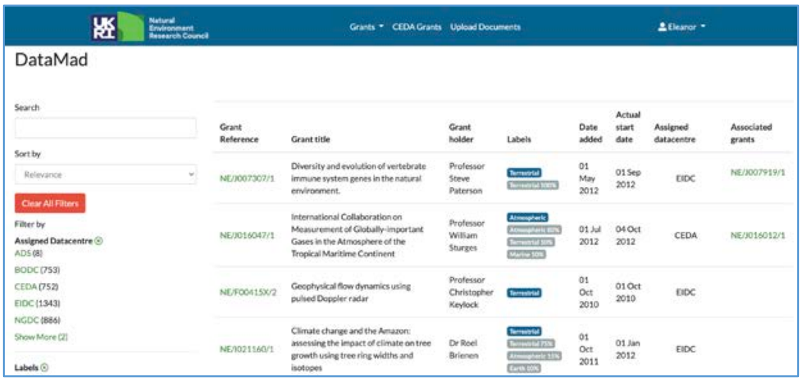
1. HIGHLIGHTS

1.1. WORKING WITH THE NERC ENVIRONMENTAL DATA SERVICE

Sam Pepler, Elle Smith, Poppy Townsend

The five NERC data centres were combined into the Environmental Data Service (EDS), in April 2018. There is now a strong emphasis on working together in an integrated manner to increase efficiency and share expertise. Below are two examples of key integration activities that CEDA have been involved in.

Sharing information in a common tool: DataMAD is a tool that is used by all data centres in the EDS that allows us to share information about NERC grants producing data so that we can coordinate data management. CEDA has developed the new version of this tool which will help with streamlining data management across the NERC data centres. This upgrade is needed as the current system is limited in the number of grants it can handle and it is difficult to use. The current tool also encourages diverging workflows between the data centres.



DataMad

Grant Reference	Grant title	Grant holder	Labels	Date added	Actual start date	Assigned datacentre	Associated grants
NE/J007307/1	Diversity and evolution of vertebrate immune system genes in the natural environment.	Professor Steve Paterson	Terrestrial Terrestrial 100%	01 May 2012	01 Sep 2012	EIDC	NE/J007919/1
NE/J016047/1	International Collaboration on Measurement of Globally-important Gases in the Atmosphere of the Tropical Maritime Continent	Professor William Sturges	Atmospheric Atmospheric 80% Terrestrial 10% Marine 10%	01 Jul 2012	04 Oct 2012	CEDA	NE/J016012/1
NE/F00415X/2	Geophysical flow dynamics using pulsed Doppler radar	Professor Christopher Keylock	Terrestrial	01 Oct 2010	01 Oct 2010	EIDC	
NE/I021160/1	Climate change and the Amazon: assessing the impact of climate on tree growth using tree ring widths and isotopes	Dr Roel Brienen	Terrestrial Terrestrial 75% Atmospheric 15% Earth 10%	01 Oct 2011	01 Jan 2012	EIDC	

Figure 2: Screenshot of new DataMAD user interface

The new DataMAD will contain all the used features from the old tool but will also allow data centres to transfer grants between data centres, as well as making it easier to find a specific grant and its related documents.

Ongoing discussions and feedback sessions with the NERC data centres means that all five data centres have agreed to use a common issue tracking system for their data management. This has allowed us to integrate this tracking system into the new DataMAD. As a result, workflows will become more alike and the transfer of grant information will be more efficient. The rollout of the new DataMAD is expected to be in Autumn 2020 (Fig. 2).

Archiving data from other NERC data centres: The CEDA Archive handles the largest datasets within the EDS. Satellites and climate models are renowned for their ability to consume storage. However, big data is not just a problem for the CEDA Archive; the other EDS centres have large datasets that simply do not fit into their infrastructures. To alleviate this problem CEDA is starting to act as the storage component for some of the data managed by the other data centres.

This year CEDA has adapted its data arrivals service to accommodate data from any of the other centres. It is currently in beta testing, with a handful of datasets stored at CEDA but catalogued in the National Geoscience Data Centre. In the coming year we will start ingesting data from British Oceanographic Data Centre as well.

Future work: There are several areas within the EDS that will be the focus for future integration activities. Some examples include: using common workflows for data management issues across EDS, agreeing standard guidelines for citing data, and how to store model code. Improvements to these areas could increase efficiency across the data centres.

The direction of future work may change depending on the outcome of the NERC Digital Solutions funding programme. This programme will work alongside the EDS to help facilitate collaboration with commercial partners and other government departments and is likely to accelerate integration activities.

1.2. TALKING ABOUT CLIMATE MODEL DATA

Martin Juckes, Ruth Petrie

JASMIN hosts close to 5 Petabytes of data from dozens of Earth System Models (ESMs) which provide a critical resource for climate scientists. These data arise from, in the main, climate model intercomparison projects which play a prominent role in the evaluation and development of climate models.
For the benefit of those managing and exploiting the data, it is essential that the data are archived with consistent metadata and data quality.
With data flowing from dozens of independent science teams, working within their own institutional and national constraints, achieving consistency and clarity in the documentation of the data products requires substantial communication at many levels. CEDA contributes to this and plays a leading role in several areas.

Committees and Conversations

The Coupled Model Intercomparison Project (CMIP) is the flagship project for the global climate modelling community. It is coordinated within the World Climate Research Programme (WCRP) through the Working Group on Coupled Models (WGCM). Two panels manage the implementation of CMIP: the CMIP Panel dealing with scientific priorisation and coordination and the WGCM Infrastructure Panel (WIP) overseeing the technical infrastructure. Several groups responsible for specific areas of work report to the WIP. Of these, CEDA contributes to CDNOT, ES-DOC and the Data Request (See Figure 3).

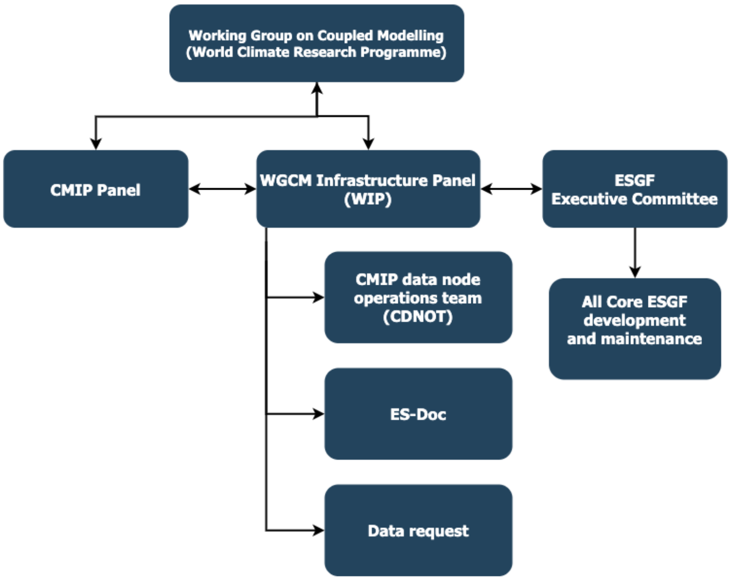

Working Group on Coupled Modelling (World Climate Research Programme)
CMIP Panel
WGCM Infrastructure Panel (WIP)
ESGF Executive Committee
CMIP data node operations team (CDNOT)
All Core ESGF development and maintenance
ES-Doc
Data request

Figure 3: Schematic of the WCRP's climate data related panels and activities

CEDA is also involved in:

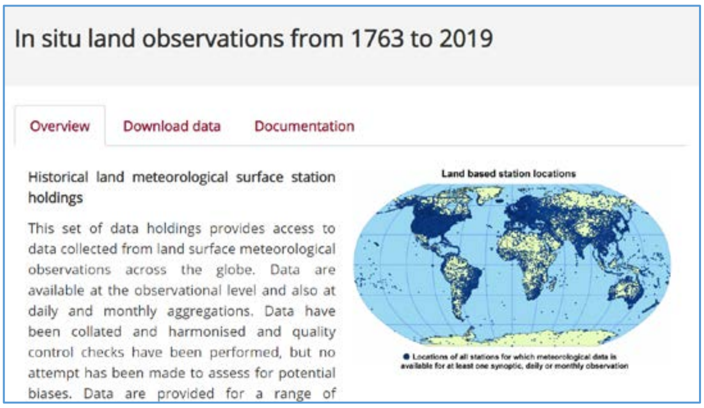
- The coordination of the exploitation of climate model data by the Intergovernmental Panel on Climate Change (IPCC) through the IPCC Task Group on Data (TG-DATA);
- Establishing data standards for climate data through the CF (Climate and Forecast) metadata conventions and membership of the CF Standard Names Committee;
- Coordinating service development and delivery in Europe through the ENES (European Network for Earth System Modelling) Data Task Force;
- Coordination and development of technology for distribution of CMIP data through the Earth System Grid Federation (ESGF) Executive Committee.

Participation in these international discussions is an important strategic activity for CEDA, which in the long term supports our aims to curate and disseminate climate data and generate societal impact.

1.3. SERVING TERABYTES OF GLOBAL OBSERVATIONS TO COPERNICUS USERS (GLAMOD)

Ag Stephens, William Tucker

The Global Land and Marine Observations Database (GLAMOD) is a large climate dataset containing over 26 billion individual observations of temperature, rainfall, pressure, etc. The GLAMOD database, hosted by CEDA, is accessible via the C3S Climate Data Store (CDS) - whose aim is to provide an interface to historical and future climate datasets (Fig. 4). The Copernicus Climate Change Service (C3S) aims to deliver authoritative information about the past, present and future climate, as well as tools to enable climate change mitigation and adaptation strategies by policymakers and businesses. This work is an important part of the CEDA contribution to the C3S. Through multiple contracts, we provide climate data to a broad range of users in both the academic and decision-making sectors. In addition, we are acquiring new knowledge, and developing systems, for handling weather observations at a scale not found in our community.



In situ land observations from 1763 to 2019

Historical land meteorological surface station holdings

This set of data holdings provides access to data collected from land surface meteorological observations across the globe. Data are available at the observational level and also at daily and monthly aggregations. Data have been collated and harmonised and quality control checks have been performed, but no attempt has been made to assess for potential biases. Data are provided for a range of

Land based station locations
● Locations of all stations for which meteorological data is available for at least one synoptic, daily or monthly observation

Figure 4: The GLAMOD land dataset entry in the CDS catalogue

How to serve 26,000,000,000 observations?
Unlike most data at CEDA, the GLAMOD project decided to use a PostgreSQL relational database in order to store and search its data holdings. This makes good sense because the dataset consists of small records, but it represents a "big-data" challenge because there are so many records. The technical solution has evolved significantly during the course of the project and we have learnt a great deal along the way. Here, we highlight some of the issues that have been overcome in order to create a working solution.

Bespoke hardware: Small (<100GB) databases work well on standard JASMIN hardware. However, loading multiple TBs of data proved to be a bottleneck and extraction times were also slow. It was necessary to buy optimised hardware because the database will exceed 10TB.

Parallelising the workflow: The initial workflow ran sequentially in a single process and did checks at the level of each observation before data was loaded. This was not sustainable. The new system breaks the workflow into several parts that can each be run in parallel: (1) Fix and restructure input files (of which there are ~235,000 for the land domain); (2) Write SQL loading scripts; (3) Load data using bulk loading SQL techniques.

Re-structuring the database and query interface: The full GLAMOD database structure consists of a comprehensive set of relational tables. Testing revealed that querying this structure was slow and complex. The project simplified the structure into a flat database table. In order to improve performance, the main table was partitioned into ~1000 separate tables split by domain (land/marine), frequency and year. The initial delivery service used a Web Feature Service served using GeoServer. Due to the API requirements of the CDS, it was necessary to replace GeoServer with a bespoke python web-application that translates the input parameters into a set of SQL "Select" statements that are sequentially submitted to the database. These are retrieved and are further processed and merged. End-users receive a zip file containing a pipe-separated text file of up to 1GB of observation records. This format can be read into desktop packages or standard software libraries.

The services developed in GLAMOD will provide cutting-edge delivery of a multi-TB database which will serve users in the research, public and private sectors.

1.4. FAIR DATA FOR THE IPCC 6TH ASSESSMENT REPORT

Charlotte Pascoe

At CEDA we aim to follow the FAIR (https://www.go-fair.org/fair-principles/) data principles, these are described below:

- Findable – data are easy to find for both humans and computers and have rich metadata describing the resource in an easily searchable platform.
- Accessible - data have known access mechanisms and support the export of structured metadata.
- Interoperable - data conform to recognised standards and formats (for data, metadata and software) and use defined vocabularies, common workflows and development protocols.
- Reusable - have well described data and metadata for provenance, traceability and reproducibility (e.g. input data, metadata, diagnostics, tool version) and implement relevant standards for file formats.

These principles are especially important for high value datasets, such as those tackling global issues like climate change. We are working closely with the Intergovernmental Panel on Climate Change (IPCC) Technical Support Unit (TSU) for Working Group 1 to make the data for the figures in the IPCC 6th Assessment Report (AR6) available as a FAIR digital resource. We are doing this by integrating the TSU's Figure Manager database with the CEDA data catalogue.

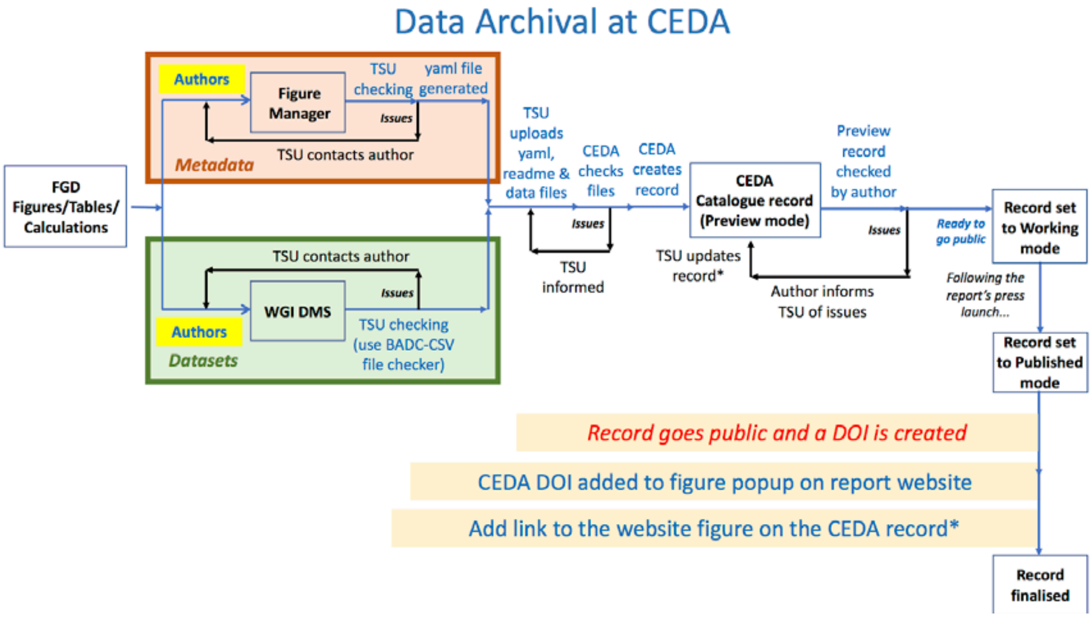

Data Archival at CEDA
Authors
Figure Manager
TSU checking
yaml file generated
Issues
TSU contacts author
Metadata
FGD Figures/Tables/ Calculations
TSU contacts author
Issues
WGI DMS
Authors
TSU checking (use BADC-CSV file checker)
Datasets
TSU uploads yaml, readme & data files
CEDA checks files
CEDA creates record
Issues
TSU informed
CEDA Catalogue record (Preview mode)
TSU updates record*
Preview record checked by author
Issues
Author informs TSU of issues
Ready to go public
Record set to Working mode
Following the report's press launch...
Record set to Published mode
Record goes public and a DOI is created
CEDA DOI added to figure popup on report website
Add link to the website figure on the CEDA record*
Record finalised

Figure 5: The integration of structured metadata from the CEDA catalogue (MOLES) into the IPCC WG1 Figure Manager

The integration of these two systems avoids duplication of effort as it allows for the same metadata to be used both to write figure captions in the assessment report and to draft catalogue records for the CEDA Archive, which help users find the data they need. The process is shown schematically in Figure 5. The protocols developed with the TSU for IPCC Working Group 1 will also be shared and applied to other IPCC working groups where applicable.

The CEDA catalogue records will "go live" when each IPCC Working Group report is published, the final integration step being to integrate the CEDA issued Digital Object Identifiers (DOIs) back into the Figure Manager, thus providing a direct link from the assessment report figures to the underlying data in the CEDA Archive.

This work will help ensure that data used in this and future IPCC assessments follows FAIR principles, adding to the transparency and credibility of the underpinning science.

1.5. CLIMATE CHANGE INITIATIVE: KNOWLEDGE EXCHANGE

Richard Smith, Alison Waterfall

The Climate Change Initiative (CCI) Knowledge Exchange project is part of the European Space Agency's CCI Programme, which is developing climate quality datasets of Essential Climate Variables from historic satellite data. The Knowledge Exchange project started in 2019, and involves multiple organisations to provide a website, educational content, data processing tools and search and download tools (See http://climate.esa.int). Within the project, CEDA has responsibility for the CCI Open Data Portal (ODP), which provides a free and open central point of access to the CCI data. This activity is a continuation of work in a prior CCI Open Data Portal project, where CEDA provided the infrastructure, data archival and data download services.

Within the CCI Knowledge Exchange project, we have taken the opportunity to make changes to improve the data portal service: the new portal uses the OpenSearch standard (with best practice from Committee on Earth Observation Satellites - CEOS) to provide an interface to the CCI datasets as well as providing faceted search. This simplifies the stream of information used by the CCI Toolbox data processing tools and the website search and discovery user interface.

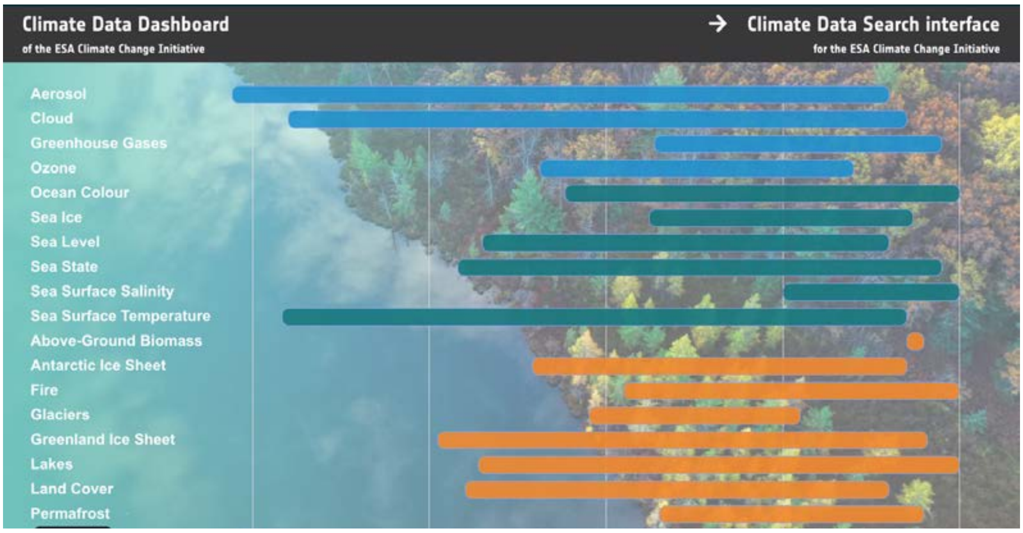


Figure 6: ESA Climate Change Initiative Data Dashboard

The previous data portal used a combination of Earth System Grid Federation (ESGF) search and CEDA's Catalogue Service for Web (CSW). The ESGF publication process meant that only netCDF data could be published to many of the portal services, and publication was a lengthy and complicated process. Moving to OpenSearch, we now use an Elasticsearch index which contains metadata about all the datasets and our own publication pipeline which can handle non-netCDF data. This means that all the available CCI data can be made available through all the portal search interfaces, and to downstream services such as the CCI Toolbox. Control of the whole data publication pipeline also allows us to streamline and automate, where appropriate, and speeds up the publishing process.

As a result of the move to OpenSearch, new types of data, such as vector data, or formats such as GeoTiff, are now displayed in the CCI portal (Fig. 6). This particularly improves the access to data from the Greenland Ice Sheets, Antarctic Ice Sheets, and Glaciers projects which were not publishable through ESGF search.

In summary, CEDA's work within the Knowledge Exchange project has redesigned the CCI publication process for the Open Data Portal from the ground up, which will reduce time-to-publication as well as enabling more datasets to be made more widely available. This will be of significant benefit during the remainder of the project, as we continue to archive the many new datasets planned by all the CCI project teams, but also to inform future developments in CEDA's core services.

1.6. SPEEDING UP THE TIME TO SCIENCE: ANALYSIS-READY SATELLITE DATA

Ed Williamson, Philip Kershaw, Victoria Bennett

CEDA are supporting Defra and the Joint Nature Conservation Committee (JNCC) by providing access to processing and archival facilities to enable the creation of the new Sentinel Analysis Ready Data (ARD) for the UK. These derived data products, from Sentinel 1 and 2 imagery, have been produced to support land use applications, such as habitat mapping. Sentinel data provided by the Copernicus Programme works well for land use applications by virtue of the frequent revisit time, high spatial resolution and open access arrangements for the data.

The concept of ARD has developed around a recognition of the need to provide standardised pre-processed data products in a common form ready for analysis thus reducing duplicated effort and the potential for inconsistencies in data preparation. It is estimated that access to ARD products could save up to 70% of project time. The use of JASMIN has provided the computing resources to facilitate large-scale ARD production and lowers the barrier for user access by providing a large centralised store to host and disseminate the data from the CEDA Archive. The data currently covers England, Scotland and Northern Ireland.

ARD data applications

Generation of these products supports use of satellite data for UK public sector environmental applications. However, there are many other applications and the data are freely available to access from the CEDA Archive under the open government licence. Some example applications of where this data has been used are habitat mapping, forest change detection and crop mapping.

Ingesting Sentinel ARD to the CEDA Archive

The data are provided to CEDA via two routes: data for England are processed by the Defra Earth Observation Data Service (using public cloud), while data for Scotland and Northern Ireland are processed by JNCC on JASMIN. These Sentinel ARD products go back in time to September 2018. Over 30 TB of ARD is now in the CEDA archive, increasing by roughly 2.5TB per month – an example is shown in Figure 7.

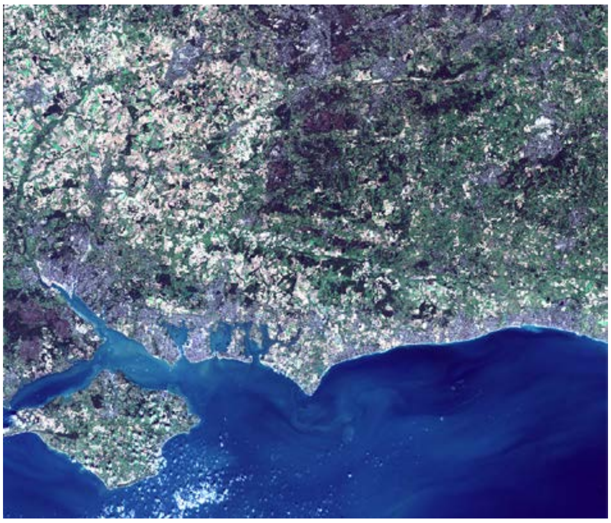


Figure 7: Sentinel 2B ARD processed image taken over Portsmouth. Contains modified Copernicus Sentinel data, processed by Defra Earth Observation Data Service

Future plans

CEDA will add the ARD products into the Satellite Data Finder (http://geo-search.ceda.ac.uk/) so that users can search for products over areas of interest. This will allow the Sentinel ARD data to be geographically searched alongside other satellite datasets from the CEDA archive. JNCC are planning some webinars on the ARD data to make users aware of the potential uses of the data and how to access the data via CEDA. They also plan to release a script repository including, for example, Python code to work with the ARD products held by CEDA.

The work with Defra and JNCC is a great example of a partnership benefiting both UK public sector users and the NCEO science community through a shared effort providing data in a common format to a broader user base.

1.7. BEING OPEN ABOUT THE BRITISH WEATHER

Graham Parton, Ag Stephens

The British are known the world over for one thing more than any other... no, not queuing, the Queen, or even our penchant for a nice cup of tea... but talking about the weather. However, getting hold of UK weather data, and even more so in an easy to use format, has sadly been historically quite hard to do. 2019 saw a significant change to this, with CEDA supporting the UK Met Office to make their long-term store of weather data openly available to all.

Weather data in the UK have been recorded for well over 150 years (Figure 8), with the Met Office being the primary custodian of both historic and recent data from hundreds of different stations: 'MIDAS' is the Met Office's historic weather observations database. Although a national asset, gaining access to these stores of data hasn't been straightforward. Barriers around licensing, data ownership, access control and available resources to support wider access have meant that providing access to these data was only possible for academics through the CEDA Archive.

Figure 8: Met Office weather station site in the UK (Image courtesy of the Met Office)

Thanks to the close collaboration between CEDA and the Met Office over the years and CEDA's experience in managing long-term archiving, an 'open' version of the Met Office's data within MIDAS were made available via the CEDA Archive under the UK's Open Government Licence (OGL) in 2019. Significantly, though, this wasn't just about making these data available under a licence that allowed all types of use - commercial to government use; personal to academic use, etc - but providing data that has been actively curated to ensure ease of access, use and understanding.

By adopting a well structured archive and an easy to use file format, users can readily find the data they want and download to use in a wide range of packages in line with a broad range of user skills - from those using spreadsheets through to GIS users and programmers.

Yearly updates to this already valued dataset collection continue to build on this work. These updates incorporate and document the significant improvements the Met Office makes in their data store, such as quality control changes or inclusion of recently-digitised historical data. This work ensures that these data remain relevant and useful for many applications.

Finally, not only can we all enjoy talking about the British weather every day, but now everyone can easily go back and revisit our weather from yesteryear; whether for commercial use, teaching or just general nostalgia - these data are now open and available for use.

1.8. NEW EARTH OBSERVATION DATASETS ARCHIVED AT CEDA

Ed Williamson, Steve Donegan

CEDA continues to support the National Centre for Earth Observation (NCEO) by adding new Earth Observation data to the CEDA Archive: these data are used in a wide range of research projects where they are processed into geophysical products and analysed to better understand the Earth system. The EO data our science community use originate from several international space agencies: CEDA transfers the NCEO's priority datasets into the CEDA archive where they can be accessed via the JASMIN computing infrastructure. This prevents duplication of effort, so that the scientists don't need to individually download and handle what can be up to several petabytes of data.

This year, one of the key activities has been expanding the selection of Sentinel satellite data products that we archive - we now hold almost 8 petabytes of Sentinel data. For these large datasets, CEDA makes use of its Near Line Archive (NLA) system. This system moves older data onto tape to ensure there is space on the disk archive for latest data products. It also allows users to temporarily request data back to disk for processing.

Many new EO data products have been added to the CEDA Archive in the last year and are available to users. These include the following:

- Sentinel 1A and B: OCN products (Interferometric Wide Swath (IW), Wave (WV)) radar data
- Sentinel 3A and B: SRAL altimeter data
- Sentinel 5P: L2 CO, L2 NO2, L2 NP, L2 SO2, L2 CH4 atmospheric composition data
- Sentinel Analysis Ready Data (ARD) from Sentinel 1 & 2 (See Highlight 1.6 in this report, and Figure 9)
- TERRA & AQUA: Various MODIS products from the LAADS DAAC, NASA's data distribution service
- SPEI Africa: High resolution Standardized Precipitation Evapotranspiration Index (SPEI) dataset for Africa
- TLS: Terrestrial Laser Scanner data from NERC funded "Weighing trees with lasers" project
- EUMETSAT: we acquire a range of EUMETSAT data for NCEO research groups
- Data generated in EC H2020 funded projects EUSTACE, FIDUCEO and BACI

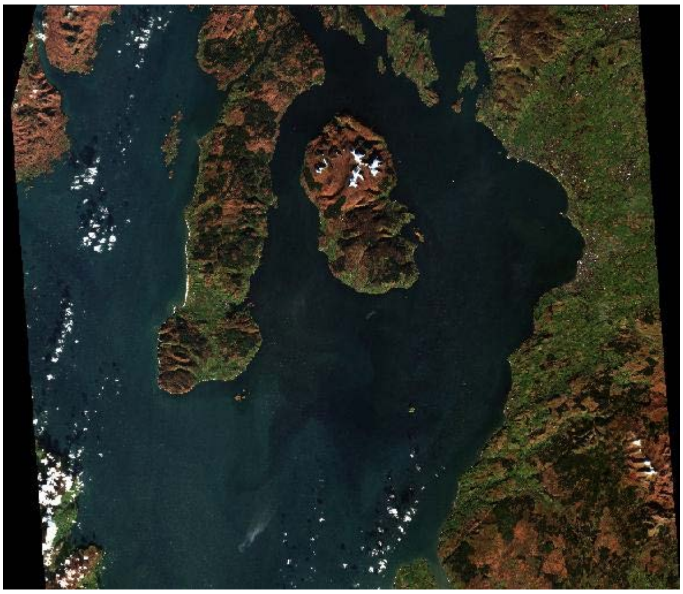


Figure 9: Sentinel 2 ARD Image of Isle of Arran, Scotland, processed on JASMIN by JNCC Simple ARD Service

CEDA have also implemented a new system for the retrieval of Sentinel 3 SLSTR Near Real Time (NRT) products. This allows appropriate products to be retrieved with short turnaround time to a JASMIN shared workspace which is accessible to interested users for a short period. This system benefits users as they do not have to wait for the full retrieval and archive deposit system - it provides near real time data more quickly for users. It has already been used extensively with positive feedback. This system could be rolled out to other datasets if required.

CEDA holds a large number of datasets to enable NCEO science, and ensure the long term availability of data products resulting from NCEO research. These can be found at https://catalogue.ceda.ac.uk/ where users can search for and download data.

1.9. LICENCE CLASSIFICATION – WHAT'S THE USE?

Graham Parton

You're just about to start on a new, exciting project and want to find some useful data. You spend a few hours looking around online and eventually start drilling down to select some tasty looking datasets, raring to go, 'Great!', you think… Then it hits you… the licence doesn't allow commercial use!

Figure 10: Licence details can be complex and frustrating

Sadly, this is not an unfamiliar story and chimes with other licencing detail woes. Organisations want an easy way to see the 'impact' their datasets have; archives need an easy way to know what applications are in line with licencing for access control; and service developers want a convenient way to check that they have permission to bring their software and data resources together - all without spending hours checking licence text and cross referencing themselves. If only there was a better way…

In 2019, CEDA sought this better way by examining the 80+ licences used in our archives. Out of this came a 'permitted-use' classification scheme. The idea was simple: read the licence and determine which types of use are permitted: personal, commercial, policy, academic, teaching, etc. The scheme also acknowledges that licences aren't always clear cut - specific clauses may need flagging, as do those which are too ambiguous; all whilst acknowledging that the licence details themselves remain intact and users will, ultimately, still need to read them.

CEDA are now in the process of adding in this classification scheme to our access control and licencing system which will allow a new, top-level search facet for the data catalogue. Through this CEDA will enable any user to quickly find data they can make use of… or list all datasets that a funder has aided which are available for commercial use… and give CEDA staff a quick and easy look-up to check dataset applications against.

However, whilst we think this is a neat idea, we also recognise that it quickly reaches limits in its applicability as such schemes aren't used elsewhere. During 2019-20, though, we've taken opportunities to talk to others, engaging the wider data management community where there is a ready interest in applying this approach more widely. Let's see what the next year brings.

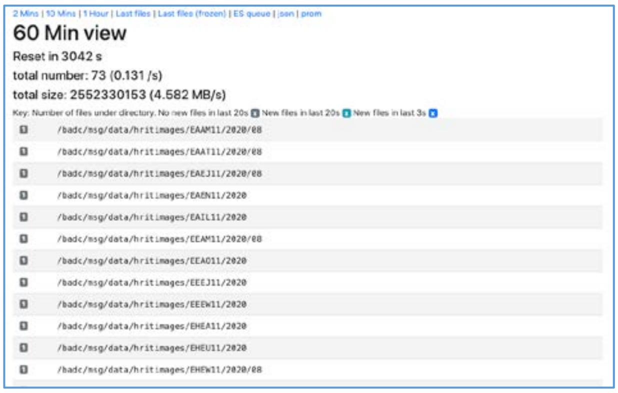
1.10. MONITORING DATA INGESTION

Sam Pepler

Significant work was done last year to improve the internal deposit system that allows data to be ingested into the CEDA Archive. This made the system more distributed and robust. However, as the efficiency of the system has increased it has been harder to see how it is operating as a whole. To combat this we have developed and added to our range of tools for monitoring data ingestions so that we can quickly diagnose problems, and verify that things are working as they should.

There are two key stores of information we tap into for these tools.

Firstly, we have engineered the new deposit system to broadcast messages for each file delivered to the archive (Figure 11). We have created a tool to track the most recently ingested files so that data scientists can watch files as they are deposited, giving them confidence that their ingest processes are working as expected. This also allows us to see the overall deposit rate across the whole system and shows which other data streams are competing for attention.



2 Mins | 10 Mins | 1 Hour | Last files | Last files (frozen) | ES queue | json | prom

60 Min view

Reset in 3042 s

total number: 73 (0.131 /s)

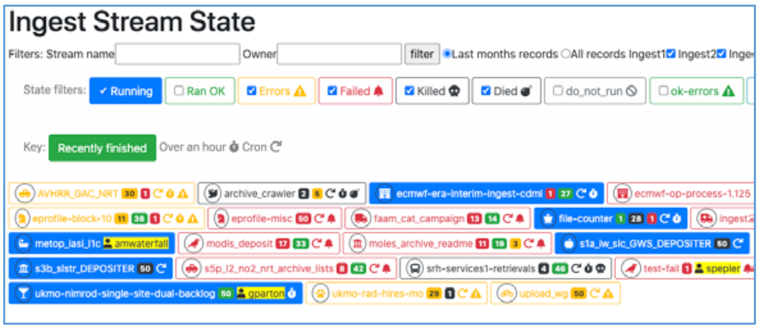
total size: 2552330153 (4.582 MB/s)

Key: Number of files under directory. No new files in last 20s 0 New files in last 20s 0 New files in last 3s 0

```
1    /badc/msg/data/hritimages/EAAM11/2020/08
1    /badc/msg/data/hritimages/EAAT11/2020/08
1    /badc/msg/data/hritimages/EAEJ11/2020/08
1    /badc/msg/data/hritimages/EAEN11/2020
1    /badc/msg/data/hritimages/EAIL11/2020
1    /badc/msg/data/hritimages/EEAM11/2020/08
1    /badc/msg/data/hritimages/EEAO11/2020
1    /badc/msg/data/hritimages/EEEJ11/2020
1    /badc/msg/data/hritimages/EEEW11/2020
1    /badc/msg/data/hritimages/EHEA11/2020
1    /badc/msg/data/hritimages/EHEU11/2020
1    /badc/msg/data/hritimages/EHEW11/2020/08
```

Figure 11: View of recently ingested files

The second information store holds the records of every ingest job performed, its output and whether it failed or succeeded. These records help us debug broken processes, and visualise which data streams are currently blocked (Figure 12).



Ingest Stream State

Filters: Stream name [ ] Owner [ ] filter ● Last months records ○ All records Ingest1 ☑ Ingest2 ☑ Inge

State filters: ✓ Running | □ Ran OK | ☑ Errors ⚠ | ☑ Failed 🔔 | ☑ Killed | ☑ Died | □ do_not_run ⊘ | □ ok-errors ⚠

Key: Recently finished | Over an hour | Cron ↻

AVHRR_GAC_NRT 30 1 ↻ ⚠ | archive_crawler 2 5 ↻ | ecmwf-era-interim-ingest-cdml 1 27 ↻ | ecmwf-op-process-1.125

eprofile-block-10 11 38 1 ↻ ⚠ | eprofile-misc 50 ↻ 🔔 | faam_cat_campaign 13 14 ↻ 🔔 | file-counter 1 28 1 ↻ | ingest

metop_iasi_l1c amwaterfall | modis_deposit 17 33 ↻ 🔔 | moles_archive_readme 11 10 3 ↻ 🔔 | s1a_iw_slc_GWS_DEPOSITER 50 ↻

s3b_slstr_DEPOSITER 50 ↻ | s5p_l2_no2_nrt_archive_lists 8 42 ↻ 🔔 | srh-services1-retrievals 4 46 ↻ | test-fail 1 spepler 🔔

ukmo-nimrod-single-site-dual-backlog 50 gparton | ukmo-rad-hires-mo 29 1 ↻ ⚠ | upload_wg 50 ↻ ⚠

Figure 12: View of status (success/fail) of ingest streams

The ingest system connects to other systems that allow us to deliver the data to users and upload from data providers.

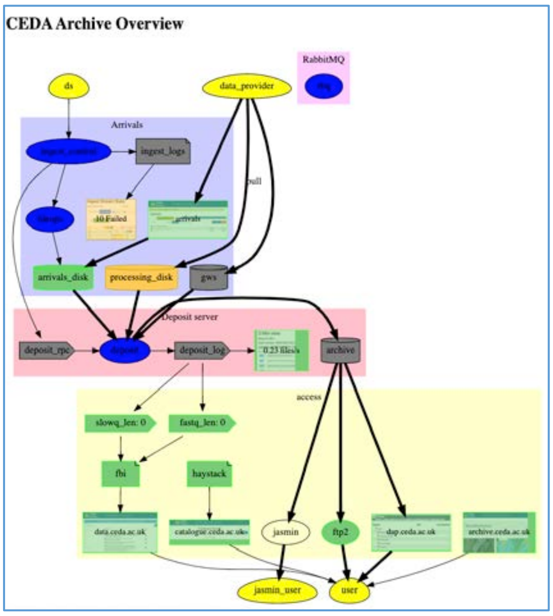


Figure 13: CEDA Archive interactive systems overview

We have created an interactive view of these interconnected systems so that we can see what is working in a more holistic way (Figure 13).

As the volume, and variety of data increases we need to develop more complex and finer tuned services to keep pace.

But, this also means we need to keep pace creating monitoring tools that let us understand what is going on, otherwise we will not know why our fancy, new, turbo-charged data ingester refuses to start!

1.11. CEDA SERVICES INVENTORY, HELPING US KEEP TRACK

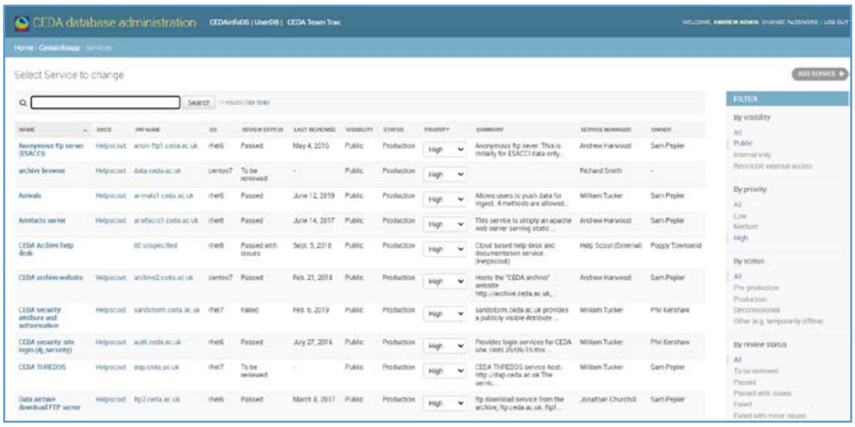
Andrew Harwood

As CEDA has grown, the management of the services that we operate has become increasingly complicated. No one can remember how to start, stop or fix every service we offer, there are just too many to deal with (over 130 at the last count)!

In order to help us with this we have developed an inventory of our services. Amongst other things this allows us to:

- See an overview of the services we run
- See who is responsible for maintaining each service and who ‘owns’ the service
- Provide detailed documentation to help in the maintenance of the service.

The inventory (Figure 14) is implemented using a Postgres database with Django providing a simple web interface. Additional detailed documentation for each service is recorded in our internal documentation site. CEDA team members record basic information about the services they are responsible for in the database and provide additional information in the linked internal documentation.



NAME	DOCS	VM NAME	OS	REVIEW STATUS	LAST REVIEWED	VISIBILITY	STATUS	PRIORITY	SUMMARY	SERVICE MANAGER	OWNER
Anonymous ftp server (ESACCI)	Helpscout	anon-ftp1.ceda.ac.uk	rhel6	Passed	May 4, 2016	Public	Production	High	Anonymous ftp sever. This is initially for ESACCI data only...	Andrew Harwood	Sam Pepler
archive browser	Helpscout	data.ceda.ac.uk	centos7	To be reviewed	-	Public	Production	High		Richard Smith	-
Arrivals	Helpscout	arrivals1.ceda.ac.uk	rhel6	Passed	June 12, 2019	Public	Production	High	Allows users to push data for ingest. 4 methods are allowed...	William Tucker	Sam Pepler
Artefacts server	Helpscout	artefacts1.ceda.ac.uk	rhel6	Passed	June 14, 2017	Public	Production	High	This service is simply an apache web server serving static ...	Andrew Harwood	Sam Pepler
CEDA Archive help desk		OS unspecified	rhel6	Passed with issues	Sept. 5, 2018	Public	Production	High	Cloud based help desk and documentation service (Helpscout)	Help Scout (External)	Poppy Townsend
CEDA archive website	Helpscout	archive2.ceda.ac.uk	centos7	Passed	Feb. 21, 2018	Public	Production	High	Hosts the "CEDA archive" website http://archive.ceda.ac.uk,...	Andrew Harwood	Sam Pepler
CEDA security: attribute and authorisation	Helpscout	sandstorm.ceda.ac.uk	rhel7	Failed	Feb. 6, 2019	Public	Production	High	sandstorm.ceda.ac.uk provides a publicly visible Attribute ...	William Tucker	Phil Kershaw
CEDA security: site login (dj_security)	Helpscout	auth.ceda.ac.uk	rhel6	Passed	July 27, 2016	Public	Production	High	Provides login services for CEDA site. Until 20/09/15 this ...	William Tucker	Phil Kershaw
CEDA THREDDS	Helpscout	dap.ceda.ac.uk	rhel7	To be reviewed	-	Public	Production	High	CEDA THREDDS service host: http://dap.ceda.ac.uk The servic...	William Tucker	Sam Pepler
Data archive download FTP server	Helpscout	ftp2.ceda.ac.uk	rhel6	Passed	March 8, 2017	Public	Production	High	ftp download service from the archive, ftp.ceda.ac.uk. ftp1...	Jonathan Churchill	Sam Pepler

Figure 14: Screenshot of CEDA Services inventory

Reviews of the inventory are performed to ensure that the information is up to date and to maintain awareness of the services we run. In addition, services which are no longer required are identified and are shut down or passed to others to maintain.

With the scale of services CEDA offers, it's important to organise and tidy not just the data, but the services we operate too.

1.12. REPEATABLE SERVICES USING ANSIBLE

Richard Smith, Matt Pryor, William Tucker

CEDA currently manages over 100 different services, which presents a major challenge around maintenance and timeliness of updates. In recent years, we have moved to using Ansible Playbooks (Fig 15) as a way to describe the desired state of a service when deployed. This allows for automated and repeatable deployments. Once the use of playbooks became more widespread in the team, it was clear that many of the services were structured in a similar way. This led to the development of a set of common patterns which can be used to deploy our most common architectures.

Figure 15: Ansible Playbooks are used to manage CEDA's software deployments

Ansible uses roles to separate out different components in a deployment. The CEDA development group have built a suite of roles called the JASMIN Ansible Roles which describe common deployments, e.g. installing a Python app or running a Django web service with an Nginx web server as a proxy.

As many of our services are based on Django, deploying a Django application is now as simple as setting a few configuration options and running a script. This cuts down on the time from development to deployment as the process is repeatedly tested and updated where needed. This means that services are deployed the same way which makes locating log files and configuration files easy as it is the same across services.

The JASMIN Ansible Roles also allows us to make changes across the estate. If an issue is identified, we can make a change in the role and then update all the affected services. A centralised repository for these roles means that all the developers can suggest improvements and allows for enhanced knowledge sharing between team members. Having a common base and standardising configurations also improves security as when we become aware of security issues, we can make a change in one place which can then be applied to all the relevant services.

In summary, using common patterns has improved the quality of our deployments and standardised the structure for similar services. This reduces the time to deploy and update services, improves reaction time to security concerns and makes bug finding easier.

1.13. MIGRATION TO CENTOS7 ON JASMIN

Fatima Chami and Matt Pritchard

The operating system is a vital part of any IT infrastructure. Until this year, the operating system used across the many hundreds of servers in JASMIN has been RedHat Enterprise Linux (RHEL6). When STFC became part of UKRI, its eligibility for favourable site licensing arrangements for RHEL was lost, motivating a change to a free, open-source offering. In addition, the "version 6" generation of RedHat/CentOS operating systems was approaching end-of-life (November 2020). The decision was taken to migrate to CentOS7, so it's been a busy year updating the operating system throughout all of JASMIN's services - all whilst minimizing disruption to users.

What needed migrating?

- Servers used to provide user access to JASMIN (login, xfer)
- Servers enabling users to work interactively (sci)
- Many hundreds of nodes within the LOTUS batch processing cluster
- Several hundred servers used to host web-based services (for JASMIN itself, for the CEDA Archive and project-specific services)
- Software stack: needed to be updated to include versions of applications compatible with the new operating system
- Batch scheduler and related updates to LOTUS: replacing the LSF scheduler with SLURM, adding new node types and introducing a new MPI implementation (OpenMPI replacing LSF platform MPI)

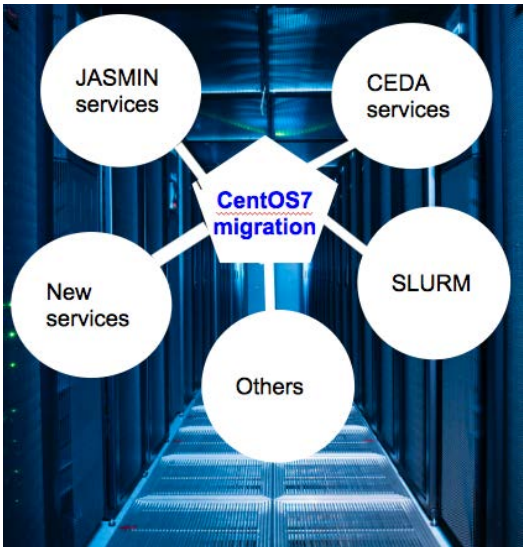


Figure 16: Overview of JASMIN services migrated to CentOS7

The migration plan

An audit of all the services within the infrastructure (Fig.16) was undertaken to decide which resources would need to be updated or decommissioned. A series of communications via email, news items and social media helped get the message out, and engaged users with what we were doing, gathering feedback about using the new systems and software, and making sure everyone was aware of the planned changes and timescales.

JASMIN user documentation was updated to refer to the new resources, alongside new training materials developed to help users make use of the new software stack and batch scheduler in particular.

Service redeployment

Most of CEDA and JASMIN's services are hosted on virtual machines, so the task of re-deploying services onto new "virtual" hardware is easier than installing new physical machines, although that's only part of the story. Each service was assessed as to whether it was still needed, then a new virtual machine created for it to be migrated to, so that the old one could simply be deleted once the migration was completed (although having twice the number of virtual machines in existence for some of the time did place a strain on resources!). However, many services are web applications built on a stack of framework and application software which all needed to be redeployed with versions compatible with the new operating system. The CEDA Development Group's use of "playbooks" to script the (re)deployment of services helped to manage the re-building of virtual machines and service dependencies in a robust and repeatable manner.

Migration to the new analysis compute environment

A phased transition was put in place, with two systems initially running in parallel (RHEL6 & CentOS7), across both interactive and batch compute systems. Although this increased the workload for support teams, it was necessary to introduce users to the new environment and allow for testing and tuning of the new system before withdrawal of the old. Gradually, nodes were converted to the new operating system and moved in batches from one scheduler to the other, alongside new nodes which had been waiting in the wings as part of the Phase 6 procurement.

The majority of CEDA and JASMIN user-facing services have now been migrated to the new CentOS7 operating system with disruption to users minimized as much as possible.

1.14. HELPING USERS GET THE MOST OUT OF JASMIN

Fatima Chami, Matt Pritchard

JASMIN is a large and multi-faceted research facility which continues to grow in response to demand from growing numbers of users from broad user communities (see Figure 17). Operated jointly by CEDA and STFC's Scientific Computing Department, it draws on a range of expertise within those groups to keep the infrastructure running and to help users make the most effective use of it. This is increasingly important with growth not only in the demand for resources but diversity of the challenges that scientists and researchers are trying to address.

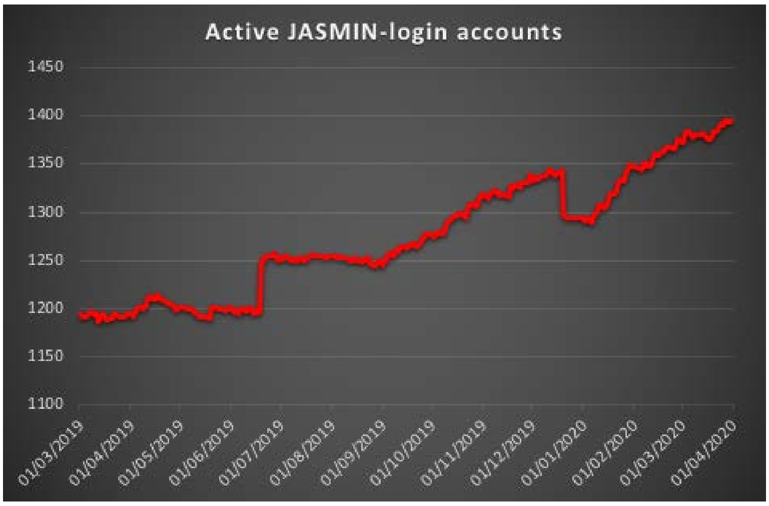


Figure 17: Increase in number of JASMIN login-accounts approved from March 2019 to March 2020

Multiple types of **storage** are available to users, as listed below. They are suitable for different purposes and require different levels of oversight and support from the JASMIN teams.

Home area: All users are provided with a home directory of 100 GB on Solid-State Drive (SSD) storage. This is ideal for storing small configuration files, code and compilation tasks, Python environments and other items. This is the only area on JASMIN that is systematically backed up by the JASMIN service.

Group Workspaces: These are collaborative workspaces provided to particular groups or scientific projects, normally with allocations of disk and tape storage, so that data can be moved between types to make best use of available capacity. Requests for space go to a Consortium Manager who holds an overall allocation for a particular science domain/community, and is responsible for tensioning requests for the finite storage between projects. Each Group Workspace has a manager who is responsible for organising the usage among their own users. We produce fortnightly volume reports summarising the usage and contents of each GWS for the GWS managers.

Scratch: Users also have access to a large, shared, scratch area, for use as intermediate storage during processing. In response to growing user numbers, and in order to maintain sufficient space, a new system to "police" the area has been set up, with automatic jobs to clear out data older than 28 days, and users are asked to make sure that they clean up after their jobs have finished.

JASMIN provides both interactive and batch **compute** for use in analysis work. Typically, users develop and test code and workflows interactively before moving processing workflows to the batch cluster, LOTUS.

LOTUS gives users access to 14,000 cores for efficient execution of large scale computing workloads. Compute resources in the LOTUS cluster are managed by the scheduler with a "fair share" policy, taking into account parameters such as job duration, memory usage and number of cores. To help users better estimate their requirements, a test queue can be used for small-scale test runs prior to a production run on the cluster.

The **scientific analysis servers** provide the interactive compute resource. These servers rely on users being sensible with usage patterns to maintain a stable, performant environment for all users to familiarise with the JASMIN system and do ad-hoc processing. However, the sharing of these resources has become problematic with excessive usage of CPU and memory resulting in slow performance, perhaps with some workflows which should be moved to the LOTUS cluster continuing to be run on interactive nodes.

Manual monitoring and communication with users to manage usage of these interactive resources is time consuming, and not scalable to the growing number of users we support. This “shared-by-all” model is soon to be replaced by “tenancy sci” machines: scientific analysis servers deployed within community-based JASMIN cloud tenancies, grouping users with similar workflows and enabling communities to manage their own computing resource. This model has been trialled successfully on part of the JASMIN cloud platform but with increased cloud capacity now available, the plan is to roll this out more widely.

It is an ongoing task for the JASMIN team to help users to understand what resources are available and how to use them efficiently. We do this by providing documentation, training workshops, webinars and user community events, and support via the JASMIN Helpdesk.

1.15. ENABLING USER WORKFLOWS ACROSS HETEROGENEOUS STORAGE

Neil Massey

Over the past three years, JASMIN has moved away from a traditional storage architecture, due to issues with cost and scalability, towards a more heterogeneous storage environment (Figure 18). Rather than having a large capacity of parallel file system hard disks, with tape as a backup medium, JASMIN now has scale-out filesystem, solid-state devices, object storage, as well as parallel file system. Additionally, tape is now used in a much more dynamic fashion, rather than purely as a backup medium. The new storage systems are also more suitable for interacting with cloud services.

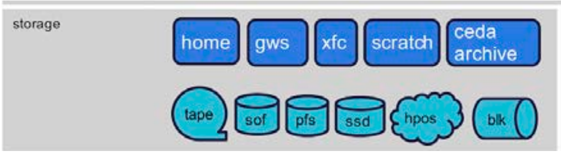


Figure 18: JASMIN storage services (dark blue) and storage types (light blue)

Each of these different storage systems has its own properties and best methods of working with it. Crucially, object storage and tape have very different user interfaces to parallel file system, and very different lag times. The time it takes to request data from an object store to receiving the data can be from a few milliseconds, if the program reading or writing can stream the data to memory, to minutes, if the program has to download or upload the entire file first. For a tape this time span could be many hours, regardless of the access method. With a parallel file system, this interaction is near instantaneous from a user perspective, and all programs support direct reading and writing from / to the disk. These differences mean that users will either need to adapt their workflows to the various storage types - or that CEDA provides the relevant services and support to make this easier for users.

To mitigate the differences, three projects have been completed, or are currently in development at CEDA:

- The **Near-line archive (NLA)** system moves some of the data held in the CEDA Archive so that the only copy is on tape. A user can then request data to be retrieved to a disk, and the NLA system makes a link to the original location in the archive.
  This system has been operational for a number of years and has become a well known and well used tool by JASMIN users, especially for Sentinel data.

- **Joint Data Migration Application (JDMA)** allows users to migrate and retrieve data from their Group Workspace to either Elastic Tape or Object Storage, using the same user interface for both. User space on their Group Workspace is limited, but they are allocated the same amount of space on Elastic Tape. Therefore, Elastic Tape can be used as a backup of Group Workspace, or more dynamically by fetching data that is going to be analysed next in their workflow.
  JDMA is extensible, so that it can be made to work with other storage systems by writing a plug-in. It also manages the upload and download of data to the storage system on the user’s behalf, and provides information about the user’s migrations and retrievals in a well catalogued and user friendly way.
  JDMA became operational in November 2019, and has been used to transfer over 700TB of user data to

tape, mostly from the decommissioned Research Data Facility on ARCHER. Feedback from users has been positive, especially with regard to the user interface.

- **S3netCDF** is a Python library developed by CEDA which enables the reading and writing of netCDF files to Object Storage, using the same Python interface as the standard netCDF4 library. This enables users to have minimal changes to their programs and workflows when working with netCDF data stored on Object Storage. S3netCDF allows the reading and writing of very large datasets, even on machines with limited memory, by subdividing large netCDF files into smaller netCDF files, which are self-describing. This removes the problem other subdividing file formats have, where losing the control file results in the rest of the data becoming unreadable.
  In 2019-2020, s3netCDF has had a complete rewrite, which has improved its performance greatly by using the asyncIO functionality in Python 3.7 and re-engineering the metadata format for the sub-divided files. An intelligent file and memory management sub-system has also been added. Many other improvements have been made and a release to the wider community is imminent.

As JASMIN continues to grow (Figure 19), and its storage architecture becomes more heterogeneous, the team will need to continue developing new ways to provide users with the best experience possible.

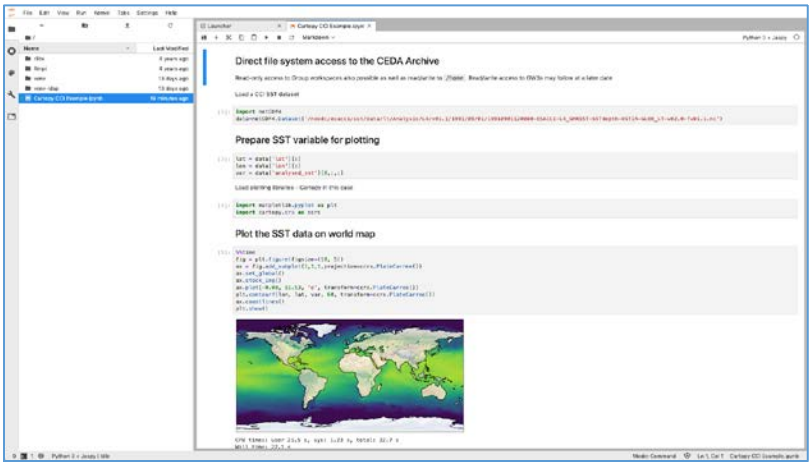
Figure 19: JASMIN storage and compute servers in machine room

1.16. JASMIN NOTEBOOK SERVICE

Matt Pryor

What is a Jupyter Notebook?

A Jupyter Notebook is an interactive document containing live code and visualisations that can be viewed and modified in a web browser. These documents can be shared, often using GitHub, and many projects distribute example code as Jupyter notebooks. Users interact with their notebooks using the open-source Jupyter Notebook server application.



Direct file system access to the CEDA Archive

Read-only access to Group workspaces also possible as well as read/write to /home. Read/write access to GWSs may follow at a later date

Load a CCI SST dataset

```
import netCDF4
data=netCDF4.Dataset('[illegible]')
```

Prepare SST variable for plotting

```
lat = data['lat'][:]
lon = data['lon'][:]
var = data['analysed_sst'][0,:,:]
```

Load plotting libraries - Cartopy in this case

```
import matplotlib.pyplot as plt
import cartopy.crs as ccrs
```

Plot the SST data on world map

```
%%time
fig = plt.figure(figsize=(10, 5))
ax = fig.add_subplot(1,1,1,projection=ccrs.PlateCarree())
ax.set_global()
ax.stock_img()
ax.plot(-0.08, 51.53, 'o', transform=ccrs.PlateCarree())
plt.contourf(lon, lat, var, 60, transform=ccrs.PlateCarree())
ax.coastlines()
plt.show()
```

```
CPU times: user [illegible], sys: [illegible], total: [illegible]
Wall time: [illegible]
```

Figure 20: JASMIN Notebook user session

The JASMIN Notebook service

Whilst it has been possible to use Jupyter Notebooks on JASMIN before now, doing so has never been officially supported and the process was cumbersome and error-prone. Our new Notebook Service aims to simplify access to Jupyter Notebooks on JASMIN by using the open-source software JupyterHub to manage multiple Jupyter Notebook servers for different users. The JASMIN Notebook service can be accessed at https://notebooks.jasmin.ac.uk, and access to the service is requested and approved using the JASMIN Accounts Portal.

Once access is approved, a user gets access to their own notebook server. The notebook server runs with the same permissions as the authenticated user, so users can access their home directory and CEDA Archive data from a notebook as they would from a scientific analysis server. Group Workspaces are also available, but are only accessible in read-only mode - this is a conscious decision, as we still want to encourage heavy processing and data production to use the LOTUS batch cluster. The JASMIN Notebook service then allows users to produce visualisations and interact with them in a much faster feedback loop than was previously possible using the scientific analysis servers (Fig. 20).

Jupyter has support for many languages including Python, R, Scala and Julia, which are implemented by plugins known as "kernels". The JASMIN Notebook Service currently provides one kernel - Python 3.7 with the latest Jaspy software environment installed. This environment is active by default, so there is no need for the module commands that are required on the sci-analysis servers and LOTUS.

In the six months since it was launched in March, the JASMIN Notebook service has seen rapid uptake with over 150 users registering to use the service. In this time, the service has been incredibly reliable.

1.17. PROVIDING MULTIPLE PYTHON ENVIRONMENTS WITH JASPY

Ag Stephens, Alan Iwi

Multiple, reproducible software environments on JASMIN

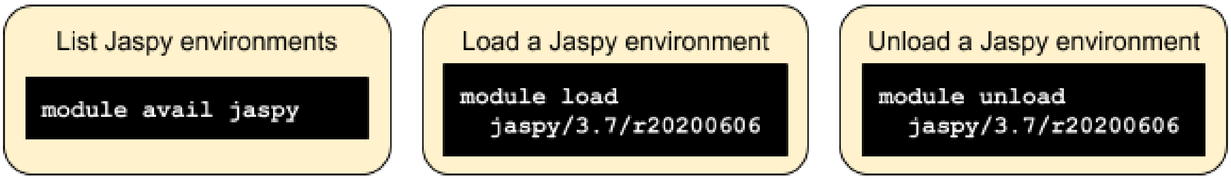
JASMIN users have a range of software requirements, from very basic python scripts through to full-blown data analysis packages. Jaspy, developed by CEDA, is a toolkit for managing and deploying software environments that include both python and non-python packages. Jaspy environments have been introduced because they provide a more flexible approach for describing, deploying and reproducing multiple environments on a single platform (i.e. JASMIN: LOTUS nodes and "sci" analysis servers). This allows us to retain previous environments and provide new ones simultaneously. Jaspy is particularly useful for (i) providing Python 2 and Python 3 access, and (ii) scientists undertaking long-running studies that require a consistent software environment over time.

Requirements for scientific software environments

Requirement	Details	Jaspy solution
Reproducibility - scientific work requires reproducible results	Generate a specific set of packages and versions from a generic set of requirements. Maintain access over time so that two equivalent installations result in the same software versions.	Jaspy makes use of Conda and its powerful package-management workflow. Conda sources (channels) may change over time, so Jaspy saves a copy of the binaries to a local, backed-up server and indexes them in a local channel.
Documentation - to guide users on software and how to access it	Provide an appropriate level of documentation detailing which software packages exist in each release.	We use Conda "environment files" to build the environments. These files list the packages and versions and are stored in public GitHub repositories, so each environment is documented as a collection of packages/versions.
Multiple environments - users need different packages	Allow multiple, but separate, software environments to co-exist on a single operating system.	Conda is designed to allow multiple environments to co-exist. Within Jaspy, it is possible to document each environment. This enables Python 2 and Python 3 environments to co-exist, and previous environments can be maintained for legacy use.
Management - given limited staff resources	Provide tools to easily construct, test, deploy, document and reproduce software environments.	Jaspy builds upon a set of excellent Conda command-line tools that simplify the package management process.

Using Jaspy

The "module" command gives users a simple interface to interact with Jaspy environments.



```
module avail jaspy
```

```
module load
  jaspy/3.7/r20200606
```

```
module unload
  jaspy/3.7/r20200606
```

Figure 21: Using the "module" command for Jaspy

Jaspy provides a flexible approach to software management that is suitable for the complex big-data analysis that is now required by scientists working on platforms such as JASMIN.

1.18. DELIVERING HANDS-ON WORKSHOPS TO OUR USER COMMUNITY

Poppy Townsend

On Wednesday 26th June, we welcomed 28 JASMIN users to our first ever hands-on interactive training event. The workshop consisted of a short introduction, followed by nine scenario-based exercises that encourage good practice for using JASMIN. Participants left the day with a set of common examples of how to use JASMIN efficiently which allows for easy adaptation to suit their individual workflows as appropriate.

The day was jam-packed with content; exercises were presented by Matt Pritchard (JASMIN Operations Manager), Fatima Chami (JASMIN User Support) and Ag Stephens (Head of Partnerships), whilst a range of other members of the JASMIN team were also on hand to provide 1-1 support with the exercises. All participants had access to a laptop so they could try out the exercises for themselves - which could have been a recipe for disaster! Luckily, the team was well equipped to deal with any difficulties faced on the day.

Figure 22: JASMIN workshop, October 2019

We received lots of positive, constructive feedback from the participants and have since incorporated this into the notes and worksheets. The workshop resources are freely available at: https://github.com/cedadev/jasmin-workshop .

Since the first workshop, we have delivered the training to over 100 attendees at four additional events:

- Internal session (consisting of staff from across RAL Space) on 11th July 2019
- RAL on 3rd October 2019 (Fig. 23)
- NCAS Leeds on 13th November 2019
- Met Office on 30th January 2020

The types of exercises users are learning at our workshops are:

- How to transfer data to/from JASMIN
- How to build and run workflows on our parallel cluster (LOTUS)
- How to choose which storage type is best for their work
- … amongst much more

Spending time providing interactive hands-on training to our user community is an essential part of the work we do. We believe well-trained users will put less pressure on the helpdesk team, spend less time on ‘getting started’ tasks: so they can spend more time doing their science, and will be less likely to inadvertently ‘break’ the services due to inexperience.

We will endeavour to continue improving and updating our provision of training opportunities in order to best support our user community.

PART 2: CEDA AND JASMIN SUMMARY INFORMATION 2019-2020

The Centre for Environmental Data Analysis (CEDA) exists to support the atmospheric, Earth Observation and near-Earth environment research communities in the UK and abroad through the provision of data management and access services. CEDA enhances this role through the development and maintenance of tools and services to aid data preservation, curation, discovery and visualisation; all of which add value for the world-wide user community.

The JASMIN data analysis facility provides petascale data-compute capabilities for the UK and wider environmental research communities. This section of the annual report presents summaries of CEDA Archive and JASMIN usage.

2. USAGE OF CEDA DATA

CEDA delivers Data Archive services for the National Centre for Atmospheric Science (NCAS) and the National Centre for Earth Observation (NCEO). In addition, CEDA delivers the NERC/STFC funded UK Solar System Data Centre (UKSSDC) and the IPCC Data Distribution Centre for the Intergovernmental Panel on Climate Change (IPCC).

Annual CEDA Archive Usage: April 2019 to March 2020	
Total number of users	22,051
Total data downloaded	885.4 TB
Total number of accesses	25,572,174
Total days activity	134,810

Table 2.1: Summary figures for usage by CEDA consumers during the reporting year

These figures can be broken down by month showing how usage has varied somewhat, but not dramatically, during the year (Table 2.2 and Figure 23). It is important to note that, with the continuing increase in processing *in-situ* in the JASMIN environment, users are less likely to download data to their own computers. Direct access to data on JASMIN is not captured in these figures.

Date	Users	Datasets	No. of accesses	Size	Activity days
2019/04	1,942	1,360	1,319,105	48.26	8,236
2019/05	1,962	1,538	1,366,318	52.19	8,986
2019/06	2,025	1,638	2,717,665	70.85	11,884
2019/07	2,140	1,350	2,511,313	55.2	11,425
2019/08	2,194	1,961	2,764,317	108.21	10,085
2019/09	2,121	1,381	1,798,582	54.3	8,746
2019/10	2,100	1,703	1,620,732	62.26	10,288
2019/11	2,459	2,675	4,321,662	81.78	16,141
2019/12	2,470	2,624	2,793,904	68.55	15,760
2020/01	3,081	1,829	1,579,361	80.38	13,804
2020/02	2,968	1,310	1,167,920	93.86	8,728
2020/03	3,252	1,596	1,611,295	109.53	10,727
Totals	**22,051**	**4,459**	**25,572,174**	**885.4 TB**	**134,810**

Table 2.2: Monthly summary figures for usage by CEDA consumers during the reporting year

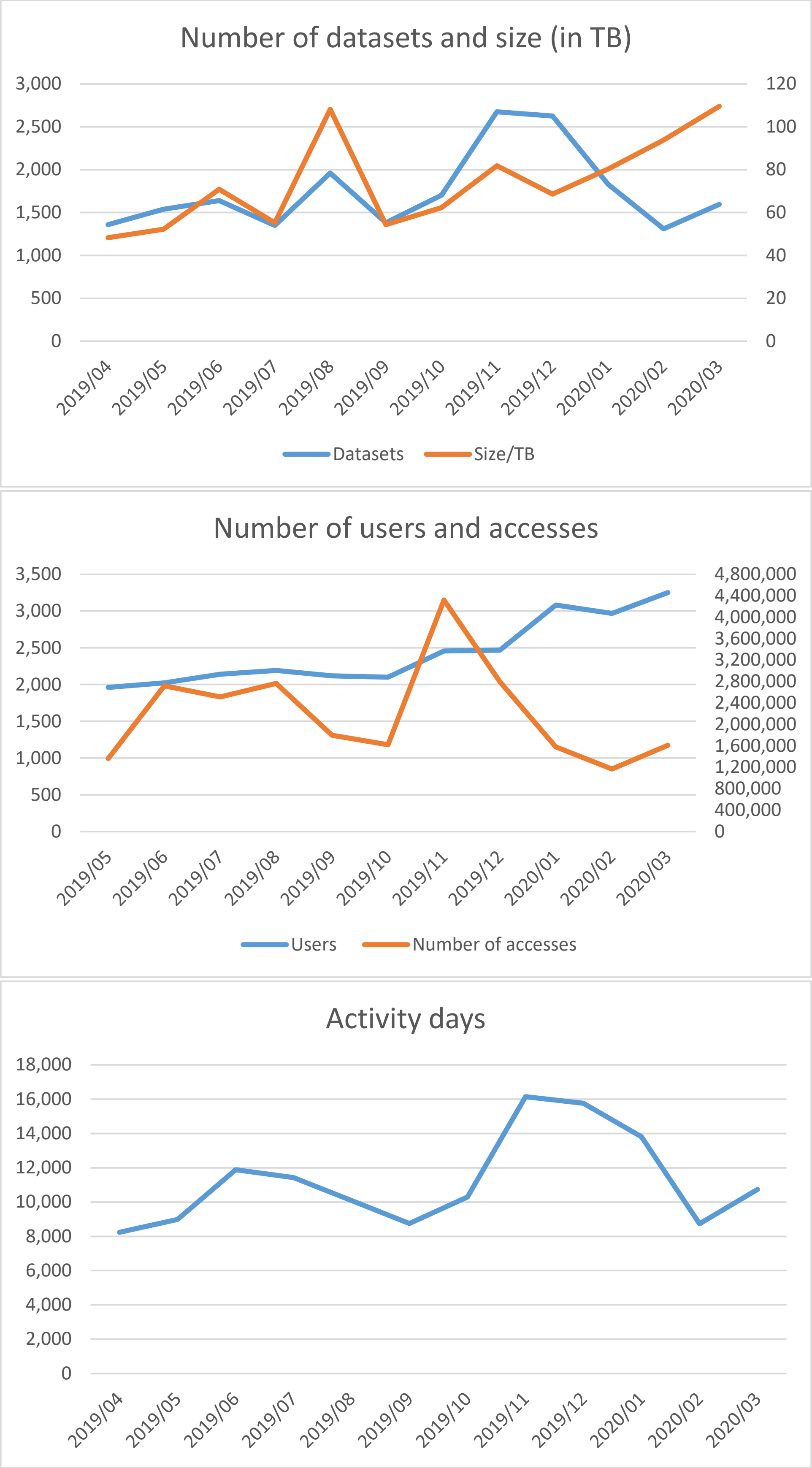


Figure 23: Breakdown of CEDA Archive usage by month

3. JASMIN

JASMIN is a globally-unique data analysis facility. It provides storage and compute facilities deployed in the combinations needed to enable highly data-intensive environmental science. Over 1500 users are currently supported exploring topics ranging from climate change and oceanography to air pollution, earthquake deformation and analysis of wildlife populations. Centred more around storage and data analysis than a "traditional" supercomputer, JASMIN provides more flexibility for a wide range of data-intensive analysis workflows.

Tens of petabytes of storage are combined with several types of compute resource: managed interactive and batch compute for building and executing large workflows, and a “community cloud” offering projects and communities a set of service components with which to build and manage their own computing needs.

JASMIN is designed, integrated and operated by Science & Technology Facilities Council (STFC) on behalf of the Natural Environment Research Council (NERC).

Architected jointly between STFC's Scientific Computing Department and the Centre for Environmental Data Analysis within RAL Space, it is operated and supported by a small but innovative team with expertise in computer science, research software engineering and environmental informatics.

Phase 6 of JASMIN’s upgrade programme this year used £2.0M of capital investment by NERC to refresh JASMIN’s batch compute hardware, alongside a number of other improvements. Meanwhile, JASMIN continued to provide a cutting-edge research environment for NERC-related environmental science while adapting its capabilities to cater for a new generation of scientific workflows.

3.1. USER SUPPORT AND OUTREACH

This year saw the development and delivery of a new introductory training workshop aimed at helping new users get the best out of JASMIN. The material was put together by members of the JASMIN team, drawing on their experience of creating and maintaining JASMIN user documentation and of helping users with their issues via the JASMIN helpdesk. The one-day workshop, designed around 9 interactive exercises, takes users through the basics of connecting to JASMIN and accessing compute and storage resources, through to more advanced topics such as managing a multi-step workflow. All the course materials are available online (https://github.com/cedadev/jasmin-workshop) and the course has already been delivered several times (at RAL, NCAS Leeds and the Met Office) to groups of 30-50 attendees, with plans to make it a regular feature of JASMIN’s user engagement activities, details of which now appear on the CEDA Events page (https://www.ceda.ac.uk/events/).

Members of the JASMIN team were also present for the first time at STFC’s Computing Innovations UK (CIUK) in the “research zone” in December 2019, representing JASMIN among other research and HPC facilities, with several talks from members of the team to an audience drawing from academia and industry.

In order to handle JASMIN-related queries more efficiently, a project is underway to create a dedicated JASMIN helpdesk, managed via the same system but as a separate “instance” from the CEDA helpdesk. Work has focussed on defining JASMIN’s own set of services and problem types, which enables better categorisation and triage of queries from JASMIN users, making sure that they reach the right person who can provide the most appropriate help.

3.2. UPGRADES

JASMIN’s new 30-petabyte Scale-out-filesystem storage cluster was brought into operation just before the start of this period and was already half full by Autumn 2019. Meanwhile, as well as replacement of block storage (used for virtual machine and cloud storage resources) the main focus of the Phase 6 upgrade was a significant refresh of compute hardware: over 130 new nodes, each with 1TB RAM, were procured and are now in the process of being integrated into the cluster. This signalled a major change for the JASMIN platform, however, with a new operating system (CentOS7), new batch scheduler (SLURM) and a new software environment (JASPY) all due to be introduced across the infrastructure.

In addition to the JASMIN Notebooks service (see highlight), a new graphical desktop service was introduced in response to user feedback, to improve performance when remotely accessing graphical applications on JASMIN.

An additional cloud platform was made available for the JASMIN Community Cloud, to run in parallel with the existing platform and between them cater for a larger set and wider range of tenant applications and workflows, but also to provide a cost-effective solution for the future. Once fully integrated, the additional platform will provide more capacity host tenancy “sci” machines: it is planned that this will become the default model for providing communities with interactive compute resources, in preference to the current sci machines which, shared by all, become overloaded at times. Newly-developed Cluster-as-a-Service components provide building blocks with which tenants can now assemble their own compute and application infrastructure in the Cloud. Several large collaborative projects are now exploiting this way of building infrastructures to support their own communities from within JASMIN. This has significantly increased uptake of the community cloud compared to the “bare bones” offering of infrastructure-as-a-service, as it enables tenants to get up and running with their applications more quickly.

Planning for Phase 7 is already underway, with the need to replace and expand other sections of JASMIN’s storage infrastructure and plans to expand the provision of object storage for cloud-based workflows and, following a successful proof-of-concept project, to establish a GPU cluster to provide a compute capability to cater for AI-type workflows. Underpinning all of this will be upgrades to JASMIN’s network infrastructure, a priority to complete before new storage and compute hardware can be installed, but also providing new 100G capability from within JASMIN to the capability now available via the RAL site to SuperJANET and the rest of the UK e-infrastructure.

4. COLLABORATIONS

CEDA continues to support the international climate modelling community through its interactions with the large global collaboration to deliver an Earth System Grid Federation. In Europe, we collaborate with partners in the ENES (European Network for Earth System Modelling) grouping, to efficiently distribute high-volume climate model simulation data.

Other major international collaborations include the European Space Agency's Climate Change Initiative Programme and the Data Distribution Centre (DDC) of the Intergovernmental Panel on Climate Change.

CEDA works closely with STFC's Scientific Computing Department to deliver the JASMIN infrastructure.

4.1. MAJOR COLLABORATIONS

In 2019/2020, significant national and international collaborations have continued. On the national scale, CEDA itself reflects a collaboration between the earth observation community, the atmospheric sciences community (via NCEO and NCAS) and the space weather community.

Additionally, CEDA is:

1. Working closely with the other NERC Environmental Data Centres, as part of the NERC Environmental Data Service.

2. Operating and evolving the Earth System Grid Federation (ESGF) in partnership with the US Programme for Climate Model Diagnosis and Intercomparison and a range of global partners in support of the sixth Coupled Model Intercomparison Project (CMIP6).

4. Working with the wider UK atmospheric science and earth observation communities, via a range of projects, with NCAS and other NERC funding.

5. Working with the European Space Agency on projects such as the ESA Climate Change Initiative (CCI) Open Data Portal.

7. CEDA is part of the UK Collaborative Ground Segment for Sentinel data (with UKSA, Airbus, Satellite Applications Catapult) with the role to provide Sentinel data mirror archives and data processing capability for the UK academic community.

8. CEDA works with ECMWF to provide EO scientists with the high resolution atmospheric analyses they need to process satellite observations.

9. With partners in Germany and the USA, CEDA provides data services on behalf of the IPCC (Intergovernmental Panel on Climate Change) through the Data Distribution Centre.

10. Supporting the Climate and Forecast Metadata (CF) Conventions with partners in University of Reading, UKMO, and multiple US research institutions.

12. With 20+ partners in the European Network for Earth System Modelling, CEDA is working to develop software and services for climate model data archives.

13. Working with academic partners in the UK Research and Innovation UKRI Cloud Working Group to share best practice, knowledge and strategy for use of cloud computing in the research domain.

5. FUNDING AND GOVERNANCE

In addition to supporting the National Centres of Atmospheric Science and Earth Observations (NCAS and NCEO, research centres of the Natural Environment Research Council, NERC), CEDA also delivers major projects with funding from a range of other bodies, including work for the European Space Agency (ESA), EC Copernicus Climate Change Service, BEIS, DEFRA and others, as well as participating and coordinating major European projects.

5.1. ANNUAL TOTAL FUNDING

Financial Year	12-13	13-14	14-15	15-16	16-17	17-18	18-19	19-20
NCAS income	935	829	829	808	808	808	808	808
NCEO income	445	392	390	393	393	393	402	393
Other NERC	287	272	600	621	825	816	733	883
Other income	1283	1486	1394	1505	1092	1280	1458	1377
Total income	2950	2979	3213	3327	3118	3297	3401	3461

Table 5.1: Overall funding for CEDA for financial years 2012 — 2013 to 2019 — 2020 (in £k)

Most of this funding comes to CEDA via a service level agreement (SLA) between the Natural Environment Research Council (NERC) and the Science and Technology Facilities Council (STFC). This SLA now covers both CEDA and JASMIN support explicitly.

5.2. EXTERNALLY FUNDED PROJECTS FOR THE YEAR 2019-2020

The table below shows CEDA's externally funded projects which were active during the reporting year.

Name	Description	Funder	Start date	End date	Value (£k)
FIDUCEO	JASMIN and data support for Fidelity and uncertainty in climate data records from Earth Observations	H2020	01/02/2015	31/08/2019	102.5
PRIMAVERA	JASMIN and archive support for new generation of global climate models	H2020	1/1/2015	31/10/2019	141.0
ESA CCI Open Data Portal	Data archive, catalogue and download services for ESA Climate Change Initiative	ESA	1/4/2014	31/3/2020	350.0
ESA CCI Knowledge Exchange	Data archive for ESA Climate Change Initiative as part of	ESA	5/9/2019	15/10/2022	445.0

	wider activity including outreach and education				
EA Air Quality Secondment	Develop architecture and roadmap for Air Quality Data management	Environment Agency	26/10/2019	31/01/2020	30.0
C3S ESGF Data Node (CP4CDS)	Operational ESGF data node for C3S	C3S	1/9/2016	31/5/2020	1043.0
Pest Risk Modelling in Africa (PRISE)	JASMIN support for UKSA IPP project	UKSA	1/12/2016	31/12/2020	117.1
BEIS IPCC DDC	UK component of IPCC Data Distribution Centre	BEIS	26/9/2018	31/03/2020	225.0
C3S CORDEX4CDS	Regional Climate Projection data for C3S	C3S	01/05/2017	30/04/2021	184.6
C3S_34e CDS WPS Services	Designing an interface between CDS toolbox and remote processing using WPS	C3S	01/01/2020	30/06/2021	359.7
C3S_434 CDS to ClimateAdapt	Transferring information between Climate Data Store and European Environment Agency’s portal	C3S	01/01/2020	31/12/2020	215.0
C3S_34g CMIP6	Including CMIP6 data in C3S Climate Data Store	C3S	01/02/2020	30/04/2021	168.3
MOHC Data Pipeline 2018-21	Supporting CMIP climate model data movement from the Met Office to the CEDA archive and ESGF	BEIS	04/06/2018	31/03/2021	450.0
UKCP18 Services 19-20	To provide data services to support access to the next generation of climate projections for the UK	Met Office/Defra	01/04/2019	30/04/2020	137.5
Support for Ensembles 18-19	CEDA support to multi-model climate ensemble archives	Met Office	01/01/2019	31/03/2020	100.0
C3S GLAMOD – Phase 2	In-situ observations for Copernicus	C3S	01/01/2019	28/02/2021	101.9

UKSA DAP Support 19-20	Funding Esther Conway to attend ESA Data Access and Preservation WG for UKSA	UKSA	01/04/2019	31/03/2020	14.0
IS-ENES3	Phase 3 of the distributed e-infrastructure of the European Network for Earth System Modelling	H2020	01/01/2019	31/12/2022	670.3
C3S Oceans Data Archival	Data archival for C3S Oceans project (U. Reading)	C3S	01/01/2019	30/06/2021	19.5
ESA EPs Common Architecture	Consultancy for ESA Exploitation Platforms Common Architecture	ESA	06/11/2018	31/10/2020	85.2
JASMIN for ESA SST CCI+	JASMIN Support for ESA CCI+ Sea Surface Temperature processing	ESA	1/7/2019	30/6/2022	30.0
JASMIN for CCI WV	JASMIN Support for ESA CCI Water Vapour processing	ESA	01/01/2019	31/12/2020	10.0

Table 5.2: Externally funded projects for 2018-2019 (non-core NERC)

5.3. GOVERNANCE

The CEDA/JASMIN board reflects our funding arrangements: CEDA and JASMIN are NERC funded facilities based at STFC Rutherford Appleton Laboratory. NERC funding includes both central provision and support from NCAS and NCEO. Thus, all three, as well as the host organisation, are represented on the board.

Membership of the CEDA/JASMIN board (March 2020) is:

Bryan Lawrence, NCAS (Chair)
Poppy Townsend, CEDA, STFC (Secretary)
Victoria Bennett, Division Head CEDA, STFC
Chris Mutlow, Director RAL Space
Stephen Mobbs, Director NCAS
John Remedios, Director NCEO
Martin Wooster, Divisional Director EOIF, NCEO
Frances Collingborn, NERC
Beth Greenaway, Head of EO, UKSA
Tony Hey, SCD, STFC
Tom Griffin, SCD, STFC

The Board aims to meet at least annually, and up to semi-annually when necessary. This year, the CEDA board met in February 2020, with the next meeting planned in June 2020.

For CEDA’s role in the NERC Environmental Data Service, there is additional governance and reporting, through the NERC Information Strategy Group and the NERC NC (National Capability) evaluation process.

PART 3: METRICS AND PUBLICATIONS

6. ADDITIONAL DATA CENTRE METRICS

CEDA is required to provide metrics quarterly in a number of categories. Some additional metrics to those provided in Chapter 1 are provided here.

Note that a considerable amount of use of CEDA is by users on JASMIN, who would not be measured in most of these statistics because the data is directly available on the file system (and we are currently unable to gather these metrics).

6.1. ACCESS RELATED METRICS

We can break down the users accessing registered datasets by geographical origin and institute type.

Area	Q1		Q2		Q3		Q4	
UK	2252	65.9%	2194	66.3%	2163	67.2%	2150	67.4%
Europe	718	21.0%	607	20.5%	363	11.3%	354	11.1%
Rest of the world	380	11.1%	376	11.4%	633	19.7%	614	19.6%
Unknown	66	1.9%	60	1.8%	60	1.9%	63	1.9%

Table 6.1: Users by area

Institute Type	Q1		Q2		Q3		Q4	
University	2537	74.3%	2434	73.6%	2371	73.7%	2358	73.9%
Government	472	13.8%	459	13.9%	453	14.1%	446	14.0%
NERC	187	5.5%	147	4.4%	136	4.2%	129	4.0%
Other	148	4.3%	196	5.9%	193	6.0%	192	6.0%
Commercial	40	1.2%	40	1.2%	40	1.2%	42	1.3%
School	30	0.9%	29	0.9%	24	0.7%	22	0.7%

Table 6.2: Users by Institute type

6.2. DATA HOLDINGS

Data Centre	Q1	Q2	Q3	Q4
CEDA	4118	3809	5553	6150

Table 6.3: Number of dataset discovery records held in the NERC data catalogue service.

	Q1	Q2	Q3	Q4
Datasets	5760	5618	6212	6252
Collections	632	576	624	631

Table 6.4: Number of dataset collections and datasets identified by CEDA and displayed via CEDA catalogue.

6.3. HELP DESK RESPONSIVENESS

	Q1	Q2	Q3	Q4
Received	952	805	919	1064
Closed	940	796	912	1039

Table 6.5: Help desk queries received and closed by quarter, including the three-day closure rates. These queries cover all aspects of data support except dataset access issues.

	Q1	Q2	Q3	Q4
Received	389	406	421	487
Closed	379	404	416	496

Table 6.6: Help desk queries specifically about access authorisation for restricted CEDA datasets and services received and closed by quarter, including the three-day closure rates.

7. PUBLICATIONS AND PRESENTATIONS

CEDA staff are in bold

Bennett, Victoria, Esther Conway and **Alison Waterfall**, "EO Data and Services at CEDA and JASMIN", presentations at NCEO data day, London, 8th Jan 2020

Conway, Esther, S. Hebden, M. Odgers, B. Archer, Poster, "Machine Learning activities at the CEOS Working Group on Information Systems and Services", Machine Learning workshop British Antarctic Survey 17-18th June 2019

Conway, Esther, Data Science Journal Special Collection (Editor) - January 2020, https://datascience.codata.org/collections/special/pv2018-conference-special-collection/

Conway, Esther, Presentations at CEOS Meetings:
CEOS WGISS – "UK ARD, Downstream and Calibration Data" 9th Oct 2019
CEOS WGISS – "UKSA Agency Report" 10th Oct 2019
CEOS WGISS - "PV2020 Conference" 10th Oct 2019
CEOS: "Persistent Identifiers Best Practices"; Data Stewardship Interest Group; Dec 2019

Eggleton, Francesca, Poster, 'Research Data Management at the Centre for Environmental Data Analysis (CEDA)', NCAS Climate Modelling Summer School, September 2019 Cambridge

Garland, Wendy, Poster, "The CEDA Archive and JASMIN: Combining environmental data repositories with an active research platform" at the International Digital Curation Conference (IDCC) conference, Dublin 17-20 Feb 2020

Juckes, M., Taylor, K. E., Durack, P. J., **Lawrence, B.**, Mizielinski, M. S., **Pamment, A.**, Peterschmitt, J.-Y., Rixen, M., & Sénési, S. (2020). The CMIP6 Data Request (DREQ, Version 01.00.31). Geoscientific Model Development, 13(1), 201–224. https://doi.org/10.5194/gmd-13-201-2020

Juckes, Martin, Ruth Petrie, Andras Horanyi (C3S): Poster "C3S 34a Lot1 : Global Climate Projections: data access, product generation and impact of frontline developments. Lot 1: Provision of support to one ESGF node in Europe.

Juckes, Martin, talk: "Air Quality Data at CEDA". Meeting of the STFC Air Quality Plus network, York, September 2019.

Juckes, Martin, talk: "Collaborative frameworks and best practices", IPCC WGI Training on Data and Software Development, Munich, 2019.

Kershaw, Philip, Matt Pryor, Alan Iwi, Ag Stephens, Martin Juckes, Ruth Petrie, Stephan Kindermann, Carsten Ehbrecht, Sébastien Denvil, Sébastien Gardoll, and Bryan Lawrence, EGU2019-18192 | Orals | ESSI3.2; Delivering resilient access to global climate projections data for the Copernicus Climate Data Store using a distributed data infrastructure and hybrid cloud model ; (Presentation) EGU General Assembly 2019, Vienna, Austria, 7–12 April 2019

Kershaw, Philip, Victoria Bennett, Jonathan Churchill, **Matt Pritchard**, Bryan Lawrence, JASMIN Challenges and Growth, CEOS WGISS-47 Meeting, (Presentation), 30 April 2019

Kershaw, Philip, Future data access architecture with Nginx, OpenID Connect, OpenSearch and Kubernetes, ESGF Future Architecture Webinar, (Presentation), 23 March 2020

Kershaw, Philip, Future Architecture for ESGF, (Presentation), IS-ENES3 Project General Assembly, 25-27 March 2020

Kershaw, Philip, Bryan Lawrence, Jonathan Churchill, **Matt Pritchard**, **Victoria Bennett**, JASMIN and the evolution of cloud-hosted data analytics platforms for the environmental sciences, STFC Computing Insight UK Conference, (Presentation), Manchester, 5-6 December 2019

Kershaw, P., Halsall, K., Lawrence, B. N., **Bennett, V., Donegan, S., Iwi, A., Juckes, M.,** Pechorro, E., **Petrie, R**., **Singleton, J., Stephens, A., Waterfall, A.**, Wilson, A., & Wood, A. (2020). Developing an Open Data Portal for the ESA Climate Change Initiative. Data Science Journal, 19, 16. https://doi.org/10.5334/dsj-2020-016

Kershaw, Philip: JASMIN Cluster-as-a-Service for Climate Data Analysis, NCEO Conference, Nottingham, 2-5 September 2019

Massey, Neil, "JASMIN: collaborative computing for environmental science", Big Data in the Geosciences Workshop, Lancaster University, June 19th to 20th 2019

Massey, Neil, "Semantic storage of climate data on object store", Special Interest Group in High Performance IO Workshop, Reading University (virtual), April 23rd 2020

Merchant, C.J., Embury, O., Bulgin, C.E., Block, T., Corlett, G., Fiedler, E., Good, S.A., Mittaz, J., Rayner, N.A., Berry, D., Eastwood, S., Taylor, M., Tsushima, Y., **Waterfall, A**., Wilson, R., Donlon, C., Satellite-based time-series of sea-surface temperature since 1981 for climate applications. Sci Data 6, 223 (2019). https://doi.org/10.1038/s41597-019-0236-x

Pamment, Alison: 'NCAS Contributions to the CF (Climate-Forecast) Metadata Conventions', Poster, NCAS staff conference, Birmingham, 1-2 July 2019.

Parton, Graham, Ag Stephens, Richard Smith and Joe Singleton; EGU2019-16513 | Posters | ESSI2.9 | Presentation; "Long-term Archive Challenges: Enhancing Data Discovery via Multilevel Metadata Aggregations At Scale"; EGU General Assembly 2019, Vienna, Austria, 7-12 April 2019

Parton, Graham Alexander, & Pepler, Sam. (2019, November). Permissible Closed-Use General licences: filling the gap between Open and Restrictive Data Licences. Presentation; The 14th International Conference on Open Repositories (or2019), Hamburg, Germany, June 10-13, 2019. Zenodo. http://doi.org/10.5281/zenodo.3554229

Parton, Graham, Kate Winfield, Sam Pepler, Centre for Environmental Data Analysis (CEDA), STFC, UK; Can we enhance data discovery by standardising licence classification? Invited Presentation; Research Data Alliance 14th Plenary (RDA 14), Helsinki, 23-25th October 2019

Parton, Graham, Kate Winfield, Sam Pepler, Centre for Environmental Data Analysis (CEDA), STFC, UK; Can we enhance data discovery by standardising licence classification? Poster; Research Data Alliance 14th Plenary (RDA 14), Helsinki, 23-25th October 2019

Parton, Graham, Diary of a data scientist website; What do data scientists do: The data science diaries; Solicited blog entry; Institute and Faculty of Actuaries (IFoA): https://www.actuaries.org.uk/learn-and-develop/lifelong-learning/data-science-actuarial-viewpoint/what-do-data-scientists-do-data-science-diaries

Pascoe, C., Lawrence, B. N., Guilyardi, E., **Juckes, M**., & Taylor, K. E. (2020). Documenting Numerical Experiments in Support of the Coupled Model Intercomparison Project Phase 6 (CMIP6). Geoscientific Model Development, 13(5), 2149–2167. https://doi.org/10.5194/gmd-13-2149-2020

Pascoe, Charlotte, Bryan Lawrence, Eric Guilyardi, Mark Greenslade, David Hassell, and Chris Blanton; EGU2019-9749 | Posters | ESSI2.1; Comparison of Earth system models in large multi-model ensembles ; EGU General Assembly 2019, Vienna, Austria, 7–12 April 2019

Pechorro, E., **V. Bennett**, P. Cipollini, F. Paul, R. Hollmann, "Climate Stories through CCI Data", Presentation, Agora session, ESA Living Planet Symposium, Milan, 13-17 May 2019

Pepler, Sam, Keeping research data long term and making it accessible. Invited Presentation; Manage your data before it manages you! Government Digital Service, The National Archive, Kew, March 2019.

Pepler, Sam, Poppy Townsend, The Environmental Data Service, Poster; NCAS Staff Meeting and Atmospheric Science Conference, Birmingham, 1-2 July / 2-3 July 2019

Popp, Thomas, Michaela I. Hegglin, Rainer Hollmann, Fabrice Ardhuin, Annett Bartsch, Ana Bastos, **Victoria Bennett**, Jacqueline Boutin, Carsten Brockmann, Michael Buchwitz, Emilio Chuvieco, Philippe Ciais, Wouter Dorigo, Darren Ghent, Richard Jones, Thomas Lavergne, Christopher J. Merchant, Benoit Meyssignac, Frank Paul, Shaun Quegan, Shubha Sathyendranath, Tracy Scanlon, Marc Schröder, Stefan G. H. Simis, Ulrika Willén, Consistency of satellite climate data records for Earth system monitoring, Bull. Amer. Meteor. Soc., https://doi.org/10.1175/BAMS-D-19-0127.1 (2020)

Rossi, C. and **V. Bennett**, 'Where to access earth observation satellite data', presentation at UK Space Conference, International Convention Centre, Wales, 24-26 September 2019

Sellar, A.A., et al. including **Alan Iwi, Ruth Petrie, Ag Stephens**, "Implementation of UK Earth system models for CMIP6", February 2020, Journal of Advances in Modeling Earth Systems 12(4), DOI: 10.1029/2019MS001946

Townsend, Poppy, Bryan Lawrence, Rosalyn Hatcher, and **Victoria Bennett**; EGU2019-17093 | Orals | EOS10.1/AS5.25/BG1.59/GI1.8/OS4.34/SM5.8 | Presentation; "Gathering impact stories from JASMIN users" ; EGU General Assembly 2019, Vienna, Austria, 7–12 April 2019

Waterfall, A., V. Bennett, E. Pechorro, "Data Standards for the ESA Climate Change Initiative", Poster, ESA Living Planet Symposium, Milan, 13-17 May 2019

Williamson, E., S. Donegan, V. Bennett, A. Waterfall, P. Kershaw, "Earth Observation Data and Services at CEDA", Poster, ESA Living Planet Symposium, Milan, 13-17 May 2019

Williamson, Ed, Steve Donegan: CEDA and JASMIN Services (Poster), NCEO Conference, Nottingham, 2-5 September 2019

Winfield, Kate, Poster 'Data Scientist in CEDA', COSINE STFC graduate conference at RAL, 13th November

Winfield, Kate, Poster 'Data Scientist in CEDA', NCAS staff conference, Birmingham, 1-2 July 2019

Winfield, Kate, Poster 'Overview of the CEDA archive', Royal Met Student conference, Birmingham, 4-5 July

Winfield, Kate, talk 'Data Management', NCAS Summer School May 2019 Isle of Arran