Centre for Environmental Data Analysis (CEDA)

Annual Report 2019

(April 2018 to March 2019)

Victoria Bennett,
Poppy Townsend (Editors)

CONTENTS

1. INTRODUCTION

The Centre for Environmental 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. Starting in April 2018, CEDA is now a component part of the NERC Environmental Data Service, bringing together the five NERC data centres into a single service commissioned by NERC as National Capability.

JASMIN is the data intensive supercomputer which provides the infrastructure upon which CEDA and the CEDA 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 a considerable investment in JASMIN Phase 4, which was fully deployed this year, enabling us to significantly increase capacity, in storage and computing, but also in the cloud 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 year has seen data from CMIP6 (Coupled Model Intercomparison Project, Phase 6) begin to arrive, at volumes dwarfing the data from its predecessor project, CMIP5. CEDA has developed tools and systems to handle the "deluge", in order to ensure access for UK researchers. In parallel we have continued to acquire regular data, around 10 Terabytes per day, of Earth Observation data from the European Sentinel satellites, creating a unique archive for climate research on JASMIN. Other highlights this year include delivering a new and successful series of user training webinars, a major activity to migrate petabytes of data onto new JASMIN storage hardware, and a range of external projects which rely on our experience and knowledge in data management, data standards and data services.

As in previous years, our key partnerships include our sister department in STFC, Scientific Computing Department, with whom we deliver the JASMIN infrastructure, and internationally the European Network for Earth Simulation (our collaborators in delivering the European component of the Earth System Grid Federation), and many other project collaborators.

This annual report presents key statistics for the year past (2018- 2019) as well as a series of highlights reports showcasing a cross section of our activities. Key metrics are also provided. I hope it provides an interesting insight into another successful year at 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. NERC ENVIRONMENTAL DATA SERVICE

Sam Pepler, Victoria Bennett

From April 2018, the five NERC data centres have been commissioned as a single Environmental Data Service (EDS), commissioned by NERC. The combined service includes:

- British Geological Society (BGS) National Geoscience Data Centre (NGDC)
- Centre for Ecology and Hydrology (CEH) Environmental Information Data Centre (EIDC)
- National Oceanography Centre (NOC) British Oceanographic Data Centre (BODC)
- National Centre for Atmospheric Science (NCAS) and National Centre for Earth Observation Centre for Environmental Data Analysis (CEDA), and
- British Antarctic Survey (BAS) Polar Data Centre (PDC)

For the individual components of the service, work has largely continued as business as usual, carrying out data management and dissemination and delivering data services for the target user communities, but the emphasis on working together in an integrated manner has increased.

Some services were already shared, e.g. the NERC Data Catalogue (Figure 1). In the new integrated service we will be improving existing services, and rolling out others where it is appropriate, e.g. the NERC Data Labs developed by CEH to support data analysis in a user friendly environment.

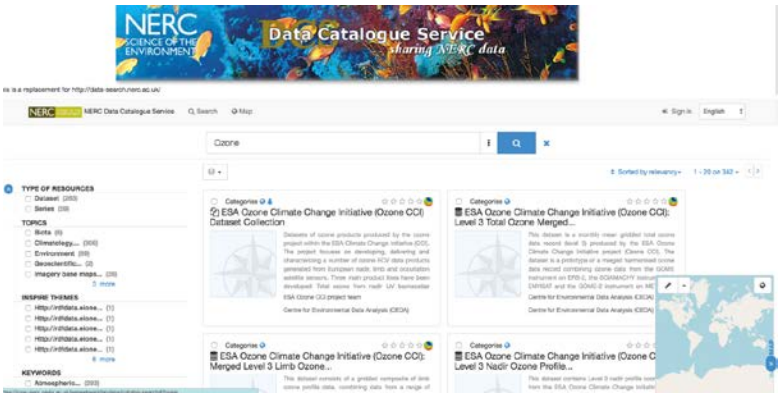


Figure 1: NERC Data Catalogue Service, interface to data from all five NERC data centres

Another area where expertise will be shared relates to controlled vocabularies: BODC has significant experience and existing systems for managing vocabularies (the NERC Vocabulary Service, figure 2): these allow data to be indexed in a consistent way. A common approach across the EDS will help provide reliable information into integrated services, e.g. data search across all centres. For example, a search for "nitrogen" would at present not return all relevant records (e.g. "nitrate" or "ammonium") if different indexing approaches and terminologies have been used.

Figure 2: Sample content from NERC Vocabulary Service, hosted by BODC

Where it makes sense, we can use shared compute: storage and other resources. We are starting work on a shared deposit service for large datasets, which will use JASMIN storage.

There is work to do to improve navigation between component services, and to look more "joined-up" to the user community. These integration tasks will continue into the coming years, to underpin the use of NERC data in the science community, but also to support data innovation and exploitation of NERC data more widely.

1.2. MANAGING NERC SCIENCE DATA: A GROWING TASK!

Kate Winfield

CEDA is responsible for curating atmospheric and earth observation data from research that has been funded by the Natural Environment Research Council (NERC), ensuring long-term preservation and re-use of data. CEDA has seen a steadily growing number of complete, ongoing and new NERC projects. In 2017, CEDA was the designated archive for data from about 250 NERC projects, which has grown to about 380 in 2019.

Curating these data isn't just a case of simply 'uploading' data files to the CEDA Archive. CEDA staff ensure that these data remain of long-term value by ensuring the data are acquired in well-supported formats and have accompanying documentation to describe them. A key role of CEDA data scientists is to liaise with data providers in universities and research institutes for the entire lifecycle of their NERC grants.

For each project, the CEDA science support team work closely with the project participants to draft and agree a Data Management Plan (DMP). For all new NERC projects, a DMP needs to be agreed between CEDA and the data provider within the first 3-6 months of the project start date. This is to ensure a project sufficiently plans how the data is preserved for the long term. We in CEDA have seen an increase in the number of DMPs that we have developed, and that have been accepted by the Principal Investigators (See figure 3).

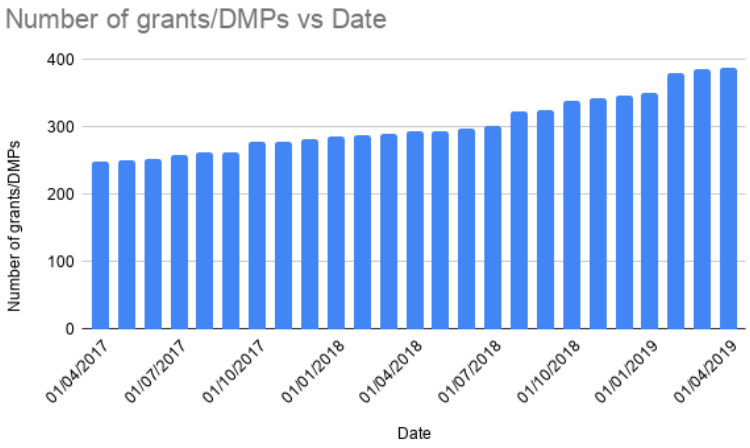


Figure 3: Number of CEDA grants with Data Management Plans since 2017

The increase in interaction with NERC projects certainly keeps us busy and it's a trend that is likely to continue. Funding bodies such as NERC, and scientific journals, increasingly encourage or require data to be made available via a recognised data centre - so we anticipate this growth to continue in the future.

As the number of projects increases, we add more NERC datasets to our catalogue and archive, for long term preservation, and dissemination to a wide user community.

1.3. FACILITATING ACCESS TO MODIS AND OTHER EARTH OBSERVATION DATASETS

Alison Waterfall

CEDA continues to facilitate access to high volume Earth Observation (EO) datasets by providing direct access to specific datasets on JASMIN and for download via the CEDA Archive. This includes access to many satellite data products that are essential to the EO community, including from the Copernicus Sentinel programme, EUMETSAT meteorological satellites, and US agencies NASA and NOAA. This is important, as the high volumes of these datasets provide a challenge for individual scientists to store in their home institutions. Providing a centralised point of access improves efficiency in both the storage of the data and allows scientists to focus on exploiting the data rather than spending valuable time simply obtaining the input datasets. Datasets have been selected according to priorities from the National Centre for Earth Observation (NCEO) and other strategic projects. Recently, MODIS data (figure 4) was identified by the National Centre for Earth Observation (NCEO) as a high priority dataset for use on JASMIN.

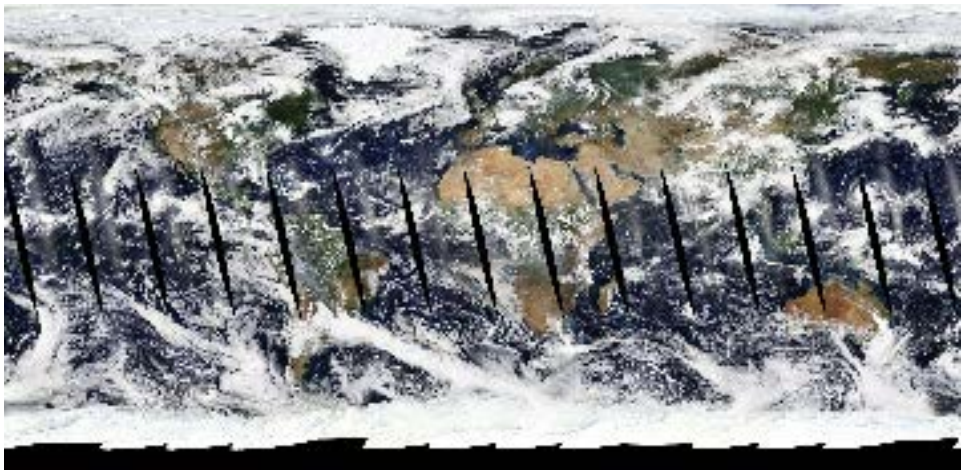


Figure 4: MODIS –Aqua example image from the NASA NEO website for 1st September 2018 (Imagery from the Land Atmosphere Near real-time Capability for EOS (LANCE) system and services from the Global Imagery Browse Services (GIBS), both operated by the NASA Earth Observing System Data and Information System (EOSDIS)

MODIS (Moderate Resolution Imaging Spectroradiometer) is a satellite instrument flown on two NASA satellites (Terra and Aqua) providing global observations in 36 spectral bands at resolutions of between 250m and 1km depending on the band. The data has been processed by NASA and a wide variety of products (over 100) are available, from Level 1B radiance data to a wide variety of products covering the atmosphere, land, ocean and cryosphere domains. It is not possible, or necessary, to hold the complete suite of MODIS products on CEDA, but effort has been made to acquire the key NCEO requested products, over 20 so far. This has been a major exercise, and in the year 2018-2019, over 500 TB of MODIS data was added to the archive (figure 5), and much more will be added in the coming months.

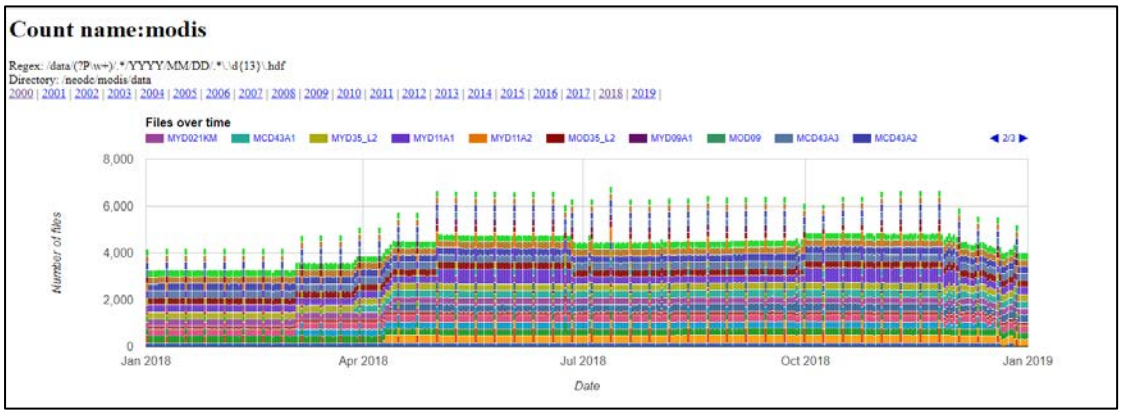


Figure 5: Graph (kept up to date and made available to users via CEDA website) shows number of files added to the CEDA Archive for each MODIS product

1.4. DATA PIPELINES FOR CMIP6

Ruth Petrie, Alan Iwi and Ag Stephens

The 6th Coupled Model Intercomparison Project (CMIP6) will deliver a multi Petabyte archive of climate model projections. CMIP6 is made up of 23 separate sub-intercomparison projects, involving about 40 modelling groups running around 250 experiments with over 100 datasets expected to be archived and published.

The role of CEDA within CMIP6 is to be the UK "node" for the Earth System Grid Federation (ESGF). The ESGF is an international collaboration for the software that powers most global climate change research, notably assessments by the Intergovernmental Panel on Climate Change (IPCC) providing tools and interfaces for data management, discovery and download. CEDA will receive CMIP6 data from the UK Met Office Hadley Centre, UK collaborative projects, and replica data from contributing modelling centres around the world utilising the ESGF network. CEDA will archive, manage and publish all received CMIP6 data and provide analysis capabilities through JASMIN. The volume of data requires that the ingestion into the CEDA Archive and publication to ESGF be fully automated and flexible enough to work for all incoming data sources. Details of the approach taken to manage these data flows are given below.

Met Office Hadley Centre pipeline

CEDA is responsible for archival and publication (to ESGF) of all model experimental data from the Met Office Hadley Centre (MOHC) contributions to CMIP6. Once a dataset has been fully prepared and verified by the MOHC then it is written to the MASS tape store and a message is sent to the CEDA RabbitMQ service. The CREPP service (see below) reads the message and will then process the dataset. The first stage involves extraction from tape using the MASS client available on JASMIN. On completion the message is consumed to indicate success.

UK Collaborative model pipeline

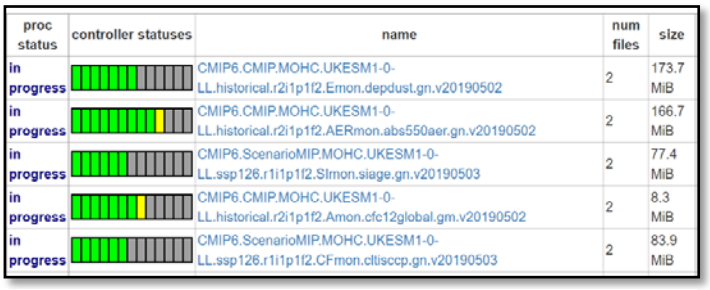
A number of simulations are being performed using the UK Earth System Model which is a collaboration between the Met Office and NERC. These simulations are performed across multiple HPC sites in the UK and abroad (such as the Korean Met Agency). The pipeline for these data is initiated when the scientists have brought their outputs to JASMIN and prepared them for submission to CMIP6. The scientist runs a command-line script to inform CREPP that the dataset can be published to ESGF.

International data replication pipeline

International replicated data are copied from the global ESGF federation. Timely data replication is necessary for UK scientists who will undertake model inter-comparisons and analyses to contribute to the next IPCC scientific assessment report. The total volume of CMIP6 data from all international contributors will be at least 15PB. This minimum estimate is too large for CEDA to hold a full copy on disk so we will prioritise data to replicate and seek alternative solutions such as tape storage as needed. Replication will continue throughout the lifetime of CMIP6.

CEDA REceive-to-Publish Pipeline (CREPP) Tool

The CEDA REceive-to-Publish Pipeline (CREPP) tool (figure 6) automates the archival and publication process of all CMIP6 data pipelines making the data available on both the ESGF and JASMIN. CREPP will receive a notification that a dataset is ready, one of the CREPP "controllers" will then begin processing the data. The early stages include some preparation then the data is ingested to the CEDA Archive, at this stage data are available on JASMIN. Once in the CEDA Archive the dataset can be passed through a number of publication controllers before becoming visible through the standard ESGF search interfaces.



proc status	controller statuses	name	num files	size
in progress		CMIP6.CMIP.MOHC.UKESM1-0-LL.historical.r2i1p1f2.Emon.depdust.gn.v20190502	2	173.7 MiB
in progress		CMIP6.CMIP.MOHC.UKESM1-0-LL.historical.r2i1p1f2.AERmon.abs550aer.gn.v20190502	2	166.7 MiB
in progress		CMIP6.ScenarioMIP.MOHC.UKESM1-0-LL.ssp126.r1i1p1f2.SImon.siage.gn.v20190503	2	77.4 MiB
in progress		CMIP6.CMIP.MOHC.UKESM1-0-LL.historical.r2i1p1f2.Amon.cfc12global.gm.v20190502	2	8.3 MiB
in progress		CMIP6.ScenarioMIP.MOHC.UKESM1-0-LL.ssp126.r1i1p1f2.CFmon.cltisccp.gn.v20190503	2	83.9 MiB

Figure 6. Screenshot of the CREPP data pipeline web interface. The figure shows a number of Met Office datasets in the process of being archived and published. A green controller status indicates successful completion and yellow is in-progress.

1.5. TRAINING OUR USERS: THE CEDA WEBINAR SERIES

Poppy Townsend

After the success of our first webinar last year, we have continued to produce a range of training materials presented in a webinar format. Four webinars have been presented since April 2018, covering the following topics; Git and Github basics , how to archive data, LOTUS basics and Further use of Git and GitHub. These webinars were predominantly guided by user community demand - and as such were highly attended - with a combined total of over 250 registered attendees from 21 different countries (figure 7). All videos and slides are freely available on our website and YouTube channel - with 320 video views combined (as of May 2019).

A range of other training topics are planned to take place in 2019, as requested by the user community. Proposed topics include; Further use of LOTUS, Python basics, Software on JASMIN. These will be shared freely on our website here.

The need for CEDA to provide additional training has intensified in recent years as we continue to receive an increasing number of user queries on the helpdesk. These webinars have already helped to diversify our documentation/training materials, reduce pressure on helpdesk staff and increase our engagement with the community.

Figure 7: Map showing the location of attendees that registered for webinars 4 and 5 (Intro to Git and GitHub, Further use of Git and GitHub)

1.6. DEVELOPMENTS IN DATA CITATION

Sarah Callaghan

In order to promote the sharing and reuse of data, CEDA mint DOIs (Digital Object Identifiers) as permanent identifiers for datasets that are complete and well documented, and ready to become part of the scientific record. This allows these datasets to be cited as if they were a journal article, and their usage tracked through those citations.

Over the calendar year 2018 the five NERC Data Centres (of which CEDA is one) minted 445 new DOIs for datasets. On the 15th January 2019, representatives from all the NERC Data Centres got together to share their experiences with minting DOIs, and to ensure that the processes we'd set up several years before were still fit for purpose, which was discovered to be the case. Figure 8 shows the increase in DOI uptake since 2011.

In the wider community, data citation uptake is progressing slowly but surely, helped by initiatives such as Make Data Count (https://makedatacount.org/), along with their event viewer (https://api.datacite.org/events), which gives information on downloads and citations for a given DOI-ed dataset.

Commercial services such as the Data Citation Index (DCI) (https://clarivate.com/products/web-of-science/web-science-form/data-citation-index/) are also in operation, though the DCI is a closed, for-pay service and is therefore not easily accessed. Some NERC datasets are indexed in the DCI, and plans have been made to index more of these - focussing on the DOI-ed datasets in our archives.

Data citation is an important part of our service, not only for tracking the impact of our data, but also for providing credit to the dataset creators for their hard work in creating the datasets in the first place. We are working with national and international partners to change the global scientific culture to make data be considered a first class research output, and make data citation the norm.

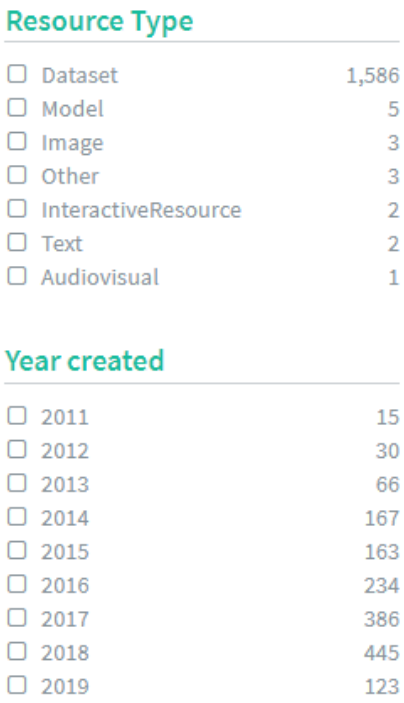


Figure 8: statistics of DOIs minted by the NERC Data Centres (as of 17th April 2019)

1.7. DEVELOPING NEW ARCHIVE TOOLS TO MEET DEMAND

Sam Pepler

It's easy to characterise the action of adding a file to the CEDA Archive as simply copying, in actual fact the reality is a lot more involved. In order to undertake this process, a tool called the 'deposit server' is used - think of this as the input funnel for the Archive. It takes requests to deposit files and checks a few basics, like file names having no forbidden characters and that the destination dataset is still open for receiving more files. It then copies the file and logs that it has been done. It has to do this many times a second, for large volumes, and it all needs to be done reliably.

A major redevelopment of the deposit process was carried out this year. The old deposit server was implemented as a local web service which made it difficult to spread the load and it struggled to keep up with the volume of requests. The new server is implemented via a queuing protocol. This enables load sharing, both across machines and by spreading peak loads over time. It also allows a robust system where single deposit servers can be taken down without affecting the overall service.

Figure 9: "Input funnel" for the CEDA Data Archive

The service is now operational and ingestion of over 25TB per day is not uncommon; the old service would have struggled with this rate of deposit. The new service enables additional features to be added in the future, for example: the latest generation of storage used, Quobyte, has introduced a number of extra checks that could be done to increase robustness, and the new queueing architecture offers new possibilities for monitoring. Overall the new service is performing well and will enable even higher rates of deposit in the future. With this input funnel open wide, the most likely problem now will be running out of storage space.

1.8. GOOGLE DATASET SEARCH

Sarah Callaghan, Madeleine Russell

In September 2018 Google launched a new search tool to find datasets made publicly available on the web, which aims to help scientists, policy makers and other user groups more easily find the data required for their work. CEDA was a key part of these efforts, not only by providing relevant tagged catalogue pages for the data to be harvested, but also by providing consultancy and user testing for the first iterations of the search interface.

The Dataset Search is similar to how Google Scholar works, in that it lets users find datasets wherever they're hosted, whether it's a publisher's site, a digital library, or an author's personal web page. Metadata about the dataset is harvested from the dataset's landing page through the use of schema.org tags, which provide a structured and standardised way of sharing that metadata.

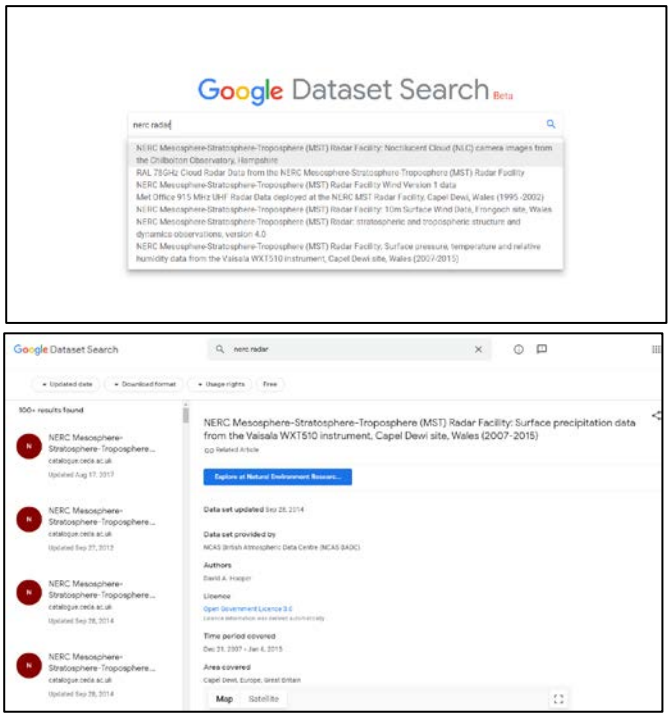


Figure 10: Example search and results, for NERC radar data, in Google Dataset Search tool

To create Dataset Search, Google developed guidelines for dataset providers to describe their data in a way that search engines can better understand the content of their pages. These guidelines include salient information about datasets: who created the dataset, when it was published, how the data was collected, what the terms are for using the data, etc. This enables search engines to collect and link this information, analyse where different versions of the same dataset might be, and find publications that may be describing or discussing the dataset. The approach is based on an open standard for describing this information (schema. org). Many CEDA datasets for environmental data are already described in this way and are particularly good examples of findable, user-friendly datasets.

The new Google Dataset Search (figure 10) offers references to most datasets in environmental and social sciences, as well as data from other disciplines including government data and data provided by news organisations. It has the potential to revolutionise how users find and discover data from a wide variety of sources, and will open up usage of CEDA's datasets to new communities and interested parties.

1.9. UK CLIMATE PROJECTIONS DATA SERVICE

Ag Stephens, Antony Wilson, Charlotte Pascoe and Neil Massey

The government-funded UK Climate Projections (or UKCP18) provide the most up-to-date assessment of how the UK climate may change during this century. Developed by the Met Office, the projection data sets include information to support scientific studies, climate change risk assessments and adaptation plans. UKCP18 uses cutting-edge climate science to deliver updates to land and marine climate scenarios out to 2100, including: probabilistic land projections, high-resolution spatially-coherent land simulations and coastal projections related to sea-level rise and storm surge.

CEDA plays a pivotal role in UKCP18 by providing both the data archive and tools for interfacing with the data sets. Following lessons learnt from the predecessor project (UKCP09), it was agreed that all data should be provided in the standard CF-netCDF format and be distributed on an Open Government Licence (OGL). This allowed CEDA to manage the data provision using its existing infrastructure: the data is fully discoverable through the CEDA Catalogue (Figure 11) and is accessible via web download services and on the local JASMIN file system.

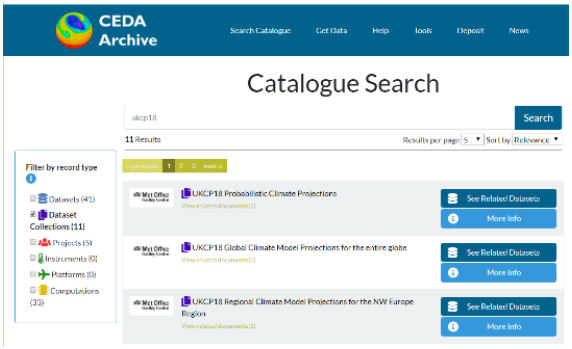


Figure 11: UKCP18 records in the CEDA Catalogue

The web-tool to support the projections needed to provide an intuitive interface for selecting data products, tailoring the outputs and enabling batch extractions where required. The emergent solution, the UKCP18 User Interface (UI), draws together a set of pre-existing technologies in a responsive and scalable web-interface (Figure 12). The architecture builds on the OGC Web Processing Service (WPS) standard to manage the client-server interactions using a well-described API that can be used independently of the UI. This enables scientific workflows and third-party software to connect directly to the service rather than interacting with a web-interface.

Figure 12: Example product from the UKCP18 UI: Percentile maps showing future change UK precipitation over administrative regions

In order to encourage innovation, collaboration and code-sharing, a significant component of the software has been released as an open-source licence as the "UKCP Data Processor" package. This Python library can be used to subset, visualise and convert UKCP18 data sets. We hope to engage with a community of developers to extend this code for wider application and greater re-use.

The service is already experiencing substantial usage with over 150 new users registering and over 3,000 data requests processed each month. Users come from a range of sectors with particular interest from those working in water resources, flood management and coastal issues. As the community gets to grips with these new products, additional (high-resolution) data are added and developers start to exploit the web-API we can expect to see usage statistics rise steadily. In preparation for this anticipated increase, we have built in scalability so that new servers can be easily added.

CEDA has an ongoing commitment to provide access to the UKCP data sets via the catalogue, the user interface and the web-API. Additionally, we will continue to add new data sets to the projections as required by the Met Office.

1.10. DATA MIGRATION

Fatima Chami, Matt Pritchard, Matt Jones and Sam Pepler

JASMIN continues to grow as a unique collaborative analysis environment for an expanding community of scientists. The ever-growing demand for storage space and the increasing diversity of scientific workflows were the drivers to start moving JASMIN towards a tiered storage approach. With four types of storage of varying capacity, cost and performance characteristics to be deployed, a data migration plan was established. A total of 6.2 PB of Group Workspace (GWS) and archive data were migrated this year to newer hardware storage, in this case QuoByte Scale-out-Filesystem (SOF).

Data migration plan

With the JASMIN2 PanFS storage reaching end of life, the CEDA Archive and GWS residing on it required migrating to newer hardware. The archive and around 70 GWSs were migrated to new QuoByte storage, a total of approximately 6.2PB, of which 5 PB GWS, and 1.2 PB archive data. The number of files in each GWS varied from hundreds of files, to tens of million files, and the volume varied from a few TB up to 275TB as shown in Fig. 13. Around 30 LOTUS compute nodes were dedicated to the migration operation to enable it to progress in parallel.

Figure 13: Graph shows the size (in TB) of group wokspaces migrated to the new QuoByte storage (Background image of the new JASMIN hardware)

Data from GWSs was migrated over a period of time of days to weeks depending on the size of the GWS volume and the resources available.

The migration plan for archive data was different to that for GWS data because the archive has a regular and controlled structure made up of filesets and volumes. In some cases these were rearranged during the migration in order to improve the archive structure, GWSs however contain data of a more "freeform" structure created by distinct communities of users to suit their own project needs, so as a result tend to be much harder to manage when migrations are required. An algorithm was used to create migration sub-tasks with evenly distributed numbers of files and overall size, to spread the migration task evenly across the compute nodes used for doing the data transfer.

Outcome & future plans

In providing storage services to its users, the JASMIN team has the task of trying to get the maximum useful life out of expensive hardware assets while minimising maintenance costs during the hardware lifetime and trying to make any changes to the system in as smooth a way as possible. As the old hardware in this case reached the end of its life, there was some urgency to move data off it before maintenance costs increased, however the new destination storage for the migration was not available until much later than originally anticipated (October/November rather than May). Hence the time available for the migration of 6.6 PB to new SOF storage was compressed to some 3-5 months. This put considerable pressure on the team and processes involved, at a time when SOF storage was still "bedding in" to operational use at scale within JASMIN, from both the operations team and user perspectives. It highlighted the need for dedicated small-file storage for groups to use for their software environments alongside their bulk data storage, since in its currently-deployed form, SOF storage does not host these efficiently. A solution has now been put in place using new SSD storage purchased specifically for this purpose, and has been rolled out successfully with initial adopters. This is part of the evolving tiered-storage ecosystem which all who use JASMIN will need to embrace: selecting appropriate storage types for different components and stages of scientific workflows, to enable each of those component storage types to be exploited to its individual strengths.

1.11. ESIWACE: STORING AND MOVING LARGE CLIMATE DATASETS

Neil Massey

ESiWACE (https://www.esiwace.eu/) is the Centre of Excellence in Simulation of Weather and Climate and is funded by the European Commission Horizon 2020 scheme to substantially improve efficiency and productivity of numerical weather and climate simulation and prepare them for future exascale systems. ESiWACE is a highly collaborative research programme with partners at 16 research institutions, including CEDA, and commercial companies across seven European Union member states.

In 2019, funding for a continuation, ESiWACE-2, was secured from the European Commission. This will extend the project for another four years, with CEDA once again being a beneficiary of the funding.

There are two major projects for ESiWACE that CEDA are working on. The first is the Joint Data Migration Application (JDMA) - which aims to provide a unified interface for users accessing data storage systems at facilities such as JASMIN. Storage systems have become increasingly heterogeneous as new and improved hardware slowly replaces old storage. For example, JASMIN Phase 4 introduced a new type of storage – object stores, in addition to the existing tape storage system and POSIX file system. Each of these three storage systems have different interfaces for users to interact with, and different latencies for the migration and retrieval of data. CEDA's work on the JDMA project offers a unified interface to the user to enable them to store and retrieve data on all three different data storage types.

Figure 14: European Commission ESiWACE project

The second project is S3netCDF. This aims to exploit the parallelism in transfer of object stores by splitting netCDF files up into smaller subarray files, and then distributing these subarray files across an object store, tape or POSIX filesystem. The location on object store, disk or tape and the position in the original netCDF array of each of the subarray files is stored in a master array file, in which the CFA conventions (Climate and Forecast Aggregation) devised by David Hassell at NCAS CMS (Computational Modelling Services) are used as a description for the location and position of the subarray. S3netCDF is presented as an extension to the standard netCDF4 Python library, and enables users to read and write S3netCDF files using exactly the same interface, with exactly one line of code change.

After considerable development work by CEDA staff, a test deployment of JDMA was made in March 2019. Beta testing of the JDMA software commenced in January 2019, with users from NCAS CMS using JDMA to store and retrieve climate model data that are held in their JASMIN group workspace. The data have been stored on both object store and tape. Some bugs have been found and fixed, particularly around the handling of symbolic links in user data, and JDMA will be available for use by all JASMIN users before the end of 2019.

The successful release of JDMA will provide a unified interface for users accessing different data storage systems on JASMIN, allowing improved service for users.

1.12. GATHERING EVIDENCE OF IMPACT FROM CEDA SERVICES

Poppy Townsend

CEDA increasingly need to demonstrate how our services enable impactful science that benefits wider society. This is fast becoming an essential requirement from our funders - and without continued funding CEDA services would cease to exist. Most of the time we do not know what impactful science CEDA has enabled - we just rely on users telling us about it. Previously we had no formal way of gathering evidence of impact information from the user community. However as the requirement to gather impact evidence has increased, it became clear that we needed to identify a better process. This was achieved by a Masters dissertation with UWE Bristol by a member of CEDA staff. The results identified ways in which CEDA could collect impact information in the most efficient way for us and the user community.

Poppy Townsend, CEDA communications manager, undertook an MSc dissertation in 2018, with UWE Bristol, to investigate how CEDA could collect impact stories in the most efficient way. This consisted of a user survey (n = 520) and focus groups (n = 26). The results suggest that whilst there was a high degree of willingness to provide impact information to CEDA there remains confusion around what 'impact' is. Users are keen to share impact in ways which utilise existing processes (don't invent new ways to collect the same information), and at times which make sense to both the research and the impact, whilst also understanding the need and purpose for sharing that information. You can read Poppy's MSc project report here and a research paper is currently in preparation.

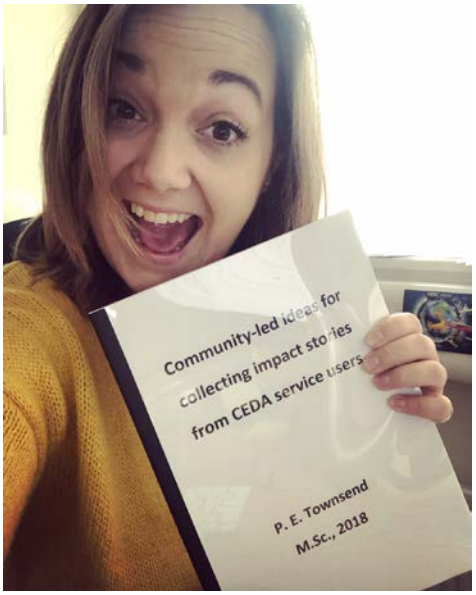


Figure 15: Poppy Townsend with Masters Dissertation, ready for submission

Since the completion of the MSc project (figure 15), CEDA have trialled a yearly collection process, based on the findings from the MSc, starting with a collection period between January and March. The first survey gathering impact information from JASMIN users was completed in January 2019, with over 50 submissions. The next steps in this project will be to start writing 'one-pagers' (or short case studies) based on any appropriate survey submissions and share these via our various communications methods (such as our annual report, the CEDA website and social media). We will also review how the collection process went and adapt where necessary for future years.

1.13. GRADUATE PLACEMENT: PROVIDING ANALYSIS ENVIRONMENTS AND VISUALISATION FOR DIAMOND DATA

Richard Smith

At STFC, graduates get the opportunity to spend three months on placement with a different department. This is to get experience in another area, learn new skills and share knowledge with other departments. I chose to go to the Visualisation Group in SCD (Scientific Computing Department) where I had two tasks: create a visualisation tool and a data analysis environment. The goal was to create somewhere scientists could view and perform initial analysis on DLS (Diamond Light Source) datasets without having to copy the data to their own PC or install any software. Prior to starting my placement, a student contacted DLS wanting to analyse a 100TB dataset but there was no easy way to access that data.

To create the visualiser, I took an existing visualiser built using PyQt and VTK; fixed bugs and extended it (figure 16). The finished application enabled the user to scroll through the slices in the dataset, render a 3D model and perform simple segmentation.

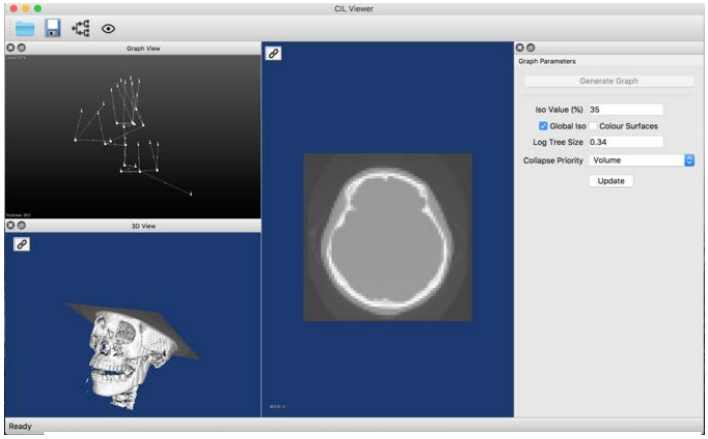


Figure 16: Visualisation tool built with PyQt and VTK

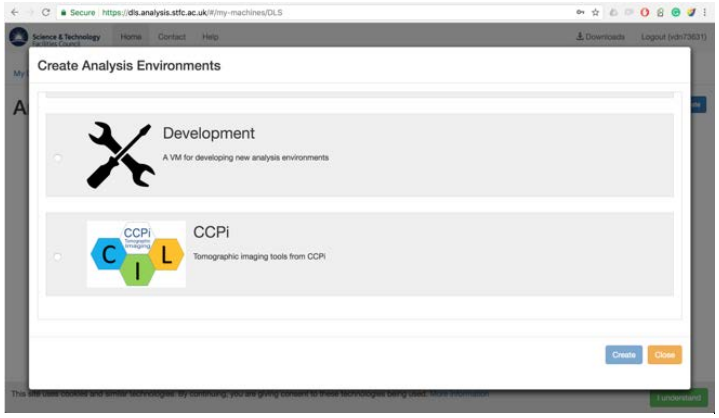


Figure 17: Screenshot of the web platform, created by the DAaaS project, to create analysis environments

The analysis environment built on work done by SCD into Data Analysis as a Service (DAaaS). Using my knowledge of Ansible, I created a script which could be plugged into the existing infrastructure. This launched a virtual machine with all the tools needed to get started with data analysis; including command line access to a batch compute cluster. DAaaS enabled users to login using their normal STFC credentials and create an analysis environment (figure 17). They could then access the desktop of this machine in their browser, removing the need for technical knowledge or extra software.

The main challenge I faced was access to the data. Existing systems for retrieving datasets were slow and not scalable to the volumes required to explore the data (e.g. the use case asked for 100TB).

The outcomes:

- A repeatable desktop environment with Conda and relevant software pre-installed
- Browser access
- Command-line access to a batch compute cluster
- Contributed to the Ansible modules built by SCD

I felt the placement was useful because it allowed me to learn new skills using pyQT, conda packaging and learning how to read documentation written for other languages. I was also able to apply some skills from my normal work such as using Ansible to create repeatable deployments. This was something that my supervisor had not used before. My placement also allowed me to see some of the other scientific challenges which are being solved by staff at STFC.

1.14. CEDA STAFF DELIVER TRAINING FOR NCEO SCIENTISTS

Poppy Townsend

In March 2019, CEDA provided a two day Researchers Forum on behalf of the National Centre for Earth Observation (NCEO). 26 delegates attended the event which covered the theme of preparing data for the CEDA Archive and using JASMIN to process data. The delegates were predominantly early careers researchers studying for PhDs in fields related to Earth Observation. The training was provided by CEDA staff and included presentations, a practical workshop and tours of JASMIN and RAL Space facilities.

Day one started with an overview about CEDA presented by Victoria Bennett (Head of CEDA), followed by two minute talks by the delegates about their science and data challenges. CEDA staff then ran two sessions covering the topics; 'Preparing your data for the CEDA Archive (Wendy Garland and Ed Williamson)' and 'Getting started with JASMIN (Fatima Chami)'. The JASMIN session was split into two talks depending on attendees experience of using JASMIN, this enabled us to appropriately tailor the talks for the audience. Next up were tours around JASMIN and other RAL Space facilities; many delegates described these tours as their favourite part of the forum. To wrap up the first day, a panel/Q&A session was held where delegates could discuss with the CEDA team how to make better use of CEDA services.

Day two was opened by Helen Brindley (NCEO Divisional Director) where she gave an update about current NCEO science. This was then followed by a more hands-on style workshop session where delegates were given a warm up task created by CEDA staff (Poppy Townsend, Sam Pepler, Ed Williamson). This consisted of a 'Top Trumps' style game as delegates were given a piece of paper each with a 'dataset' on it with different attributes. They then had to order themselves in a line from most 'valuable' to least 'valuable' - this involved lots of discussion and arguments amongst the group (which was the point of the game!). The warm up task was created in order to show that the aim of data management is working out what you're supposed to do with it - and that this isn't always easy!

Figure 18: Group photo of attendees at the NCEO forum held at CEDA's home institute STFC, in Oxfordshire

The delegates were then given some time to prepare a short group presentation focussing on a specific dataset and planning how they would: produce a data management plan, make the data more accessible to users, publish and market the data, think about impact and application. The group presentations wrapped up the end of a successful couple of days - and left the CEDA team with several new ideas to think about! CEDA presentation content delivered during the forum can be found here. Figure 18 shows the workshop attendees and CEDA staff involved.

Events like the NCEO forum, are a key part of CEDA's role in supporting the environmental science community. Training and engaging with the next generation of environmental scientists allows us to share best practices, assess the needs of the community and identify what areas we need to improve upon or develop further in the future.

1.15. PV2018 CONFERENCE

Esther Conway

The Centre for Environmental Data Analysis hosted the PV2018 conference in May 2018 on behalf of the UK Space Agency, and jointly organised by STFC, NCEO and the Satellite Applications Catapult. For its ninth edition, the conference series moved to the Rutherford Appleton Laboratory, part of the Harwell Space Cluster in the UK. The conference addressed prospects in the domain of data preservation, stewardship and value adding of scientific data and research related information. This was a highly successful event which saw attendance grow to over 200 participants from Europe, Asia, Africa and America (figure 19). Participants attended workshops, tours and special events in addition to the main conference.

CEDA's Esther Conway was chair of the scientific programme committee, conference co-ordinator and editor of the PV2018 special edition in the Data Science Journal. In addition, a number of CEDA staff supported in the lead up to, and during the event: chairing sessions, welcoming visitors and preparing the abstract booklet.

We were extremely proud to have welcomed a broad range of delegates from around the world including representatives from NASA, University of Botswana, Chinese Meteorological Agency, NOAA, ESA, CNES, DLR and Industry. A total of 77 abstracts were submitted for presentation at the conference. Following the peer review process by members of the conference programme committee, 48 papers were selected for oral presentation. They were complemented by 22 poster presentations (for a total of 310 distinct co-authors with affiliations distributed over different countries from all continents).

Figure 19: Attendees at PV2018, at STFC Rutherford Appleton Laboratory

The conference presentations and proceedings are now available here and consist of a collection of 43 short papers and 20 abstracts corresponding to the oral and poster presentations delivered at the conference. They are organized in sections according to conference sessions followed by the contributions that were presented during the poster session. We also anticipate the publication of a PV2018 special edition of the Data Science Journal in 2019.

Further to the oral and poster contributions, the conference was fortunate to receive keynote lectures and discussion panel addressing a variety of data science topics of interest to PV2018

1. The work of the CEOS WGISS group by Mirko Albani (ESA ESRIN)
2. The EVER-EST project by Rosemarie Leone (ESA ESRIN)
3. Data value and curation across countries, across domains discussion lead by Katrin Molch (DLR)
4. The European Open Science Cloud by Juan Bicarregui (STFC)
5. Copernicus C3S planned service by Carlo Buontempo (ECMWF)
6. Open science data management by Rachel Bruce (JISC)

Feedback from conference attendees was extremely positive. Organising the conference and its complex logistics was challenging at times, but a highly successful way to raise the visibility of CEDA in the UK and international data preservation community.

PART 2: CEDA AND JASMIN SUMMARY INFORMATION 2018-2019

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 2018 to March 2019	
Total number of users	15,338
Total data downloaded	841.6 TB
Total number of accesses	13,573,433
Total days activity	104,391

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

These figures can be broken down by month showing that usage in general has remained fairly stable during the year (table 2.2 and figure 20). It is important to note that, with the advent of 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	Methods	Datasets	No. of accesses	Size	Activity days
2018/04	1,911	11	1,193	1,275,568	61.78	9,507
2018/05	1,877	11	1,151	1,547,141	66.54	10,074
2018/06	1,722	11	1,106	778,618	108.76	7,785
2018/07	1,751	11	1,208	674,695	94.49	8,163
2018/08	1,633	11	1,133	883,588	46.44	7,813
2018/09	1,780	12	1,377	959,507	68.67	10,484
2018/10	1,959	13	1,236	1,462,957	107.82	9,320
2018/11	1,982	14	1,177	1,210,170	50.26	8,036
2018/12	1,807	11	1,122	1,101,235	38.68	7,404
2019/01	1,953	12	1,426	1,528,839	47.51	9,375
2019/02	1,734	12	1,289	603,883	57.3	6,976
2019/03	2,049	12	1,524	1,547,232	93.33	9,454
Totals	**15,338**	**14**	**3,386**	**13,573,433**	**841.57 TB**	**104,391**

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

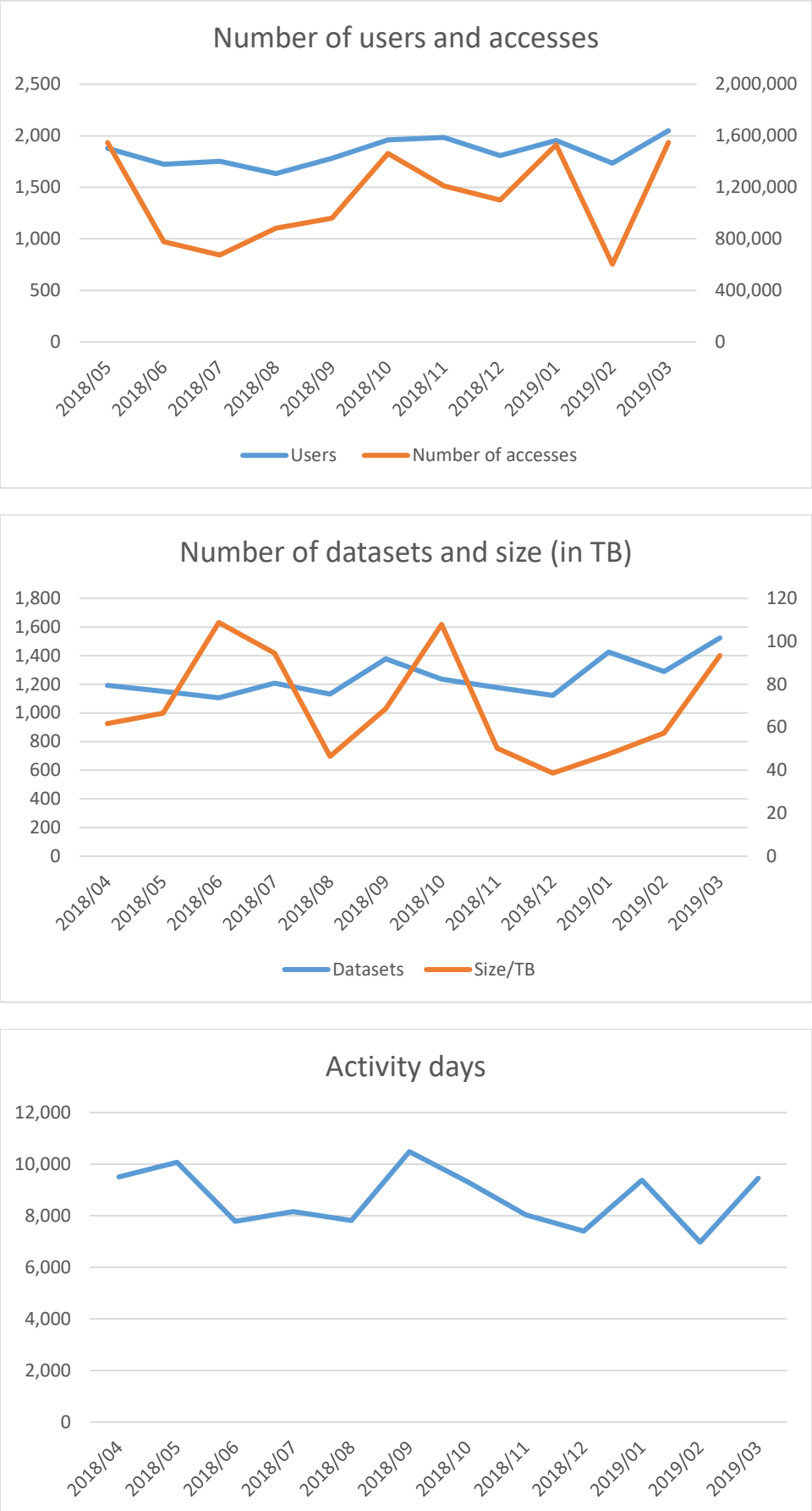


Figure 20: Breakdown of CEDA Archive usage by month

3. JASMIN

JASMIN is a globally-unique data analysis environment, designed, integrated and operated by STFC on behalf of the Natural Environment Research Council (NERC). It provides storage and compute facilities deployed in the combinations needed to enable highly data-intensive environmental science. 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.

During this period, JASMIN received a £1.7M capital investment by NERC to provide JASMIN's Phase 5 upgrade, which successfully refreshed JASMIN's tape storage capability alongside improvements to compute, networking and cloud systems. Meanwhile, JASMIN continued to provide a cutting-edge research environment for NERC-related environmental science.

3.1. STORAGE

The main focus of the Phase 5 upgrade was to procure a new tape library and associated media (figure 21) to maintain and increase capacity for near-line storage, which is expected to maintain importance among JASMIN's range of storage types. This is now in operational use for JASMIN's Elastic Tape service, and for the StorageD service used for both backup and Near-Line Archive (NLA) functions of the CEDA Archive.

Meanwhile, following the roll-out of QuoByte Scale-out-Filesystem (SOF) storage, of which 30PB was purchased in Phase 4 (2017-2018), archive and group workspace data volumes were migrated from older hardware to new SOF volumes, enabling the decommissioning of Phase 2 hardware: resulting in the physical removal of some 7 tonnes of equipment after many years of service.

These migrations, and the adoption of the new storage in user workflows, were not without problems: these and other issues resulted in some user disruption. The SOF storage initially suffered from problems associated with inadvertent parallel-write operations, which it does not support, but subsequent updates to both client and server software have since improved the ability of the system to "fail gracefully". Some adaptation of user workflows to the current range of storage types (in particular, their available quantities and suitability for particular tasks) is required and is the focus of new documentation and training materials developed by the JASMIN team.

The Caringo high-performance object store (HPOS) was also brought into operation, providing a new type of storage which will play a large part in future workflows. Being considerably cheaper than traditional (POSIX) file system storage, it offers the possibility of significantly improved data access into and out of JASMIN cloud tenancies, so will feature prominently in JASMIN's storage mix in the future. A smaller deployment of QuoByte (which can function as either POSIX filesystem or HPOS), initially used as a HPOS proof-of-concept, was re-tasked as an additional 1.2 PB of SOF storage.

Development work commenced on a "Joint Data Migration App" (JDMA), conceived as a tool to move data between a variety of storage backends, including disk, object store and (elastic) tape. A new "Transfer Cache" (XFC) service was also developed and introduced: this enables users to self-provision large but temporary areas of disk space for initial transfer and data manipulation tasks.

At the start of this reporting year there were 164 Group Workspaces with allocations totalling 9.5 PB, serving a total community of 1815 users. By the end of the period this had increased to 202 group workspaces in use with allocations totalling 12.7 PB, serving a total community of 2517 users.

Figure 21: The Spectra Tape library which was procured and installed as part of the JASMIN Phase 5 upgrade.

3.2. MANAGED COMPUTE INFRASTRUCTURE

The deployment of new hardware from the Phase 4 procurement added significant capacity to the LOTUS batch processing cluster, addressing the utilization level which was regularly topping 70% before the upgrade. In November 2018 this had settled to 62% following the introduction of the new nodes, despite over 3.7M jobs from 168 distinct users of LOTUS during that month, continuing the upward trend of usage.

Two further scientific analysis servers were deployed as virtual machines, alongside an additional higher-memory physical machine to add to the pool of “sci” servers for interactive compute tasks. Concern about the resource usage and stability of these machines has highlighted the need to move to a model of using managed tenancies in the JASMIN community cloud to enable communities to deploy their own sci machines, with their own manager responsible for controlling access, user behavior and resource usage. It is planned to roll out this model in the following period.

The "JASMIN Analysis Platform" (JAP) software stack, deployed as common software across the managed compute infrastructure, continued to be maintained but the main focus of work turned to developing an alternative packaging and deployment method. “JASPY” emerged as the candidate Conda-based environment for relevant Python packages, with a plan to replace the JAP at the same time as moving the managed compute estate from the RedHat Enterprise Linux operating system to CentOS7 later in 2019.

3.3. NETWORK AND SUPPORTING INFRASTRUCTURE

JASMIN’s “Science DMZ” or “Data Transfer Zone” continued to provide the preferred route for high-performance data transfers for cases where large volumes of data need to be moved efficiently. Several key data movement

services use this zone, and increases in traffic reached a level soon after this period which required bringing forward plans to upgrade its link to the recently-upgraded RAL site connection to the outside world. In addition to providing services for individual JASMIN users, work continued in collaboration with international partners in the ESGF infrastructure, helping to prepare CEDA for participation in large-scale institutional data transfers as part of the anticipated CMIP6 data replication activity.

Work associated with the Phase 5 upgrade included adding a local fibre connection across the Harwell Campus between JASMIN's new "super-spine" internal network and an additional data centre earmarked for possible future expansion.

3.4. JASMIN CLOUD

The JASMIN cloud portal continued to provide tenants of JASMIN's community cloud – both managed and external tenancies – with a web-based interface enabling them to deploy and manage virtual machines at will within their tenancies.

Development work provided the basis for "cluster as a service", which will enable tenant projects to deploy transient batch processing capability within their own tenancies. Work continues with this and other projects including a containerized Notebook service.

3.5. USER SUPPORT AND OUTREACH

The third JASMIN user conference was held in June 2018, hosted over two days. The event aimed to bring together members of JASMIN's user community with opportunities to learn about recent developments of the infrastructure and service, to share experiences of using JASMIN with other users and with the JASMIN team. A day of plenary sessions at a local venue with presentations from the JASMIN team and current projects was followed on the second day, at RAL Space, with training workshops about key parts of the system. As with previous events, tours of the JASMIN infrastructure laid on by the Scientific Computing Department proved very popular with attendees. The event was also used to collect information from users and projects about "impact stories": how using JASMIN had helped their projects achieve their scientific goals and the benefits brought about by the outcomes of those projects.

Following a successful first webinar "Getting Started with JASMIN" held in March 2018, a further session "Introduction to LOTUS" was held in October 2018, with further events planned in due course, among other training provided by the CEDA and JASMIN teams.

Work continues on improving and updating user and internal documentation and training materials as the JASMIN system evolves and in response to user queries and feedback.

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 2018/2019, 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	11-12	12-13	13-14	14-15	15-16	16-17	17-18	18-19
NCAS income	883	935	829	829	808	808	808	808
NCEO income	419	445	392	390	393	393	393	402
Other NERC	527	287	272	600	621	825	816	733
Other income	1099	1283	1486	1394	1505	1092	1280	1458
Total income	2928	2950	2979	3213	3327	3118	3297	3401

Table 5.1: Overall funding for CEDA for financial years 2011 — 2012 to 2018 — 2019 (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 2018-2019

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)
ESA CCI Data Portal Extension	Archive and data services for ESA CCI programme	ESA	17/12/2017	31/12/2018	100.0
EUSTACE	EU Surface Temperature for All Corners of Earth	H2020	01/01/2015	30/06/2018	142.9
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
BACI	JASMIN support for BACI (Biosphere Atmosphere Change Index)	H2020	01/04/2015	31/10/2018	25.7

PRIMAVERA	JASMIN and archive support for new generation of global climate models	H2020	1/1/2015	31/10/2019	141.0
ESA Sentinel Data Hub Relay	STFC as part of network of ESA relay hubs for Sentinel data	ESA	1/5/2015	31/12/2018	311.2
UKCP09 User Interface 18-19	UK Climate Projections platform, user interface development	EA	01/04/2018	31/03/2019	19.8
CDS Zone	Support for ECV processing	UKSA/NCEO	01/08/2016	15/03/2019	400.0
C3S ESGF Data Node (CP4CDS)	Operational ESGF data node for C3S	C3S	1/9/2016	14/12/2019	898.3
Pest Risk Modelling in Africa (PRISE)	JASMIN support for UKSA IPP project	UKSA	1/12/2016	31/12/2019	104.8
C3S ECVs production SST	JASMIN and archive support for C3S SST ECV processing	C3S	01/02/2017	31/10/2018	63.9
BEIS IPCC DDC	UK component of IPCC Data Distribution Centre	BEIS	26/9/2018	31/03/2020	148.5
C3S CORDEX4CDS	Regional Climate Projection data for C3S	C3S	01/05/2017	30/04/2020	141.2
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 18-19	To provide data services to support access to the next generation of climate projections for the UK (2018)	Met Office/Defra	02/05/2018	31/03/2019	140.0
Support for Ensembles 18-19	CEDA support to multi-model climate ensemble archives	Met Office	01/01/2018	31/03/2019	50.0
C3S GLAMOD	In-situ observations for Copernicus	C3S	01/07/2017	31/03/2019	95.0
C3S GLAMOD – Phase 2	In-situ observations for Copernicus	C3S	01/01/2019	28/02/2021	101.9

UKSA funding for LTDP WG and PV conference	Funding Esther Conway to attend ESA LTDP WG and organise PV conference in UK	UKSA	01/04/2018	31/03/2019	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
Copernicus Data Support SLA	Acquire and archive Copernicus Sentinel data from all satellites	UKSA	01/01/2019	31/07/2019	169.3
ESA EPs Common Architecture	Consultancy for ESA Exploitation Platforms Common Architecture	ESA	06/11/2018	31/10/2020	85.2
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 2019) is:

Bryan Lawrence, NCAS (Chair)
Sarah Callaghan, 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

The board aims to meet at least annually, and up to semi-annually when necessary. This year, the CEDA board met on the 15th June 2018 and 5th of February 2019; the next board is planned in February 2020.

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	2431	62.3%	2383	63.8%	2365	64.3%	2311	65.1%
Europe	446	11.4%	422	11.3%	415	11.3%	404	11.4%
Rest of the world	945	24.2%	861	23.0%	829	22.5%	767	21.6%
Unknown	79	2.0%	72	1.9%	70	1.9%	67	1.9%

Table 6.1: Users by area

Institute Type	Q1		Q2		Q3		Q4	
University	2840	72.8%	2723	72.8%	2702	73.4%	2625	74.0%
Government	578	14.8%	549	14.7%	530	14.4%	501	14.1%
NERC	170	4.4%	169	4.5%	158	4.3%	152	4.3%
Other	226	5.8%	213	5.7%	205	5.6%	191	5.4%
Commercial	44	1.1%	48	1.3%	50	1.4%	44	1.2%
School	41	1.0%	34	0.9%	32	0.9%	34	1.0%

Table 6.2: Users by Institute type

6.2. DATA HOLDINGS

Data Centre	Q1	Q2	Q3	Q4
CEDA	4459	4450	4450	4450

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

	Q1	Q2	Q3	Q4
Datasets	5501	5571	5620	5709
Collections	599	607	612	625

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	611	691	692	776
Closed	576	703	642	771

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	403	448	501	497
Closed	407	441	497	511

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

Bennett, V., P.Kershaw, R.Smith and B.Lawrence, 'JASMIN: Managing variety in a climate data community platform', Proceedings of the 2019 conference on Big Data from Space (BiDS '19). 19-21 February 2019, Munich, Germany. ISBN 978-92-76-00034-1 ISSN 1831-9424 doi: 10.2760/848593

Bennett, V., P. Kershaw, R. Smith and B. Lawrence, 'JASMIN: Managing variety in a climate data community platform', Proceedings of the 2019 conference on Big Data from Space (BiDS '19). 19-21 February 2019, Munich, Germany. ISBN 978-92-76-00034-1 ISSN 1831-9424 doi: 10.2760/848593

Bennett, Victoria, Fay Done, Clive Farquhar, Kevin Halsall, Phillip Kershaw, Alison Waterfall, Antony Wilson, Alex Wood, 'The ESA CCI Open Data Portal', Poster, UK National Earth Observation Conference, Birmingham, 4-7 September 2018, (http://www.rspsoc.org.uk/images/ukneoc2018/Abstract_Booklet_20180829.pdf)

Chami, F., 'Introduction to JASMIN', Oral presentation, NCEO Research Forum, Harwell March 2019.

Chami, F, 'Introduction to LOTUS: JASMIN's batch computing', webinar https://www.ceda.ac.uk/events/introduction-to-lotus-jasmins-batch-computing-cluster-webinar/, October 2018

Chami, F., 'Interactive computing on JASMIN', Oral presentation, JASMIN2018 User Conference, Harwell, June 2018 http://www.jasmin.ac.uk/jasmin2018/

Conway, E., 'Proceedings of the 2018 conference on adding value and preserving data', 2018, https://epubs.stfc.ac.uk/work/37981055

Donegan, Stephen, Edward Williamson, Victoria Bennett, 'CEDA and JASMIN Services', Poster, UK National Earth Observation Conference, Birmingham, 4-7 September 2018, (http://www.rspsoc.org.uk/images/ukneoc2018/Abstract_Booklet_20180829.pdf)

Garland, Wendy, Ag Stephens, Richard Smith, 'EUFAR Flight Finder & CEDA Satellite DataFinder'; Poster, PV2018 Conference, RAL, 15-17 May 2018

Garland, W., R. Smith, A. Stephens, 'EUFAR Flight Finder & Satellite Data Finder', Poster, NCAS Staff Meeting 3-4 July 2018

Juckes, M., J.A. Pamment, C. Pascoe, A. Stephens, 'Building an Infrastructure for Climate Model Archives' pp 143 - 146, Proceedings of PV2018 Conference, RAL, 15-17 May 2018.

Juckes, Martin N, 'The EU Copernicus Programme and a New Generation of Data Services: a Perspective from the Centre for Environmental Data Analysis'. In: Marine-Earth Informatics 2018, July 12th, 2018, Tokyo. http://cedadocs.ceda.ac.uk/1381/

Juckes, Martin N and Taylor, Karl and Mizielinski, Matthew and Pamment, Alison and Denvil, Sebastien and Senesi, Stephane (2018) 'Status and Outlook for the CMIP Data Request'. In: ESGF Face-to-Face 2018, December 4-7, 2018, Washington D.C., USA. http://cedadocs.ceda.ac.uk/1382/

Juckes, M., Taylor, K. E., Durack, P., Lawrence, B., Mizielinski, M., Pamment, A., Peterschmitt, J.-Y., Rixen, M., and Sénésis, S.: 'The CMIP6 Data Request (version 01.00.31)', Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2019-219, in review, 2019.

Juckes, Martin N. 'CMIP6 Data Request: Status and Outlook'. In: WGCM22, March 26-29, 2019, Barcelona. http://cedadocs.ceda.ac.uk/id/eprint/1385

Kershaw, Philip, Victoria Bennett, Steve Donegan, Alan Iwi, Martin Juckes, Ruth Petrie, Joe Singleton, Ag Stephens, Alison Waterfall and Antony Wilson, 'Developing an Open Data Portal for the ESA Climate Change Initiative', Presentation, PV2018 Conference, RAL, 15-17 May 2018

Kershaw, Philip, Neil Massey, Matt Pritchard, Bryan Lawrence, Jonathan Churchill, Athanasios Kanaris, Cristina Del Cano Novales and Matt Pryor, JASMIN, 'Providing a platform to support Data Analysis and the Creation of Virtual Research Environments', Presentation, PV2018 Conference, RAL, 15-17 May 2018

Kershaw, Philip; Bennett, Victoria; Stephens, Ag; Juckes, Martin; Waterfall, Alison; Wilson, Antony John; Petrie, Ruth E; Singleton, Joe; Donegan, Steve; Iwi, Alan, 'The application of ESGF to develop an Open Data Portal for ESA Climate Change Initiative', Presentation, Earth System Grid Federation Face-to-Face Meeting, Washington DC, 2018

Kershaw, Philip; 'New Capabilities and the evolution of services on JASMIN', Presentation, JASMIN Conference, Abingdon, June 2018

Kershaw, Philip, Victoria Bennett, Jonathan Churchill, Bryan Lawrence, Neil Massey, Matt Pryor, Matt Pritchard, 'JASMIN a platform for the development of downstream EO applications and services', Presentation, UK National Earth Observation Conference, Birmingham, 4-7 September 2018, (http://www.rspsoc.org.uk/images/ukneoc2018/Abstract_Booklet_20180829.pdf)

Kershaw, Philip; Evans, Ben; 'Rethinking the ESGF Architecture', Presentation, Earth System Grid Federation Face-to-Face Meeting, Washington DC, 2018

Kershaw, Philip; 'ESGF Identity, Entitlement and Access Management Working Team Update', Presentation, Earth System Grid Federation Face-to-Face Meeting, Washington DC, 2018

Massey, N., D. Hassell, M. Jones and B. Lawrence, 'Semantic storage of climate data on object stores', Poster, NCAS Staff Meeting 3-4 July 2018

Massey, Neil, David Hassell, Matt Jones, Philip Kershaw and Bryan Lawrence, 'Semantic storage of climate data on object stores', Presentation, EGU2018, 8-13 April 2018

Pamment, A., D. Hassell, A. Heaps, A. Iwi, S. Pepler, C. Roberts, A. Stephens, D. Walker, 'Essential IT Skills, Training for Environmental Scientists', Poster, NCAS Staff Meeting 3-4 July 2018

Parton, G., A. Stephens, J. Singleton, W. Tucker, T. Cairnes, 'Bringing NCAS data to the masses: CEDA Catalogue developments and service integrations in 2017', Poster, NCAS Staff Meeting 3-4 July 2018 http://cedadocs.ceda.ac.uk/1444/

Pascoe, Charlotte, The Data Distribution Centre of the Intergovernmental Panel on Climate Change, Presentation. PV2018 Conference, RAL, 15-17 May 2018

Pepler, Sam, Using JASMIN as the Storage infrastructure for the CEDA Archive, Presentation. JASMIN conference, Abingdon, June 2018 (http://www.jasmin.ac.uk/jasmin2018/)

Pepler, S., M. Juckes, S. Callaghan, 'Joining up the NERC Data Centres', Poster, NCAS Staff Meeting 3-4 July 2018

Pepler, Sam, Bit preservation processes in the Centre for Environmental Data Analysis Archive, Poster, PV 2018, RAL, 15-17 May 2018

Petrie, Ruth, 'Global and regional climate projections for the Copernicus Climate Data Store', Poster, C3S General Assembly, Berlin, 24-28 September 2018

Pritchard, Matt, ‘JASMIN Storage improvements as of Phase 4’, Presentation, JASMIN Conference, Abingdon, June 2018 (http://www.jasmin.ac.uk/jasmin2018/)

Townsend, P., ‘Weighing up big data’, Poster, NCAS Staff Meeting 3-4 July 2018

Watson-Parris, Duncan, Nick Schutgens, Zak Kipling, Phil Kershaw, Bryan Lawrence, and Philip Stier, ‘The Community Intercomparison Suite (CIS): an open-source toolbox’, Poster, EGU2018, Vienna, 8-13 April 2018