Centre for Environmental Data Analysis (CEDA)

Annual Report 2017

(April 2016 to March 2017)

Victoria Bennett, Sarah Callaghan,
Poppy Townsend (Editors)

CONTENTS

1. INTRODUCTION

The Centre for Environmental Analysis (CEDA) is based in the Science and Technology Research Council (STFC)'s RAL Space department. CEDA operates data centres and carries out related project work, primarily for the Natural Environment Research Council (NERC). CEDA's mission is to deliver long term curation of scientifically important environmental data at the same time as facilitating the use of data by the environmental science community.

CEDA was established in 2005, when the activities of two of the Natural Environment Research Council (NERC) designated data centres were merged: the British Atmospheric Data Centre, and the NERC Earth Observation Data Centre. Consolidated annual reports have been produced since 2009. This annual report presents key statistics for the year past (2016- 2017) as well as a series of snapshots of activity, expressed as short highlights and short reports. Key metrics are also provided.

JASMIN is the data intensive supercomputer which provides the infrastructure upon which CEDA and the CEDA services are delivered. Usage of JASMIN in the community continues to grow and the role of CEDA staff continues to evolve to include support for "big data" tools and algorithms. With uncertainty around the next large-scale capital investment, interim funding this year enabled capacity to be maintained, and new technologies to be trialled ahead of JASMIN Phase 4.

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.

Ensuring access for UK researchers to the unprecedented volumes of Earth Observation Data from the European Sentinel satellites continues to be a major focus for CEDA. Similar, if not larger, volumes of climate model data for CMIP6 (Coupled Model Intercomparison Project, Phase 6) will follow soon, and preparations for this are well underway. Other highlights this year include JASMIN's fifth year of successful operations, and CEDA's engagement activities with the public and our users.

Over the years we have reported our key partnerships, and as before, these revolve around our neighbours on the Harwell site (including the Satellite Applications Catapult, with whom we share delivery of the facility for Climate and Environmental Monitoring from Space, CEMS), the European Network for Earth Simulation (with whom we share the delivery of the European component of the Earth System Grid Federation), and many other project collaborators.

Victoria Bennett, Head of CEDA, RAL Space

PART 1: HIGHLIGHTS AND MAJOR ACTIVITIES

This first section provides a selection of descriptions of key activities and highlights from the year. It has two chapters: one with short highlights selected to showcase some CEDA activities supported through different funding streams, and a second describing a range of key areas of focus for CEDA staff this year:

1. HIGHLIGHTS

Highlight topics this year focus on CEDA's climate model data archival work, and preparations for the upcoming flood of data from CMIP6 (Coupled Model Intercomparison Project, phase 6), celebrating 5 successful years of JASMIN operations, and public engagement activities undertaken by CEDA staff.

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

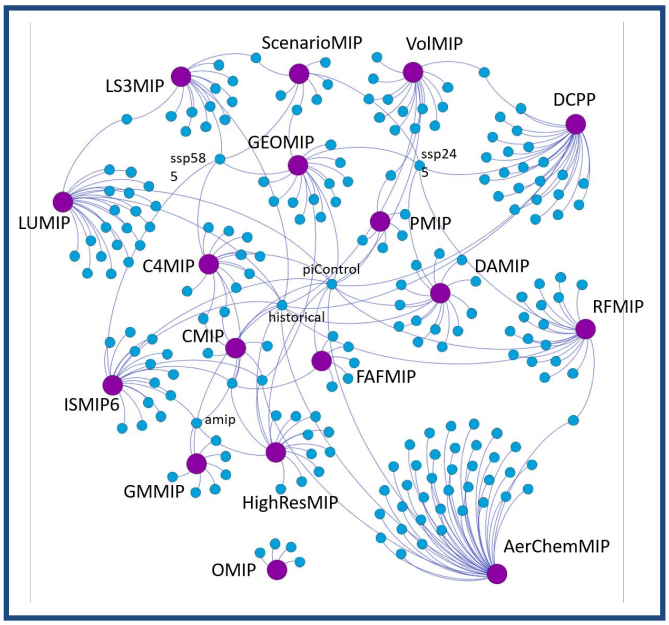
CEDA Highlight

Building an archive of climate model data

Martin Juckes, Alison Pamment, Charlotte Pascoe, Ag Stephens

Alison Pamment Environmental Data Scientist specializing in metadata conventions and development of training courses

The CMIP (Coupled Model Intercomparison Project) has been coordinating climate model experiments involving multiple international modelling teams since 1995. The current sixth phase of the CMIP process (CMIP6) consists of a suite of common experiments, and 21 separate CMIP-Endorsed Model Intercomparison Projects (MIPs) making a total of 244 separate experiments. CEDA is making a major contribution to archiving and providing user access to CMIP6 data, working in collaboration with an international network of data service providers.



Above: A schematic diagram showing the 21 Model Intercomparison Projects (MIPs) in CMIP6. The large purple dots represent the individual MIPs. Each smaller blue dot represents a climate model experiment and the lines show how the experiments contribute to the MIPS – some experiments, such as control runs, are important to several MIPS and therefore have many connections.

CEDA Contributions to CMIP6

CEDA staff are using their expertise in collecting and organising metadata (precise descriptions of data contents) to gather information about the MIPs and the individual model experiments and making this available to users. This work is carried out using technology developed as part of Earth System Documentation Project (ES-DOC). It allows end users of the climate model archive to ask questions such as "Which MIPs make use of experiment A?“

A second major contribution being made by CEDA is the coordination of the CMIP6 Data Request, a compilation of the output data requirements from all the MIPs, into a consolidated technical document. Over 2000 modelled variables have been specified and these in turn are linked to the experiment specifications provided in ES-DOC. Each variable has to be carefully defined and added to a ‘controlled vocabulary’ (glossary) of terms.

Access to the CMIP6 Archives

CEDA is a major contributor to the Earth System Grid Federation (ESGF), a collaboration producing the software that enables most global climate change research, notably assessments by the Intergovernmental Panel on Climate Change (IPCC). Petabytes of high-profile climate simulations can be archived and replicated across the globe and it is this technology that will be used to provide end user access to the CMIP6 data.

This work was funded by Natural Environment Research Council and the European Commission.

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

CEDA Highlight

Happy 5th Birthday JASMIN!

Poppy Townsend

Poppy Townsend
Data Scientist

Five years ago, on 9th March 2012, an initial 4.5 petabytes of high-performance storage was brought online, signalling the 'birth' of JASMIN and revolutionising access to environmental science data for NERC. JASMIN is half super-computer and half data-centre and provides a globally unique platform hosted at STFC, for computing and storing UK and European environmental science data.

A bustling hub of scientific collaboration, JASMIN has nearly 150 NERC-related science projects using its group workspaces and the CEDA archive. These bring together researchers from across the domains of data-intensive environmental science.

Below: The machine room where JASMIN is hosted at STFC

Around 1000 virtual machines and 5000 cores of computing power are woven together with a fast internal network; which efficiently moves petabytes of data around, into, and out of JASMIN every month. Today, JASMIN's storage capacity has tripled to over 16 petabytes - as a comparison, 1 petabyte of average length songs would equate to approximately 2000 years of continuous play!

Already, JASMIN has established its own 'data gravity' - the ability of data to attract additional applications and services. Research projects benefit (and in many cases have only become feasible) from locating their analysis activities alongside the CEDA archive and other resources, so that they can share their work efficiently with colleagues and other projects alike. Meanwhile the JASMIN cloud has evolved to enable projects to showcase their work via outward-facing services which they maintain themselves.

Above: The disks used for JASMIN storage

JASMIN now has over 1200 users from a variety of scientific backgrounds and levels of expertise. Over 100 (mostly doctoral) students register to use JASMIN every year; with around 2000 more registering each year to exploit data via the CEDA download services.

Over the last five years, JASMIN has enabled a vibrant and dynamic UK research community to do new and exciting science. Significant but exciting challenges lie ahead, as the NERC community's demand for JASMIN's data storage and computing continues to grow.

Find out more about this and other CEDA activities our website or Twitter: @cedanews

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

CEDA Highlight

Public Engagement at CEDA

Poppy Townsend

Poppy Townsend
Data Scientist

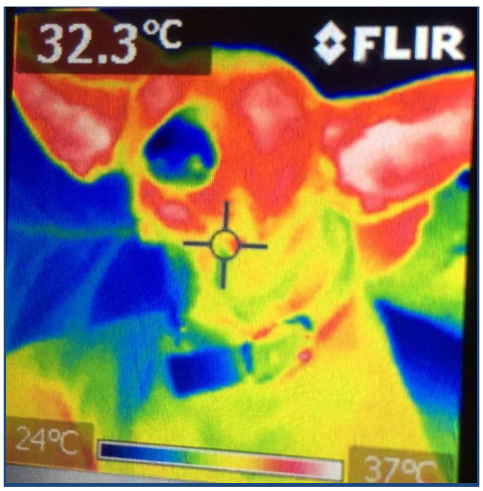
Public engagement is a term used for describing activities and benefits of scientific research to a non-specialist audience. We do this in a variety of ways, such as radio interviews, school visits, or public open days. Here at CEDA we believe it is essential to engage with the wider public and show them the benefits of environmental data and science and how the work that we do is relevant to them.

Some examples of recent activities

- May 2016 - Charlotte Pascoe, a CEDA Senior Data Scientist, was interviewed by Paul Hudson for the Weather Show on BBC local radio for Yorkshire and Lincolnshire. She talked about Soapbox Science, her PhD, potential vorticity, CMIP6 and big data.

- October 2016 - CEDA staff* ran a joint stand with NCEO at the NERC public showcase called 'Into the Blue' in Manchester. They used an Infrared camera to talk about measuring climate change and satellites. (*Poppy Townsend, Ed Williamson, Graham Parton and Fatima Chami)
- November 2016 - The IR camera made yet another appearance, as did some common kitchen ingredients to make exothermic and endothermic reactions for a local guide group. These were provided by Poppy Townsend and RAL Space colleague, Ed Polehampton.

Below: We use an Infrared camera for many activities to talk about sea surface temperature. Here is an example of it being used on my dog!



Many members of the public benefited from these activities including; families, teachers, school children, and radio listeners!

Above: CEDA staff at NERC Into the Blue showing a family some thermochromic paper

What are the plans for the future?

CEDA will continue to provide public engagement activities as we believe it to be beneficial (and enjoyable) for both our staff and members of the public.

If you wish to collaborate with us, or request any public engagement then please get in touch!

Find out more about this and other CEDA activities our website or Twitter: @cedanews

2. SHORT REPORTS

2.1. THE ESA CCI OPEN DATA PORTAL

Alison Waterfall

The CCI Open Data Portal is a project to develop a central portal for the European Space Agency's (ESA) Climate Change Initiative (CCI) programme. The ESA Climate Change Initiative is producing climate quality datasets from historic satellite measurements for many Essential Climate Variables (ECV's), and consists of 14 different project consortia involving many scientists across Europe. More information can be found at www.cci.esa.int.

The Open Data Portal project was initially a 3 year project, starting in 2015, and is led by Telespazio Vega. Other project partners included CGI and University of Reading. CEDA is providing the backend services to the portal, hosting the data within the CEDA archive, and also providing JASMIN tenancies to host the front end and CCI website. The front end of the portal has been developed by CGI, but the information on the data is supplied to them in a real-time manner via CEDA's services. The front end of the portal was officially launched in September 2016, and can be accessed at www.cci.int/data.

A key intention of the portal was to provide a diverse range of data download methods and to open up the data to the widest possible audience. Data is available to be downloaded by anonymous ftp alongside the standard CEDA download services, and suitable data is also available via http, OPeNDAP, a Web Mapping Service (WMS) and Web Coverage Service (WCS). These latter make use of the Earth System Grid Federation, and CEDA expertise from other climate projects.

The portal front end consists of a number of different ways of downloading the data; a dashboard that provides a graphical overview of the CCI products; a search option allowing faceted search, and a visualisation option. Although the front end itself was developed outside of CEDA, it relies on CEDA services to provide the metadata that is used to populate the service in real time, and to host the data itself. The portal harvests the necessary metadata from CEDA's CSW (Catalogue Service for the Web); MOLES records for the data to be displayed in the portal are tagged with an ESACCI keyword, and this is searched for by the portal in real time. This allows us to add, update or remove datasets as needed in an automated way.

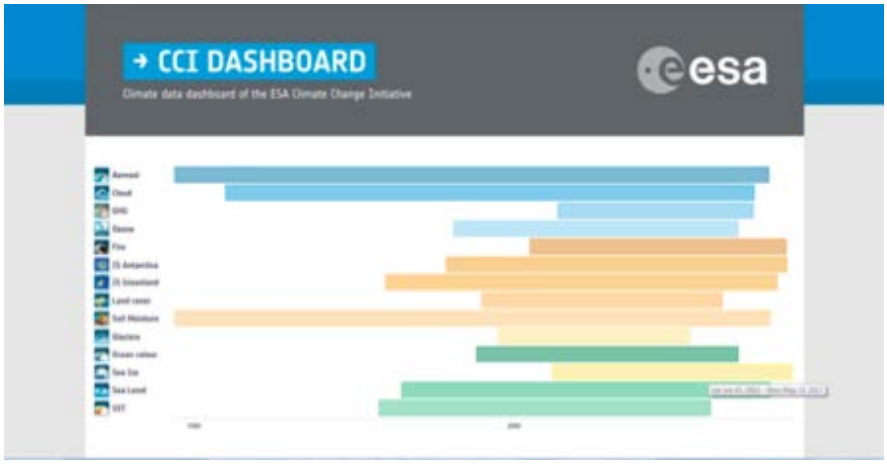


Figure 1: The CCI Open Data Portal dashboard. This provides an overview of the time periods covered by the datasets in the portal. Clicking on each bar allows users to drill down and discover individual datasets and their download links

In order to ensure that data can be searched and accessed by the widest possible audience, a key task in the initial phase was the development of a SKOS representation of the CCI vocabularies. This enabled the datasets to be tagged with vocabulary terms to allow faceted search via the CCI Open Data portal and indexed within ESGF. The CCI portal, and the data archive at CEDA, are key to ensuring wide uptake of the CCI's data products. Usage is increasing as the resource is publicised in new user communities and more data continues to be added.

2.2. SATELLITE DATA HOLDINGS AT CEDA

Steve Donegan, Ed Williamson, Sam Pepler, Richard Smith & Ag Stephens

CEDA provides access to almost 2 petabytes of Earth Observation data. Satellite data from the ESA Sentinel missions, EUMETSAT (METOP) and NASA/USGS (Landsat, MODIS) and aircraft data from NERC, FAAM and EUFAR continue to be retrieved. This data is accessible for easy download or processing on JASMIN by users. CEDA staff continue to improve the performance of the archive to ensure timely retrieval and ease of access to data.

CEDA has retrieved some 1.8 petabytes of ESA Sentinel data during 2016/17 for the CEDA archive. Much of this is from Sentinel 1A and 2A, which been in operation for more than 2 years. In addition, this year Sentinel 1B and Sentinel 3A data have started to flow into the CEDA archive. In order to cope with the ever increasing volume of data from the growing number of Sentinels in space, CEDA has continually updated and adapted the mirror archive data retrieving software. To mitigate against JASMIN disk filling up, new tools are in development to move older data from the JASMIN fast disk storage onto a near line tape archive (NLA). The NLA will allow data to be retrieved temporarily via an interface so users can access this data alongside the most recent data. Developments such as these are vital to allow the ongoing retrieval of the most recent data to the CEDA archive especially considering that further Sentinels are expected in the period 2017/2018.

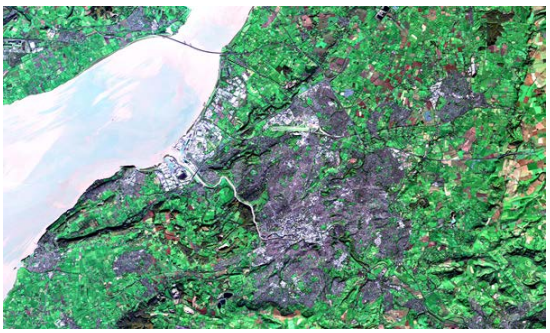


Figure 2: Sentinel 2 image of Bristol contains modified Copernicus data (2017), processed by CEDA

ESA disseminates Sentinel data through a series of data hubs. In order to make data transfers between the various hubs more efficient and to load balance large scale transfers ESA have created a network of relay hubs. CEDA has won a contract to run such a relay for ESA, and work began in 2016 to test the integration of this relay on CEDA servers with the wider network. During the reporting period, CEDA have run the relay as part of a wider test programme.

CEDA recognises that the advent of the Sentinel Missions has brought about a wealth of new opportunities for the exploitation of "big" satellite data. In order to support the requirements of the research community, both public and private sectors, CEDA have started to develop a web-tool for searching the main Earth Observation (EO) datasets held in the archive. The new tool will provide spatial, temporal and keyword searches to allow users to quickly locate the most relevant satellite products to their needs (see Figure 2).

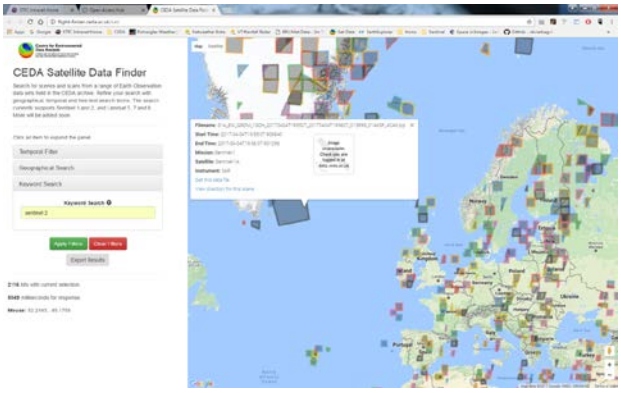


Figure 3: Mock-up of planned Satellite data finder tool.

CEDA have supported projects, such as the ESA Forestry-TEP, with the dedicated retrieval of Sentinel2 and Landsat scenes over specific locations. CEDA also maintains the ESA SNAP software within the JASMIN infrastructure to allow easy processing of Sentinel products by users (see Figure 3). CEDA continues to work closely with the UK EO community and is a partner in the National Centre for Earth Observation (NCEO). CEDA exploits these links to ensure that the data retrieved to the archive matches its users' requirements.

2.3. AUTOMATING THE INGESTION PIPELINE FOR CMIP6

Ag Stephens

The 6th Coupled Model Intercomparison Project (CMIP6) will deliver a multi Petabyte archive of climate projections beyond the scale of anything we have seen so far. The role of CEDA within CMIP6 is to act the UK "node" for the Earth System Grid Federation (ESGF). 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). ESGF provides tools and interfaces for data management, discovery, search, browse, extract, subset and download to a diverse range of sectors interested in climate change and impacts.

The design for CMIP6 is made up of 23 separate sub-intercomparison projects, involving about 40 modelling groups running around 250 experiments. The expected number of data sets to be ingested and published will be in the 10s to 100s of 1000s. As the UK "node" for ESGF, CEDA is responsible for ingesting data from the Met Office, NCAS and other contributors to CMIP6. Table 1 indicates the complexity of the process: there are a number of stages that each need to be run and verified in order to complete a "data set publication".

Table 1. Stages in ingestion/publication of CMIP6.

Stage	Description
1	Message queue
2	Ingestion
3	CEDA catalogue record generation
4	Map file generation
5	ES-DOC record generation
6	Quality Control
7	ESGF publication: DB
8	ESGF publication: TDS (without re-init)
9	ESGF publication: TDS main catalogue / reload
10	ESGF publication: index
11	Consume RabbitMQ Message

In order to automate this multi-stage workflow we are developing the CEDA REceive-to-Publish Pipeline (CREPP) tool. CREPP uses a simple client-server architecture in which a common database is accessible across multiple parts of the CEDA infrastructure so that a range of task-specific "controllers" can carry out their work. For example, the ingestion and catalogue record controllers (stages 2 and 3) will run on the standard CEDA ingest servers and once completed will trigger the next stage. Stages 4 to 10 will run on an ESGF-specific server in order to carry out the relevant publication steps.

This flexible design ensures that existing roles and permissions can be respected whilst allowing the entire workflow to be automated. Figure 4 shows a schematic of how each controller is continually polling the CREPP database to find out if it has more tasks to perform.

CREPP will be an essential part of the operational environment required to manage the UK component of CMIP6 publication. However, it will also be used to manage the international component of "replicated" data that is copied from international nodes across the federation. Timely replication is necessary to give access to UK scientists who will undertake inter-comparisons

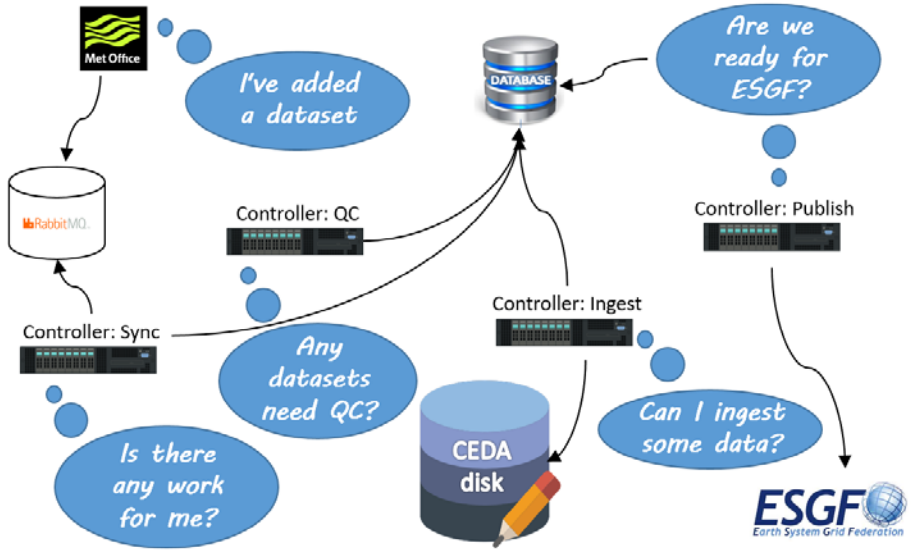


Figure 4: Schematic of CREPP in action: each controller asks the database whether it has a next task to perform

and analyses to contribute to the next IPCC scientific assessment report.

In this context, CEDA and JASMIN have a very important role in supporting global climate science and helping to understand the impact of climate change on our planet.

2.4. CLIMATE DATA FROM SPACE ZONE

Victoria Bennett

Funding was announced by the UK Space Agency (UKSA) and the National Centre for Earth Observation (NCEO) in April 2016, to create the "Climate Data From Space" (CDS) computing zone on JASMIN. The CDS-zone, managed by CEDA, provides dedicated infrastructure and software support to the UK Climate Data from Space community and is intended to support the computational needs of expert groups in the UK producing climate-quality data products from Earth Observation data. The funding (£500k) will support the CDS-zone for an initial period of two years.

Over the past years, several UK-led teams have developed world-leading methods to derive Essential Climate Variables from satellite data. This work has been funded as a research activity (e.g. by NERC or the European Space Agency Climate Change Initiative). The European Commission's Copernicus Climate Change Service (C3S) is funding these teams to make their data production chains more reliable and robust, as an operational data stream. To underpin this work, a sufficiently sized computing environment is required, with direct access to the input data. JASMIN is able to support these requirements through the CDS-zone. In detail, the Climate Data from Space Zone comprises:

- 700 TB of fast disk storage
- 200 processing cores
- 500 TB tape storage
- Community software, including a "dashboard" showing the status of each data product

The first users of the CDS-zone are producing datasets of Sea Surface Temperature (Reading, Met Office), greenhouse gases (Leicester), aerosol (STFC-RAL) and clouds (STFC-RAL).

The CDS-zone initiative stemmed from the UK's Climate Data from Space Stakeholder Group (CDSSG), who have worked to establish the UK's "seamless supply chain for climate data from space". This seamless supply chain is one of the elements required to position the UK at the leading edge of EO data exploitation for Climate, a long term goal stated in the UK's Space Innovation and Growth Strategy Plan.[1]

More projects are expected join in the coming years, as C3S incorporates further climate data production streams. The aim is that the CDS zone, with its data processing facilities, community tools and software to allow regular climate data production, is a key step towards stimulating growth and innovation in the UK's climate services industries.

Figure 5: Space Innovation and Growth Stratgy Report, 2013

[1] http://www.ukspace.org/wp-content/uploads/2013/11/Space-IGS-Space-Growth-Action-Plan-2014-2030-Nov-2013.pdf

2.5. CLIMATE INFORMATION PLATFORM FOR COPERNICUS

Martin Juckes

The European Commission funded the Climate Information Platform for Copernicus (CLIPC) project to develop a demonstration platform for the Copernicus Climate Change Service. CEDA led an international consortium of 22 partners which developed an attractive portal providing a range of advanced services, powered thorough a platform with a flexible layered architecture enable by extensive use of detailed data standards.

The consortium exploited a wide range of expertise:

- Climate data from models, Earth observation and in situ;
- Standards for climate data and data services;
- Information technology to support distributed web services;
- Processing of data to climate change impact indices;
- User engagement.

The expertise in user engagement was particularly beneficial in holding the broad range of partners to a common story line for the development and deployment of the CLIPC platform.

The CLIPC platform allows data to be added and automatically integrated into all services provided that they adhere to the required standards. The standards were expressed with sufficient flexibility to support the full range of CLIPC data products, spanning all climate observations as well as climate impact indicators which include additional environmental or socio-economic data (such as economic impact of flooding).

The platform exploits the Climate and Forecast Convention (CF) and a suite of domain specific profiles (e.g. for global climate projections, regional climate projections, in situ observations, etc.) to support more detailed information in a structured manner, making the information accessible to users browsing through a multi-tiered service infrastructure. Each domain specific profile contains a collection of interrelated vocabularies. Where possible, underlying terms were exposed through Simple Knowledge Organisation System (SKOS) vocabularies, to exploit the robust scalability of this World Wide Web consortium standard. A glossary of terms, imported from the IPCC Data Distribution Centre and extended with contributions from a range of European consortia, was used to provide, via a SKOS service, clear and authoritative definitions of all technical terms used in the CLIPC portal.

Figure 6 shows a screenshot of the visualisation of a very high resolution environmental impact indicator (resolving spatial scales down to 20m) and a high resolution indicator of economic impact (resolving spatial scales down to 10km).

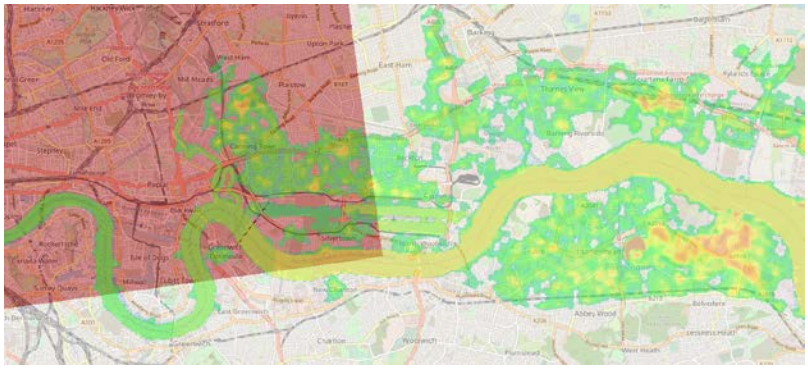


Figure 6: 100 year return value of projected flooding in 2030 at 25m resolution, superimposed on lower resolution estimate of economic impact and street map of East London

CEDA had key roles in CLIPC activities on data, standards, communication and outreach, and management and coordination.

2.6. JASMIN CLOUD AND ACCOUNTS PORTAL

Philip Kershaw[1], Jonathan Churchill[2], Matt Pryor[1], Cristina Del Cano Novales[2]

[1] CEDA
[2] Scientific Computing Department, STFC

The cloud and accounts portal are two applications which have a key role in making resources available to users on JASMIN. A core design principle for JASMIN is to make computing resources available alongside data, and a Cloud computing paradigm fits closely with this model. JASMIN runs its own *Community Cloud* to enable users to effectively borrow chunks of the overall pool of computing resources offered by JASMIN and use them for their own purposes. In practice, this means the ability to create their own virtual computers and install their own software as part of a *tenancy* - their share of the overall cloud's resources. Cloud users have used this capability on JASMIN in all kinds of novel ways, from running software for virtual training courses, to operating websites, portals and providing virtual computer desktops.

Cloud computing is better known though for the commercial or public providers of these services such as *Amazon Web Services*, *Microsoft Azure* and *Google Cloud Platform*. However, even a quick look at the web interfaces reveals the range and complexity of functionality offered by these services. This is also true for the underlying cloud technology used for JASMIN. A custom web portal was developed, with the goal of simplifying access, making it friendlier to use and more intuitive for users who may not be familiar with cloud computing. From the portal, users can easily select the kind of virtual machine they would like and provision it themselves. Once created, they can access the new machine and customise as needed installing their own applications and services. To date, users of the cloud have used for hosting applications ranging from virtual research labs for teaching to various portals for accessing data.

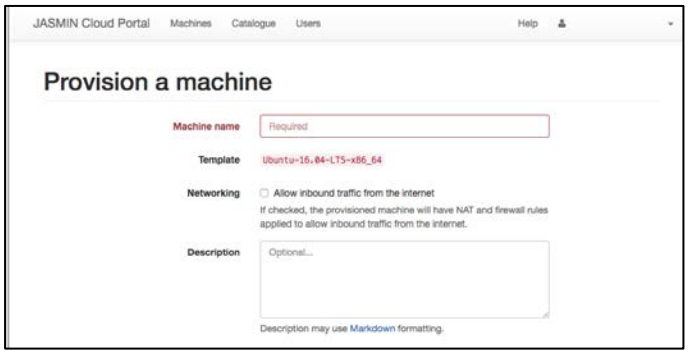


Figure 7: Cloud portal user interface

Figure 8: JASMIN Cloud Portal welcome page

Another example of the integration between the infrastructure and JASMIN's user-facing services is the Accounts Portal. This is a user's first view of JASMIN, providing a web interface to register for an account. Once a user is approved for access, it provides a means to apply for access to various JASMIN services, including group workspaces and cloud tenancies. The portal delegates the authority of granting access to 'service owners', so that the CEDA helpdesk team does not need to be directly involved - projects or groups of users can effectively manage access between themselves.

A primary motivator for the JASMIN Accounts Portal was the fact that there is an increasing number of users of JASMIN that do not require access to the CEDA archive - they only use Group Workspaces or a Cloud tenancy. The decision was taken to have a separate user account for access to JASMIN, where previously this access was managed via a user's CEDA account. MyCEDA, used to manage your CEDA user profile, but is now only responsible for managing access to data in the CEDA archive while access to compute and storage resources on JASMIN is managed by the JASMIN Accounts Portal.

2.7. 2016 JASMIN 1ST CONFERENCE

Fatima Chami and Matt Pritchard

JASMIN has existed for five years as a national facility to enable petabyte-scale data-intensive computing in environmental sciences. JASMIN's strength as a data analysis environment for large, cross-disciplinary datasets has attracted users from a wide range of scientific communities and fostered innovative collaborations.

In its role managing the JASMIN Service to the user community, CEDA hosted the first ever JASMIN user conference to help strengthen the interaction between service provider and the scientific user community, whilst also offering an opportunity for networking and exchanging knowledge among current and potential JASMIN users. The conference was held at the STFC Rutherford Appleton Laboratory, JASMIN's host institution, which gave many members of the community the unique opportunity to tour the facility and learn about the scale and innovative design. JASMIN is a joint collaboration between STFC's Scientific Computing Department and CEDA, in RAL Space. The cross-departmental team has worked closely throughout JASMIN's evolution, in specifying, architecting, procuring and operating the service on behalf of NERC.

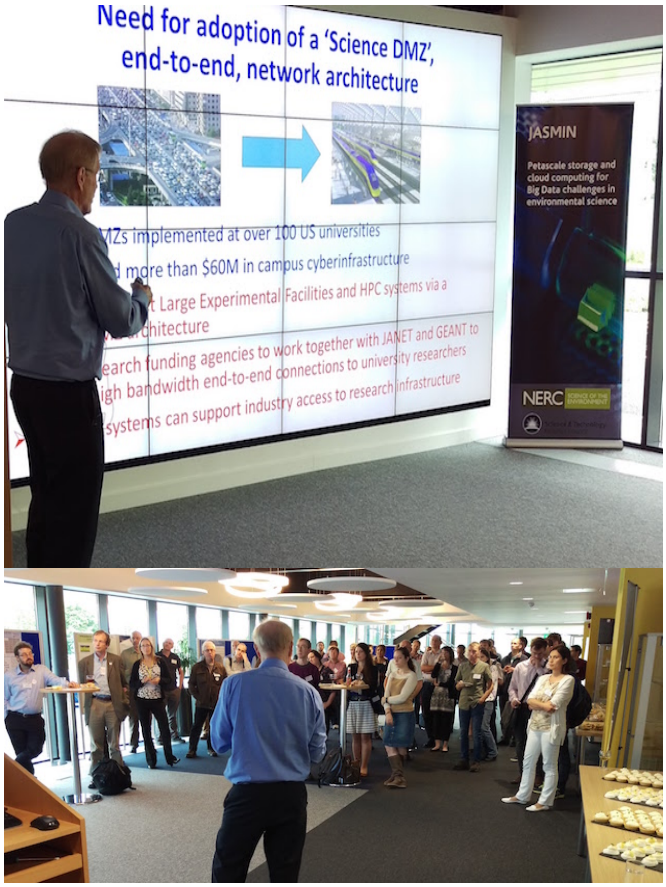


Figure 9: Prof Tony Hey (STFC Chief Data Scientist) delivers his presentation on Data Intensive Science at the evening reception of the JASMIN User Conference 2016

The conference was held over two days in June 2016. Information about the current scale and capabilities of JASMIN was presented by the JASMIN team followed by presentations from a number of current data-intensive science projects on JASMIN. Workshop sessions helped new and existing users learn how to get the best out of JASMIN's processing and data transfer capabilities, and how scientific collaborations might go about managing their own computing infrastructures within the JASMIN Cloud. Discussions focussed on how scientific software might be managed by a community for its own and wider benefit, and learned about particular software tools developed around the JASMIN platform which help distil best practice and enable efficient scientific workflows. Delegates also discussed other environmental science challenges which could be addressed by using JASMIN.

The number of attendees at the conference was 74, with 22 visiting the JASMIN facility during the event. A wide range of scientific disciplines across the environmental sciences were represented at the conference. Evaluation feedback from delegates has been excellent and recognised that the event gave delegates the opportunity to explore the capabilities of JASMIN across a diverse range of workflows, exposing the flexible and powerful analysis platform which JASMIN provides. The majority of delegates expressed a wish to see similar events hosted annually while others suggested additional training events which are now in preparation. CEDA plans to host a variety of future events to encourage closer engagement with and among JASMIN user community and ensure that JASMIN continues to be a world-leading facility for data-intensive environmental science.

2.8. JASMIN CLOUD HOSTING FOR ESA SUMMER SCHOOL

Phil Kershaw

Earlier this year, JASMIN provided the computing resources for around 50 students at a Data Assimilation workshop, part of an Earth Observation Summer School held at the European Space Agency ESRIN centre near Rome. A number of technical developments made this event possible, not to mention the hard work of those involved in preparing the course.

The course took advantage of the popular open source software Jupyter aka IPython Notebook, to deliver the training materials. IPython Notebooks provided an interactive environment for running Python code embedded with text and plots accessible via a web browser. The text, code and plots can be saved and shared as notebooks.

The notebook environment has been developed into a hosted service running on JASMIN's Cloud through the ESA-funded OPTIRAD and Synergy projects. The combination of Notebook and cloud hosting neatly fit OPTIRAD's goal of providing a shared virtual research environment for researchers working on the development of algorithms for land-data assimilation.

Figure 10: Attendees at ESA Summer School, ESRIN. Credit: ESA

The course itself, has demonstrated two key benefits of Cloud computing: the ability to access the computing resources from anywhere (the course was in Rome but the computing resources were hosted from JASMIN in Oxfordshire) and secondly, elasticity - the ability to grow computing resources to meet a given demand. For the duration of the course, the portion of the cloud allocated to the project was expanded temporarily to support the large number of students accessing the environment.

Work is underway to expand the system developed for OPTIRAD into a Notebook service available for the whole JASMIN community. The development team at CEDA are using software tools such as *Docker*, *Kubernetes* and *Ansible* to make it flexible and to automate its deployment and make it easily maintained into the future.

Acknowledgements: Jose Gomez-Dans, UCL, Nicola Pounder, Assimila Ltd. and teams at respective organisations; Matt Pryor, CEDA and the support of NCEO. This research was funded by ESA OPTIRAD and Synergy projects and JASMIN core funding.

2.9. CEDA INTERNAL DEVELOPMENTS

Richard Smith, William Tucker

Several important internal CEDA projects have been carried out by our developers to help support and improve the work done by our data scientists and ultimately improve user experience. Three activities are described here:

The **Data Management Plan (DMP) Tool** is an internal web application which is used to manage projects that have data suitable for archival at CEDA, such as grants funded by the Natural Environment Research Council (NERC). The task was to streamline the workflow for the data scientists and reduce the work required to track the status of each project. Several areas for improvement were identified; including making sure notes contained necessary information, creating reminders for project milestones and reducing the need for copy/paste operations. These were addressed through collaboration between CEDA data scientists and CEDA developers. The project began late 2016 and the first iteration was completed early 2017. The solutions implemented include automatic creation of reminders when creating a project and a "to-do" list which allows the data scientists to quickly identify which projects need their attention first. To allow more accurate tracking, integration with Google Drive has been implemented, utilising templates that automatically create a pre-filled project specific DMP from information within the tool.

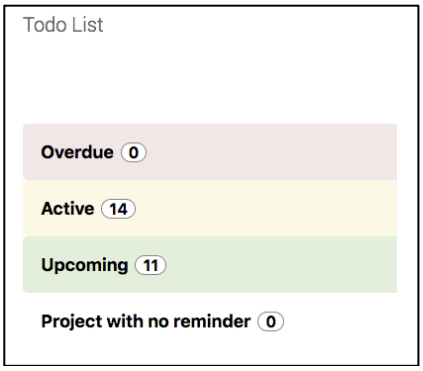


Figure 11: DMP tool to-do list for data scientists

On the 28th November, we updated our **CEDA data catalogue site** resolving a long-standing problem related to login and authorisation: users were experiencing an inescapable loop which denied them access to certain data. This problem had caused our helpdesk staff to spend a lot of time resolving issues from users who were trapped in this loop. Thanks to a recent security upgrade to our login service, we had access to new, encrypted browser cookies which stored information about a user's data permissions; from these, we were able to engineer a solution to the access problem. We also added extra functionality to the catalogue pages, improving usability by allowing visitors to see which datasets they have permission to access and how to apply for access to restricted data.

During December, some development effort was shifted towards long-overdue improvements to **CEDA's data browser service**, a website providing direct access to data from our archive. The motivation for this work was to improve the look-and-feel of the site, as well as to offer better support for download capabilities that were in-demand amongst our users. The developer and data science teams worked together to determine the most requested features and tried to incorporate them into the design of a new data browser service. On the 1st March 2017, we released the first version of the new CEDA data browser. It was an updated version of an older CEDA service, PyDAP, which incorporated a widely-used framework for providing fine-grain access to scientific data called OPeNDAP. The addition of this functionality, as well as some general improvements to the web-interface and updated software, made this service much more useful and secure than the sites that had preceded it. The success of this release also allowed us to retire the old data browser and PyDAP sites, substantially reducing our maintenance workload.

Internal projects, like these, are vital for the smooth running of our services and shows the collaborative nature of CEDA's work between the developer, data scientist and helpdesk teams.

2.10. CEDA CATALOGUE EXPORT SERVICES

Steve Donegan

CEDA provides access to petabytes of Atmospheric, Model and Earth Observation data to the UK academic community and beyond. It is vitally important that CEDA maintains a metadata catalogue of its data holdings that record key aspects of the archived data, such as location, measurement period, data originator as well as important peripheral information as ownership and access constraints. This metadata catalogue is used to link searches made by CEDA users to the data in the archive. The CEDA catalogue also provides a basis to publish metadata for export, so that external software clients can access this information programmatically for consumption in dedicated themed portals.

CEDA's catalogue is populated by both scanning data held in the archive as well as manual input from CEDA data scientists. These aspects include information on e.g. the instrument, the platform on which it is deployed and the project the data is associated with. There is a searchable web interface on the CEDA homepage (http://www.ceda.ac.uk) to the data catalogue, with results pages displaying links to the download and the archive access applications. The data catalogue forms the basis of all metadata transformations prior to export.

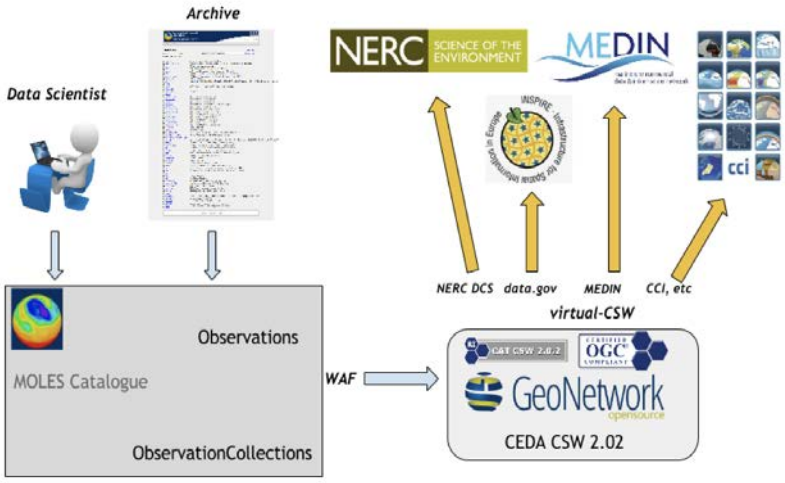


Figure 12: How CEDA publishes its metadata. Metadata is generated from MOLES content where it is harvested and placed in an OGC CSW from where clients can access the metadata in UK Gemini 2.2 format

In some cases simple access to the CEDA homepage and data catalogue is not sufficient. CEDA generates metadata based on the data catalogue, and publishes it for consumption to a number of web portals and systems that provide data discovery services to a much wider global community.

In order to conform with EU INSPIRE legislation, the UK national portal (data.gov.uk) specified a metadata profile based on the ISO19115 schema for geospatial data and services. CEDA has put in place a system to transform MOLES records into this profile (UK Gemini 2.2) and publish them. These records are the metadata format used for publishing to services such as the NERC Data Catalogue Service (DCS) (http://data-search.nerc.ac.uk/), the MEDIN Discovery Portal (http://portal.oceannet.org/), data.gov.uk, and various other projects such as the CCI (ESA Climate Change Initiative) portal. They are also published to the Global Earth Observation Program portal (GEOSS).

CEDA uses an open source Geonetworks OGC Catalogue Server for the Web (CSW) 2.0.2 (http://geonetwork-opensource.org/) to both harvest the UK Gemini records from CEDA's catalogue, as well as make them available to the wider community via published CSW endpoints. The CSW is a specification that makes metadata content searchable to client servers and returns information on available datasets and services at a range of content levels. In order to meet obligations to different projects and thematic consumers CEDA uses a system of defined keywords to provide a system of "virtual" CSW's.

These metadata records and services ensure visibility of CEDA data at both national and global levels.

2.11. CEDA'S HELPDESK: BIGGER, BETTER & CHEAPER

Hayley Gray

In October 2016, CEDA upgraded to a new helpdesk system called Help Scout. CEDA chose this helpdesk service as an easily accessible, more reliable cloud based service that is cheaper than CEDA's previous system, which required an expensive update. The development of JASMIN (CEDA's data intensive computing facility) has dramatically increased the number of queries to the helpdesk and CEDA anticipates that this trend will increase, therefore having an efficient reliable helpdesk service was considered vital. Moving to this service was a smooth transition with no interruptions to CEDA's users.

Help Scout provides companies worldwide with an easily accessible helpdesk system. This email type system has integrated documents, allowing users and staff to access a combined system providing an expanded user-friendly experience.

Help Scout provides a service to support CEDA's users via three features, as shown in Figure 13. Firstly, the 'Docs' are full of information to help users with any queries they may have. A couple of examples are: systematic guidance on how to set up an account for CEDA, and how to apply for data in CEDA's data catalogue. This documentation provides easy access to information and should provide an answer to a user's issue before having to contact the helpdesk. Ultimately, fewer queries should be sent to the helpdesk meaning time saved for CEDA staff allowing greater focus on other tasks.

Secondly, the 'Beacon' (customised help form) is situated on the CEDA sites allowing an easy search within the 'Docs'. If no suitable results are found, the user can still contact the helpdesk via the Beacon. Doing this captures the page URL link, previous searches and user's browser configuration. This gives CEDA staff valuable context to get a better idea of how to help the user and offer a more rapid solution.

Finally, the 'Help Desk' is the traditional method for user queries via an email system. The structure of the new helpdesk is in an understandable format with clear grouped conversations for each user query and provides easily accessible documents within the same system. This helps both CEDA and its global research community of over 6500 users who annually contact the helpdesk. As a result of these features, CEDA staff are spending less time answering queries thus making it better value for money without cutting any corners.

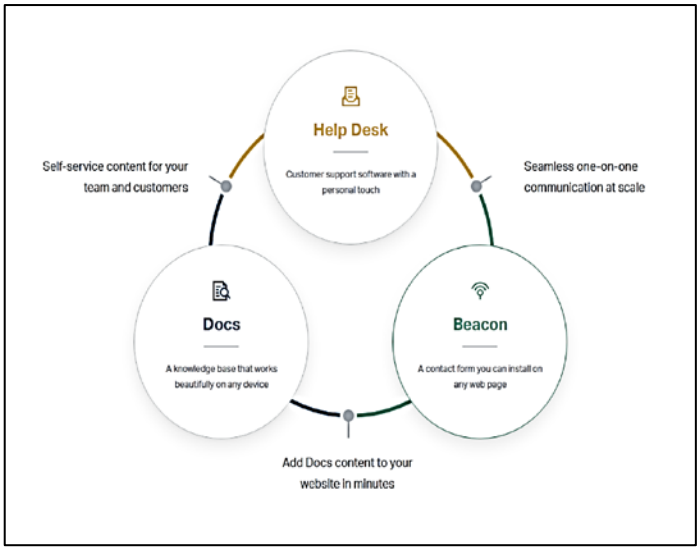


Figure 13: The new Help Scout user support mechanisms

Another benefit of Help Scout is the reports tab. This enables the helpdesk operator to produce statistics by monitoring workload and user satisfaction. CEDA can also spot trends in queries and failed document searches to resolve them at source by enhancing documentation, thus improving access to self-service support, to address common issues.

Help Scout enables us to do more with a cheaper service, by allowing us to provide an enhanced provision for the users with the use of the integrated help documents. The beacon and the reports tab act as extra support mechanisms, along with the access to a direct query email direct to the helpdesk.

CEDA staff successfully set up the new system, migrated the existing helpdesk into Help Scout, and provided internal training for staff involved in triaging and responding to user queries.

2.12. SHARING EXPERTISE AND ENGAGING WITH THE RESEARCH COMMUNITY: GRADUATE PLACEMENT AT FAAM

Poppy Townsend

One key provider of data to CEDA is the Facility for Airborne Atmospheric Measurements (FAAM); a flying laboratory that is used by the UK atmospheric science community on campaigns across the world. CEDA archive all core data, with parameters such as temperature, humidity and altitude, immediately after each flight (>1000 flights to date). This requires CEDA and FAAM to work together closely so that all data is archived appropriately. To encourage and develop this relationship further, Poppy Townsend a Graduate Data Scientist spent 3 months at FAAM on a placement.

Figure 14: Poppy Townsend and the FAAM aircraft

Poppy's work at CEDA focuses on data at the end of a project's life cycle when it is ready for archival, whereas at FAAM she saw first hand the data collection stage. Her work helped to facilitate post-flight validation and quality control to improve the flow of core data to CEDA. This was done by creation of a Quality Assurance-Quality Check (QA-QC) report using Python programming that could be generated immediately after a flight; allowing researchers to have a 'quick look' at all FAAM core data collected.

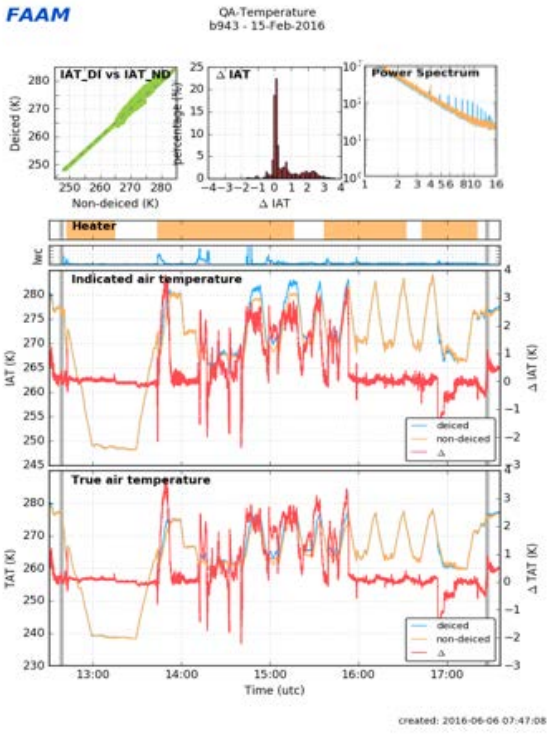


Figure 15: Example of core instrument data in the QA-QC report

The report contains 18 pages of instrument plots including; temperature, humidity, liquid/total water, turbulence, aerosols, carbon monoxide, irradiance, dew point and mixing ratios, pressure and altitude. Graphs of these core parameters help instrument scientists and Principal Investigators (PIs) to quickly identify instrument issues or interesting data features in the flight data. This report is currently generated after each flight but there are plans to create it in near-real time.

Figure 16: View from onboard the FAAM aircraft on Flight B951

This sharing of expertise between CEDA and FAAM has developed a better understanding of how each organisation works. Engaging with core research communities is important to CEDA and this placement has facilitated this, whilst also providing opportunities for future collaboration.

PART 2: CEDA AND JASMIN SUMMARY INFORMATION 2016-2017

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 environment is increasingly providing data-compute capabilities for the UK and wider environmental research communities. This section of the annual report presents summaries of CEDA and JASMIN usage.

3. 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 Usage: April 2016 to March 2017	
Total number of users	20,592
Total data downloaded	1049.4 TB
Total number of accesses	16,772,624
Total days activity	115,046

Table 3.1. : Summary figures for usage by CEDA consumers during the reporting year

These figures can be broken down by month showing that while usage in terms of number of users and accesses has remained static, there is a general trend upwards in the amount of data downloaded (table 3.2 and figure 3.1)

Month	Users	Methods	Datasets	Number of accesses	Size (TB)	Activity days
201604	2116	10	1,142	1,079,315	53.1	6,135
201605	1485	9	821	2,005,106	90.7	4,707
201606	1760	10	820	1,091,630	158.7	5,509
201607	1972	11	1,108	4,031,672	99.9	6,700
201608	1393	11	971	2,494,343	237.2	5,529
201609	1775	11	2,307	851,782	57.7	8,088
201610	1755	11	1,471	551,780	49.7	7,636
201611	1929	9	1,098	1,088,442	26.3	6,608
201612	1538	10	1,300	720,093	29.7	6,770
201701	1584	10	908	746,489	93.3	5,407
201702	1749	10	1,610	1,200,679	60.0	18,701
201703	1536	10	1,476	911,293	93.1	33,256
Totals	**20,592**	**122**	**15,032**	**16,772,624**	**1049.4**	**115,046**

Table 3.2: Monthly summary figures for usage by CEDA consumers during the reporting year

Note that additional metrics also appear in the data centre metrics section, chapter 7.

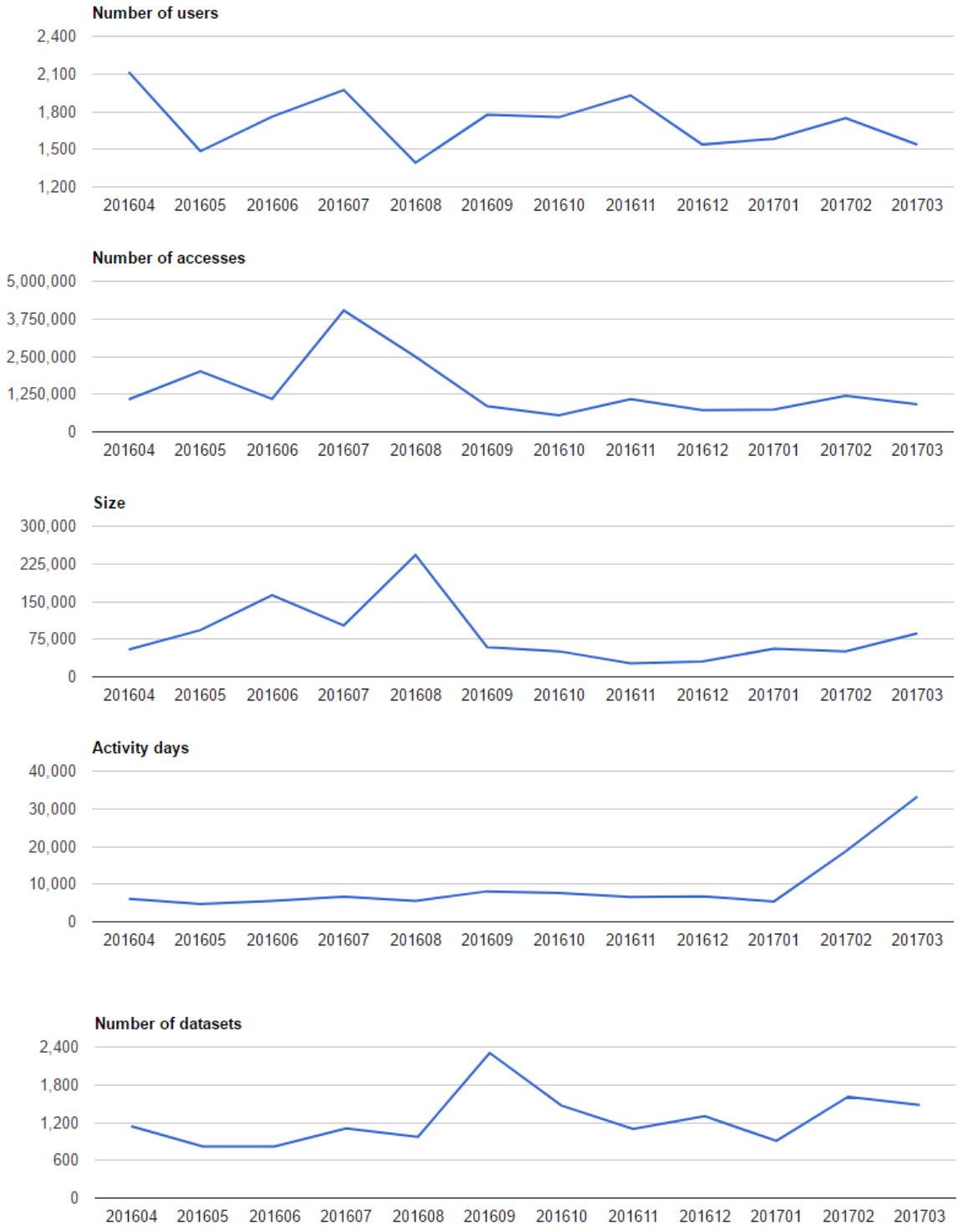


Figure 3.1: Breakdown of CEDA usage by month

4. JASMIN

JASMIN is a globally-unique data intensive super-computer, designed, integrated and operated by STFC providing four main functions: storage, batch computing, managed infrastructure compute and a private cloud to provide Infrastructure-as-a-Service (the JASMIN “un-managed cloud”).

During this period, JASMIN continued to provide a cutting-edge research environment for NERC-related environmental science. While larger-scale funding capital funding was still unclear, interim funding was secured for a limited procurement to maintain capacity and capability across key areas while undertaking strategic proof-of-concept activities to inform future technology choices for Phase 4 and beyond.

4.1. STORAGE

A procurement deal was reached to add a further 1PB of Panasas parallel storage, with support arrangements to continue operating JASMIN's phase 1 and 2 storage: the phase 1 storage (approx. 5PB) for an additional year beyond its planned end-of-life.

With JASMIN's NetApp storage due for retirement at the end of this period, new iSCSI storage was procured to support JASMIN's cloud environment. Meanwhile further storage hardware was procured to create a Proof-of-Concept for two alternative storage technologies based on software-defined storage: Object Storage and Scale-out Filesystem. These potentially offer vastly scalable storage at a reduced cost, but proof-of-concept work needs to evaluate their functionality and performance at scale before larger investment decisions can be made in future phases. The hardware was used to evaluate a number vendors' software offerings in this arena, before being turned over for use as scale-out filesystem storage in the next reporting period.

At the start of this reporting year there were 113 Group Workspaces across JASMIN and CEMS with allocations totalling 7 PB, serving a total community of 914 users. By the end of the period this had increased to 147 group workspaces in use with allocations of 8.6 PB respectively, serving a total community of 1304 users.

The Elastic Tape service, a near-line tape storage facility available to JASMIN and CEMS Group Workspace managers, continued to provide a flexible and scalable complement to “fast disk” storage, enabling GWS managers to optimise their use of online disk resource. A further 5PB of tape was procured to increase the capacity of this service.

4.2. BATCH COMPUTING

The Phase 3.5 procurement of enabled LOTUS to grow by 1120 cores to over 6000 cores in total: 56 2x10 core servers each with 256 GB RAM were added to the cluster in Autumn 2016.

LOTUS utilization regularly topped 70% before this upgrade, meaning that nearly all jobs need to queue before starting. By the end of the period, utilization had stabilized around 70% (even with that new capacity) and the number of active LOTUS users per month was 141.

Host groups to identify hosts of similar CPU generation and RAM, were introduced during this period to help users with particular hardware requirements for their code.

4.3. MANAGED COMPUTE INFRASTRUCTURE

Additional services were added to JASMIN’s “Science DMZ” or “Data Transfer Zone”: this is an area of network situated as close as possible to the RAL site border routers, with high-performance hardware and appropriate security policies to support high-throughput data transfer services. Work during this period has added servers to act as production and test gridftp endpoints for use within the ESGF infrastructure, helping to prepare CEDA for participation in large-scale CMIP6 data. In addition, a relay hub for ESA Sentinel data was deployed to increase

the efficiency of CEDA's activities in ingesting and disseminating Sentinel data, with data flows of 10TB/day expected once all Sentinel satellites are operational.

In response to requests for better support for graphical/desktop applications, X2GO server was deployed on login and sci machines and is being trialled with some users. This provides a local (JASMIN-side) X server which can be used for X11-based graphical applications more efficiently than sending X11 graphics over the wide-area network.

The "JASMIN Analysis Platform" software stack continued to evolve with a new release (1.1-28) deployed across the managed infrastructure and LOTUS in May 2017.

4.4. JASMIN CLOUD

The JASMIN cloud portal continued to provide tenants of JASMIN's unmanaged cloud with a web-based interface enabling them to deploy and manage virtual machines at will within their tenancies. This is now the default interface for new cloud tenancies as VMWare vCloud Director is now deprecated.

A 3-node proof-of-concept cluster was established for the OpenStack, paving the way for eventual migration away from VMWare as JASMIN's cloud management infrastructure to OpenStack. If successful, this should provide improved capability for cloud management as well as cost savings.

100Gbit networking hardware was procured as part of Phase 3.5, believed to be the first instance of this within STFC. This is targeted towards improved accessibility and throughput of data between the unmanaged cloud and the managed infrastructure. Once deployed, it is intended that a high-performance OPeNDAP service will be deployed over this network, visible to tenants within the unmanaged cloud.

4.5. USER SUPPORT AND OUTREACH

June 2016 saw the first JASMIN User Conference, attended by around 70 delegates and hosted by RAL Space. This event provided attendees with opportunities to learn about recent developments to the JASMIN infrastructure and about some of the myriad scientific projects using it, and interact with STFC staff as well as other users, sharing experiences of using JASMIN and how it can be used in particular workflows. The event was held over 2 days and included some training sessions as well as the chance for users to visit the machine room where JASMIN is hosted in STFC's Scientific Computing Department.

A major re-write of JASMIN (and CEDA) documentation was undertaken during this year, with new documentation helping to guide users through the process of getting started with JASMIN and how to register and access services within the JASMIN environment. A YouTube channel was set up, initially with tutorial videos about how to use the JASMIN accounts portal but will be used for further training content.

5. COLLABORATIONS

CEDA continues to support international climate modelling community through its interactions with the large global collaboration to deliver an Earth System Grid Federation under the auspices of the Global Organisation for Earth System Science Portals (GO-ESSP).

Other major international collaborations include participation in the infrastructure projects (originally IS-ENES and now IS-ENES2) to support the European Network for Earth system Simulation (ENES), and on Earth Observation data, the European Space Agency's Climate Change Initiative (CCI).

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

5.1. MAJOR COLLABORATIONS

In 2016/2017, 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 Data Operations Group.

2. Operating and evolving the Earth System Grid Federation 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).

3. A leading partner in many major European projects including IS-ENES 2 (developing an InfraStructure for a European Network for Earth system Simulation and CLIPC (Climate Information Portal for Copernicus).

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

6. Working with commercial and academic partners and the Satellite Applications Catapult, to support both academic research and opportunities for commercial applications and downstream services from EO and Climate data.

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

6. 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.

6.1. ANNUAL TOTAL FUNDING

Financial Year	08-09	09-10	10-11	11-12	12-13	13-14	14-15	15-16	16-17
NCAS income	970	866	906	883	935	829	829	808	808
NCEO income	378	389	450	419	445	392	390	393	393
Other NERC	788	481	341	527	287	272	600	621	825
Other income	461	710	1144	1099	1283	1486	1394	1505	1092
Total income	2597	2446	2841	2928	2950	2979	3213	3327	3118

Table 6.1: Overall funding for CEDA for financial years 2008 — 2009 to 2016 — 2017 (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.

6.2. EXTERNALLY FUNDED PROJECTS FOR THE YEAR 2016-2017

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)
CEMS support to ESA CCI SST	JASMIN support for CCI SST data processing	ESA	01/05/2014	31/03/2017	72.9
CLIPC	Climate Information Portal for Copernicus	FP7	12/12/2013	31/01/2017	703.9
ESA CCI Data Portal	Archive and data services for ESA CCI programme	ESA	01/05/2015	31/03/2018	195.3
ESA Optirad	OPTIRAD (OPTImisation environment for joint retrieval of multi-sensor RADiances) hosted data assimilation environment	ESA	01/04/2014	31/01/2018	22.0
EUFAR-2	The European Facility for Airborne Research (EU-FAR), coordinating the operations of the European fleet of	FP7	01/02/2014	31/01/2018	174.0

	aircraft in the field of environmental research				
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	21/02/2019	102.5
BACI	JASMIN support for BACI (Biosphere Atmosphere Change Index)	H2020	01/04/2015	31/10/2018	25.7
IS-ENES2	InfraStructure for the European Network for Earth System Modelling	FP7	01/04/2013	31/03/2017	551.0
SPECS Data Archive	Data archive for the SPECS FW7 seasonal prediction	Institut Catala De Ciencies Del Clima	01/08/2013	31/10/2016	121.0
Forestry TEP	ESA Thematic Exploitation Platform	ESA	01/01/15	31/07/17	125.2
Monsoon Overflow 2	Providing support for MONSooN users on JASMIN	Met Office	01/04/14	31/03/17	199.6
Optirad-2 SY-4Sci Synergy	Extension of Optirad hosted data assimilation project	ESA	01/04/15	31/12/16	10.9
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	30/10/2017	311.2
UKCP09 User Interface 15-18	UK Climate Projections platform, user interface development	EA	01/04/2015	31/03/2018	86.5
C3S CDS Infrastructure	Consultancy role for C3S Climate Data Store project	C3S	01/07/2016	30/06/2017	14.7
UKSA CDS Zone	Support for ECV processing	UKSA/NCEO	01/08/2016	15/03/2019	155.0

DECC Clim Arch Supp Extension	Climate Model archive	DECC	1/4/2016	31/03/2017	520.4
MEDIN Metadata Services	Metadata services for MEDIN	MEDIN	1/4/2015	28/02/2018	30.7
Support for Ensembles Met Office 16-17	Support for Ensembles project	Met Office	1/7/2016	31/3/2017	36.3
C3S ESGF Data Node (CP4CDS)	Operational ESGF data node for C3S	C3S	1/9/2016	31/8/2017	309.0
JASMIN Met Office Support	JASMIN and archive support	Met Office	1/6/2015	31/3/2017	70.0
Pest Risk Modelling in Africa (PRISE)	JASMIN support for UKSA IPP project	UKSA	1/12/2016	31/03/2019	97.7
C3S ECVs production SST	JASMIN and archive support for C3S SST ECV processing	C3S	01/02/2017	31/10/2018	63.9

Table 4.2: Externally funded projects for 2016-2017 (non-core NERC)

6.3. GOVERNANCE

In response to organisational changes in STFC RAL Space, and also in NCAS, some changes to CEDA's management and governance were implemented during 16-17. A CEDA Steering Board has been set up, with representation from CEDA's main stakeholders, NCAS, NCEO, NERC and STFC. After 17 years as Director of CEDA, Professor Bryan Lawrence has moved full time to the University of Reading but, as NCAS Director of Models and Data, retains roles as Chairman of the CEDA Steering Board, and PI of JASMIN. Victoria Bennett now leads the CEDA Division, as Division Head, within STFC's RAL Space Department.

PART 3: METRICS AND PUBLICATIONS

7. 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.

7.1. ACCESS RELATED METRICS

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

Area	Q1		Q2		Q3		Q4	
UK	1673	40.8%	2950	61.8%	2996	62.0%	2517	60.1%
Europe	610	14.9%	558	11.7%	550	11.4%	499	11.9%
Rest of the world	1693	41.3%	1176	24.6%	1186	24.6%	1096	26.2%
Unknown	126	3.1%	91	1.9%	98	2.0%	76	1.8%

Table 7.1: Users by area

Institute Type	Q1		Q2		Q3		Q4	
University	2947	71.8%	3297	69.0%	3346	71.8%	3075	73.4%
Government	610	14.9%	815	17.1%	813	16.8%	613	14.6%
NERC	101	2.5%	286	6.0%	282	5.8%	160	3.8%
Other	292	7.1%	278	5.8%	283	5.9%	246	5.9%
Commercial	73	1.8%	55	1.2%	57	1.2%	47	1.1%
School	79	1.9%	41	0.9%	46	1.0%	45	1.1%

Table 7.2: Users by Institute type

7.2. DATA HOLDINGS

Data Centre	Q1	Q2	Q3	Q4
CEDA	3198	3205	3208	3203

Table 7.3: Number of “dataset” discovery records held in the NERC data catalogue service.

	Q1	Q2	Q3	Q4
datasets	3387	3507	3536	3602
collections	390	401	401	403

Table 7.4: Number of “dataset collections” and “datasets” identified by CEDA and displayed via CEDA catalogue.

7.3. HELP DESK RESPONSIVENESS

	Q1	Q2	Q3	Q4
Received	392	397	560	581
% Closed within 3-days	helpdesk: 66.23 % (88.33 % resolved after 3 days)	70.28 (100 % resolved after 3 days)	358 users helped over the quarter 50% resolved on first reply	350 users helped over the quarter-to-date (to 15/03) 56% resolved on first reply
Closed	394	478	508	590

Table 7.5: 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. (Note that in Q2 – the % resolved figure is skewed as closed queries were moved to new helpdesk service. In Q3 it became no longer possible to obtain the “Percent queries dealt with in 3 working days statistic” due to the new helpdesk system.)

	Q1	Q2	Q3	Q4
Received	1059	823	1023	849
% Closed within 3-days	85.93 % 93.39% resolved after 3 days	94.90 99.3% resolved after 3 days	358 users helped over the quarter 50% resolved on first reply	350 users helped over the quarter-to-date (to 15/03) 56% resolved on first reply
Closed	999	889	980	889

Table 7.6: Help desk queries specifically about access authorisation for restricted CEDA datasets and services received and closed by quarter, including the three-day closure rates. (In Q3 it became no longer possible to obtain the “Percent queries dealt with in 3 working days statistic” due to the new helpdesk system.)

8. PUBLICATIONS AND PRESENTATIONS

Bennett, V.L. and CEDA Team, 'CEDA and JASMIN Services' , Poster, NCEO Annual Science Conference, University of Warwick, July 2016

Chami F.; Oral presentation on processing on LOTUS at the Met Office workshop on big data processing, 14 September 2016

Chami F.; Oral presentation on JASMIN at NERC scientific computing meeting, Daresbury 16 May 2016

Clarke, H., E. Pechorro, V. Bennett, C. Farquhar, J. Blower, 'CCI Open Data Portal', Living Planet Symposium, Proceedings of the conference held 9-13 May 2016 in Prague, Czech Republic. Edited by L. Ouwehand. ESA-SP Volume 740, ISBN: 978-92-9221-305-3, p.374

Donegan, S.; Bennett, V.; Waterfall, A.; Kershaw, P.; Williamson, E.; 'Sentinel, Climate & EO datasets and the JASMIN "super-data" cluster at CEDA' ; ESA Living Planet Symposium, Prague, 9-13 May 2016, Paper 1713

Garland, W.E. 'CEDA Developments'; Presentation 18/01/2017. NCAS-FAAM-MetOffice Meeting on FAAM data. University of Cranfield.

Juckes, M.N., Rob Swart, Lars Bärring, Annemarie Groot, Peter Thysse, Wim Som de Cerff, Luis Costa, Johannes Lückenkötter, Sarah Callaghan, Victoria Bennett; 'A Climate Information Platform for Copernicus: managing the data flood' EGU General Assembly Abstracts 2016/4 Volume 18, pages 15396

Juckes, M.N., Rob Swart, Lars Bärring, Annemarie Groot, Peter Thysse, Wim Som de Cerff, Luis Costa, Johannes Lückenkötter, Victoria Bennett, Sarah Callaghan; 'Communicating across the disciplines to support climate services: the CLIPC portal', EGU General Assembly Conference Abstracts, 2016/4, Volume 18, Pages16255

Juckes, M.N., Phil Kershaw, Ruth Petrie, Rob Thomas, Antony Wilson; 'Vocabulary services to support climate research and climate data exploitation', Poster for RMetS NCAS Conference 2016, 6-8 July 2016. https://www.rmets.org/sites/default/files/Book%20of%20Abstracts%20-%20main%20conference%20Final.pdf

Kershaw, Philip, Jonathan Churchill, Bryan N Lawrence, Matt Pritchard, 'JASMIN and the adoption of cloud-native architecture for managing data and compute at scale', Presentation, 16th Workshop on Meteorological Operational Systems, ECMWF, 1-3 March 2017

Kershaw, Philip, Jonathan Churchill, Bryan N Lawrence, Matt Pritchard; 'The JASMIN Data Analysis Facility for the Environmental Sciences Community and the Role of Data-as-a-Service', Presentation, 97th AMS Annual Meeting, 22-26 January 2017

Kershaw, Philip, Victoria Bennett, Steve Donegan, Martin Juckes, Graham Alexander Parton, Ag Stephens, Alison Waterfall, Antony John Wilson, Ruth Petrie; 'Building an Open Data Portal for the European Space Agency Climate Change Initiative based on an Iterative Development Methodology and Linked Data Technologies', Poster, AGU 2016, 12-16 December 2016

Kershaw, Philip; 'Experiences and Lessons Learnt with Collaborative e-Research Infrastructure and the application of Identity Management and Access Control for the Centre for Environmental Data Analysis', Presentation, AGU 2016, 12-16 December 2016

Kershaw, Philip, Victoria Bennett, Jonathan Churchill, Ag Stephens, Matt Pritchard, Matt Pryor, Kleanthis Tsaousis; 'Data discovery and access services for JASMIN - key components in delivering a platform for Open Science', Presentation, ESA EO Open Science 2016, 12-14 September 2016

Kershaw Philip; 'JASMIN Services', JASMIN User Conference, June 27-28 2016

Parton, G.A.; 'Making FAAM Flights Discoverable'. Presentation 28/04/2016. NCAS Data Meeting on archiving and visualising observations. University of Reading.

Parton, G.A.; 'CEDA Developments'. Presentation 07/11/2016. NCAS Data Meeting on archiving and visualising observations. University of Leeds

Pepler, S., 'Fixity Checking a Large Climate Data Archive', Poster for RMetS NCAS Conference 2016, 6-8 July 2016. https://www.rmets.org/sites/default/files/Book%20of%20Abstracts%20-%20main%20conference%20Final.pdf

Pritchard, M., and J Churchill, 'JASMIN: petascale storage and terabit networking for environmental science'. Oral presentation at JISC Networkshop43, 31 March 2015 2 April 2016. https://www.jisc.ac.uk/events/networkshop43-31-mar-2015/programme

Pritchard, M., J.Churchill; 'Enabling high-performance access to big data from space' Oral presentation at ESA 2016 conference on Big Data from Space 15-17 March 2016. https://ec.europa.eu/jrc/en/publication/proceedings-2016-conference-big-data-space-bids16

Pritchard, M.; 'JASMIN Overview (2016)', JASMIN User Conference, June 27-28 2016

Pritchard, M.; 'Enabling efficient movement of data into & out of a high-performance analysis environment' JISC Workshop on Campus Network Engineering for Data Intensive Science, London, October 2016

Townsend, P.; 'Data Science takes off… literally!'; poster at Student Royal Met Soc Student Conference, Manchester, 4-6 July 2016

Waterfall, A., V. Bennett, S. Donegan, M. Juckes, P. Kershaw, R. Petrie, A. Stephens, and A. Wilson; 'Big Data Challenges indexing large-volume, heterogeneous EO datasets for effective data discovery'; Living Planet Symposium, Proceedings of the conference held 9-13 May 2016 in Prague, Czech Republic. Edited by L. Ouwehand. ESA-SP Volume 740, ISBN: 978-92-9221-305-3, p.29

Williamson, E.; 'Sentinel and CEDA services' – poster for NCEO Researchers Forum, 27-28 March 2017, Leicester University

Williamson, E.; 'Sentinel and CEDA services' – poster for RSPSoc Wavelength Conference, 13-15 March 2017, Kingston University

Williamson, E.; 'Sentinel data deluge', Poster at STFC YEScon (graduate conference), Harwell, 17 November 2016