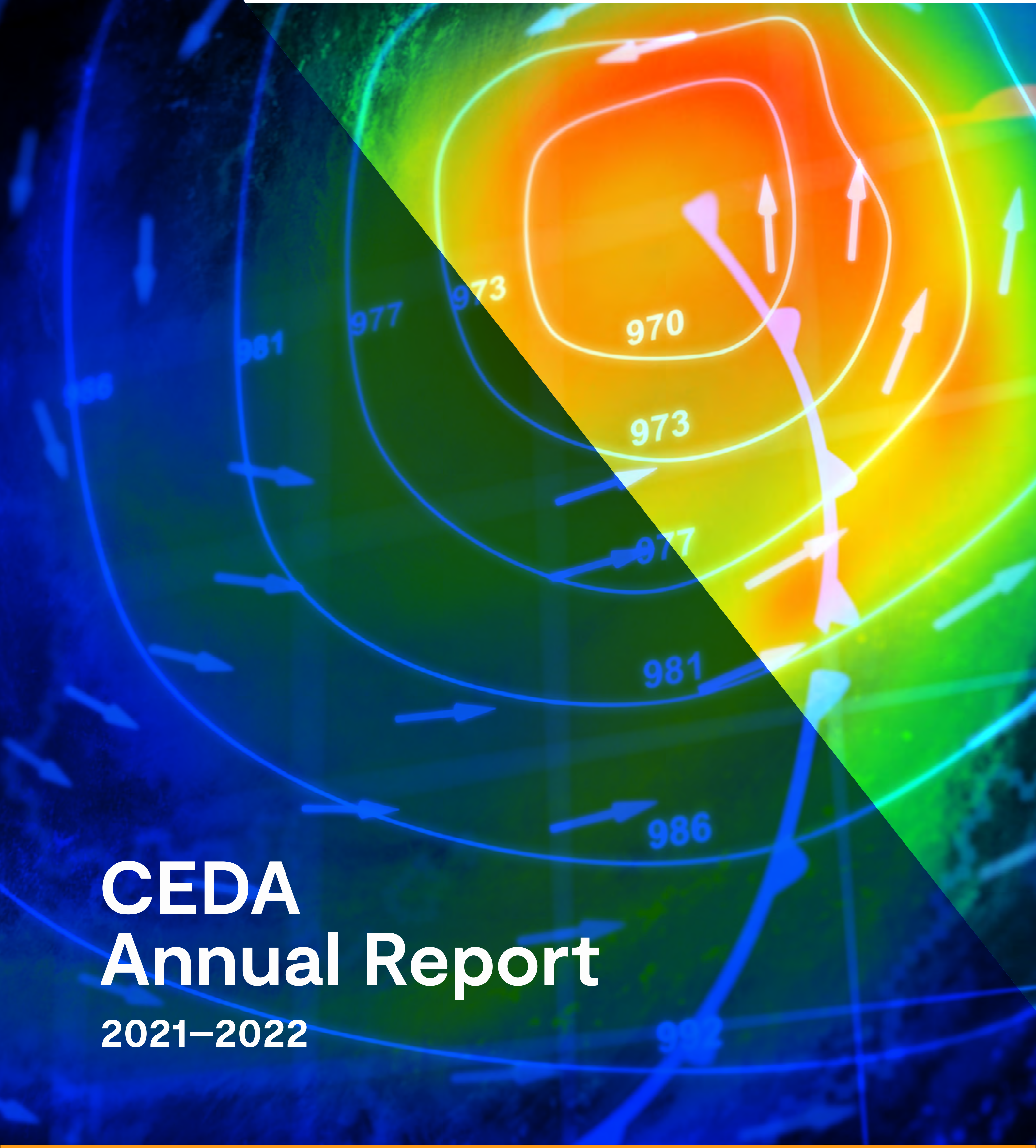
UKRI

Science and Technology Facilities Council

Natural Environment Research Council

Centre for Environmental Data Analysis

SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL



CEDA Annual Report

2021–2022

JASMIN

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

Contents

Editors: Poppy Townsend, Philip Kershaw

Introduction

I'm delighted to introduce the latest annual report showcasing the work of CEDA. It's been a time of change with new starters and new roles – amongst others for me as the incoming head of CEDA. We also welcome Adrian Hines who fills the new position of JASMIN Director. I want to acknowledge and thank those who have moved onto roles outside the organisation for their contributions to make CEDA what it is.

Looking back, 2021-22 was a year of uncertainty and disruption for many with the impact of the pandemic. It's great then to see the strengths in our continued work over this period engaging externally with the user community and with partner organisations in our international collaborations. Besides these themes, the report is organised into topics ranging from open access to data and the application of new technologies. As an engineer by background I'm naturally drawn to the latter and how we can address the challenges of working with data at scale and bring transformation to our services. However, ultimately it's about the people that make up CEDA and their contributions serving our users. There is much hard work that is done behind the scenes in user support and maintaining a complex and innovative set of services and underlying infrastructure. Looking forward, JASMINx – broadening the scope of JASMIN to new communities – and the NERC EDS (Environmental Data Service) will be important in shaping our future plans.

Philip Kershaw
Head of CEDA

About CEDA

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

CEDA Archive

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

JASMIN

JASMIN is the data intensive supercomputer which provides the infrastructure upon which the CEDA archives and services are delivered. Increasingly, JASMIN provides flexible data analysis capabilities to a growing community, who benefit from high performance compute and a private cloud, co-located with petascale data storage.

Meet the team

Philip Kershaw
Head of CEDA

Jennifer Bulpett
Senior Project Manager

Katie Cartmell
Project Manager

Fatima Chami
JASMIN User Support

Esther Conway
Senior Data Scientist: Earth Observation

Steve Donegan
Senior Data Scientist: Earth Observation

Rhys Evans
Software Engineer

Wendy Garland
Senior Data Scientist: Aircraft

Hayley Gray
User Support and Metadata Services Operator

Adrian Hines
Director of JASMIN

Andrew Harwood
Infrastructure Manager

Alan Iwi
Senior Software Engineer

Matt Jones
Senior DevOps Engineer

Martin Juckes
Head of CEDA-Atmosphere and deputy head of CEDA

Jack Leland
Software Engineer

Alex Manning
Research Software Engineer/Operations Support for JASMIN

Neil Massey
Senior Software Engineer

Alison Pamment
Data Scientist: metadata standards

Charlotte Pascoe
Senior Data Scientist: Models

Graham Parton
Senior Data Scientist: Observations

Sam Pepler
Head of Curation

Matt Pritchard
JASMIN Operations Manager

Elle Smith
Software Engineer

Ag Stephens
Head of Partnerships

Poppy Townsend
Communications Manager

William Tucker
Software Engineer

Alison Waterfall
Senior Data Scientist: Earth Observation

Matthew Wild
Senior Data Scientist: UKSSDC

Ed Williamson
Data Scientist: Earth Observation

International activities and collaborations

In the past 20 years, CEDA has built up a comprehensive network of relationships with other organisations on a national, European and global level. We have helped forge the direction and development of key activities, including ESGF - Earth System Grid Federation - a globally distributed data infrastructure - and IS-ENES, as well as providing significant input into the Copernicus Climate Change Service programme. This section of the report provides examples of work carried out at CEDA in support of internationally recognised and supported activities. It also includes some collaborations that were carried out in support of UK science. This section demonstrates that much of our work takes place within the wider global context.

Publishing IPCC datasets: being FAIR with climate data

ELLIE FISHER, CHARLOTTE PASCOE

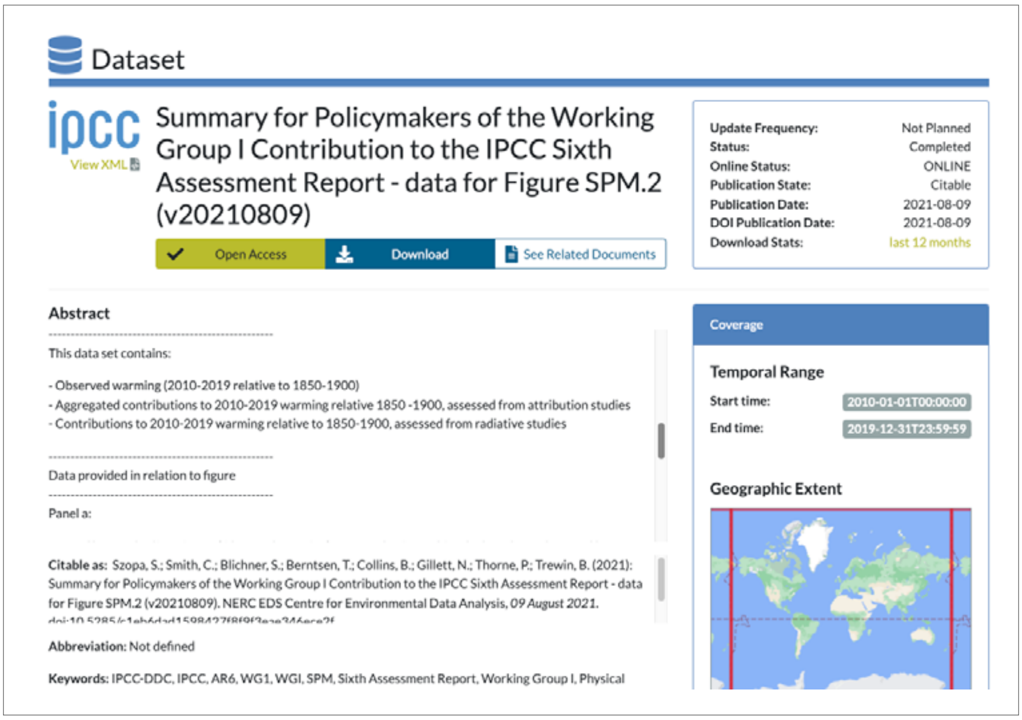
CEDA has been working with the Intergovernmental Panel on Climate Change (IPCC) Technical Support Unit (TSU) for Working Group 1 since December 2019. The latest IPCC publication is the 2021 Sixth Assessment Report (AR6), which outlines the present state of the climate system, and future projections for climate trajectories along with their likely impacts on components of this system. Various datasets have been used to produce scientific figures for chapters of the report, drawing on model data from the 6th Coupled Model Intercomparison Project (CMIP6), historical observations and other sources of climate information.

CEDA is the primary archive for these figure datasets, and has been a key part of the process of preparing them for archival, long-term storage and dissemination to users (through the CEDA catalogue). Charlotte Pascoe (Senior Data Scientist) is the lead on this work, assisted over the course of the project by Ellie Fisher (Summer Placement Data Scientist) and Kate Winfield (Environmental Data Scientist). Close collaboration with the TSU and figure authors has enabled the publication of 74 datasets, each with a 'landing page' in the form of a metadata record. This record is the portal through which the stored data in the archive are accessed and provides a direct link to the data from the AR6 figure.

A driving force for this has been guidance from the IPCC Data Distribution Centre, which emphasises that data must be Findable, Accessible, Interoperable, and Reusable (FAIR).

- **Findable** – The catalogue enhances findability by recording key information which can be easily searched, and fully signposting each record by linking to the wider project (AR6) and providing relevant documentation such as the Digital Object Identifier (DOI) for the code, supplementary information and GitHub repository.
- **Accessible** – Scientific language is understandable, acronyms and specific terminology fully explained in metadata, such that it is completely interpretable.
- **Interoperable** – Archived data are from CMIP6 models and observational data, which is shared between projects. These data can be reused, shared and adapted elsewhere, with credit, under a Creative Commons Attribution 4.0 license. Variables follow CF conventions and have standard names.
- **Reusable** – There is code associated with each figure, which allows users to reproduce the figures from the report independently. The code is presented in the form of Jupyter notebooks or MATLAB/Python scripts, making it open-source.

This is a new paradigm for IPCC report publication which not only supports open science but also provides a citable platform to acknowledge the work of IPCC authors.



Dataset

ipcc View XML

Summary for Policymakers of the Working Group I Contribution to the IPCC Sixth Assessment Report - data for Figure SPM.2 (v20210809)

✓ Open Access | Download | See Related Documents

Update Frequency:	Not Planned
Status:	Completed
Online Status:	ONLINE
Publication State:	Citable
Publication Date:	2021-08-09
DOI Publication Date:	2021-08-09
Download Stats:	last 12 months

Abstract

This data set contains:

- Observed warming (2010-2019 relative to 1850-1900)
- Aggregated contributions to 2010-2019 warming relative 1850 -1900, assessed from attribution studies
- Contributions to 2010-2019 warming relative to 1850-1900, assessed from radiative studies

Data provided in relation to figure

Panel a:

Citable as: Szopa, S.; Smith, C.; Blichner, S.; Berntsen, T.; Collins, B.; Gillett, N.; Thorne, P.; Trewin, B. (2021): Summary for Policymakers of the Working Group I Contribution to the IPCC Sixth Assessment Report - data for Figure SPM.2 (v20210809). NERC EDS Centre for Environmental Data Analysis, *09 August 2021.* doi:10.5285/c1eb6dad1598427f8f9f3eae346eca2f

Abbreviation: Not defined

Keywords: IPCC-DDC, IPCC, AR6, WG1, WGI, SPM, Sixth Assessment Report, Working Group I, Physical

Coverage

Temporal Range

Start time: 2010-01-01T00:00:00

End time: 2019-12-31T23:59:59

Geographic Extent

User view of the CEDA catalogue record for dataset SPM.2, showing assessed contributions to observed warming, with metadata relating to report figure.

Notebooks for Earth Observation: best practice and capacity development

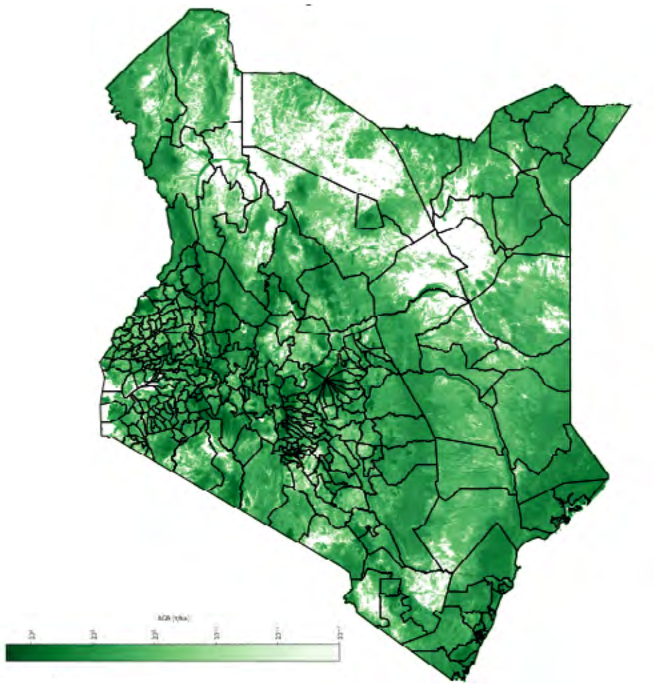
ESTHER CONWAY

A Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualisations and narrative text. Uses of Jupyter Notebooks within the Earth observation community include data cleaning and transformation, numerical simulation, statistical modelling, data visualisation, machine learning, and much more.

We are at a particularly exciting time with this technology where many archives are deploying notebook services. These services allow unprecedented access to petabytes of data allowing users from any part of the globe to engage with Earth observation data in a very powerful way. Jupyter Notebooks produced during a research project can very often be the best starting point for new users to engage with data deposited with an archive, however, this raises unique challenges. While notebooks can be a valuable resource, there are issues surrounding input data/processing/technical dependencies and quality. Poor quality notebooks with hidden dependencies may cause new users a lot of problems.

To deal with these issues, CEDA is leading work on Jupyter Notebooks with the Committee for Earth Observation Satellites (CEOS). This has involved conducting a number of surveys and running webinars on Jupyter Notebooks to gain a better understanding of the Earth observation community needs. We have partnered with various international organisations – CSIRO, NASA, JAXA, EUMETSAT, ESA, University of Alaska and Digital Earth – to demonstrate the broad range of notebook applications in the area of capacity development and training. We showcased a range of data analysis services on various different platforms:

- using the JASMIN Notebook Service
- exploring Open Data Cube on Google Earth Engine
- generating time series on the ESA PDGS (European Space Agency -- Payload Data Ground Segment)
- processing large volumes of data on the Earth Analytics and Interoperability Lab



This image is above-ground biomass in Kenya in tonnes per hectare. The data is on the CEDA Archive. The Jupyter Notebook reads data directly from the Archive and reprojects the tiff file, overlaying it with a shapefile containing Kenyan administrative regions to produce this image. This process makes it possible to extract key statistics on levels of above-ground woody biomass for various regions and provide reports to key government departments in Kenya.

Having engaged with over 500 people from over 50 countries, two core needs for the wider community became evident. The first was the need for a Jupyter Notebook's best practice document to support the creation and preservation of high-quality reusable notebooks. The second was the need for basic training to get the next generation of researchers ready to engage with emerging services. Going forward CEDA is leading work within CEOS to develop a best practice document and support training initiatives.

Mission planning for collaborative data services

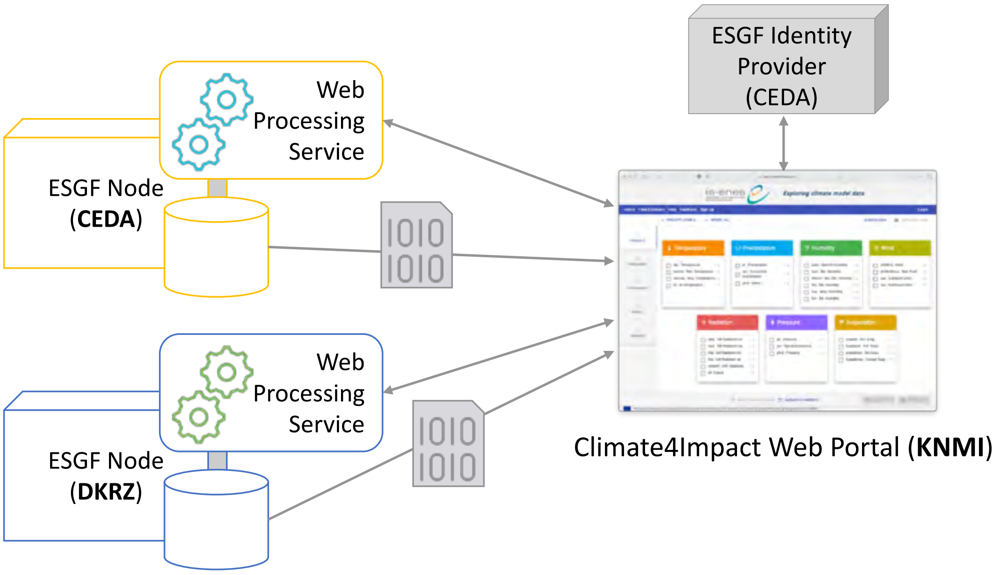
MARTIN JUCKES

CEDA led the development of the mission and objectives white paper of the European Network of Earth System Modelling (ENES) Climate Data Infrastructure (CDI). This white paper sets out the plans for provision of collaborative European data services to support the curation and dissemination of World Climate Research Programme (WCRP) climate model simulations. The paper was written by experts from within the ENES consortium over a period of 12 months, with frequent online meetings to maintain momentum and resolve any differences of opinion over scope and priorities.

The development of the white paper has been important to build a consensus for the partners in the consortium and provides the foundation for efforts to develop sustainable funding for the ENES CDI.

Web-based data analysis with European Climate Data Infrastructure

PHILIP KERSHAW



Our Dutch partners in the IS-ENES3 project – KNMI - are providing a web portal which integrates web processing services running at two different sites – one operated at CEDA and a second at project partner DKRZ in Germany. These specialised services greatly assist users by enabling them to access only the subset of data they require.

CEDA together with other European partners collaborate in the development and operation of a distributed software infrastructure to support the dissemination of climate model data from key projects such as CMIP6 and CORDEX. The ENES (European Network for Earth System Modelling) CDI (Climate Data Infrastructure) has evolved over a series of EU funded projects culminating in the latest, IS-ENES3*.

Through the course of the current project we decided to initiate a new activity to bring together some new developments and demonstrate new capabilities. Firstly, the ESGF software upon which the ENES CDI builds has had a major upgrade (see article Upgrade for the Earth System Grid Federation). Second, we wanted to take advantage of the subsetting services being utilised for C3S which we have described in the section above. Finally, in the wider context of the climate / Earth sciences research community new capabilities are coming to the fore with the use of public cloud (e.g. Amazon Web Services, Microsoft Azure, Google cloud) and web-based platforms for data analysis such as Jupyter Notebooks.

We therefore designed an activity to bring together all these new capabilities into an integrated demonstration system.

The figure illustrates this. On the left the focal point is the Climate4Impact web portal. This enables users to search for and select climate model data which can be staged to public cloud storage for further analysis with Jupyter Notebooks. On the left, it shows subsetting services hosted at participating sites CEDA and DKRZ. When the user selects data for analysis, instead of all the data needing to be copied to the cloud, the subsetting services cut out only the pieces of data that are needed. This minimises the amount of data that needs to be transferred and reduces the amount of cloud storage needed. Both CEDA and DKRZ services use the new version of the ESGF software developed.

* This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°824084

Subsetting service for C3S climate datasets

AG STEPHENS, ELLE SMITH

CEDA has developed a subsetting service for the Copernicus Climate Change Service (C3S) that underpins the operational C3S Climate Data Store. Accessed by services and scientists from around the world, the CEDA-built service exposes CMIP and CORDEX climate simulation data via a remote web processing service. CDS users can select a subset from a large ensemble of simulations – by choosing a time range, vertical levels or a spatial region. The underlying framework, known as 'roocs', is a versatile set of services and libraries written in the Python programming language.

Roocs is also being employed with ENES European research infrastructure and will be extended for use within the Earth System Grid Federation in the near future.

Championing data stewardship in the Research Data Alliance

GRAHAM PARTON

CEDA staff have taken part in research data management for the environmental sciences for over 25 years, providing dedicated long-term curation services and support to the scientific community both in the UK and globally. This has often seen CEDA at the forefront of developing services, tools, standards and practices around research data management, giving us a depth of experience, knowledge and skills well-respected in the global community.

We're also keen to share such experiences with the wider research data management community, engaging in shared projects and presenting at conferences, plenaries and workshops. One important community is the Research Data Alliance (RDA), 'a community-driven initiative … with the goal of building the social and technical infrastructure to enable open sharing and re-use of data.'

Over the past year, two CEDA staff members have been contributing to a number of RDA groups. Via the 'Professionalising Data stewardship' Interest Group, CEDA have been furthering the standing of research data management as a whole and aiding career pathways for staff, specifically leading on the Models of Data Stewardship task group. This task group ran a community survey in October 2021, releasing its report and dataset as valuable community resources for review this year. Meanwhile, via the 'InteroperAble Descriptions of Observed Property Terminologies' (or I-ADOPT) Working Group, we have been aiding the I-ADOPT recommendations to aid interoperability between cross-domain terminologies… see page 17 for more on this!

Through these present activities and a raft of previous interactions with RDA and other communities, CEDA remains not only engaged with the wider research data management community but continues to enhance those conversations, drawing on a wealth of expertise for the common good.

Engagement with our core user communities

Environmental science, and the climate crisis in particular, requires increasingly multidisciplinary research. Our core communities have historically been predominantly UK researchers funded by the Natural Environment Research Council (NERC) - usually working within atmospheric or Earth observation disciplines. However, our data and services are increasingly being used by people from a diverse range of research areas with varying degrees of knowledge about our data and services.

This section of the report provides some examples of how we have engaged with key members of our communities, improved services or workflows based on feedback, and offered training and support over the past year.

Workflows on JASMIN: engaging with earth observation users

AG STEPHENS, FATIMA CHAMI AND MATT JONES

In early 2012, with JASMIN newly commissioned and operational, the first users exercised this new resource via its LOTUS batch computing environment. One such user was stunned to find that processing that would normally have taken weeks to complete on their legacy system could be completed in a matter of days. Since then the service has seen huge growth with a myriad of different processing activities having been conducted on LOTUS from users across the NERC community and their partners in national and international collaborations. Alongside the intensive use of this computing resource there has been a succession of changes and enhancements to the system as new hardware and software systems have been introduced.

With the growth in LOTUS and its user base over the years there has been an ongoing need to analyse and understand users' workflows in order to better support the community. One such project was initiated in early 2022 with this express purpose.

The NCEO Workflows Analysis Project was a three month activity led by CEDA with the aim of engaging directly with scientists from the National Centre of Earth Observation (NCEO) who use LOTUS and could help us identify some key patterns in their workflows. We also wanted to identify any current and historical problems and concerns and develop ideas and prototypes for overcoming problems.

The initial phase of the project was focussed around initiation and engagement. We did this by planning and scheduling a series of meetings and deciding on how to share and manage information. A GitHub project board and repository were set up and updated by CEDA. Meetings were hosted virtually on Zoom and brought together scientists working on different EO research projects and based at different institutions such as Reading University, King's College London, Leicester University and RAL Space Remote Sensing Group (RSG) at the Harwell campus.

The second phase was concerned with analysing these workflows and identifying common patterns and priority issues to be addressed. The scientists were asked to describe their workflows, in terms of inputs, outputs, and software. All meeting notes and project progress were managed on the private GitHub repository by CEDA. This way information was shared and enabled structured discussion.

Most of the workflows could be classified as simple parallel, with no dependency or need for communication between parallel tasks. Some were complex and included multiple sub-workflows. The common issues identified were related in part to the scheduler's response to a workflow and in part to the data size and movement (being input or generated).

Of particular importance is the storage type and location used for hosting the data. This can impact the read/write (known as IO) rate and efficiency during the runtime. The NCEO workflows often lacked an efficient mechanism for re-running jobs/workflows and fault tolerance.

Additional meetings with the NCEO scientists were carried out to review the findings and recommendations for future work, investigating different approaches for bringing data to JASMIN and for moving output data around on JASMIN. CEDA also developed a prototype benchmarking tool during the project. Re-use and co-development of libraries and scripts for interacting with the scheduler were considered and it was recognised that support is required for those re-running large workflows. Ideas for Net Zero targets and increasing efficiency were also discussed.

Based on our interactions, we gained some valuable insights and compiled a set of recommendations - investigating different approaches for bringing data to JASMIN from outside and also how data can be efficiently moved around the different storage classes within the system.

An important area for us to focus on for the future is the benchmarking of workflows so that we have the information needed to analyse what is happening and resolve any blocks or take advantage of where efficiency gains can be made.

The NCEO workflow analysis project was considered a successful initiative to engage with our users. It improved our understanding and we intend to embed the findings in new training and help materials to benefit other users.

Ideas for future engagement and consultation with other communities to improve JASMIN and shape its future are of particular interest with the growing number of users and diverse communities.

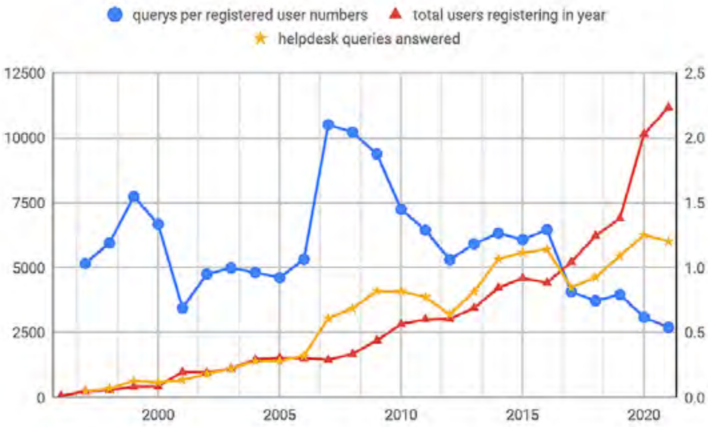
CEDA helpdesk over the decades

GRAHAM PARTON

CEDA's origins stretch back to around 1996 with the emergence of the British Atmospheric Data Centre (BADC) from its precursor the Geophysical Data Centre (GDF). Ever since these early days, a key part of CEDA's service has been to provide user support. The scale and nature of that support has changed over the years (see timeline) – by the end of 2020 CEDA had answered over 71700 user queries in total.

The graph shows the number of queries has increased over the years, inline with the growth of our user community. The evolution of this service began with a shared mailbox and associated PERL scripting before moving to using dedicated commercial software – locally installed and managed initially but now making use of a commercial cloud based service to give added resilience.

Support has also evolved to greater integration with documentation and 'help beacons' to provide in-context self-service user support, but always underpinned by a dedicated team of CEDA staff available via email or phone. Where we'll be in another 5 years' time remains to be seen, but one thing we do know is that user support will remain at the heart of what we do.



Graph showing the growth in numbers of queries and registering users per year (left axis) and query-to-registering-numbers ratio (right axis).

TIMELINE

1996
BADC is formed - support by shared email box, evolves to use PERL scripting in bespoke tool

2003
Move to commercial 'Footprints' helpdesk system

2005
'CEDA' formed by the merger of BADC and NEODC

2007
September BADC and NEODC helpdesk merged into combined 'CEDA helpdesk'

2011
CEDA wins contract to operate CEMS helpdesk (CEMS is a JASMIN like service)

2013
CEMS helpdesk merged into CEDA helpdesk

2016
Switch to Help Scout cloud based helpdesk system with integrated documentation site

2017
Creation of dedicated 'data management' helpdesk to manage correspondence with data providers, CEDA helpdesk serves end-users and JASMIN service users

2020
dedicated JASMIN helpdesk created to split out support from main CEDA helpdesk

2022
CEDA starts to provide support for a central NERC Environment Data Service helpdesk

Sharing data deposit services

SAM PEPLER

The Arrivals Service is a key component in CEDA's infrastructure providing the entry point for users to deposit data in the archive. As part of the Environmental Data Service, we are starting to 'share' this service with the other data centres, for example the National Geoscience Data Centre. This year we have added the ability for the other data centres to review data deposited and trigger different ingest workflows. Next year we will try to improve the look and feel of our service so that it fits in more naturally with how the other data centres work.

Introduction to Scientific Computing Course

ALISON PAMMENT

CEDA staff have led the annual National Centre for Atmospheric Science (NCAS) Introduction to Scientific Computing (ISC) course since its inception in 2014. From their beginnings as a pilot presentation for a 3 day course at the University of York, the training materials have been refined and developed into a 5 day course covering:

- an introduction to using the Linux bash shell
- an introduction to programming in Python
- using Python to work with environmental science data
- software version control using git and GitHub
- an introduction to NCAS, CEDA and JASMIN
- the importance of data standards to data sharing

From 2016 the course has been held each November as a classroom-based event hosted by the NCAS Operations Group at the University of Leeds. Unfortunately, the lockdown necessitated by the Covid-19 pandemic led to the cancellation of the 2020 ISC course. In 2021, CEDA took the decision to convert the ISC training materials to a form that could be presented virtually. We were very keen that this should not lead to a reduction in the standard of training in terms of either the course content, or the student experience.

A major strength of the ISC is its practical element. Each topic is introduced with a presentation lasting approximately 20-30 minutes, immediately followed by a series of exercises undertaken by the students to put their learning into practice. Much of the work in converting the course to a virtual event was the rewriting of the practical exercises using Jupyter Notebooks, enabling the students to work within a standard software environment running on the JASMIN system. The presenting team undertook practice Zoom sessions to ensure smooth running of the course for a large group. We also used the Slack instant messaging system to allow students to type individual questions and make the answers visible to all participants, thus avoiding the need for trainers to answer similar questions repeatedly.

Some of the students and course organisers who attended the first virtual ISC course.

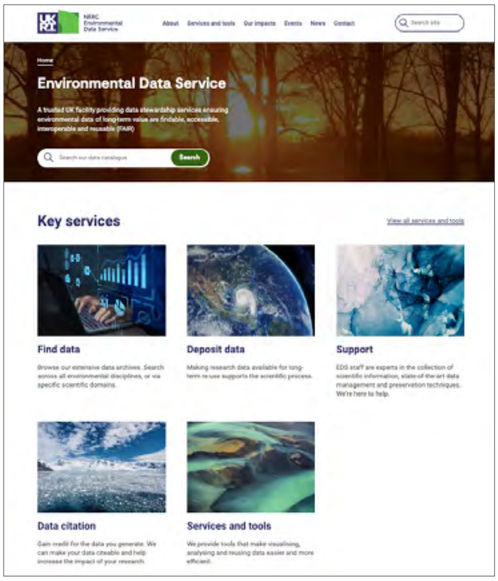
The first presentation of the virtual ISC took place in November 2021 with 36 students in attendance. Many were PhD students from UK universities, and the virtual format facilitated the attendance of overseas students from locations including Hong Kong and Mexico. Student feedback from the course was encouraging, with many attendees saying they had learnt useful skills and that the course itself was well organised. The feedback will also be used to inform continued development of the virtual ISC, which will be presented again in November 2022.

Raising awareness about the Environmental Data Service

POPPY TOWNSEND

The NERC Environmental Data Service (EDS) provides a focal point for scientific data and information spanning all environmental science domains. The EDS is made up of a network of distributed data centres across NERC, each with domain specific expertise; we (CEDA) are the experts in atmospheric, Earth observation and solar and space physics.

Over the last year, we have been working closely with EDS colleagues to raise awareness and communicate about the work we do. A five year stakeholder engagement strategy and implementation plan was developed. Various activities have been supported or led by CEDA staff - including writing content for the new website, developing impact stories, sharing help desk abilities, and coordinating how we collect reporting metrics more efficiently.


UKRI
NERC Environmental Data Service
About
Services and tools
Our impacts
Events
News
Contact
Search site
Home
Search our data catalogue
Search
View all services and tools

Environmental Data Service

A trusted UK facility providing data stewardship services ensuring environmental data of long-term value are findable, accessible, interoperable and reusable (FAIR)

Key services

Find data

Browse our extensive data archives. Search across all environmental disciplines, or via specific scientific domains.

Deposit data

Making research data available for long-term re-use supports the scientific process.

Support

EDS staff are experts in the collection of scientific information, state-of-the-art data management and preservation techniques. We're here to help.

Data citation

Gain credit for the data you generate. We can make your data citeable and help increase the impact of your research.

Services and tools

We provide tools that make visualising, analysing and reusing data easier and more efficient.

The development of a new website which was launched in early 2022 was a key activity we were involved with.

Efficient data movement at scale: ARCHER2 to JASMIN

MATT PRITCHARD

Users of the ARCHER2 supercomputer generate large volumes of data that need to be moved quickly in order to enable the next user to make best use of expensive high performance computing (HPC) time. The JASMIN team help ensure that data transfers between ARCHER2 and JASMIN can happen as efficiently as possible.

With fast networks connecting UK DRI (Digital Research Infrastructure) facilities like the ARCHER2 supercomputer and the JASMIN data analysis facility, moving data around should be relatively easy these days. However, as with any task, it's important to choose the right tool.

Familiar data transfer tools like scp, sftp and rsync are widely used and convenient, but while they work well for small amounts of data within a local network (e.g. within an organisation or site), they have in-built limitations making them poor choices for long distance data transfers. Even if the network connection has a high bandwidth, they are not equipped to make best use of it, particularly as distances increase.

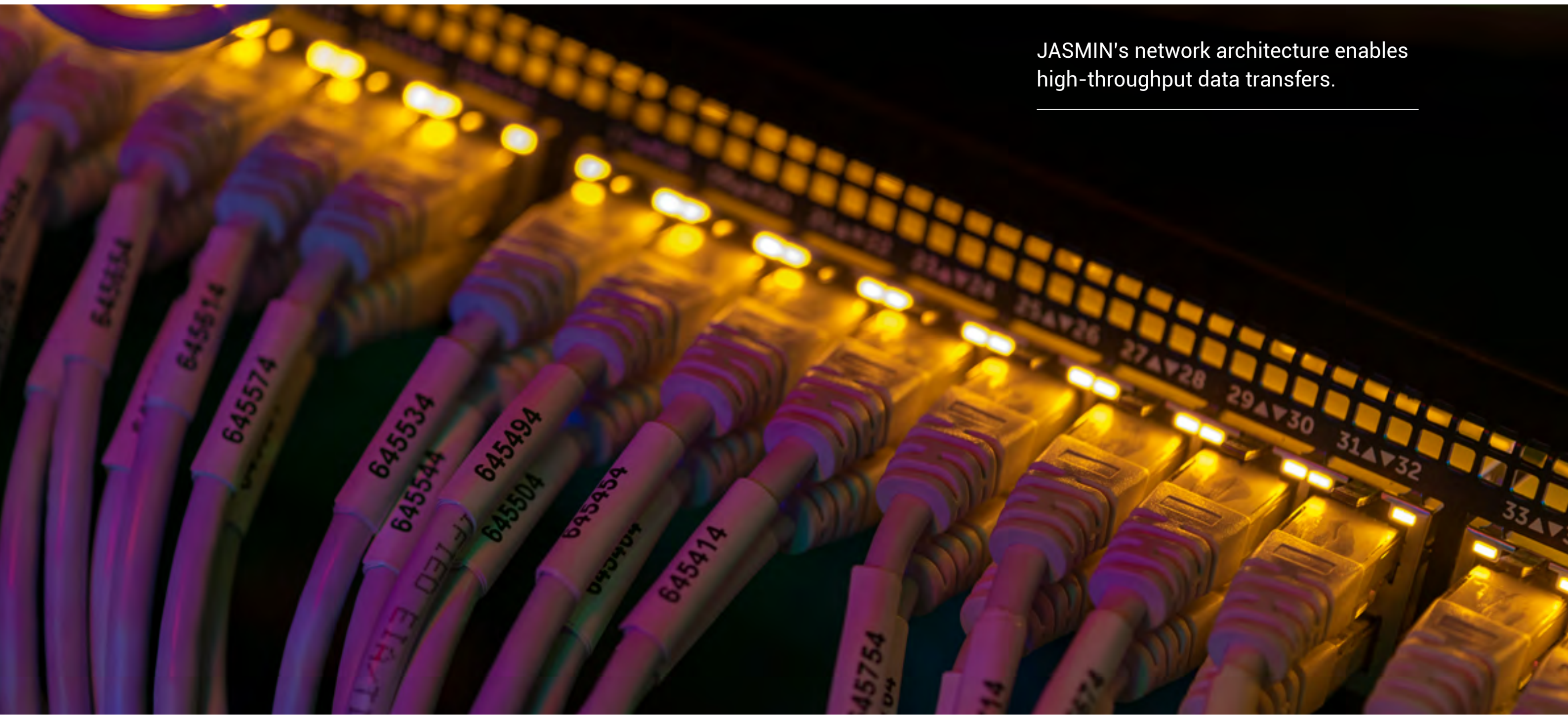
Parallel data transfer tools like Gridftp and Globus have been around for a while. It's an open secret among those who move big data around that if you're willing to invest the time in learning how to use them, these tools will save you time overall. They have features that enable multiple files to be transferred at once, but also the ability to use multiple network streams for each file, parallelising the transfer and making better use of the available bandwidth.

The JASMIN team worked with members of the NCAS Computational Modelling Support (CMS) team as well as those at Edinburgh Parallel Computing Centre (EPCC) involved in operating ARCHER2, to ensure that the right tools were in place and that the configuration at both ends supported how users wanted to use them in their workflows.

The tool of choice in this case was Gridftp, but the challenge was how to enable transfers to be spawned throughout long-running workflows on ARCHER2 (sometimes over the course of a month: much longer than the user would remain logged-in), while still maintaining security for users of both systems.

A solution was found whereby a temporary digital certificate could be granted to JASMIN users of ARCHER2, but the certificate lifetime could be set to a longer period. When used in a long-running workflow, this pre-authorised the 'user-not-present' transfers to take place at various points during the workflow. Meanwhile, the team looked at the options used with the Gridftp client, optimising the choices for the particular route and destination file system on JASMIN. This involved choosing the number of parallel network streams, how many files to transfer at once, and options to synchronise directories from source to destination.

The resulting advice was captured in documentation now available, for use by other JASMIN users of ARCHER2 but also applicable to other contexts. Further work is underway to create training exercises using high-performance data transfer tools and to expand the existing documentation to include some of the more sophisticated features.



JASMIN's network architecture enables high-throughput data transfers.

Ensuring the quality and accessibility of the CEDA Archive

It is vital that users of the CEDA Archive can be confident that it is being maintained to the highest standards. In June 2021, CEDA was certified as a Trustworthy Data Repository by the CoreTrustSeal Standards and Certification Board, a standards body created by the World Data System of the International Science Council.

We are working continuously to ensure that data in the CEDA Archive comply with the FAIR data principles, namely that our data should be Findable, Accessible, Interoperable and Reusable (FAIR). Here we give some examples of how our work is helping to deliver FAIR data, and thus supporting our user community to publish and share their work according to Open Science principles.

Aircraft data: providing long term access

WENDY GARLAND

Airborne research flights are an effective and flexible way to acquire direct in-situ and remote sensed atmospheric and earth observation measurements. A wide range of research is supported, including, but not limited to:

- Sampling and analysing specific air parcels and plumes after fires, such as; industrial (e.g. Buncefield), Yorkshire moors, forest fires, and after volcanic eruptions and dust events
- Observations around storms and extreme weather events to better understand the processes and development
- Meteorological and chemistry observations to improve atmospheric models and therefore forecasts
- Testing, calibration and validation for satellite instrumentation

These data are high quality, accurate and tailored to the specific scientific needs. As flights are expensive in both monetary terms and staff effort, it is essential that the data collected are used and exploited to their full potential. CEDA provides long term access to aircraft data in widely accessible formats to ensure long term reuse.

The CEDA archive holds in-situ atmospheric and remote sensing data from over 2500 flights from a range of aircraft (as shown in the table below).

Aircraft	Dates	Number of flights
Facility for Airborne Atmospheric Measurement (FAAM)	2005-present	>1300
NERC Airborne research Facility (NERC-ARF)	1981-2019	1100
British Antarctic Survey (BAS) twin-otter aircraft	2011-present	60
EUropean Facility for Airborne Research project (EUFAR)	2008-2017	169
Met Office Research Flight (MRF)	1993-2000	70

Because the scientific aim of each flight is different, the choice of aircraft (manoeuvrability, range, payload, etc), the geospatial coverage and choice of instruments operated (therefore which data files are produced), varies between flights.

Scientific data collected onboard during a flight is paired together with accurate geospatial coordinates of the aircraft platform. Additional data may be collected to support the accuracy of the data and build a more complete picture of the scientific phenomenon taking place. For example, vertical profiles made by sondes dropped from the aircraft at intervals, air samples collected for later analysis in the laboratory or corresponding ground-based measurements.

Once the flight is complete, data are downloaded from the aircraft and processed by the instrument specialists before being uploaded to the CEDA Archive. Although processed and deposited separately, all data collected during a scientific campaign are interrelated and codependent on the geospatial information. When data is deposited at CEDA - potentially by multiple different scientists - the CEDA data scientists undertake work to regroup and store all available data together as one dataset per flight. The delivery time between collection and archival can vary greatly by instrument and aircraft due to the complex processing involved.

Each flight is flown for a specific project with a different set of users and potentially different access restrictions. This all means that from CEDA's point of view, aircraft data are a complex beast to archive and curate.

To make curation of flight data workable and to meet the FAIR (findable, accessible, interoperable, reusable) principles we have strived to make our aircraft datasets as uniform as possible and we store them in well structured, open and documented archives to facilitate access and reuse. We require the data to be supplied in community agreed standard formats with a file naming convention for interoperability, and we have automated ingestion processes to put the data in the correct archive location whenever (and from whoever) it arrives. Each flight has a catalogue record linking details of the aircraft, the instruments and the project it was flown for together with geospatial and temporal metadata making it findable.

As every flight is unique and flown over a specific geographical area to fit the science aims, it is useful to display flight paths in a map application. The flight finder tool was developed at CEDA during the EUFAR project. It has a user interface using Google maps that also supports various searches such as date ranges and keyword search. Search capability is underpinned by a database using Elasticsearch. This is another way to make the data findable and links the user from the map to the data in the archive.

The benefit of these long term accessible archives is that aircraft data can be reused, for example a paper published this year uses FAAM data from 2013 to look at cloud processes that may affect accurate weather prediction. CEDA's work to make data FAIR ensures that duplication of expense and effort can be avoided – thus meaning essential science can easily be completed.

Improving data interoperability using scientific vocabularies

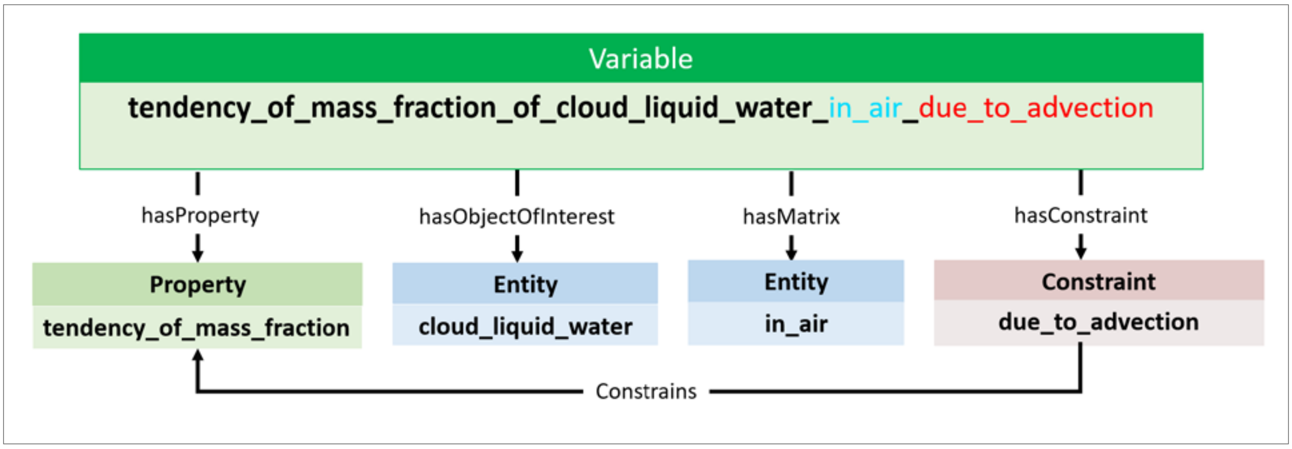
ALISON PAMMENT

Scientists searching for data related to their research often need a way of finding out which scientific variables, such as air temperature or sea ice thickness, are contained within the many available datasets. Lists known as 'controlled vocabularies' are used to standardise the labelling of these variables. Using these labels, a data user can search across datasets from different sources and find all those containing a variable of interest.

Within the atmospheric science community, the CF (Climate-Forecast) Metadata Conventions are widely used to give a detailed description of the content of data files. The CF conventions contain a controlled vocabulary, known as CF Standard Names, for variable naming. This vocabulary is maintained by CEDA staff as an important service to the international scientific community. CF standard names currently list more than 4,500 scientific variables, many of which relate to climate modelling and observations, and this number is steadily increasing.

Controlled vocabularies have also been developed within other sections of the environmental science community, covering domains such as oceanography, ecology and biodiversity. The purpose is to produce data conforming to the FAIR principles (Findable, Accessible, Interoperable, Reusable) but the use of different vocabularies by different communities can present a barrier to interoperability. CEDA staff worked as part of an international group, known as I-ADOPT (InteroperAble Descriptions of Observable Property Terminology) to develop a framework that will permit searches across datasets labelled with different controlled vocabularies.

Many controlled vocabularies for scientific variables contain similar descriptive elements. Firstly, a description of the object that is being observed or modelled, such as a tree or a wave. In the I-ADOPT Interoperability Framework, this is called the 'object of interest'. Secondly, the 'observable property' (the measurement represented by the data) for example, height or temperature. The environment containing the object of interest, such as air or sea water, is known as the 'matrix'. There may also be additional 'constraints' on the variable, for example, focussing on a process such as evaporation or convection. Variable names from different controlled vocabularies can be broken down into their elements according to the I-ADOPT framework, enabling the identification of synonyms. This facilitates the building of search tools which return results using equivalent terms across many vocabularies, thus increasing the number of search results. The diagram shows a CF standard name that has been broken down into its elements according to the I-ADOPT framework.



An example of a CF standard name decomposed into the concepts of the I-ADOPT interoperability framework. Many elements of the CF grammar correspond directly to I-ADOPT concepts.

NCAS data project: making the data pipeline FAIR

GRAHAM PARTON

> “Whether the weather be fine, or whether
> the weather be not,
>
> Whether the weather be cold, or whether
> the weather be hot,
>
> We’ll weather the weather, whatever the
> weather,
>
> Whether we like it or not.”

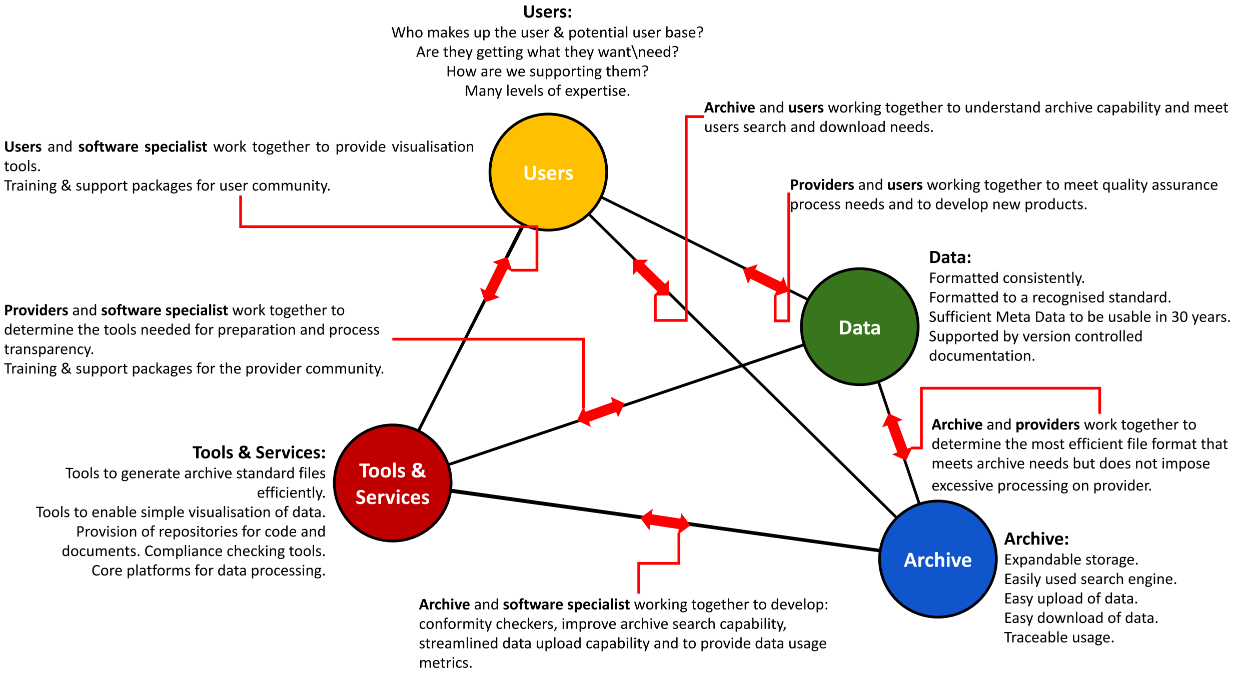
So goes the short poem by that most prolific of writers ‘Anon’… But actually knowing what the weather (and other atmospheric components) is doing, or more pertinently, has done is of key importance to the work of many environmental scientists. So much so that the National Centre of Atmospheric Science (NCAS) has a dedicated team of leading instrument scientists developing and maintaining a large suite of instruments from wind profilers to ceilometers; from airborne instruments and lab equipment, to ensure high quality data are gathered to underpin our understanding of the world around us.

However, as good as data may be, if they are not Findable, Accessible, Interoperable and Re-usable (the so called FAIR data principles) then these efforts and their potential value now and for years to come can become curtailed. Over the past few years a number of CEDA staff have been working closely with colleagues in NCAS - within the NCAS Data Project – to embed the FAIR data principles into practice along the length of the data pipeline: from instrument inlet to user download.

The NCAS Data Project has drawn together instrument scientists, IT staff, data curators and end-users to establish two key components: data standards and workflows, with their developments possible through a third key aspect: dialogue. This dialogue has ensured that as the standards and workflows are constructed they remain meeting the needs of those along the data pipeline. The data standards have built on existing metadata formats to ensure all the required details about the data are obtained and structured. This ensures the longevity and reusability of the data; whilst the workflows permit ease of data flows in a timely fashion.

To date, the project has developed new data standards for instrument and image files adding additional refinements complementing well established standards such as CF-conventions. These provide the necessary structured metadata that underpins end to end workflows from data production through to end curation and onto end-user workflows. These standards are already being implemented in the data production part of the data flows. We are working with our partners to establish the delivery and ingestion stages for final archiving with CEDA. Meanwhile, we are rounding off these efforts by building final data cataloguing and publishing workflows. All of which is enabled by the underpinning data standards.

The result: NCAS is able to ensure delivery of high quality data in a timely manner through to an archive that aids wider discoverability – in essence FAIR delivery of NCAS observations.



Users:
Who makes up the user & potential user base?
Are they getting what they want\need?
How are we supporting them?
Many levels of expertise.

Archive and **users** working together to understand archive capability and meet users search and download needs.

Users and **software specialist** work together to provide visualisation tools.
Training & support packages for user community.

Providers and **users** working together to meet quality assurance process needs and to develop new products.

Data:
Formatted consistently.
Formatted to a recognised standard.
Sufficient Meta Data to be usable in 30 years.
Supported by version controlled documentation.

Providers and **software specialist** work together to determine the tools needed for preparation and process transparency.
Training & support packages for the provider community.

Archive and **providers** work together to determine the most efficient file format that meets archive needs but does not impose excessive processing on provider.

Tools & Services:
Tools to generate archive standard files efficiently.
Tools to enable simple visualisation of data.
Provision of repositories for code and documents. Compliance checking tools.
Core platforms for data processing.

Archive:
Expandable storage.
Easily used search engine.
Easy upload of data.
Easy download of data.
Traceable usage.

Archive and **software specialist** working together to develop: conformity checkers, improve archive search capability, streamlined data upload capability and to provide data usage metrics.

The NCAS Data Pyramid – 4 key components, driven through dialogue and interactions.
Diagram courtesy of Dr Barbara Brooks, NCAS Director of Scientific Services, Facilities & Training

New techologies

At CEDA we are constantly appraising and evaluating our practices in providing storage and cataloguing of environmental data. Here we elaborate on a few different approaches, which use leading edge technologies, we've taken to using in the last year.

Software upgrade for the Earth System Grid Federation

ALAN IWI AND WILLIAM TUCKER

The Earth System Grid Federation (ESGF), is an international collaboration effort which provides a globally distributed software infrastructure for the dissemination of outputs from important climate models. Last year we reported on the future architecture, a collaboration between technical representatives of the participating organisations in ESGF to redesign the overall system to improve the maintainability and scalability of the data and its search services. ESGF software is run at sites all over the world. We are now at the point where the implementation of many of the new features has been completed and it is being installed at the first sites.

Flexible and scalable deployment

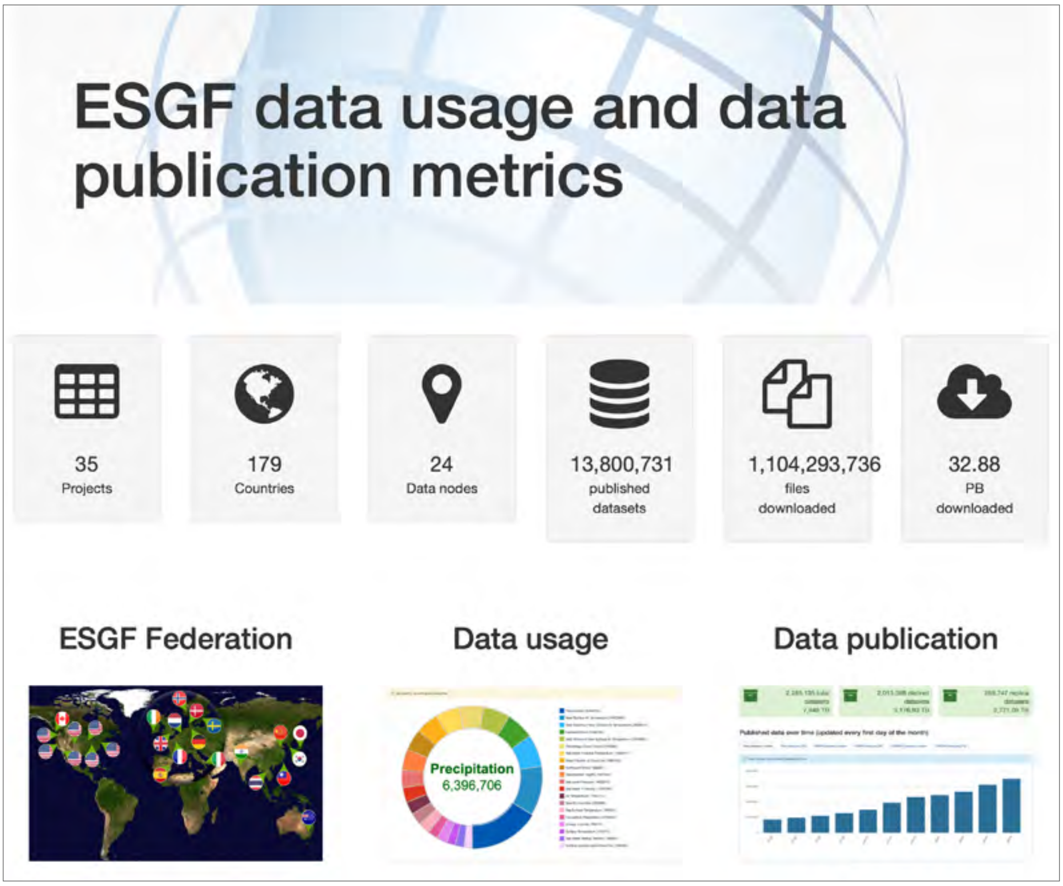
For the new system applications are deployed as Docker containers. This allows for a platform-agnostic deployment which can run on almost any operating system, or across a cluster of computers using Kubernetes. This has facilitated the first deployment of ESGF on public cloud as well as supporting the existing model of installation at partner sites. Another benefit to this design is the possibility to easily scale-out services to meet spikes in demand – for example for file downloads.

Federated access

Users will be able to login to access restricted ESGF data products using their CEDA account, or accounts from any of the partner organisations. This is made possible using a new system for single sign-on using industry standard identity management technologies such as OpenID Connect and OAuth 2.0. Together with partners in ESGF, we are using cloud services to provide centralised hosting provision and resilience.

Next steps

Most components of the new architecture are already deployed at CEDA, replacing our legacy ESGF systems. Our focus now is demonstrating the new features to other ESGF partners and helping them with the installation process at their sites. A major future activity will be the integration of the new search system based on the STAC standard (see related article – Building on STAC as CEDA's future search system).



ESGF Dashboard shows the network of sites in the federation and statistics for data.

Credit: ESGF partner CMCC – http://esgf-ui.cmcc.it

Building on STAC as CEDA's future search system

RHYS EVANS, MAHIR RAHMAN, RICHARD SMITH & AG STEPHENS

CEDA hosts over 20 Petabytes of atmospheric and Earth observation data. Sources include aircraft campaigns, satellites, automatic weather stations and climate models, amongst many others. The CEDA Archive consists of mostly netCDF data but we also have other formats including historical data where the format is not immediately obvious from the file name and extension.

Over the years, we have successfully employed a range of standards to enable us to provide search services including, amongst others, Catalogue Services for the Web (CSW) and OpenSearch. Ideally we require flexible, scalable standards which allow us to expose the bulk of the CEDA Archive via faceted search. This can then be used to build user interfaces, enhance search services, and facilitate interoperability with user tools and our partners in national and international collaborations. We aim to create a full stack software implementation for other organisations with similar desires, including an indexing framework, API server, clients and vocabulary management.

STAC (Spatio Temporal Asset Catalogue) provides a well defined, developer friendly JavaScript Object Notation (JSON) Application Programming Interface (API) which defines the minimum specification to describe and search geo-spatial assets. The standard was developed with the Earth observation community as the primary beneficiary for the API. As members of the Earth observation, atmospheric science and Earth system modelling communities as well as for our own interests, we were interested to see whether STAC could fit our needs. Over the course of the last eighteen months having reviewed and analysed the STAC standard, we have concluded with a yes to this question and have embarked on a full implementation to deliver search capability for the CEDA Archive and as part of our international collaboration with the Earth System Grid Federation (ESGF). This has been greatly assisted by existing open source implementations and an active STAC developer community supporting it.

An immediate challenge for our work was to incorporate the necessary metadata needed for a STAC catalogue. This is analogous to the way search engines crawl the Web and index content from websites so that it is discoverable by their search search service. In our case, a software package called the STAC generator has been developed. This is run over the content of the CEDA Archive ands content is fed into a database running ElasticSearch, a NoSQL solution which is scalable to meet our needs.

STAC is designed to be minimalistic with the core specification requiring only space and time on a latitude, longitude grid in a normal 365-day calendar. This core is then extensible through a range of both core and community made extensions. CEDA has developed several extensions to meet the requirements that weren't met by STAC's core and community functionality.

Many of the projects stored at CEDA have specific vocabularies of terms which categorise data. To reduce the need for new users to understand exact terms, CEDA's STAC catalogue and API will be integrated with a vocabulary server that will allow search on both data collection-specific and general vocabulary terms.

At the time of writing, we are preparing a complete baseline index of the CEDA Archive. This will form the basis for a deeper integration with CEDA's service infrastructure. Alongside this activity, for ESGF, we have prepared a complete index of some of our climate modelling data mirroring the legacy ESGF search services which our new STAC-based system will replace. Our partners, DKRZ, have rolled out a deployment of our system for their own data holdings. We have also prepared a new dedicated ESGF STAC client which builds on top of pySTAC. We are working with the ESGF community to migrate the distributed infrastructure to use the new search API.

Groundwork on future storage and data centre operation

JACK LELAND AND NEIL MASSEY

CEDA holds a large archive of environmental data and, to be of use to the community it must be: (a) easily discoverable, (b) indexed appropriately (i.e. create a list of file locations and specific content about the files) and (c) stored efficiently. Working at the multi-petabyte scale these requirements can present considerable challenges.

Data on JASMIN is now stored across a wide variety of different systems reflecting the scale of the infrastructure and range of requirements. For example, data-intensive processing ideally needs fast disk, whereas for data accessed infrequently tape storage can provide extensive capacity that is very cost-effective. Similarly, object storage on JASMIN provides a large capacity but underlying disk-based media gives much better performance.

Here we explore a number of projects, within the CEDA team, where we are tackling some of the challenges of working at scale and developing data management solutions for the archive to better support our user community.

NLDS

The Near-line Data Store (NLDS) is being developed at CEDA to give users of JASMIN greater opportunities to use tape storage, by reducing the latency of tape (i.e. the time taken between making a request and it being executed) and abstracting away the complex tape interface.

The NLDS design uses object storage as the front end to the tape. Users write data to the object store, via the NLDS interface. The data is automatically written to tape for them by the NLDS system and also kept on the object store. When the object store is about to fill up, a policy based process decides which data to delete from the object store, while keeping it on tape.

The user can then retrieve data, again using the NLDS interface. The NLDS abstracts the management of the data on the underlying media either accessing direct from object store cache or pulling direct from tape storage.

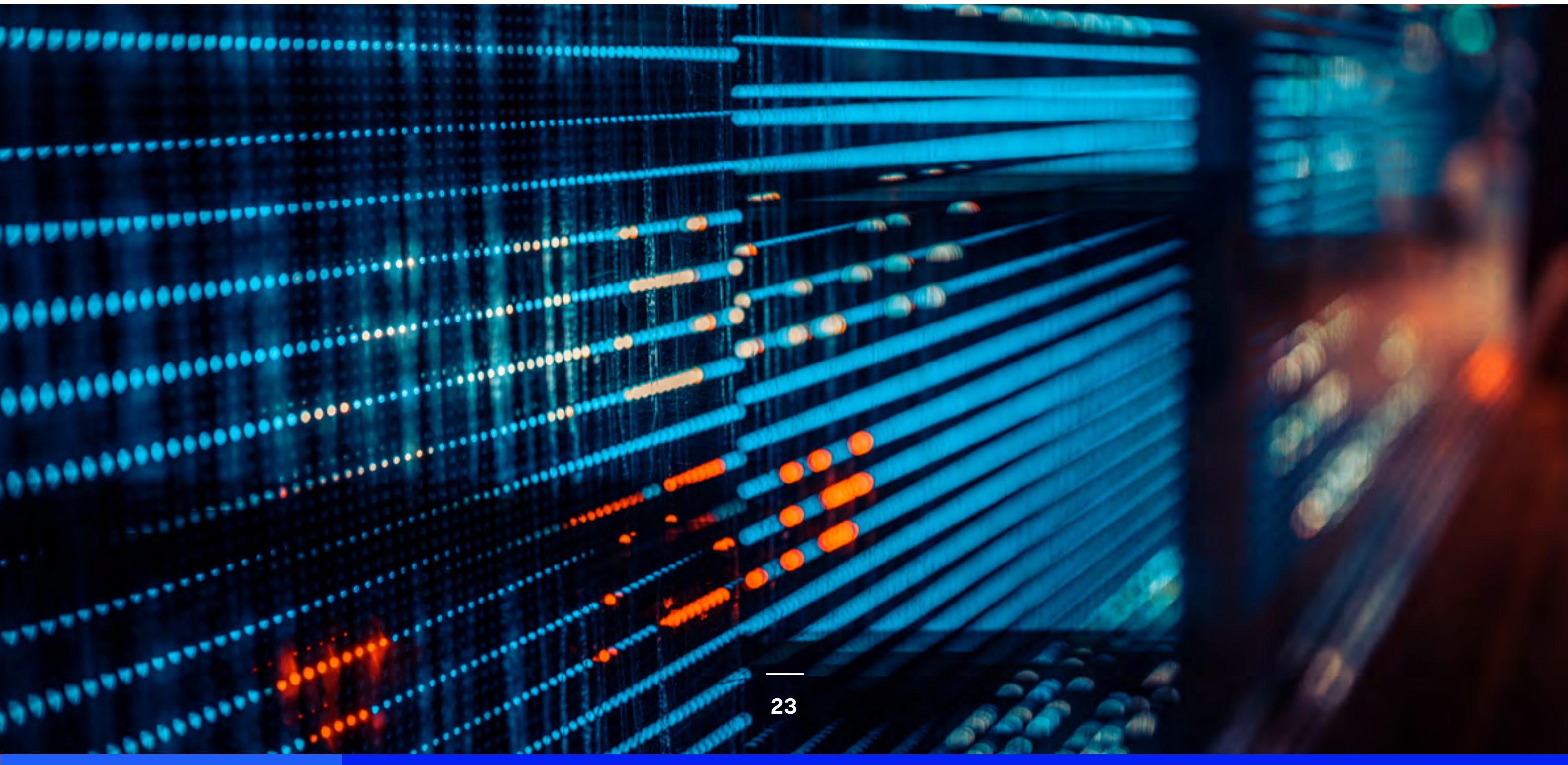
CEDA Archive App

The ceda-archive-app is a critical piece of software for the CEDA Archive which keeps an inventory of where the data is and which storage system(s) it is on. Over the last year major work has been carried out to re-engineer the software to allow it to manage data for different storage systems. This will be important as the underlying infrastructure on JASMIN grows, changes and new systems are added in the future.

CFA-C

The Climate-Forecast (CF) conventions provide a means of standardising variables and metadata for the netCDF data format, widely used for files that make up the CEDA Archive.

The Climate and Forecast Aggregation conventions (CFA), build on CF to provide a means of keeping an inventory of content across multiple files. This provides the potential to interrogate large volumes of data as a multi-dimensional hyper-cube. In the past year, in conjunction with colleagues at NCAS Computational Modelling Services, a new specification of CFA has been completed together with a reference software implementation and APIs (Application Programming Interfaces). This allows other software to easily use or integrate CFA, and importantly, the standard libraries from Unidata the creators of netCDF.



Responding to the climate emergency

UKRI has committed to reaching net zero by 2040 or sooner. CEDA is supporting a strong NERC effort to understand and inform responses to the climate emergency. This challenge will require coordination and collective action on an organisational level, across all the UKRI research councils as well as internationally.

This section describes three lines of work, dealing with science communication, monitoring and improving the energy efficiency of our own work, and leading a cross-UKRI effort to establish a robust pathway to carbon neutral digital research infrastructure.

Scoping net zero research computing

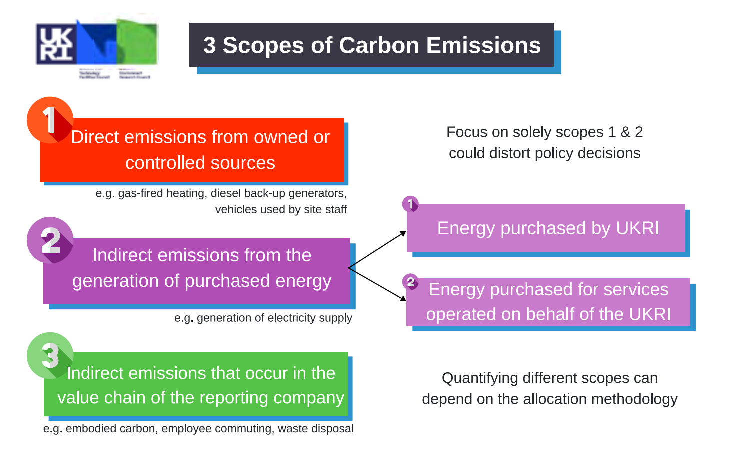
MARTIN JUCKES, AG STEPHENS, CHARLOTTE PASCOE, POPPY TOWNSEND, JENNIFER BULPETT, KATIE CARTMELL, LUCY WOODWARD, SOPHIE MOSSELMANS

Digital Research Infrastructure (DRI) refers to everything from software to data storage, from networks to the developers who maintain services. It's essential for our scientific research, but has a substantial carbon footprint due to its energy-intensive nature. Scientists across the UK research community have played and continue to play a major role in warning the world about the dangers of accelerating climate change. Now that there is a global consensus to take action, UK Research and Innovation should also play a leading role in responding to the advice of its scientists.

A new scoping project, delivered by a team in CEDA, will provide recommendations to enable UKRI's digital research infrastructure to reach net zero. The work targets all the UKRI computing facilities (owned and majority funded) and their user communities. The recommendations will be delivered to UKRI and to the DRI stakeholders – so that an appropriate action plan can be implemented.

A broad range of technical studies, surveys, community workshops and a literature review, are being used to gather evidence. A series of stakeholder meetings will be used to develop recommendations from the accumulated evidence. Other engagement work, such as an art commission, will be undertaken in order to encourage uptake and buy in from the research community.

The project started in November 2021 and will complete in June 2023. Interim results, based on the literature survey, have been published in a report called Complexity, Challenges and Opportunities for Carbon Neutral Digital Research. The report highlights the need to look beyond technical solutions and engage with users and all stakeholders to build a framework which can deliver net zero. The core team at CEDA will continue working with project partners, from 19 institutions, to deliver a comprehensive final report for UKRI.


3 Scopes of Carbon Emissions
1
Direct emissions from owned or controlled sources
e.g. gas-fired heating, diesel back-up generators, vehicles used by site staff
2
Indirect emissions from the generation of purchased energy
e.g. generation of electricity supply
3
Indirect emissions that occur in the value chain of the reporting company
e.g. embodied carbon, employee commuting, waste disposal
Focus on solely scopes 1 & 2 could distort policy decisions
1
Energy purchased by UKRI
2
Energy purchased for services operated on behalf of the UKRI
Quantifying different scopes can depend on the allocation methodology

The three scopes of emissions defined by the Greenhouse Gas Protocol form the basis of national and international reporting requirements. This is discussed in further detail in the project's interim report – emphasising that all scopes should be considered by UKRI.

Representation at COP26

POPPY TOWNSEND

The 26th UN Climate Change Conference of the Parties (COP26) was held in Glasgow from 1-12 November 2021. CEDA staff supported this important event in various ways.

COP26 hackathons

Three virtual hackathon events were run, during May and June 2021, by a range of UK Universities and the Met Office.

Over 150 attendees in various teams, investigated topics ranging from climate change to oceanography, biogeochemistry, and more. All data storage and analysis was performed on JASMIN. Access to climate data (such as CMIP6) held in the CEDA Archive, alongside the processing capabilities of JASMIN, were essential for these hackathons.

Representing UKRI in the Green Zone

Poppy Townsend, CEDA Communication Manager, supported the UK Research and Innovation (UKRI) exhibition stand in the public 'Green Zone' at COP26. She spoke to members of the public about the importance of data and digital infrastructures when tackling climate change.

Poppy Townsend standing in front of the art installation at the UKRI exhibition stand at COP26. The blue to red colours indicate global temperature increases, based on the show your stripes data.

JASMIN Power Metrics

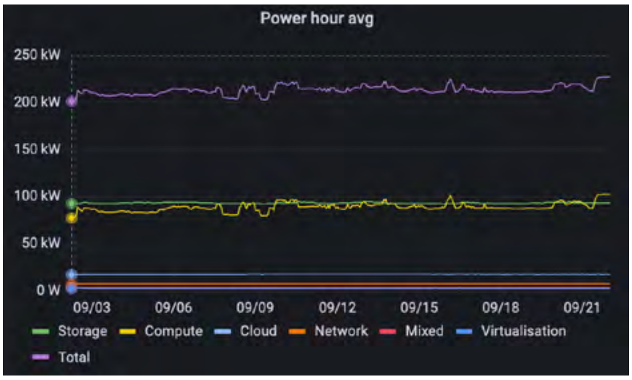
MATTHEW JONES

An important aspect of understanding the environmental impact of JASMIN is having metrics about the power consumption of the system. Especially useful for this is the ability to attribute power consumption for the different functional areas of the JASMIN infrastructure (for example: cloud, storage, and compute). Equally important would be the possibility of obtaining figures from the power distribution unit (PDU) for hardware associated with the different respective procurement phases over JASMIN's lifetime. In the last year, we have made an addition to one of our services to enable long term metrics of the power consumption of JASMIN to be stored and displayed.

This information existed in a database maintained by our colleagues in the Scientific Computing Department. However, the retention time was only about one day, important for day to day monitoring but not sufficient to build up a long-term picture. We believed it would be more useful to have a long running time series of the power consumption, especially as new phases of JASMIN are installed. In order to create a timeseries of the power metrics, which were available to view in a useful way, it required three pieces of work.

The first of these was for the technical JASMIN team to add tags to the PDUs in the database to describe the function of the servers in the rack. These tags are also referred to as roles and are: storage, compute, cloud, network, virtualisation, and mixed (where the rack contains multiple different roles).

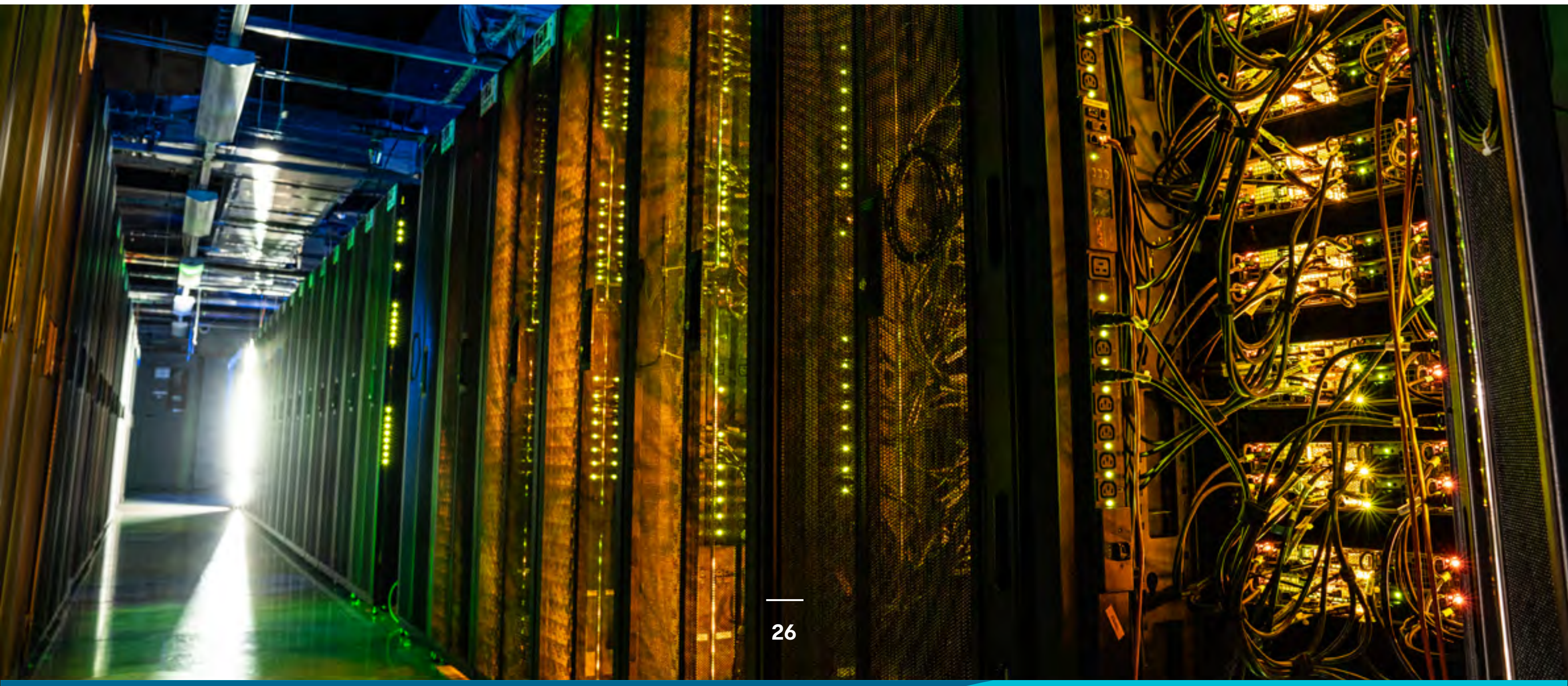
Secondly, the PDU's categories and their power consumption needed adding into a long-term storage location - so that the metrics could be easily retrieved from it when needed. This involved adding to the JASMIN Metrics service (a new service, still in development) to read from the data source for the PDU power consumption and cross-referencing the name of the PDU with another system to get the role of the PDU. The metrics that are gathered from the PDU are the instantaneous power consumption, and the average consumption over the last hour.



Example of hourly average power consumption for the different categories of JASMIN for a period in September 2022.

The final piece is to display this information in a consumable format. Currently an internal dashboard for the JASMIN Metrics system is being used to display the data. An example of the time series for the average hourly power consumption for July for the different rack roles in JASMIN is shown in the figure. The JASMIN Metrics system made the display, collection and storage of the data much simpler than it would have been had it not been in place.

The availability of these metrics has enabled a discussion about the power consumption of JASMIN in the JASMIN Management Committee and in the Net Zero DRI project. It has also highlighted the importance of having a readily available metric to monitor the power usage. These metrics are being actively used in the management of the system and are an essential contribution of our ability to monitor the environmental impact of JASMIN.



Hidden day to day operations

Many of the things CEDA does are to enable smoother operations of the group as a whole. A lot of these projects may not be obvious from an external perspective but they are critical in running a sustainable operation and can take considerable effort. Let's have a look under the hood ...

Access Instructor: managing access for 1,000's of datasets

RHYS EVANS AND SAM PEPLER

The CEDA Archive hosts a huge variety of data from many different source data providers and owners. Whilst many of our datasets have open access policies there are scenarios where data providers require licensing and various access conditions to be enforced. In these cases, we apply access control policies which may necessitate users authenticating with our services and registering and agreeing to terms of access for the data in question.

The 'Security Database', our tool for imposing access control for data held on the CEDA Archive, was very much due for an overhaul. It was based on software that had already reached end of life, meaning it was no longer actively being updated or patched. There were also a number of critical features that needed adding to allow automation and a better user experience. This was also an opportunity to simplify, as some requirements had changed since the code had first been written.

A new app was created and named the 'Access Instructor' as it tells each archive access system what the access rules are. Like the old Security Database, the Access Instructor would need to change the access to the Archive. However, we wanted this to be restricted as much as possible. It was therefore decided that the app be split in two. The instructor half would allow users to construct new and edit/remove existing rules. The other half, the Access Writer, would then execute these rules on the archive. To allow communication between the two halves a queue was used utilising the RabbitMQ messaging passing system. The Instructor writes to this queue, while the Writer reads from it decoupling the two services and allowing them to work independently. The system can then operate asynchronously and is more resistant to failures on either side. Additionally, having the Writer read from a queue rather than being directly contactable increases security.

The new app has increased the security of the system by both moving to more secure software and reducing the number of accounts with permission to change access to the Archive. The new code is also simpler and more readable. We hope this will allow any subsequent updates to be made more easily. The increase in automation has also simplified things for the staff who use this system. Permission and files are now updated and created automatically on rule creation removing the need for staff to "trigger" these changes. Finally, the Access Instructor's API allows other CEDA services to retrieve information on the licensing and permission for a given path. We aim to programmatically display this information to external users via our catalogue. This will make the access descriptions for restricted datasets more transparent and accurate.

Maintaining our services

ANDREW HARWOOD

CEDA currently runs over 80 separate public facing services and an additional 40 internal services. Managing this number of services requires a bit of organisation. We do this by maintaining a catalogue of service information, which includes information such as the staff member who 'owns' the service, plus appropriate documentation for fixing issues. We also use monitoring tools to check that our services are running and to record any downtime. If a service goes down we can use the catalogue to identify the best person to fix it so it can be restored as quickly as possible.

We regularly review our services to identify any that are no longer required so they can be closed down, releasing resources for new services.

Accounts system upgrade

WILLIAM TUCKER, RHYS EVANS AND ALEX MANNING

Building on our existing work on the JASMIN Accounts Portal, we're developing a suite of applications for managing user access to CEDA datasets and services. This new system is designed to make it simpler and more intuitive for users to request access to datasets, as well as streamline the back-end procedures around user approval and account verification.

We're also adding new single sign-on capability to CEDA accounts integrating OpenID Connect, an industry standard for such functionality. This will allow seamless interoperability between JASMIN and CEDA sites and the possibility for federated access to services.

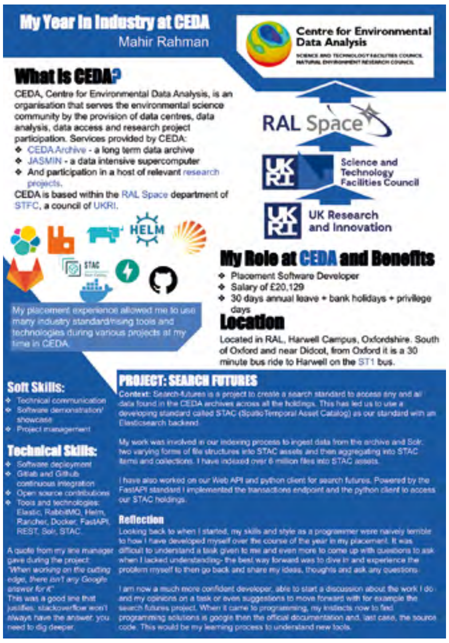
People and skills

The ability to make progress across the diverse range of activities described in this report is dependent upon having the right people, with the right skills, assigned to the right tasks. This in turn is reliant upon a combination of recruitment, retention and skills development. In response to a challenging employment market, CEDA have pursued a broad range of recruitment and engagement approaches which have been successful in attracting new talent, whilst targeted skills development has ensured that the team are well placed to make progress. CEDA benefits from the diversity of its team members, enabling a wide range of experience and perspectives to be brought to bear on the work.

A year at CEDA: the experience of an industrial placement student

MAHIR RAHMAN – UNIVERSITY OF YORK

As a Year in Industry during my degree (MEng Computer Science with Artificial Intelligence), I undertook a placement at CEDA as a software developer. My role was to work on creating code for projects, resolving GitHub issues, and any other assignments given to me by my line manager. My work contributed to collaborations working with organisations including the Met Office, European Space Agency and NASA developing software and systems to facilitate access to and analysis of environmental data.



My Year In Industry at CEDA

Mahir Rahman

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL
RAL Space
UKRI Science and Technology Facilities Council
UKRI UK Research and Innovation
HELM
STAC

What Is CEDA?

CEDA, Centre for Environmental Data Analysis, is an organisation that serves the environmental science community by the provision of data centres, data analysis, data access and research project participation. Services provided by CEDA:

- CEDA Archive - a long term data archive
- JASMIN - a data intensive supercomputer
- And participation in a host of relevant research projects.

CEDA is based within the RAL Space department of STFC, a council of UKRI.

My placement experience allowed me to use many industry standard/rising tools and technologies during various projects at my time in CEDA.

My Role at CEDA and Benefits

- Placement Software Developer
- Salary of £20,129
- 30 days annual leave + bank holidays + privilege days

Location

Located in RAL, Harwell Campus, Oxfordshire. South of Oxford and near Didcot, from Oxford it is a 30 minute bus ride to Harwell on the ST1 bus.

PROJECT: SEARCH FUTURES

Context: Search-futures is a project to create a search standard to access any and all data found in the CEDA archives across all the holdings. This has led us to use a developing standard called STAC (SpatioTemporal Asset Catalog) as our standard with an Elasticsearch backend.

My work was involved in our indexing process to ingest data from the archive and Solr, two varying forms of file structures into STAC assets and then aggregating into STAC items and collections. I have indexed over 6 million files into STAC assets.

I have also worked on our Web API and python client for search futures. Powered by the FastAPI standard I implemented the transactions endpoint and the python client to access our STAC holdings.

Soft Skills:

- Technical communication
- Software demonstration/ showcase
- Project management

Technical Skills:

- Software deployment
- Gitlab and Github continuous integration
- Open source contributions
- Tools and technologies: Elastic, RabbitMQ, Helm, Rancher, Docker, FastAPI, REST, Solr, STAC.

A quote from my line manager gave during the project: *"When working on the cutting edge, there isn't any Google answer for it"*
This was a good line that justifies, stackoverflow won't always have the answer, you need to dig deeper.

Reflection

Looking back to when I started, my skills and style as a programmer were naively terrible to how I have developed myself over the course of the year in my placement. It was difficult to understand a task given to me and even more to come up with questions to ask when I lacked understanding- the best way forward was to dive in and experience the problem myself to then go back and share my ideas, thoughts and ask any questions.

I am now a much more confident developer, able to start a discussion about the work I do and my opinions on a task or even suggestions to move forward with for example the search futures project. When it came to programming, my instincts now to find programming solutions is google then the offical documentation and, last case, the source code. This would be my learning process to understand new tools.

Poster snippet summarising Year in Industry experience.

When I started my placement at CEDA, I was working from home, so my induction took place virtually. Prior to my start, I had meetings with my would-be line manager, so I was fairly well introduced to who I would be working with. As a starter, there was an induction worksheet to introduce me to various tools and services, such as navigating and using the CEDA catalogue and using the JASMIN service.

After a number of small tasks to ease me in, the project that I substantially contributed towards and was a part of was 'Search Futures'. The premise of the project was to create a new way to search the CEDA data holdings. The solution utilises an existing open-source standard that is also relatively new and still evolving called Spatio Temporal Asset Catalog (STAC). My work within the project revolved around our use of STAC and ElasticSearch.

After spending a year at CEDA, I consider my placement to be very fortunate for its environment and shared personal ideals. The organisational structure felt very flat, and it was easy to approach anyone to ask for more information and context regarding a project. As a member of the Search Futures project, I felt I was treated like a fully qualified developer, where my thoughts were listened to, and I was given background information and autonomy related to the project. While working in my placement expedited my abilities as a programmer, I also had the opportunity to use industry-standard software and tools such as Docker and Kubernetes when it came to deploying code. Upon finishing my placement and returning to university, I have gained the adequate skills and competency that will help me excel even more in my degree. My placement at CEDA will let me take away my skills and experience gained alongside my accomplishments to apply in future work relevant to software development.

Supporting summer placement students

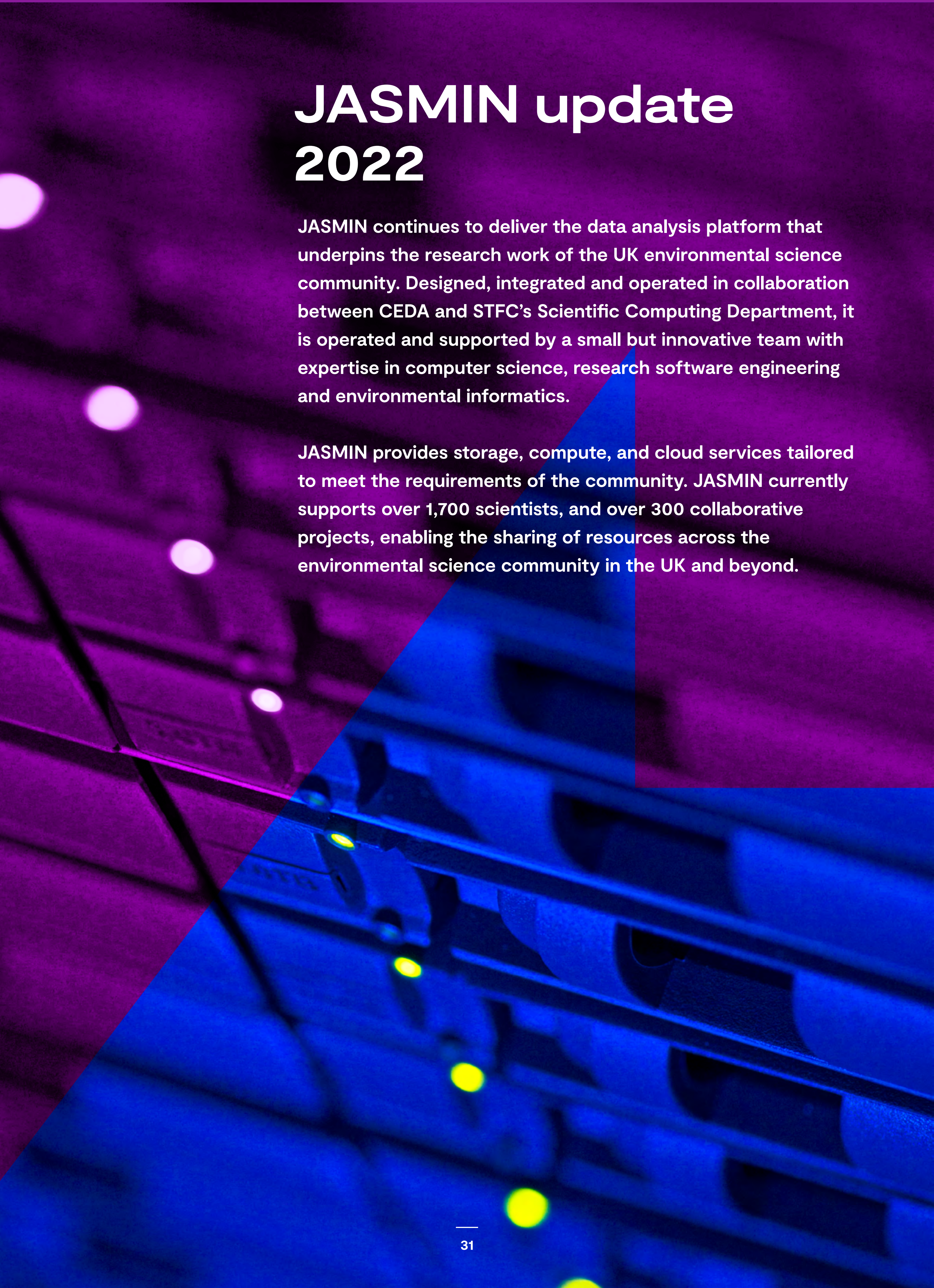
ADRIAN HINES

Harriet Gilmour used a summer break from 3rd year university undergraduate studies in Natural Sciences at Lancaster University to complete a 10 week placement within CEDA. This enabled her to gain an insight into life as an Environmental Data Scientist, and to gain experience in a field of science that she is keen to pursue in the future.

Harriet's work within CEDA was focussed on trialling the new NERC Model Metadata Catalogue and gaining a broader understanding of climate models. Recently, the NERC Environmental Data Service has been developing a new model metadata catalogue, bringing together all the information about the models into one place. The work involved testing the service out for climate model data that CEDA holds; primarily very large (many terabytes), complex climate models such as those involved in Coupled Model Intercomparison Project (CMIP6).

Harriet worked with her supervisor on field mapping, populating the new catalogue with example records, and testing options for the transfer of large amounts of information over to the new catalogue. The work was summarised in a feedback report for the creators of the new NERC Model Catalogue, as well as for internal use at CEDA. Harriet also presented the work to colleagues at CEDA.

The placement enabled Harriet to develop her Python and Linux skills, whilst also providing a valuable opportunity to speak to many other data scientists at CEDA, and find out more about what their roles involve. Harriet said that the placement was "so enjoyable and valuable", and that she was "excited to see where all these new skills I've learnt take me in the future!"



JASMIN update 2022

JASMIN continues to deliver the data analysis platform that underpins the research work of the UK environmental science community. Designed, integrated and operated in collaboration between CEDA and STFC's Scientific Computing Department, it is operated and supported by a small but innovative team with expertise in computer science, research software engineering and environmental informatics.

JASMIN provides storage, compute, and cloud services tailored to meet the requirements of the community. JASMIN currently supports over 1,700 scientists, and over 300 collaborative projects, enabling the sharing of resources across the environmental science community in the UK and beyond.

JASMIN UPDATE 2022

MATT PRITCHARD AND ADRIAN HINES

Day-to-day operation of JASMIN is a significant undertaking given the complexity and scale of the system. User support generates steady demand, with over 2,500 helpdesk queries resolved during the past year. Alongside the responsive support, the JASMIN team also delivered a series of user seminars, continued the delivery of regular training workshops (adding new 'advanced' content, with further material planned), and supported 3rd party training events including the HydroJULES summer school.

Alongside the effort required for day-to-day operation of JASMIN and user support, there have been two key drivers for progress during this period: an evolution of the service in response to user demand, and the exploration of longer-term direction.

During the year, progress was made with the following enhancements to the service:

- GPU cluster deployment: Following a successful proof-of-concept with a smaller deployment, a full-scale cluster of 2x8 NVidia A100 nodes and 14x4 NVidia A100 GPU nodes has now been integrated and is now being trialled with invited users, to help define software and configuration requirements for use with machine-learning and AI workflows.
- LOTUS upgrade: Retired CPU nodes in the LOTUS cluster were replaced with 16 x 48-core nodes each with large (1TB) memory capacity.
- Cloud upgrade: An essential upgrade of the underlying software supporting the JASMIN Cloud is now underway, bringing the version up to date to enable full vendor support to continue.
- DASK capability for the LOTUS cluster: Developments are underway to add a DASK gateway capability to LOTUS, which will enable user-friendly and dynamic parallelisation of computational tasks with minimal code modification.
- Development of a new Near-Line Data Store (NLDS) is now in progress at CEDA in collaboration with the University of Reading. This will eventually replace current tape storage interfaces with a system designed to enable seamless use of different storage types (disk, object store and tape) within user workflows, essential for better cost- and carbon-efficiency of JASMIN storage.

In preparation for the future, approval was granted for appointment of a new Director of JASMIN, providing additional leadership capacity to support the future development. Two key considerations are shaping the thinking about the longer term direction of JASMIN. Firstly, an increased awareness of the environmental impact of services, focussed through the UKRI Net Zero Digital Research Infrastructure scoping project (see page 24), has led to the production of power usage metrics for JASMIN as a first step towards understanding and working to reduce impacts. Secondly, a UKRI focus on a coherent cross-research council digital infrastructure has motivated work on the potential expansion of JASMIN to serve other communities, including a study completed by Jisc that made recommendations for future pathways to expansion.

Metrics and finance

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.

Usage of CEDA Data

CEDA delivers Data Archive services for the National Centre for Atmospheric Science (NCAS), the National Centre for Earth Observation (NCEO) and the NERC/STFC funded UK Solar System Data Centre (UKSSDC), as part of the NERC Environmental Data Service (EDS).

Annual CEDA archive usage: April 2021 – March 2022

Total number of users	149,815
Total data downloaded	913 TB
Total number of accesses	103,789,965
Total days activity	308,381

Note that a considerable amount of use of CEDA Archive data 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).

We can break down the users accessing registered datasets by geographical origin and institute type. In total, there were over 11,180 new user registrations. Note though, that many datasets are fully open so do not require user registration: those users are not captured here, unless they happen to have registered and logged in.

Users by area (%)

UK	67.1
Europe	11.5
Rest of the world	18.7
Unknown	2.6

Users by Institute type (%)

University	69.7
Government	16.8
NERC	4.5
Other	6.6
Commercial	1.5
School	0.9

1,131 new datasets have been archived and made available via the CEDA catalogue – bringing the total number of datasets to 7,914.

The Archive now holds over 21.5 petabytes of data.

Usage of JASMIN

JASMIN delivers analysis and compute capability for the environmental science community, predominantly for NCAS and NCEO, but also supports other NERC domains: oceanography, polar science, geology and earth sciences, ecology and hydrology.

As of March 2022 there are:

- 2,264 active JASMIN users
- 280 shared group workspaces, with nearly 22 petabytes of storage allocated to them.

General overview

Supporting our users

Over 6,757 queries were received by the helpdesk (covering both CEDA Archive and JASMIN services). These queries cover all aspects of data support except dataset/service applications or long term data management discussions.

1,792 applications were processed for access to restricted datasets.

Collaborations

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

In 2021-2022, significant national and international collaborations have continued to support the international climate modelling community, EO and atmospheric research. 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).
3. Working with the wider UK atmospheric science and earth observation communities, via a range of projects, with NCAS and other NERC funding.
4. Working with the European Space Agency on projects such as the ESA Climate Change Initiative (CCI) Knowledge Exchange.

5. 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.
6. CEDA works with ECMWF to provide EO scientists with the high resolution atmospheric analyses they need to process satellite observations.
7. CEDA works with the UK Met Office to disseminate climate and weather data to the research community
8. Supporting the Climate and Forecast Metadata (CF) Conventions with partners in University of Reading, UK Met Office, and multiple US research institutions.
9. With 20+ partners in the European Network for Earth System Modelling, CEDA is working to develop software and services for climate model data archives.
10. Working with academic partners in the UK Research and Innovation Cloud Working Group to share best practice, knowledge and strategy for use of cloud computing in the research domain.
11. Member of the UKRI DARE UK Scientific and Technical Advisory Group. DARE UK is a UKRI Digital Research Infrastructure Programme aiming to establish best practice for federated digital research infrastructure capability for the next generation of Trusted Research Environments.
12. Member of Oversight Committee for LSST:UK. The committee reviews and provides technical guidance for software development for the LSST:UK project, the UK's contribution to the Vera C. Rubin Observatory's Legacy Survey of Space and Time (LSST), a next-generation sky survey to be conducted on a facility under construction in Chile.

Funding and governance

In addition to supporting NCAS and NCEO, 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.

Governance

During the reporting period, our governance structure was reviewed and is undergoing changes. The CEDA/JASMIN board had its last meeting in March 2022. This is due to wider governance changes in NERC/UKRI surrounding digital research infrastructure based on the new digital strategy. New governance structures will be implemented for JASMIN and CEDA separately via a new Digital Research Infrastructure Group (DRIG). This new group will have an overview of the current NERC portfolio for digital research infrastructure, but also consider future support.

Overall funding for CEDA for financial years 2013-2022 (£k)

Financial Year	2013-14	2014-15	2015-16	2016-17	2017-18	2018-19	2019-20	2020-21	2021-22
NCAS income	829	829	808	808	808	808	808	808	808
NCEO income	392	390	393	393	393	402	393	418	393
Other NERC income	272	600	621	825	816	733	883	784	1,371
Other income	1,486	1,394	1,505	1,092	1,280	1,458	1,377	1,476	1391
Total income	**2,979**	**3,213**	**3,327**	**3,118**	**3,297**	**3,401**	**3,461**	**3,486**	**5,203**

Most of this funding comes to CEDA via a service level agreement (SLA) between NERC and STFC.

Externally funded projects

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 Knowledge Exchange	Data archive for ESA Climate Change Initiative as part of wider activity including outreach and education	ESA	05/09/19	15/10/22	445.0
Pest Risk Modelling in Africa (PRISE)	JASMIN support for UKSA IPP project	UKSA	01/12/16	31/03/22	127.3
BEIS IPCC DDC	UK component of IPCC Data Distribution Centre	BEIS	26/09/18	30/10/21	297.2
C3S CORDEX4CDS	Regional Climate Projection data for C3S	C3S	01/05/17	31/12/21	207.9
C3S_34e CDS WPS Services	Designing an interface between CDS toolbox and remote processing using WPS	C3S	01/01/20	30/06/21	359.7
C3S_34f ESGF Data Node	Maintenance of the dedicated C3S ESGF infrastructure for CMIP Global Climate Models	C3S	16/12/20	31/12/21	229.4
C3S_434 CDS to ClimateAdapt	Transferring information between Climate Data Store and European Environment Agency's portal	C3S	01/01/20	31/12/21	321.9
C3S_34g CMIP6	Including CMIP6 data in C3S Climate Data Store	C3S	01/02/20	31/12/21	272.0
UKSA DAP Support 20-21	Funding Esther Conway to attend ESA Data Access and Preservation WG for UKSA	UKSA	01/04/20	31/03/23	24.0
IS-ENES3	Phase 3 of the distributed e-infrastructure of the European Network for Earth System Modelling	H2020	01/01/19	31/03/23	780.7
C3S Oceans Data Archival	Data archival for C3S Oceans project (U. Reading)	C3S	01/01/19	30/06/21	19.5
JASMIN for ESA SST CCI+	JASMIN Support for ESA CCI+ Sea Surface Temperature processing	ESA	01/07/19	30/06/22	30.0
ESA Digital Twin Earth Precursor	Use of JASMIN's cloud to support the development of an AI surrogate model for soil moisture data calculations derived from the JULES land surface model and LAVENDAR data assimilation system.	ESA	01/02/20	31/08/21	66.1
JASMINx Phase 1	Capital funding for a new investment in JASMIN including upgrade to LOTUS batch computing system and consultancy on future user requirements.	NERC	01/04/21	31/03/22	1,180
JASMIN for JNCC ARD Service	JASMIN Support for JNCC Sentinel ARD service provision	JNCC	01/03/20	31/03/22	33.7
JASMIN for JNCC Core	JASMIN Support for JNCC Sentinel ARD processing	JNCC	01/01/20	31/03/22	37.0
ESA Data Hub Relay	Operation of a Sentinel Data Hub Relay for ESA	ESA	01/04/21	28/02/24	739.1

Name	Description	Funder	Start date	End date	Value (£k)
MOHC Data Pipeline	Continuation of the MOHC Data Pipeline that supports the transfer, cataloguing, publication and dissemination of the Met Office Hadley Centre climate simulations	Met Office	01/04/21	31/03/24	450.0
UKCP18 Services: 2021-22	Continued work on the UK Climate Projections User Interface and Data Services	Met Office	01/04/21	31/03/22	86.5
Support for Ensembles 2021-22	Supporting the services required to ingest, manage and deliver large climate multi-model ensembles.	Met Office	01/04/21	01/09/21	20
EO Data Architecture	Funding for the UKSA Call for strategy on EO Data Architecture	UKSA via Telespazio	01/03/22	30/04/22	19.9
UKRI Net Zero Digital Research Infrastructure	Scoping project to investigate how UKRI can achieve net zero computing	UKRI	01/11/21	31/06/22	60

Publications, posters and talks

Bennett, V., Presentation on research community needs at UK Space Conference, Panel Session EO Data for Climate and Environmental Research and Business, 28th September 2021

Bennett, V., JASMIN (Presentation), SPAN DAWG (Space Academic Network Data Analysis Working Group), May 2021.

Conway, E., Jupyter Notebook Best Practice (Presentation), WGISS-52 Meeting (Online), Committee on Earth Observation (CEOS), October 2021.

Donegan S., Williamson, E., CEDA Datasets and Services (Poster), National EO Conference; September 2021.

Evans, R., My apprenticeship experience, Interview for the National Apprenticeships Week, February 2022

Kershaw P. J., Stephens S., ESA Digital Twin Earth Climate Explorer Project on JASMIN (Presentation), WGISS-51 Meeting (Online), Committee on Earth Observation (CEOS), April 2021.

Kershaw P. J., JASMIN - Providing a Data Analysis Platform for the ESA Digital Twin Earth Precursors Climate Explorer (Presentation), Data Science for EO session, National EO Week Conference 2021, Sept 2021.

Kershaw, P. J., Earth System Grid Federation Future Architecture, Copernicus, Cloud and ESA, (Keynote presentation), ClimateData.ca meeting. Dec 2021.

Kershaw, P. J., Earth System Grid Federation Status and Plans - European Perspective, presentation, 24th Session of the Working Group on Coupled Modelling (WGCM), December 2021

Rand K., **Kershaw P.J.**, Radhakrishnan A., Nikonov S., Balaji V., Vahlenkamp H., Flamig Z., Durachta J., Tucker W., Abernathey R., O'Leary N., Henderson N., ESGF in the Cloud: A Community-driven Effort for Scalable Data Access and Analysis (Poster), AGU Fall Meeting 2021, December 2021

Massey, N., Object Storage at CEDA / JASMIN (Presentation), ESA DAP WG8 202, October 2021.

Pamment, A., Magagna, B., Moncoiffe, G., Devaraju, A., Stoica, M. and Schindler, S.: Improving the Interoperability of CF Standard Names in Environmental Science, AGU Fall Meeting, December 2021.

Magagna, B., Moncoiffe G., Stoica M., Devaraju A., **Pamment, A.**, Schindler S. and Huber R.: The I-ADOPT Interoperability Framework: a proposal for FAIRer observable property descriptions, EGU General Assembly 2021. https://doi.org/10.5194/egusphere-egu21-13155

Parton, G. A., The need to connect data stewardship roles across the research lifecycle and ecosystem… a domain repository perspective (Invited keynote presentation), Research Data Management Forum 2021 (RDMF-2021), July 2021, https://doi.org/10.5281/zenodo.5105972

Parton, G. A., Credit where (data) credit's due - Data publishing at STFC's CEDA Archive service, STFC Open Science Café, November 2021, https://doi.org/10.5281/zenodo.5711084

Parton G. A., Professionalising Data Stewardship (Presentation), IG session, Research Data Alliance Plenary, November 2021.

Drummond, C.; Wang, Y.; **Parton, G. A.**; Lehtsalu, L.; Rantasaari, J., Professionalising Data Stewardship | Global updates to inform RDA community on future directions (RDA-US webinar), Research Data Alliance, June 2021, https://www.rd-alliance.org/professionalising-data-stewardship-IG-june2022webinar

Pascoe, C., Video poster for the European Climate Data Explorer entry to the E-library of Projects at the European Climate Change Adaptation Conference (ECCA 2021) in collaboration with Hans-Martin Fuessel from the European Environment Agency and Sam Arnold from the Copernicus Climate Change Service.
https://www.ecca21.eu/participants/23#364049
video: https://youtu.be/5J3fHuuvaW8, June 2021.

Pepler, S., Arrivals Service - Working with Other NERC Data Centres to Share the Load, NCAS / RMetS Atmospheric Science Conference, July 2021.

Pepler S., STFC/CEDA Approach, on Data Integrity and Authenticity on Cloud (Presentation), Data Preservation and Stewardship session, WGISS-53 Meeting (Online), Committee on Earth Observation (CEOS), March 2022.

Pritchard, M., Supporting large-scale data analysis on JASMIN (Presentation), JISC Networkshop49, April 2021. https://www.jisc.ac.uk/events/networkshop49-27-apr-2021/programme#

Stephens, A., CEDA and JASMIN - the role of Open Science and a digitally-enabled environment in pursuit of Net Zero. Environmental Intelligence Conference 2021: Beyond COP26: The Road to Net Zero, December 2021.

Stephens, A., Overview of JASMIN (Presentation), NERC Constructing a Digital Environment webinar series, January 2022. https://digitalenvironment.org/cde-webinar-series/#Stephens

Noone, S., Atkinson, C., Berry, D. I., Dunn, R. J. H., Freeman, E., Perez Gonzalez, I., Kennedy, J. J., Kent, E. C., Kettle, A., McNeill, S., Menne, M, **Stephens, A.**, Thorne, P. W., **Tucker, W.**, Voces, C., Willett, K. M., Progress towards a holistic land and marine surface meteorological database and a call for additional contributions. Geosci Data J. 2021; 8: 103– 120. https://doi.org/10.1002/gdj3.109

Smith R., Balaji V., Radhakrishnan A., Evans R., Abernathey R., **Stephens A.**, **Kershaw P. J.**, Developing Community standards-based Search Tools for Earth System Model Data using STAC, AGU Fall Meeting 2021, Dec 2021