National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

JASMIN

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

CEDA Annual Report 22–23

UKRI STFC and NERC

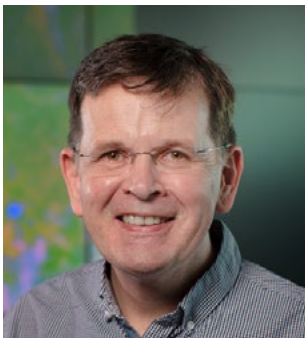
Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

Contents

Editors: Poppy Townsend, Jesse Alexander, Graham Parton, Molly MacRae

Introduction

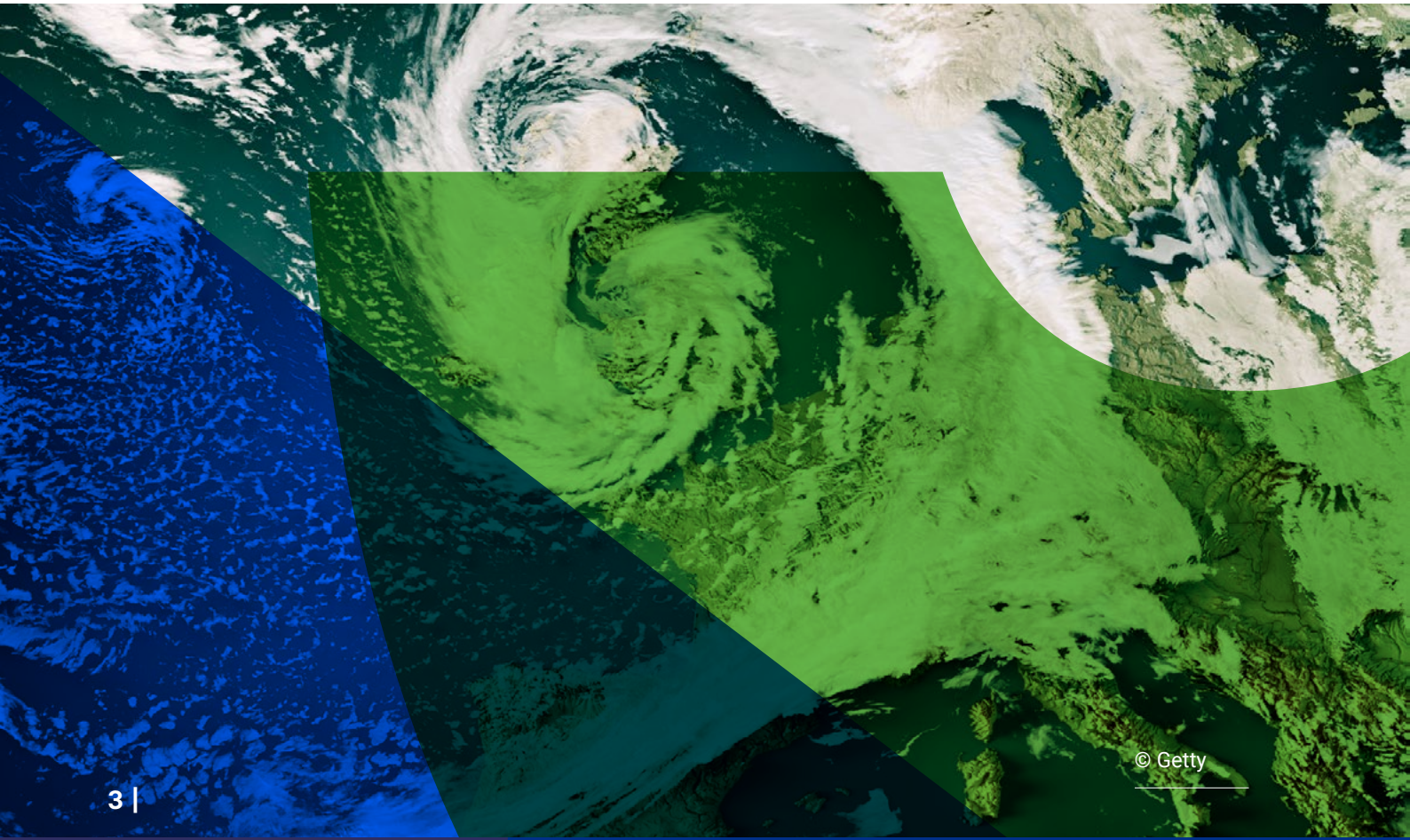
PHIL KERSHAW, HEAD OF CEDA



Often when I speak to people about the work of CEDA what comes across in their impressions is the strongly collaborative nature of the work we do: communication and networking are strong themes. These are very evident reading through the articles in the year's annual report with the Net Zero Digital Research Infrastructure Scoping project or for example the impactful work with the IPCC to make figure data available for the Sixth Assessment Report (AR6). This reporting year has also seen the completion of phase one of the Environmental Data Service in which we have been strengthening our collaborations with the other NERC data centres and are already preparing for the recommissioning process which is starting in earnest with more exciting work planned. Towards the end of the year, we also embarked on a major new activity through NCEO with the Earth Observation DataHub. Working with other public sector partners this two-year programme represents a major investment in digital infrastructure to facilitate the exploitation of EO datasets and will engage with industry through procurement of new software and services.

Our training courses also continue to be an important aspect of our work reaching out to our user community and equipping them with the computing skills they need to support their research. This links closely with our programme of technical work particularly with JASMIN. The commissioning of the new ORCHID GPU cluster marks an important milestone.

Finally, I want to extend my thanks to all our staff and to our funders and stakeholders. It has been great to see the team grow as we welcomed new starters from the STFC graduate programme. Fittingly we have also celebrated the long-term commitment of members of our team through the NCAS long service awards.



© Getty

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.

Harriett Richardson with illuminated globe © National Centre for Atmospheric Science

Meet the Team

Phil Kershaw
Head of CEDA

Emily Anderson
Graduate Environmental Data Scientist

Jennifer Bulpett
Senior Project Manager

Katie Cartmell
Project Manager

Fatima Chami
JASMIN User Support

Esther Conway
Senior Data Scientist: Earth Observation

William Cross
Apprentice Software Developer

Adrian Dębski
Software Developer Industrial Placement

Steve Donegan
Senior Data Scientist: Earth Observation

Rhys Evans
Software Engineer

Nicola Farmer
Graduate Software Engineer

Wendy Garland
Senior Data Scientist: Aircraft

Hayley Gray
User Support and Metadata Services Operator

Andrew Harwood
Infrastructure Manager

Adrian Hines
Director of JASMIN

Alan Iwi
Senior Software Engineer

Matt Jones
Senior DevOps Engineer

Martin Juckes
Head of Atmospheric Science and Research and deputy head of CEDA

Diane Knappett
Senior Data Scientist: Earth Observation

Jack Leland
Software Engineer

Molly MacRae
Graduate Environmental Data Scientist

Alex Manning
Research Software Engineer / Operations Support for JASMIN

Neil Massey
Senior Software Engineer

Alison Pamment
Data Scientist: Metadata Standards

Graham Parton
Data Management Specialist

Charlotte Pascoe
Senior Data Scientist: Models

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

Daniel Westwood
Graduate Software Engineer

Matthew Wild
Senior Data Scientist: UKSSDC

Ed Williamson
Data Scientist: Earth Observation

Insights from Data Use

Data is a cornerstone of our society. It underpins the environmental sciences, government service delivery and can aid commercial endeavours. The Natural Environment Research Council understands the inherent wealth in the vast data assets available through the data centres that make up its Environmental Data Service, which includes the CEDA Archive.

This section draws attention to a number of aspects of CEDA's work supporting and enabling data usage. Our 'What's Hot' article explores those data most often exploited by the community, which sits alongside a highlight from our Web Processing Service about facilitating wider tool usage to exploit the data we hold. We also hear from staff on engagement with our stakeholders - from researchers, trainers and the wider community - through a range of activities and through engaging with artists and the media.

Ralph Burton working on computer models © National Centre for Atmospheric Science

Reaching Out: Engaging the Wider Community via the Arts

GRAHAM PARTON, CHARLOTTE PASCOE, POPPY TOWNSEND

We live in a world with a myriad of ways to communicate with one another. Be that via social media, stories, reports and the arts; communicating is essential for sharing ideas and finding solutions to the challenges we find both on a local and global scale. Ensuring that research isn't retained within the confines of a given community but becomes accessible to all, leaving a lasting legacy and enabling change, was a driving force behind two very different engagements between artists and CEDA staff this year: Paul Millhouse-Smith as the artist in residence as part of the UKRI funded Net Zero Digital Research Infrastructure Scoping Project and Adrien Segal the US-UK Fulbright Scholar working with the University of Dundee Botanical Gardens.

Paul Millhouse-Smith: inspiring collective action

Paul Millhouse-Smith is a multi-disciplinary artist and technologist and was commissioned by CEDA for six months to create a body of work that encourages people to look at the challenge of net zero for digital research infrastructure from a fresh perspective. The resulting installation comprises physical and virtual artworks, inspired by conversations between Paul and the research and digital infrastructure community across a series of creative workshops, and 1-2-1 meetings.

Inspired by some of the world's earliest examples of narrative art and information sharing via pottery, Paul Millhouse-Smith has used cutting-edge 3D ceramic printing technology and conversations around human impacts and climate change, to make a collection of decorated ceramic vases. Paul uses clay for this commission as a metaphor for sustainability and lasting change – and the change that is required for reaching net zero digital research infrastructure.

At the art commission launch event, project stakeholders met in person for the first time during the project and members of the public were invited to attend. Approximately 30-40 members of the public spoke to the project team about the art installation and the meaning behind it.

Paul also created a virtual space to interact with a digital version of the ceramic vases.

Paul's work encourages people to look at the challenge of net zero for digital research infrastructure from a fresh perspective and inspire meaningful change. It is designed to stimulate engaging conversations across the research community and beyond.

The journey towards carbon-neutral digital research infrastructure will involve UK researchers from across disciplines and a broad transformation will be required. The arts have an important role in this transformation as a medium through which people can consider their values, shift attitudes, reflect on issues, and find connection. For further information see the consortium activities section of the UKRI Net Zero Digital Research Infrastructure Scoping Project technical report.

The ceramic 3D printed vases were all inspired by conversations Paul had with the research community. The vase in the front of the picture shows a visual representation of a sound recording taken in the JASMIN machine room.

Adrien Segal: Cyclogenesis

Graham Parton, one of CEDA's data scientists, drew on his PhD studies of extra-tropical cyclones and the Sting Jet feature that can lead to the most damaging winds we see in these storms to support Adrien Segal's work 'Cyclogenesis' at the University of Dundee Botanical Gardens. Adrien created an earthwork that commemorates recent extratropical storms Arwin and Barra during the winter of 2021-22 that dramatically changed the landscape of the garden. Manifested from materials at hand, including eucalyptus trees felled as a result of the storms, the artwork creates a space for visitors to experience the cyclonic patterns and air movements within storm development (cyclogenesis) by walking through and around the earthwork. Graham continues to advise on these aspects of such storm systems to aid Adrien as she continues developing her piece with the botanical garden as it moves towards its planting phase.

As the artwork continues its own genesis and beyond, it will also be an important outreach tool sharing the science of cyclogenesis, sting jets and reflections on climate change with the garden's visitors over the years ahead.

Cyclogenesis: An Earthwork - Process Installation by Adrien Segal in the Living Lab at the University of Dundee Botanical Garden. Various materials including willow, clay, earth, wood, bronze and found materials.
© Image Courtesy of the Artist Adrien Segal

What's Hot?

SAM PEPLER, JESSE ALEXANDER, GRAHAM PARTON

CEDA Archive holds a treasure trove of data with approximately 350 million files in 8500 datasets... However, some datasets are definitely more popular than others. But what's hot at any one time depends on how you ask the question. Here are a few highlights where we look at some of the highlights in terms of data volume (petabytes - PB), number of users or IP addresses.

Data from the Sentinel series of satellites and MODIS instruments from NASA are always top of the access request list from the Earth Observation community. These are monsters at 13PB and 2PB respectively, (1 Petabyte is equivalent to over 100,000 copies of the original Avatar film in ultra-HD!) It is perhaps unsurprising then that not many people actually download them. Instead, many researchers are accessing them by using JASMIN - our data analysis platform - where they can analyse them directly at massive scales and speeds which is key for their work.

Another titan of our archive are our CMIP6 holdings weighing in at 3.4PB. These data from the Climate Model Intercomparison Programme (CMIP) are a key element feeding into the assessment reports from the UN's Intergovernmental Panel on Climate Change (IPCC) - a source of advice and guidance on the science and impacts of climate change for governments. Again, because of its huge size, a great amount of this data's use is done on JASMIN, but it's still a popular download choice; with 231TB and 26 million files downloaded from 73 thousand IP addresses.

MIDAS Open is the UK Met Office's data from surface meteorological stations around the UK. Originally, there was only restricted access for researchers to these data but the Met Office moved to make a sub-set available under an open data licence in 2019. Whilst this has always been popular, this new, open and improved format version increases accessibility to a wider range of people. This includes commercial, government and personal users amongst more than 3700 users who downloaded 1.5TB of MIDAS Open data this year - despite the whole dataset only totalling 300GB.

ESA CCI data is climate-quality earth observation data derived from satellites and used for climate change research. There are many bespoke tools and a thriving user community driving up the use of this data - this year, over 224TB and 31 million files were downloaded from 58 thousand IP addresses.

Finally, we hold digitised 19th-century Royal Naval and colonial logs in the CORRAL dataset, important for climate science, helping us capture past weather and climate data. Although this is not a recent dataset, it remains hot, being unusually popular with over 4000 users this year alone. Whilst we don't know all of the uses this supports, anecdotally we know that it remains popular for teaching and training as well as holding interest for those researching historical events or family histories tied to these voyages.

GB = Gigabyte

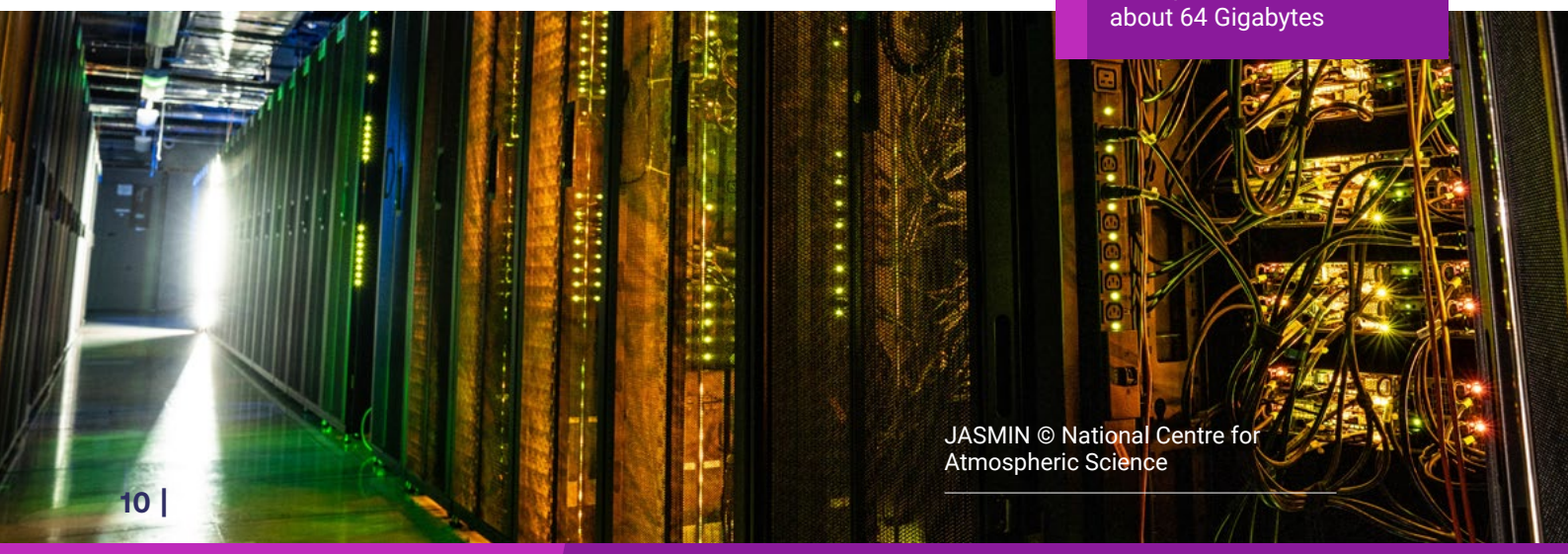
TB = Terabyte

PT = Petabyte

1000GB = 1 Terabyte

1000TB = 1 Petabyte

Most phones can store about 64 Gigabytes



JASMIN © National Centre for Atmospheric Science

The CEDA Archive now holds

Over **20.3** Petabytes of data

Nearly **345.5m** files

Organised in **8,565** datasets

This year alone

4.2 Petabytes of new data

Nearly **23m** files

Over **650** datasets

have been added to the CEDA Archive

Breaking Barriers: The CEDA Web Processing Service

ALAN IWI, AG STEPHENS

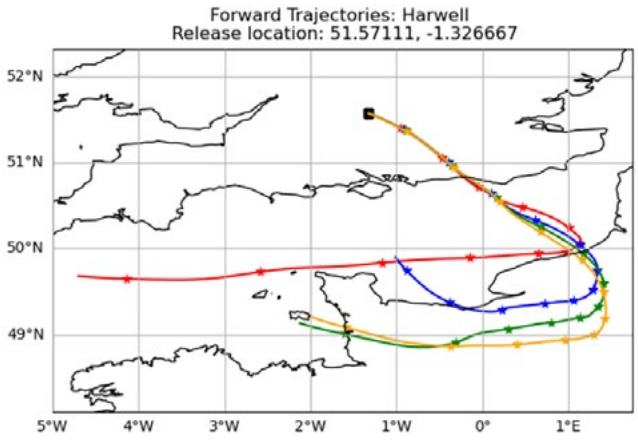
CEDA offers a variety of services to a wide range of users. Working with big scientific data sets can be complex, and not everyone interested in our data has the computational skills to do this. To make our data more accessible, we are making some of our services available on the CEDA Web Processing Service (WPS).

This year, we have added the Met Office pollution dispersion model called NAME. Originally standing for Nuclear Accident Model, it is now NAME III - Numerical Atmospheric-Dispersion Modelling Environment. The NAME model is useful to a variety of people as it can be used to predict how a polluting event could spread to surrounding areas. For example, volcanic eruptions that could affect aircraft flight paths or the air quality impact of industrial facilities.

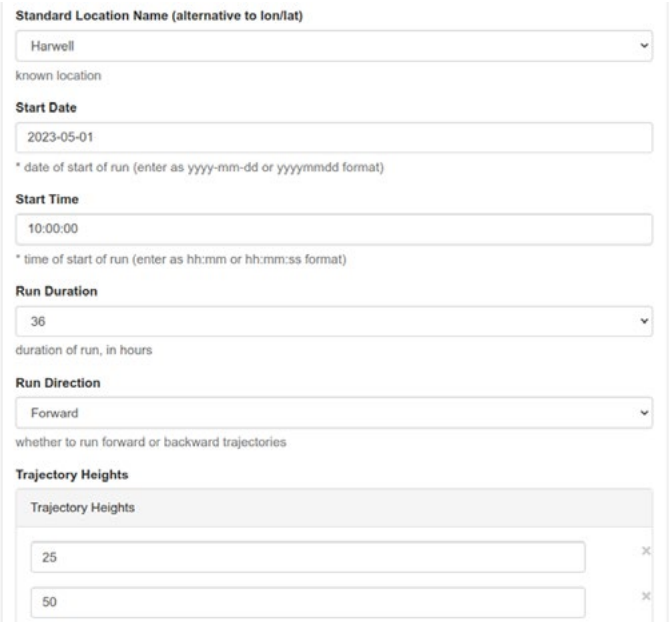
The model allows you to calculate how particles in the air disperse through the atmosphere, or it can be run in reverse to determine where the air mass at a certain time and place came from. We wanted to provide an easier way for people without a scientific background to be able to use and access this resource.

The CEDA WPS allows us to provide access to a wider range of people who may not have the time or resources to work with tools and data programmatically - instead, they can use web forms to select and run their processes. Depending on the specific process, users may not need a CEDA or JASMIN account to access services through the WPS.

The WPS also integrates with a batch processing system which parallelises work occurring on the same system, meaning that if a service is in high demand, the work is queued, scheduled and split into a pool of servers, increasing resilience. By providing access to powerful tools such as the NAME model, via the CEDA WPS, we lower the barriers to working with scientific data!



Summary plot of the trajectory run generated by the WPS showing the dispersion of atmospheric particles.



Simple online form for users to specify the parameters for a NAME trajectory run.

Storytelling with Data

CHARLOTTE PASCOE

Data has long been a fundamental part of the journalist's storytelling toolkit. From the pandemic to climate change, data has not only been used to illustrate news stories, but data itself has become the news. In this world of data-driven journalism, trusted sources of data are of vital importance. CEDA is seen as just such a resource and has become an important resource for storytellers.

One Million Years into Climate Change's Past

Climate Adam's YouTube video exploring the history of climate change. https://youtu.be/oa0ZHAcPHew

CEDA worked closely with the Intergovernmental Panel on Climate Change (IPCC) to prepare figure data publication for their report that is aimed at policymakers. CEDA published the data on the same day as the report itself - something that had never happened before. This enabled journalists to access figure data (as part of their press pack) in advance of the official publication of the report - meaning their news articles and analysis were ready to go on the morning the report was released. It also meant that anyone reading those stories could find and access the cited data immediately. In the months after publication, the IPCC figure datasets were the most visited parts of the CEDA Archive.

The long tail of CEDA's work with the IPCC is now complete. We have data for over 200 data-driven figures containing time series, mapped data, box plots, and cross sections. With information covering all aspects of climate change, from temperature change and sea-level rise to the changing composition of the atmosphere. This data has been used in creative ways to enhance climate change communication initiatives such as in Dr Adam Levy's ClimateAdam YouTube video "One million years into climate change's past" which used IPCC data at CEDA to tell the story of just how unprecedentedly rapid the climate change of the last century has been.

By making data available in just a few clicks from the IPCC reports, CEDA is helping the IPCC in their mission to open up their science to the world. Our work is fundamental for using transparency to build trust in the call for the imperative to take climate action now.

© Getty

The Future of Training for Atmospheric Scientists

ALISON PAMMENT

Besides the range of technology and services CEDA offers, our staff support the wider community by delivering a portfolio of training, designed to ensure that both our staff and users have the skills needed to perform world-class environmental research.

For almost 10 years our computing courses have been part of the training programme of the National Centre for Atmospheric Science (NCAS). In 2022, I was seconded to NCAS to review their entire professional and technical training portfolio and provide recommendations on how it should develop in the future.

The project considered several key questions, including:

- What range of training should NCAS offer to address current and future needs in atmospheric science?
- How can NCAS ensure that courses provide a good learning experience?

Key to such reviews is the participant feedback collected each time a course is run. This provides an invaluable resource for assessing existing courses and gathering additional training needs. There are many positive comments about current courses: students appreciate the professional knowledge and expertise of the trainers; the opportunities for networking; and, the practical sessions which consolidate learning through hands-on experience.

Sustainable Airborne Research and Atmospheric Science Symposium © National Centre for Atmospheric Science

A way to deliver a good training experience for both students and trainers is to ensure that clearly defined learning objectives are a focus throughout. This avoids course timetables becoming too crowded, allowing better pacing of the materials. When applying for courses, students should be given clear information about prerequisite knowledge so they can prepare adequately and thus maximise the benefits gained from attending the training.

NCAS trains atmospheric scientists in the use of many complex tools needed to carry out observational, laboratory and modelling research. Evidence from both student feedback and an NCAS staff survey showed the need for a new course on applying machine learning to the analysis of atmospheric science data. There is also a need for a research data management course that could be run in collaboration between CEDA and other NERC data centres.

Outputs from my secondment include a report detailing specific recommendations for the development of the NCAS training portfolio, the project produced other benefits. An NCAS training group was created to oversee the planning, development, and delivery of new training courses. This enables resources to be drawn together from across the organisation to help push forward new training projects in a timely fashion. A new NCAS course has already benefited from this new group - the course for radar and lidar remote sensing is under development and will be launched in the 2023-24 academic year. An NCAS trainers' forum was also initiated - this group provides a line of communication for staff wishing to collaborate and exchange views on the subject of training. It is open to all NCAS staff with an interest in training, including those who would like to become involved, and provides opportunities to share skills and best practices. If you are interested in discussing the findings from my secondment, please get in touch, I'd be happy to share further details.

Reflecting on my time seconded with NCAS, I really enjoyed working on the training portfolio project. Training the next generation of scientists is hugely important, and it's a privilege to be a part of it. The new courses that are planned, and the groups we created to nurture the training activity, will ensure that NCAS continues to be a trainer of choice in the atmospheric sciences.

FAAM aircraft tour
© National Centre for Atmospheric Science

This year our team delivered training

To over **300** people

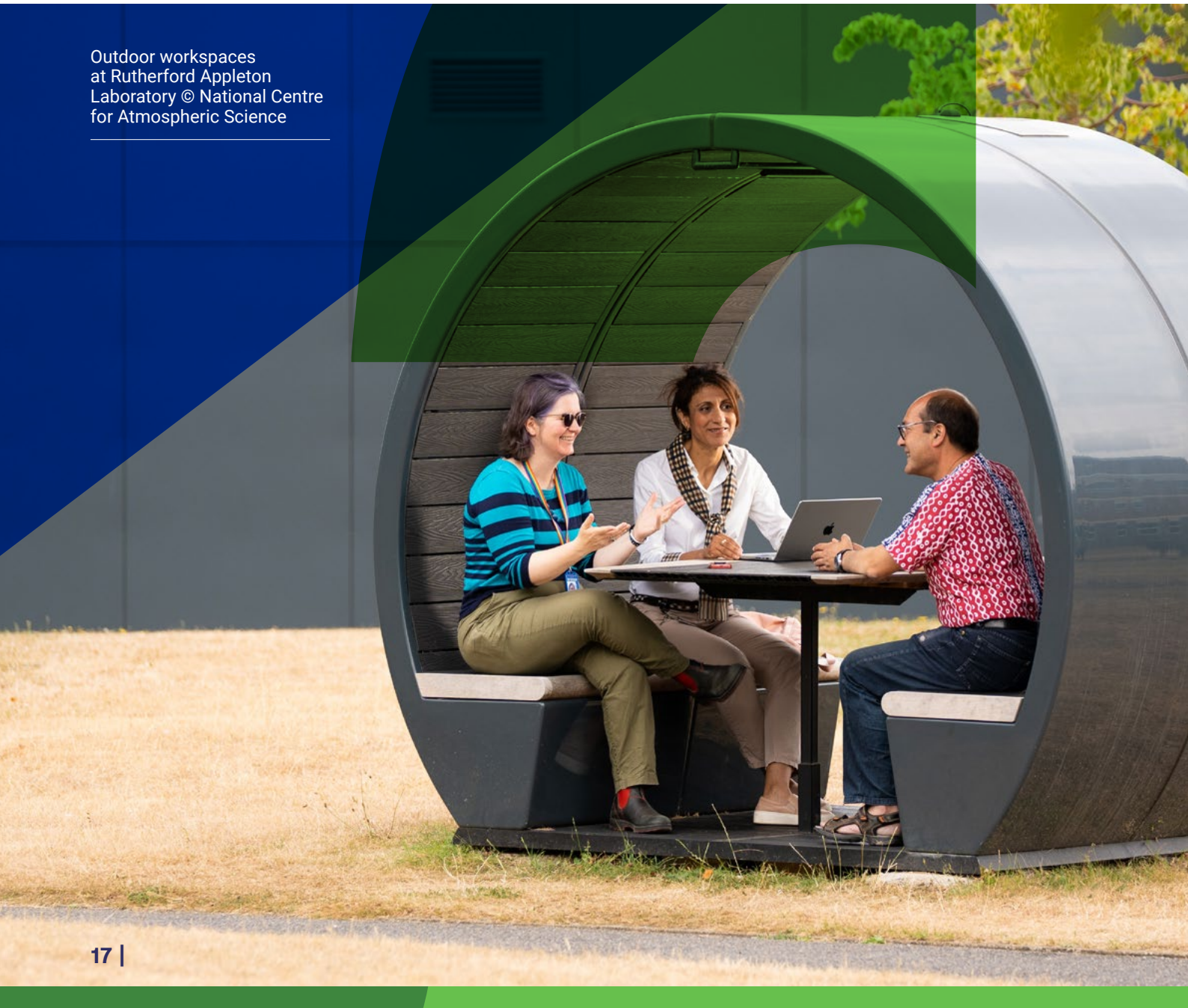
JASMIN staff delivering training to users.

Collaboration, Expertise, and History

“No man is an island, ” penned English poet John Donne in 1624, and indeed, CEDA has never operated in isolation - involved with the wider research data management and environmental science communities throughout its 25+ years of operation.

In 2022-23, CEDA staff were noted for significant contributions and engagements. The examples in this section show how CEDA operates across research councils, research domains, national boundaries, and, also, over time. The benefits from these interactions flow in both directions: CEDA enhances those projects and networks it engages with and these bring lessons that enhance our skills, understanding and working.

Whilst most of this section covers our staff and project work, the final article segues towards our longest standing asset: our archives, spotlighting a range of datasets that demonstrate breadth of data. These all demonstrate the assorted communities that CEDA has, and continues to, support over the years.



Outdoor workspaces at Rutherford Appleton Laboratory © National Centre for Atmospheric Science

The Power of Collaboration: Lessons from Net Zero and Beyond

MARTIN JUCKES, POPPY TOWNSEND, JENNIFER BULPETT, CHARLOTTE PASCOE, KATIE CARTMELL, AG STEPHENS, MOLLY MACRAE

At any given time, CEDA is working on several different collaborations and projects. As experts in data management, software infrastructure, and climate data services, we are well-equipped to contribute to many projects.

These projects often span multiple years and involve bringing together people from numerous different countries and organisations. Whilst this provides rich opportunities to learn and work towards something great, there are unique challenges that come with such large-scale collaborations. This was particularly true of the UKRI Net Zero Digital Research Infrastructure Scoping Project which established over 300 new contacts across all UKRI research domains in a relatively short project span of 18 months.

Project team and partners at the art commission launch event in Glasgow. This was a key output that would not have existed if we had not taken a multi-disciplinary approach. See the earlier report highlight for further details about this component of the project.

Here are four main lessons from this project:

1. **Learning from others is a key objective** when reaching out to new communities, and needs to be valued in its own right. It can take a long time to build different perspectives together into a synthesis but it brings many benefits in the long term. If time is taken to build a common understanding, the effort needed to resolve technical problems can be greatly reduced.
2. **Every new contact brings new ideas.** This was particularly evident during the many outreach activities held in order to reach as many sectors of the community as possible. With each event, the effort spent was rewarded with new and valuable perspectives.
3. **Multitasking can be challenging**, especially when trying to work around the schedules and objectives of multiple stakeholders, but it is also one of CEDA's strengths.
4. **Major societal challenges can benefit greatly** from a multi-disciplinary approach. When discussing the net zero challenge, people from the physical sciences tended to look more to technical solutions while those in the arts and humanities embraced the concept of behaviour change.

As we continue to navigate these collaborations, CEDA remains dedicated to creating a sustainable and impactful legacy for future generations. Our efforts persist in uniting various sectors and international communities towards innovative solutions for scientific challenges, particularly in the pursuit of a net zero future.

Examples of key projects, teams and networks that we have been involved with this year include:

- **UKRI Net Zero Digital Research Infrastructure scoping project** - CEDA led a consortium of 20 partners to review the sustainability implications of expanding digital research infrastructure
- **IS-ENES Infrastructure for the European Network for Earth System Modelling** - CEDA is a founding member and strong contributor to IS-ENES
- **Earth Observation Data Hub** - CEDA is leading this major project - see details on separate article
- **NERC Environmental Data Service** - A collaboration of NERC's 5 environmental data centres
- **Coupled Model Intercomparison Project Phase 7** - CEDA is leading the CMIP7 Data Request and contributing in a number of other important areas of this developing international mega-project
- **Horizon Europe ClimatEurope2 project** - Contributing to work on training and uncertainty
- **IPCC Data Distribution Centre** - CEDA is one of four partners providing expert advice and services
- **STFC Air Quality Network Plus** - Advice on data and data services
- **National Centre for Atmospheric Science** - CEDA is the designated data centre for this research community
- **National Centre for Earth Observation** - CEDA is the designated data centre for this research community

© Getty

10

large* collaborative externally funded projects

were actively worked on this year.

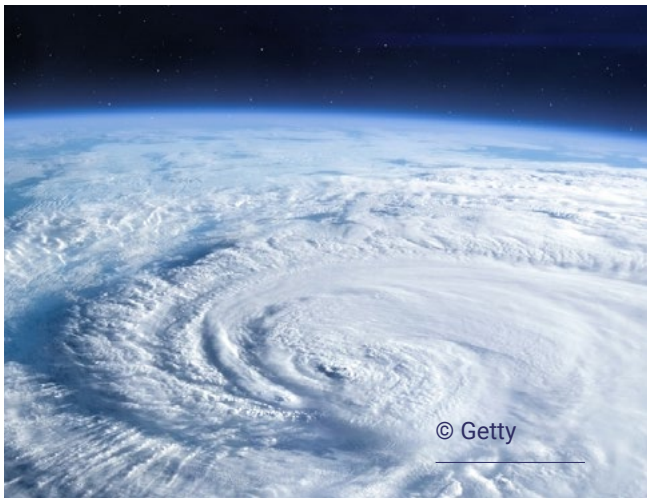
*A large project has a value greater than £15K

Operations team in social space © National Centre for Atmospheric Science

The UK Earth Observation Data Hub: CEDA Leads the Way

AG STEPHENS, PHILIP KERSHAW

The UK has a world-leading Earth Observation (EO) sector delivering around £1 billion of annual benefit to the UK across nine key civilian use areas[1]. The results of studies and user engagement activities have consistently indicated that there are barriers to the effective exploitation of EO data to its full potential in downstream applications.
Seeing this as a strategic priority, the UK government has awarded National Centre for Earth Observation (NCEO) the leading procurement role on. A programme to build the UK's Earth Observation Data Hub (EODH) - a cloud-based data, software and infrastructure platform.


© Getty

Bringing together public and private sector partners to serve the entire UK Earth Observation sector, this £10M collaborative project is led by NCEO (based at the University of Leicester) and CEDA. The service will bring together academic, commercial, and government partners to capitalise on the skills, services and infrastructure available and address issues around the fragmentation of access to EO data sources.
The project aims to deliver a central hub through which a suite of services are delivered to the community in a manner that allows them to scale as their service demand requires, but in a way that allows them to closely manage their usage of scalable workflows.

Key use cases being explored include:

- A JupyterHub capability that allows users to interactively develop, test and publish workflows for others to discover and use with their own services and tools
- A 'web-map' interface to discover and interact with a vast range of Earth Observation and climate datasets such as Sentinel, CMIP6, UKCP and outputs from the EOCIS, a sister project to the Data Hub
- A common data catalogue search capability spanning both the Data Hub and external sources
- Powerful web APIs that enable developers to build external applications, using the Hub Platform as their backend for data access, processing and storage - with three funded projects to kick-start the overall programme
- Data quality systems that give users confidence that the data they use complies with international standards

All of this is implemented in a way that can be scheduled based on events such as new satellite scenes being made available, or responding to changing conditions (for example a current natural disaster).

It's an exciting time ahead for the UK's Earth Observation community, and CEDA in particular, as we bring a wealth of skills from the community together to realise something that is greater than the sum of its parts. The programme runs until March 2025 so there is a great deal of work to do in a short timescale. For more information, take a look at the project website.

1 London Economics (2018). Value of satellite-derived Earth Observation capabilities to the UK Government today and by 2020.

Breathing New Life: CEDA's Impact in the STFC Air Quality Network

WENDY GARLAND

Air pollution is a major, global issue classified in 2014 by the United Nations Environment Programme as the 'world's worst environmental health risk'. There are therefore huge opportunities to deliver economic and societal impact from improving air quality.

The Science and Technology Facilities Council (STFC) Air Quality Network (SAQN) is building a multidisciplinary community to exploit the untapped potential of STFC capabilities in addressing air quality challenges. The 522-member strong network, supported by over £480k of research council funding, was built to cut across traditional discipline boundaries and bring together government, industry and other research communities. The network drew on meetings and fora, by funding small scoping and proof of concept studies to demonstrate the potential of new ideas, tackle air quality issues, and build collaborations amongst researchers.

CEDA has been an integral partner since the project began in 2019, providing organisational support to the management team to run the network and organise meetings and conferences. The CEDA team advised in discussion panels, mentored groups in the sandpits, and reviewed project proposals whilst reporting to the steering committee. By sharing JASMIN computing facilities with this community, we expanded our regular user base to include the medical research and air quality monitoring communities, both of which usually engage with the Environment Agency and local government.

The SAQN network has worked closely with seven other networks funded under the UK Clean Air Strategic Priority Fund and the UKRI Clean Air Champions, culminating in a joint conference in Birmingham in 2023. Over 200 air quality researchers, policymakers, industry partners and civic sector groups who work across the atmospheric, health, social and building science disciplines came together. CEDA staff also attended a high-level meeting presenting the network to the government's chief medical officer, senior research council executives, the Department for Environment, Food and Rural Affairs, and other government department representatives.

Dr Sarah Moller, Academic Lead of the SAQN says: "The level of engagement in the SAQN that we have seen from STFC staff has been brilliant. This has resulted in some exciting collaborations and new research directions that will hopefully continue to develop beyond the life of the network. CEDA has been engaged in the SAQN right from its inception, and their support for and engagement in all network activities has been really valuable."

Overall, SAQN has proved to be a great opportunity to amplify awareness about CEDA to the air quality community and share our wealth of skills, data and services. We hope that by bringing together these different groups, a vibrant community will persist beyond the conclusion of this project!

Becky Wagner with Mobile Air Pollution Laboratory © National Centre for Atmospheric Science

Dedicated for the Long-Term: NCAS Staff Awards

WENDY GARLAND

CEDA has been an integral part of the National Centre for Atmospheric Science (NCAS) since its creation in 2003, providing the essential data and computing services to enable the groundbreaking research NCAS achieves.

This year, NCAS has celebrated its 20th anniversary and has begun presenting long service awards to its staff celebrating their dedication, skills, knowledge and commitment. These awards reward people who have made long-term contributions to NCAS, through continuous or cumulative service, and were given for greater than 10, and greater than 20 years of committed service.

Seven members of the CEDA team received long-service awards in 2023. Graham Parton received a 10 year service award. Sam Pepler, Martin Juckes, Wendy Garland, Charlotte Pascoe, Alison Pamment and Matt Pritchard were all awarded 20 year service awards.

Reflecting on this, Barbara Brooks, NCAS Scientific Services, Facilities and Training Director, said: "CEDA staff have made and continue to make valued contributions to the services supporting the atmospheric science community.

I'm delighted that NCAS had the opportunity to recognise the continual dedication and specific contributions of CEDA staff through our NCAS Long Service Awards and to celebrate this with the awardees at our 2023 staff meeting.

Looking to the future, NCAS is excited to have CEDA staff's continual contribution to NCAS and its ongoing work at the forefront of climate research."

CEDA's long-service awardees: (L-R) Charlotte Pascoe, Sam Pepler, Wendy Garland, Martin Juckes, Alison Pamment, Graham Parton. (Matt Pritchard not pictured).

Variety is the Spice of Life!

GRAHAM PARTON, DIANE KNAPPETT

CEDA tackles all 4 'V's of the Big Data challenge: Volume, Velocity, Veracity and Variety on a daily basis. Managing vast quantities of data, both being streamed into and accessed from our archive, as well as maintaining the data stored within, is a challenge. Whilst other data providers deliver amazing services on large volumes of specific data, it is the sheer variety of our holdings that demonstrates the breadth of work CEDA engages with.

Within the CEDA Archive, you'll find large scale collections such as the latest satellite data and climate model data from a huge range of modelling centres around the world. But also smaller-scale datasets such as those from citizen science and 'data rescue' projects like the digitised CORRAL ship logs and the Ben Nevis weather station rescue project.

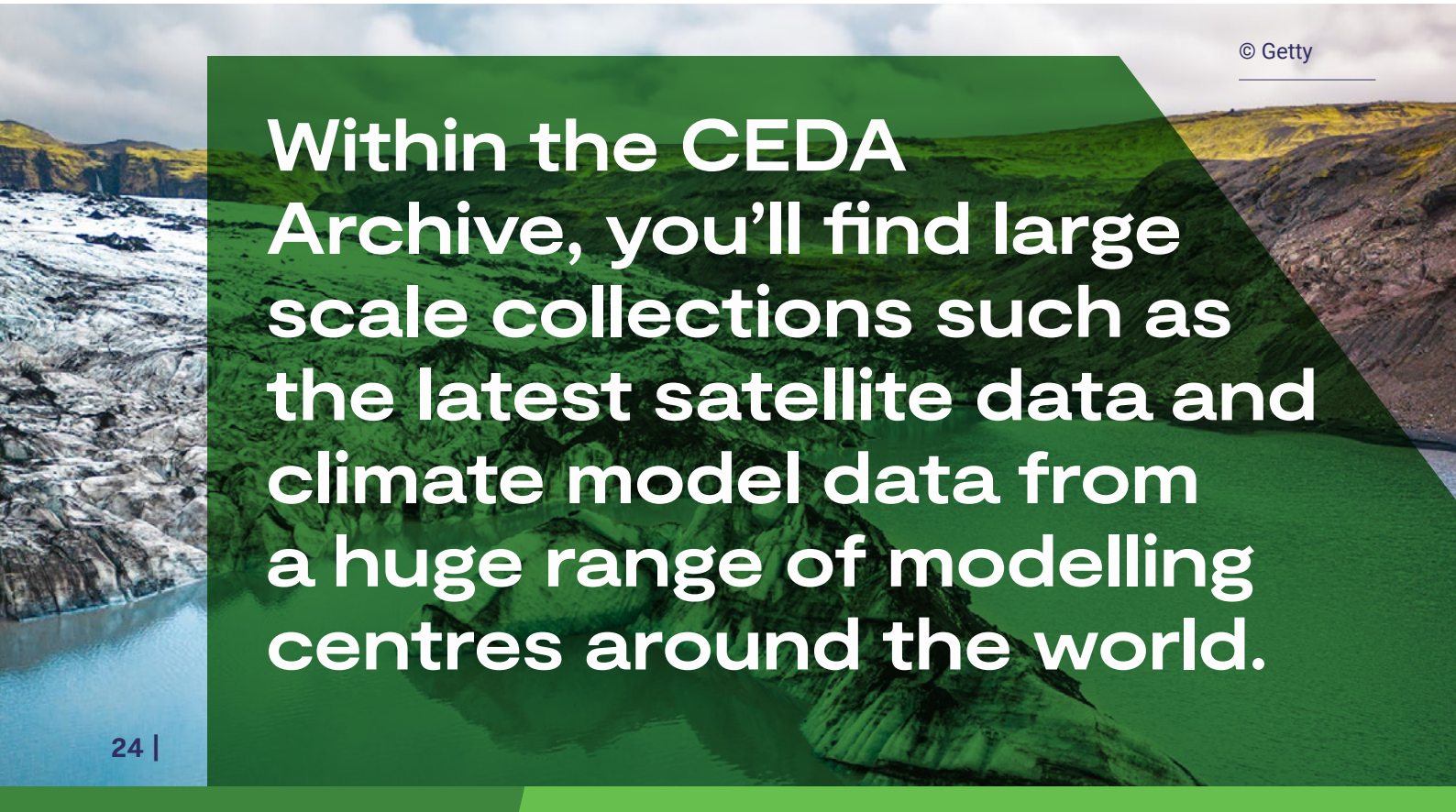
Two particular datasets, added to our holdings this year, that demonstrate the variety at scale are the ESA CCI sea ice concentration datasets. Each dataset provides a unique perspective on changes in sea ice concentration in recent decades.

The first is based on data from a historical meteorological satellite mission, Nimbus-5, and covers the period 1972-1977 - some of the earliest available sea ice satellite data. This provides information on sea ice concentration and extends 20 years earlier than previously available. This new dataset has been provided, through the ESA CCI programme, using a method known as satellite archaeology - where researchers go back through old data and reprocess past holdings of satellite images.

In contrast, the second state-of-the-art high resolution dataset provides some of the most recent sea ice data, spanning almost 30 years from 1991-2020, at a higher spatial resolution than previously available.

These contrasting two datasets push the boundaries of what's available from the satellite Earth observation era - and exemplify the diverse types of historical and new innovative data we hold.

Whether the data holdings in our archive are big or small, our skilled staff show the same levels of care and dedication to ensure these important assets are properly curated and made available for now and into the future, continuing to deliver scientific impact and aid in pushing the frontiers of our knowledge.



© Getty

Within the CEDA Archive, you'll find large scale collections such as the latest satellite data and climate model data from a huge range of modelling centres around the world.

Delivering User-Focused Service Enhancements

User needs continue to grow alongside the ever increasing volumes of environmental data and the science questions that are being explored. Just as our user communities strive to push forward the boundaries of our knowledge, CEDA also drives forward the underpinning standards, technology and services that we deliver.

This section of the report touches on a few of the developments we've delivered this year to enhance services for data quality and information flow whilst also ensuring greater resilience and scalability of our infrastructure.

Manchester Air Quality Supersite © National Centre for Atmospheric Science

Bridging the Metadata Divide

ALISON WATERFALL

Since 2015, CEDA has hosted the CCI Open Data Portal for the European Space Agency's (ESA) Climate Change Initiative (CCI). It contains a wealth of climate data, however ESA does not fund the longer term operation production of these datasets. This is usually covered by alternative services, such as the Copernicus Climate Change Service (C3S). In addition, some datasets have an earlier heritage outside of CCI, and this can lead to the availability of related, but not identical datasets within different portals. It is important for users to easily find the most up-to-date and suitable version of the data for their use, whilst still being able to access the full CCI portfolio.

While it is beneficial to have these climate datasets hosted and extended in multiple platforms, it can provide a challenge to both users and the individual platforms to identify how these datasets relate to each other and which is the most current dataset that should be used. Whilst the CCI Open Data Portal provides a central point for all the different CCI products available, a number of these datasets have become outdated or have been continued in an external service.

To tackle this issue as part of the CCI Knowledge Exchange project, CEDA has developed a 'Metadata Bridge' which provides a service detailing the relationship between different products by mapping between pairs of dataset URLs. These relationships can be complicated and vary from data that has been simply reformatted, through extensions in time or new versions, up to more complex combinations of all of these. We combatted this by using a controlled vocabulary based on the DataCite 'relationType' descriptions, as well as a free text field to allow more user-friendly descriptions.

An additional complication was that services differed on their definition of a catalogued 'dataset'. As such, identifying subproducts that could be mapped more closely to each other was required alongside determining the appropriate product level to record the mapping, e.g. between the high-level catalogue records, or individual subsets, or both. These mappings are stored in the metadata bridge service, providing a database and external API which can be queried by any service to provide them with information on any related product. We used this functionality to integrate the information into the CCI Open Data Portal.

Currently, the metadata bridge only contains information for a subset of the related CCI and C3S products but we will be updating this to include all products in the coming year, and also to include similar information from the EUMETSAT Satellite Application Facilities. However, the approach would be more widely applicable to any related datasets and may have further applications beyond CCI.

Over
10,000
new users

have registered to access the CEDA Archive this year – a record high!

In comparison, the previous 5 years combined is just over 41,000 new users – averaging ~8,200 per year.

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.

Fast and Agile: Getting our Metadata Services into Shape

GRAHAM PARTON AND SAM PEPLER

In previous annual reports, we've highlighted the work we've done to ensure a comprehensive catalogue service to aid data discovery. Whilst ensuring we've got everything detailed is great, having a service which is fast and responsive to what our users need in order to find relevant data assets is at the heart of our service delivery. To achieve this in a way that continues to scale with our ever-growing archive and changing user needs, we must also adapt our tools and approaches.

Listening to the feedback from external users and CEDA colleagues we were aware of a few issues that were hampering service delivery. Our data catalogue search lacked finesse, returning large volumes of results that didn't quite match what users wanted - and there was a lack of search options to help refine this. Meanwhile, our internal system to track our archive's 350 million (and ever-growing!) files - our so called File Based Index (FBI) - was falling behind in the race to keep up with the tens of thousands of new files ingested every day.

Both of these tools make use of an industry-standard powerful index and search tool called Elasticsearch which CEDA brought into our repertoire of tools a few years ago. This gives us incredibly fast searching functionality and provides features such as 'fuzzy' matching to cope with the idiosyncrasies of human language - all in a tool that we've found scales to what CEDA needs now and into the future. However, we also learnt that we hadn't been using these optimally for the catalogue or FBI index. In many ways, our indexing was simply out of shape for the work we needed it to do.

We then set about improving the way these two services worked and the structure of their associated Elasticsearch indexes in a way that not only delivered improved service speeds and accuracy, but also within an agile development process that helped us to deliver gains quickly as we went along.

Key to this was our desire to keep the user experience at the heart of our way of working. We worked with a range of users - both internally and externally - to gather their real-world use cases. Then with short, agile development cycles we were able to deliver improvements that better aligned with the way people use our services. These incremental steps meant we were able to pause and receive feedback from users after each stage and update our plans accordingly.

As a result, we have much-improved metadata flows from the files to the File Based Index and onwards through the catalogue and its search and, ultimately, on to our end-users. This has enabled us to provide geospatial and temporal search, and implement monitoring tools to easily gather information such as the number of files in a directory or the parameters within files and datasets. As well as the catalogue search improvements, which deliver better results and refinement options, information displayed in our main archive web browsing service (driven from the FBI contents) now shows more clearly when datasets have been superseded, helping users avoid using obsolete data.

Both these services are now in much better shape thanks to our agile approach to development meaning they are leaner and faster all whilst delivering improved experiences for our users.

If you want to take a look try https://data.ceda.ac.uk to see over 300 million files dynamically browsable and summarised.

Checking Interoperability Works!

GRAHAM PARTON, AG STEPHENS

Making data interoperable is a desire for many of our data providers, but making this happen can be hard. Over the years CEDA staff have engaged with colleagues in the National Centre for Atmospheric Science (NCAS) to develop a range of data standards to aid this interoperability for a wide range of instruments. Data from these instruments is then verified to comply with an open source tool - called Checksit - developed as part of this work.

Thanks to the ingenious way this tool is written, it is already set up to work with a range of other data, offering two different modes of operation. Firstly you can provide it with a suite of checks written against known standards, such as the NCAS Data Standards. Or you can choose any existing file as a reference file, preferably containing high quality metadata, and compare the new files against it. It can also be run in batch mode so that thousands of files can be checked to generate a single output that summarises their compliance or non-compliance to a given standard. This open source tool can be used both by providers as they work up their data and by archives to vet data before curation. The goal though is the same - high quality data that helps interoperability happen.

Navigating CEDA Towards Greater Service Resilience

WILLIAM TUCKER, ALEX MANNING

Thanks to Kubernetes - an open-source system for improving service deployment - we can now use automation pipelines to almost entirely automate the process of preparing, testing and deploying services. New features can be quickly tested in a staging environment before being made live and Kubernetes allows us to roll out these updates with zero downtime. If issues are detected at any stage, it will automatically roll back changes to prevent any disruption to the service.

Benefits of Moving the CEDA Download Services to Kubernetes

The CEDA Archive download services suite is now running on our Kubernetes cluster with a "horizontal pod auto scaler", a system that dynamically evaluates the CPU usage of the download service containers and automatically scales the deployment to meet demand. When the load increases, Kubernetes automatically creates new service containers to allow more simultaneous requests to be processed whilst scaling back resources when demand is low.

As we rolled more services on our cluster, the load balancer that handles and directs traffic became at risk of being overwhelmed; putting the availability of all services on the cluster at risk. To tackle this issue, we installed a second dedicated load balancer for high-throughput services, ensuring efficient load distribution and protection against spikes in demand or denial-of-service attacks.

Writable Group Workspaces from JASMIN Notebooks

The JASMIN Notebook service provides easy–to–use Jupyter notebooks to our users and is another service that runs on our Kubernetes cluster. This service allows direct read access to the CEDA Archive, and read-write access to user's home directories. Previously JASMIN group workspaces were read-only, but by installing the OPA (Open Policy Agent) Gatekeeper Kubernetes policy controller in our cluster we were able to overcome some security concerns and make them writable.

This is a long-requested change which greatly improves the utility of the notebook service for our users. Allowing users to write to the group workspaces enables users to better collaborate with their colleagues on the rest of JASMIN.

Token-based Authorisation for Scripted Access to CEDA Data

Lastly, we've listened to user feedback and simplified access to archive resources and CEDA data. We replaced the complex client certificate system with a token-based access system using OAuth 2.0 (a secure authorization protocol used by many websites such as GitHub and Google) which has far fewer steps, more functionality on Windows, and works with a larger variety of scripts and tools. We'll continue listening to feedback as we roll out incremental improvements to services to make it easier for users to access our data.

JASMIN Technology and Innovation

The JASMIN data analysis platform provides data storage, computational power, and cloud services that enable the environmental science community to process, use and share their data. JASMIN is established as a key element of the UKRI digital research infrastructure – funded primarily by the Natural Environment Research Council, and operated jointly by CEDA and STFC's Scientific Computing.

Alongside providing an essential facility for many research projects, the system hosts the curated datasets within the CEDA Archive, and many related services. The JASMIN team brings together experts in computing, research software engineering, and environmental informatics, and strives to strike a balance between the maintenance of current capacity and capabilities, and the development of new tools and services. This section provides some of the highlights from work on JASMIN during the year.

Fatima Chami inspects JASMIN computer servers at Rutherford Appleton Laboratory © National Centre for Atmospheric Science

JASMIN delivers analysis and compute capability for the environmental science community, including domains such as: meteorology, climatology, earth observation, oceanography, polar science, geology and earth sciences, ecology and hydrology.

JASMIN: Firm Roots and New Shoots

ADRIAN HINES


Adrian Hines, Director of JASMIN

The creation of the Director of JASMIN post recognised the need for greater leadership capacity for an increasingly large and complex system supporting a growing user community, within the context of the broader UKRI digital research infrastructure landscape. Taking up this role in May 2022, I was naturally unsure exactly what to expect. So what have I discovered during my first year as the Director of JASMIN?

First of all, I've seen the firm roots that have been established by a dedicated team of experts who cover an incredible amount of ground from user support, and developing and maintaining services, right through to managing and installing hardware in the data centre. I've seen a really effective model for collaborative working between staff in CEDA and staff in Scientific Computing which gives clarity on roles and responsibilities. I've seen a user base that is both supportive and engaged, and who really value the service that JASMIN provides. I've seen colleagues at the Natural Environment Research Council (NERC) committed to supporting and securing funding for JASMIN, both now, and in the future.

Secondly, as the articles in this part of the report highlight, JASMIN's growth continues. In some cases it's replenishing our existing infrastructure, such as the £5.3M of capital investment via NERC for storage renewal and the successful migration of our JASMIN Cloud services to the VIO7 (VMware Integrated OpenStack) platform. New shoots for JASMIN have come in the form of the delivery of the ORCHID GPU cluster, dedicated to the burgeoning science utilising the power of AI workflows in the environmental sector. In addition the development of the DASK gateway server (see highlight later in the report!), configured to work with the JASMIN notebook service now allows access to parallel computing for notebook users. Key to these successes has been a dedication to working closely with our user community and it has been great to see their level of engagement and positive feedback that we have received.

We have also taken steps to ensure the longevity and resilience of JASMIN infrastructure through commissioning a cyber-security review of JASMIN, which gave the system a clean bill of health, with just a small number of actions recommended.

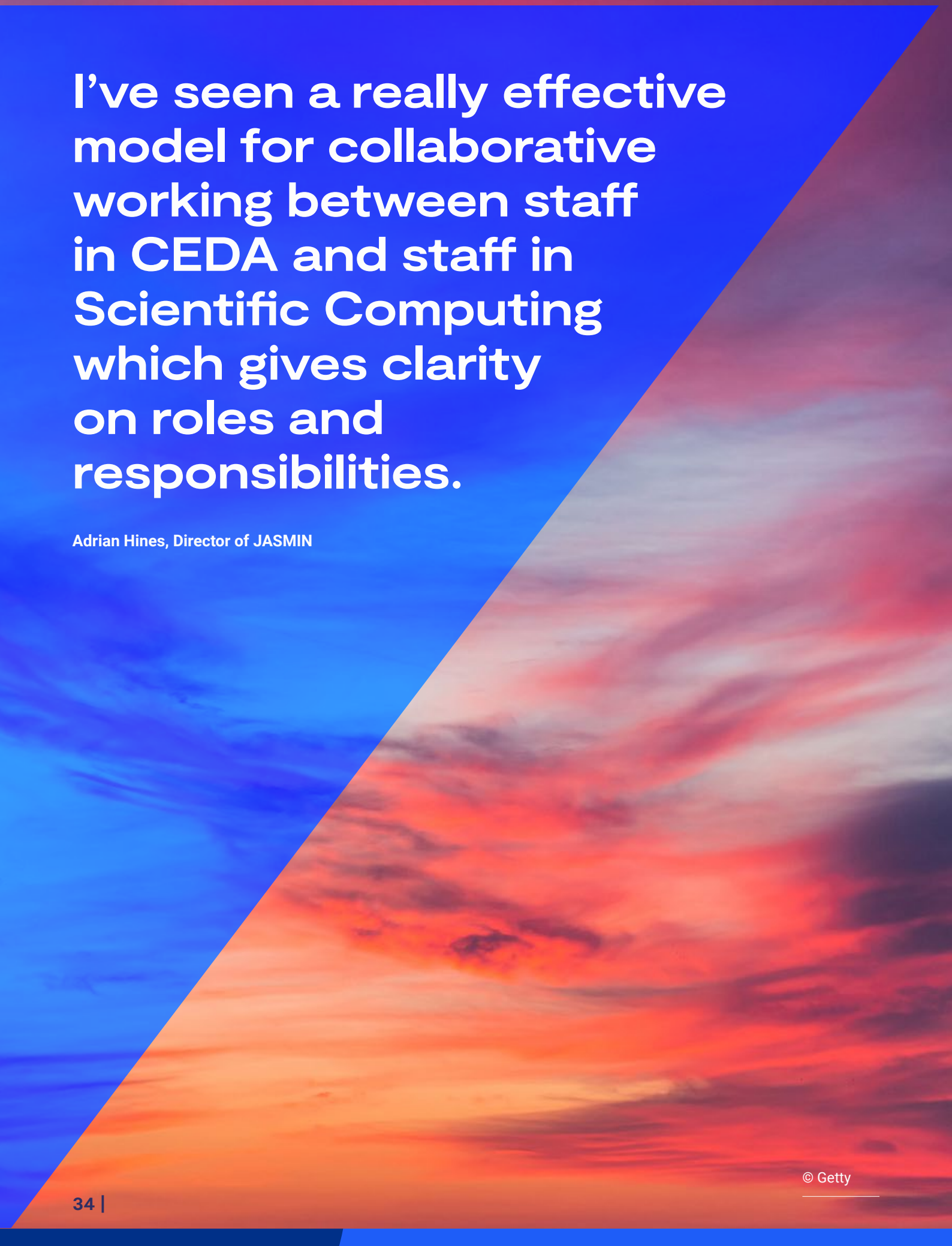
With a mind to the future, JASMIN was used as a case study helping to frame the challenges involved in moving towards net zero, and providing a basis for developing an environmental sustainability plan.

As a foundation for JASMIN's future, the governance structure has been fully established, comprising my own post as the JASMIN Director, a Programme Board, a Community Oversight Group, and a JASMIN Management Committee. This governance structure ensures representation across the key stakeholders and provides a robust framework for strategic and operational decision-making.

There are undoubtedly challenges ahead with the need to establish the position of JASMIN in an increasingly complex and crowded landscape, and the need to balance maintaining capacity and capability against pressures for future expansion. The firm roots that the team have set down over the years and the wide-ranging support means we are well placed to embrace these challenges as JASMIN, and the science it supports, continues to develop new shoots into the future.

I've seen a really effective model for collaborative working between staff in CEDA and staff in Scientific Computing which gives clarity on roles and responsibilities.

Adrian Hines, Director of JASMIN



© Getty

Boosting JASMIN's Infrastructure: A Year of Progress and Expansion

JAMES HANNAH

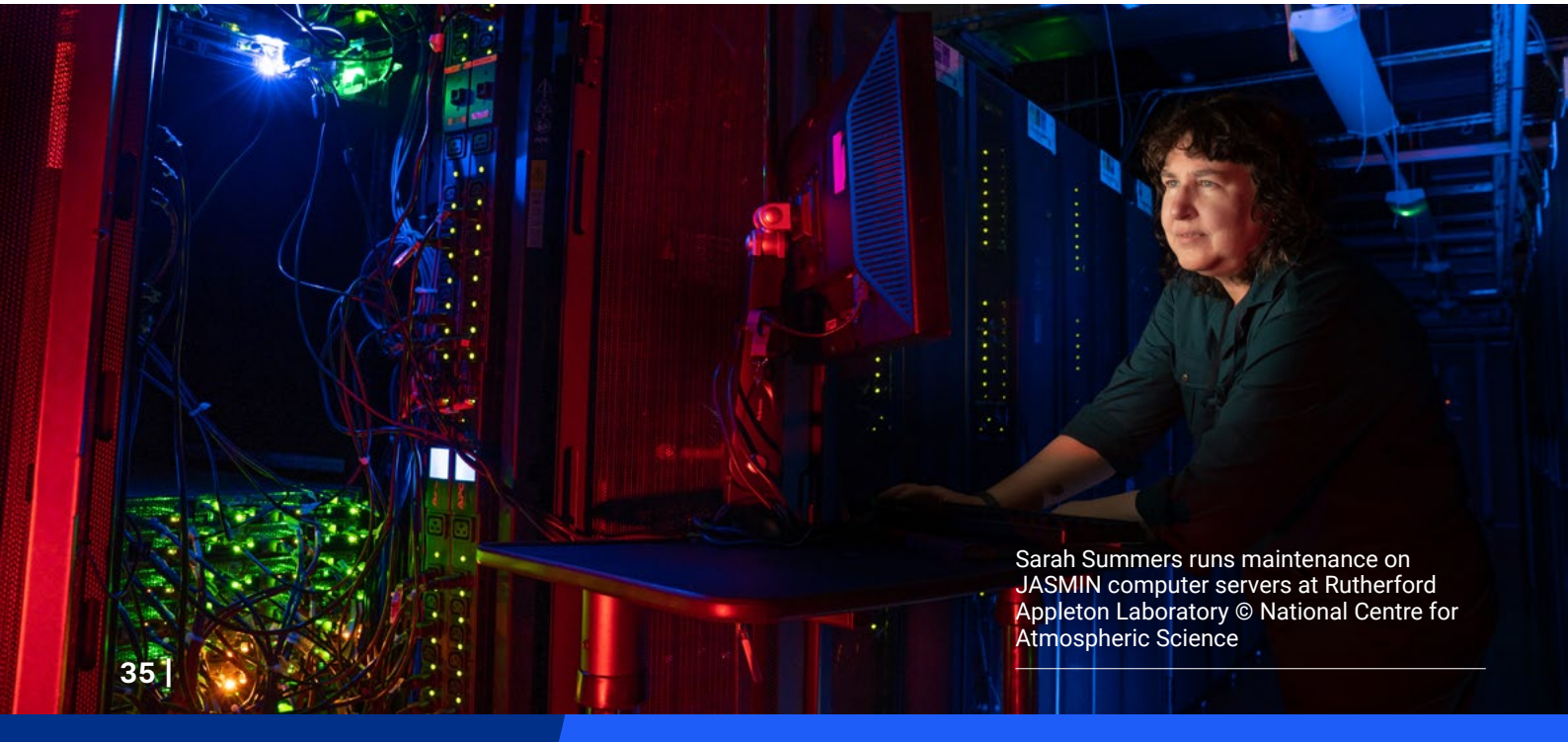
The infrastructure team from STFC Scientific Computing, who operate the underlying server infrastructure for JASMIN, have focused this year on expanding and improving the capacity and performance of the storage available to users in multiple key areas.

For users of the LOTUS batch computing environment, the main working storage is our parallel file system (PFS), which provides scratch storage and general high-performance storage for I/O heavy models. This parallel storage also equips users with numerous workspaces for specific projects that perform large numbers of parallel file read/writes and is used for high-bandwidth ingestion of remote datasets into the CEDA Archive. This year we've nearly doubled the amount of parallel filesystem space available in the cluster, which should ensure new projects can be granted the space they require, and more of this highly performant storage can be made available to existing project users.

The bulk of JASMIN's storage is provided using high-end storage servers and a software-defined "scale-out filesystem" solution. This storage is accessed whenever JASMIN users access files from the CEDA Archive and holds approximately 40 Petabytes of data, split between the Archive and group workspaces. This year we've increased the amount of physical hard-drive and solid-state disk (SSD) SOF storage available to the cluster by approximately 50% to improve the performance, capacity and reliability of the cluster. While this is not immediately usable by JASMIN users, it helps lead to the improved performance of the group workspaces and archive mounts. This allows us to move some of the oldest parts of this storage towards retirement.

The final part of storage we've improved this year is the high-performance SSD filesystem storage which provides user home directories and file shares for projects that use a large number of small files. These have different performance requirements to the other (PFS/SOF) storage platforms. We've also replaced two older flash storage devices to deliver greater performance and more storage than was previously provided.

Although the above purchases don't deliver a great deal of exciting new headline functionality, they should ultimately lead to greater performance for JASMIN users and increased platform reliability over the long term. This hardware was funded with approximately £5 million of funding delivered from NERC as part of the Digital Research Infrastructure programme.



Sarah Summers runs maintenance on JASMIN computer servers at Rutherford Appleton Laboratory © National Centre for Atmospheric Science

Accelerating Innovation: Unleashing the Power of GPUs with ORCHID

FATIMA CHAMI

Graphics Processing Units (GPUs), a mainstay of the video gaming industry, are powerful graphic cards primarily used to accelerate image analysis and other processing tasks. The architecture of GPUs, with hundreds of processors that perform specific tasks concurrently, makes them particularly suited for large data processing tasks such as artificial intelligence, image detection, and object inference. These are all techniques being explored by the environmental science community.

GPUs are invaluable in accelerating processing tasks, making them particularly useful when dealing with large datasets such as satellite images. A new dedicated GPU cluster, called ORCHID, has been procured as a new service via a £1 million investment from NERC. It offers efficient scaling and faster processing with more bandwidth.

During 2022, we began opening up access to the ORCHID cluster. A select number of users had initial access to ORCHID for 6-9 months before the official launch in January 2023. This allowed feedback and user experiences to shape our user support, configuration and software stack to permit full utilisation of this resource for Machine Learning and AI processes. We hosted an online webinar announcing ORCHID to the user community, demonstrating how to access ORCHID and its impressive capabilities. We allowed the test users to share their experiences and the wide applications of GPUs including model simulation, ocean bacteria analysis, and protein folding calculations. This was an opportunity for sharing recommendations that might benefit potential new users and it was of great interest to those embarking on research involving machine learning techniques, or who have been seeking GPU resources for use in their work.

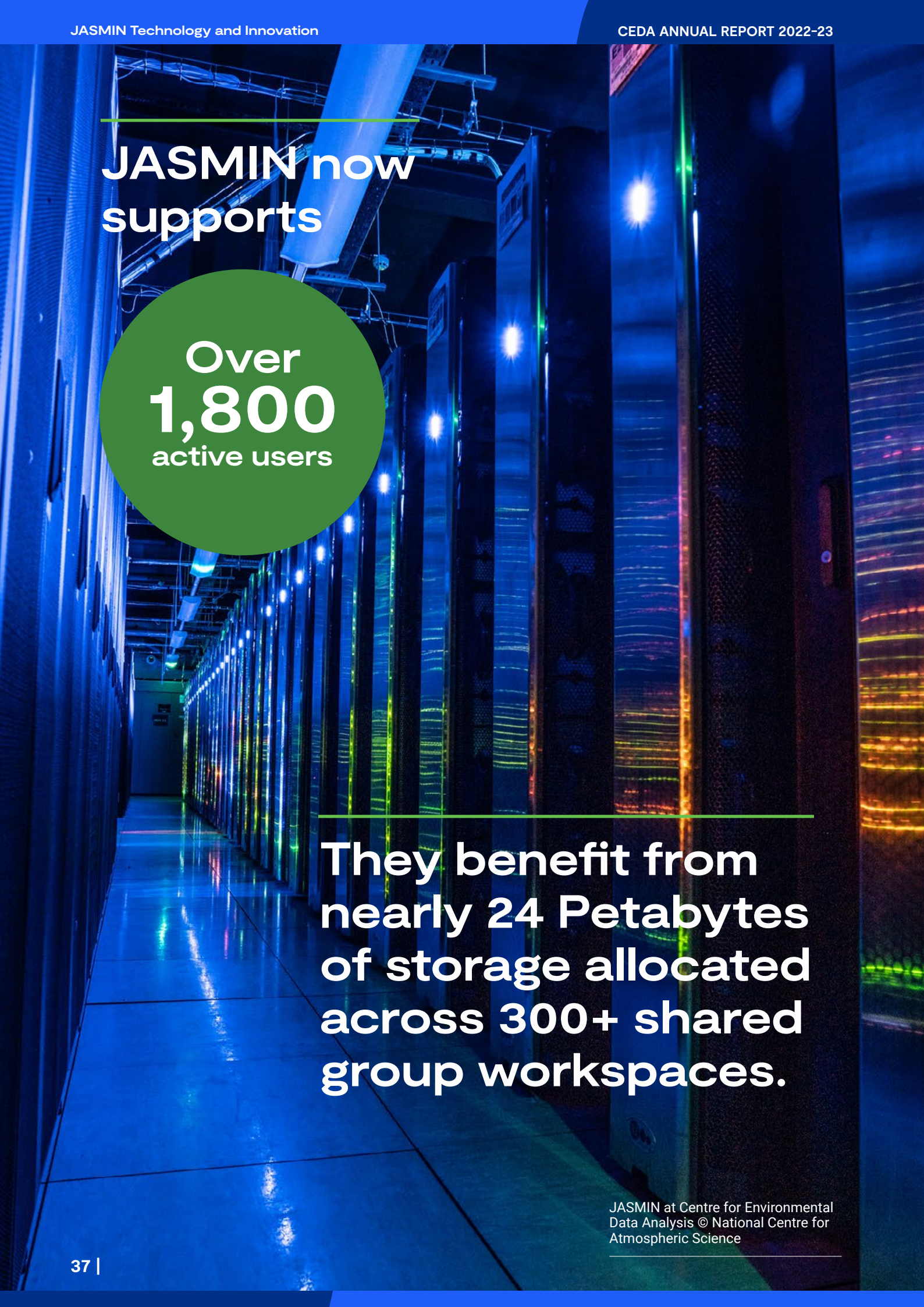
Transitioning to GPU processing may require code adjustments for users: a challenging but valuable endeavour for those wishing to harness the technology's power. We will deliver further collaborative training to our users on deep learning and computer vision analysis which harness the power of the GPUs. There is also a strategy to expand the GPU infrastructure to accommodate more users and projects that require GPU acceleration.

NERC recognises the need to feed into training in this area with a £117M doctoral training centre grant award for the AI in Environmental Sciences call in 2022 and ongoing capital grant calls. This marks an exciting time ahead for our team as we facilitate the acceleration of innovative science through the power of GPU technology.

JASMIN now supports

Over 1,800 active users

They benefit from nearly 24 Petabytes of storage allocated across 300+ shared group workspaces.



JASMIN at Centre for Environmental Data Analysis © National Centre for Atmospheric Science

Revolutionising Cloud Services: The Upgrade Journey to VIO7 and Beyond

MATT JONES

JASMIN provides users with a cloud-computing service, allowing an institution or project to provision their own virtual machines and clusters without needing to maintain associated physical infrastructure. Two different types of cloud services are available: the JASMIN External Cloud and the JASMIN Managed Cloud.

The JASMIN Managed Cloud service sits within the JASMIN firewall, meaning that users can conveniently access our storage and thus the CEDA Archive and Group Workspaces. Whilst tenants are not allowed root access and are limited to the types of virtual machines they can deploy, they are not responsible for the security or patching of these machines. In contrast, the JASMIN External Cloud sits outside of the main firewall, offers root access and the freedom to provision their own infrastructure such as web portals or remote desktop services. This has the caveat that they have full responsibility for their users and the security of their machines.

An additional aspect of the External Cloud is the availability of Cluster-as-a-Service (CaaS). This provides tenants with the opportunity to create clusters of machines that provide services including Kubernetes (see earlier highlight about the benefits of using Kubernetes). The CaaS system automatically provisions machines and installs the necessary prerequisites for each cluster. This allows users to host their own services using our infrastructure and gives them a lot of freedom and flexibility. These services can be distributed to non-JASMIN users who may not normally be eligible to use JASMIN.

The software that we used (VMware Integrated OpenStack 5 - VIO5) was being retired, meaning we had to upgrade to VIO7 in order to still provide a consistent and secure cloud service. This turned out to be a much bigger job than first anticipated and involved a huge amount of work and collaboration with our infrastructure colleagues in STFC's Scientific Computing (SCD).

The infrastructure team had to build a new cloud infrastructure and install VIO7 before we could work on it. Only then were we able to test the JASMIN cloud portal and update the CaaS system. The whole collaborative process took around 18 months in total.

A lot of time was spent trying to get the new CaaS system working with the new VIO and Ansible versions. Due to the complexity and challenge of the upgrade, we were limited to upgrading to a newer version of the existing software. In the future, we are looking to upgrade to an alternative, more robust system.

We have learned a lot from this mammoth upgrade! Equipped with our newfound knowledge, we are looking forward to an even brighter future for JASMIN cloud users.

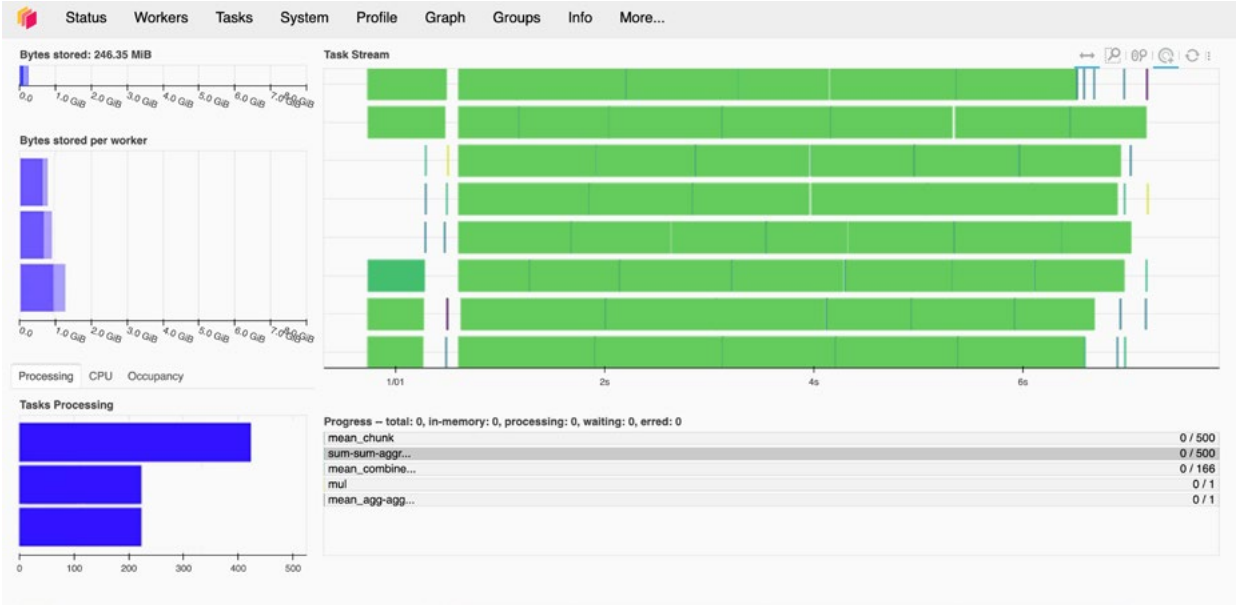
The JASMIN Cloud supports over 50 tenants, each provided with tools to build bespoke computing requirements, enabling them to exploit the power of JASMIN while reducing the overhead of infrastructure maintenance.

JASMIN hardware that is maintained by our colleagues in STFC Scientific Computing © National Centre for Atmospheric Science

Dask: Opening the Gateway to High Performance Parallel Data Processing

ALEX MANNING

This year, we have enabled parallel processing in the JASMIN notebook service, via the implemented use of the Dask parallel computing Python library. This work allows the use of the notebook service for larger and more complex data processing tasks.



The Dask gateway dashboard gives users a clear understanding of the task steam, the number of workers and how many tasks each worker is assigned

Providing easy parallelisation of Python code, allows for a huge number of workflows - from speeding up common Python tools, to deploying complex pipelines which enable the efficient processing of large data. Users can now harness the power of Dask to help them extract information from the huge datasets available on JASMIN, from inside the friendly Jupyterlab notebook environment.

In the past, it has not been possible to use Dask from the JASMIN notebook service, and running Dask on our LOTUS batch compute cluster has been technically difficult for users, as clusters have many parts which must all be able to communicate with each other.

This was particularly limiting for the notebook service since it meant users were limited to the 4 GiB of memory provided by the service, and therefore was much more difficult for them to process extremely large datasets.

To fix this problem, and allow users to use Dask with the JASMIN notebook service, we installed and configured a gateway server, which automatically creates and manages Dask clusters on their behalf. This means the users can benefit from the powerful processing capacity of Dask, without having to understand the technical details of how to interact with LOTUS.

We anticipate this added processing capability to be particularly useful for users with limited parallel computing knowledge but who wish to analyse very large datasets. Future plans will include creation of new example use cases and provision of training materials - adding to the existing resources already available.

Metrics, Finance, Projects and Outputs 2022–23

CEDA has sought to continuously deliver high-quality services to our communities throughout our operational life. Earlier articles have provided specific insights into CEDA's activities and this section rounds off the annual report by reporting our key usage metrics followed by finance and project details. It also lists our other outputs: publications, posters and talks followed by a list of training delivered by CEDA staff during 2022-23.

People working inside FAAM aircraft © National Centre for Atmospheric Science

Summary and usage information

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 worldwide 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.

Data Archive Service

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).

Data Curation

663 new datasets have been created and made available via the CEDA catalogue - bringing the total number of datasets to 8565. 145 new DOIs were issued in this period, bringing the number of DOIed resources to 647.

Annual CEDA Archive Deposits: April 2022 to March 2023

Total volume deposited	4.26 Pb
Total number of files deposited	22,952,884
Total number of accesses	103,789,965
Total dataset deposited to	2274

The Archive now holds over 20.3 Petabytes of data in 345,417,850 files.

Usage of CEDA Data

663 new datasets have been created and made available via the CEDA catalogue - bringing the total number of datasets to 8565. 145 new DOIs were issued in this period, bringing the number of DOIed resources to 647.

Annual CEDA Archive Usage: April 2022 to March 2023

Total number of users/distinct IP addresses	155,698
Total data downloaded	1.11 Pb
Total number of accesses	82,586,951
Datasets accessed	6079

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

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).

For datasets requiring users to register for access additional demographics (geographical area and institute type) are available. 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.

	New users period 2022-23	5 year period 2017-22
Registrations	10,102	41,642
% new users from UK	52	47
% new users in Universities	61	66

Table: New registered users statistics for reporting year and preceding 5 year period.

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 2023 there are:

- Over 1,800 active JASMIN users
- Over 300 shared group workspaces, with nearly 24 Petabytes of storage allocated to them.
- Over 50 tenancies in the JASMIN community cloud

General overview

Supporting our users

Over 5736 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.

1075 applications were processed for access to restricted datasets.

Collaborations

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

In 2022/2023, 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. Working with 20+ partners in the European Network for Earth System Modelling (ENES), CEDA is working to develop software and services for climate model data archives.
3. Operating and evolving the Earth System Grid Federation (ESGF) in partnership with ENES and global network of climate data service providers led by Oak Ridge National Laboratory and the US Programme for Climate Model Diagnosis and Intercomparison in support of the sixth Coupled Model Intercomparison Project (CMIP).
4. Working with the wider UK atmospheric science and earth observation communities, via a range of projects, with NCAS and other NERC funding.

5. Working with the European Space Agency on projects such as the ESA Climate Change Initiative (CCI) Knowledge Exchange.
6. Delivering part of the UK Collaborative Ground Segment for Sentinel data (with UK Space Agency, Airbus, Satellite Applications Catapult) with the role to provide Sentinel data mirror archives and data processing capability for the UK academic community.
7. Operating a Data Hub Relay (DHR) on behalf of ESA as a fully functioning member of the ESA DHR network. This supports CEDA use of this for Copernicus data retrievals for the CEDA Mirror Archive.
8. Representing the UK internationally on the inter-space agency CEOS data group (WGISS) via a CEDA staff member and being actively involved in WGISS Jupyter Notebooks Best Practice workstream, organizing and leading CEOS Jupyter Notebooks Day – Supporting Global Best Practice and Training.
9. Supporting the preparation and running of the PV 2023 conference.
10. Working with ECMWF to provide EO scientists with the high resolution atmospheric analyses they need to process satellite observations.
11. Disseminating climate and weather data to the research community, working with colleagues in the UK Met Office.
12. Supporting the Climate and Forecast Metadata (CF) Conventions with partners in University of Reading, UK Met Office, and multiple US research institutions.
13. Working with academic partners in the UK Research and Innovation Cloud Working Group and NERC Digital Research Infrastructure Group to share best practice, knowledge and strategy for use of cloud computing in the research domain.
14. Working with the CMIP Panel, the WGCM Infrastructure Panel, and a range of CMIP Task Forces to develop the collaborative framework for the next phase of CMIP and with the IPCC Task Group on Data (TG-DATA) to ensure transparency and clear lines of provenance for data cited in the IPCC Assessment Reports.
15. Working with the Research Data Alliance to develop and advance FAIR data standards and research data management practices and career pathways.

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.

Annual total funding

Financial Year	13-14	14-15	15-16	16-17	17-18	18-19	19-20	20-21	21-22	22-23
NCAS income	829	829	808	808	808	808	808	808	808	813
NCEO income	392	390	393	393	393	402	393	418	393	396
Other NERC income	272	600	621	825	816	733	883	784	1371	1789
Other income	1486	1394	1505	1092	1280	1458	1377	1476	1391	1590
Total income	**2979**	**3213**	**3327**	**3118**	**3297**	**3401**	**3461**	**3486**	**5203**	**4588**

Table: Overall funding for CEDA for financial years 2013/14 to 2022/23 (in £k)

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)
EO DataHub	Collaborative project involving public sector and industry partners to develop a UK digital infrastructure for the exploitation of Earth Observation data to support industry, public sector and the research community	DSIT	01/02/2023	31/03/2025	9700.0
ESA CCI Knowledge Exchange	Data archive for ESA Climate Change Initiative as part of wider activity including outreach and education	ESA	5/9/2019	15/10/2022	445.0
UKSA DAP Support 20-21	Funding Esther Conway to attend ESA Data Access and Preservation WG for UKSA	UKSA	01/04/2020	31/03/2023	24.0
IS-ENES3	Phase 3 of the distributed e-infrastructure of the European Network for Earth System Modelling	H2020	01/01/2019	31/03/2023	780.7
JASMIN for ESA SST CCI+	JASMIN Support for ESA CCI+ Sea Surface Temperature processing	ESA	01/07/ 2019	30/06/2022	30.0
ESA Data Hub Relay	Operation of a Sentinel Data Hub Relay for ESA	ESA	01/04/2021	28/02/2024	739.1
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	01/04/2021	31/03/2024	450.0
EO Data Architecture	Funding for the UKSA Call for strategy on EO Data Architecture	UKSA via Telespazio	01/03/2022	30/04/2022	19.9
UKRI Net Zero Digital Research Infrastructure	Scoping project to investigate how UKRI can achieve net zero computing	UKRI	01/11/2021	30/06/2023	440
ClimatEurope2	Horizon Europe project on standards for climate services	EC	01/09/2022	28/02/2027	88

Table: Externally funded projects with a value greater than £15k for 2022/23 (non-core NERC)

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 and the creation of an independent governance structure for JASMIN.

Publications list

Conway, E., CEOS Jupyter notebooks for Earth Observation: Best Practice and Capacity Development (Presentation), ESA Living Planet Symposium, May 2022.

Conway, E., WGISS Jupyter Notebooks Day and Best Practice (Presentation), CEOS WGISS 54, October 2022.

Conway, E., CEOS Jupyter Notebooks Day – Supporting Global Best Practice and Training, October 2022 (Presentations)

Evans, R., STAC at CEDA (Presentation), Cloud-Native Geospatial Outreach Event, May 2022

Garland, W., UAS data considerations (Presentation), NERC Net Zero Aerial Capability (NZArC):Uncrewed Aerial Systems (UAS) Workshop, June 2023.

Hines, A., Bah, B., Berod, D., Bouchet, V., Braconnot, P., Cao, W., Govett, M., Graham, T., Honda, Y., Jansons, E., Jean, M., Lawrence, B., Rowe, K., Luterbacher, J., Schultz, M., Visbeck, M., Duan, Y., & Li, J. (2023). WMO Concept Note on Data Handling and the Application of Artificial Intelligence in Environmental Modelling. World Meteorological Organisation, https://library.wmo.int/idurl/4/66272.

Hooper, D. A., Brooks, B. J., **Garland, W. E., & Parton, G. A.** (2022). The National Centre for Atmospheric Science Image Metadata Standard (NCAS-IMAGE) version 1.0 (1.0), Zenodo. https://doi.org/10.5281/zenodo.6368295

Juckes, M., Pascoe, C., Woodward, L., Vanderbauwhede, W., & Weiland, M. (2022). Interim Report: Complexity, Challenges and Opportunities for Carbon Neutral Digital Research. Zenodo. https://doi.org/10.5281/zenodo.7016952

Juckes, M., Cofino, A., Supporting standards for CF and Data Request (Presentation), IS-ENES3 Final General Assembly, January 2023.

Juckes, M., Preparing CMIP7: task groups/info from the WIP (Presentation), IS-ENES3 Final General Assembly, January 2023.

Juckes, M., Adloff, F., Joussaume, S., Kindermann, S., Lautenschlager, M., Lawrence, B., Levavasseur, G., Fiore, S., Pagé, C., & Spinuso, A. (2022). ENES CDI - Mission and Objectives. Zenodo. https://doi.org/10.5281/zenodo.6451925

Juckes, M., Bessembinder, J., (2023): IS-ENES3 Final Report on Innovation. https://raw.githubusercontent.com/IS-ENES3/IS-ENES-Website/main/pdf_documents/IS-ENES3_D2.3.pdf

Lawrence, B., Adloff, F., André, J.-C., Biercamp, J., Fladrich, U., Hines, A., Joussaume, S., **Juckes, M.**, Serradell, K., & Valcke, S. (2022). European objectives for Models, Tools and HPC in the future research infrastructure. https://doi.org/10.5281/zenodo.6807389

Juckes, M., Bärring, L., Cofiño, A. S., Stockhause, M., 2023. IS-ENES3 Standards Synthesis. https://raw.githubusercontent.com/IS-ENES3/IS-ENES-Website/main/pdf_documents/IS-ENES3_D3.3.pdf

Juckes, M., Bessembinder, J., Rioual, J., Valcke, S., Acosta, M., Blockley, E., Maisonnave, E., Matthews, D., Caubel, A., Meurdesoif, Y., 2023. White paper on innovation on tools, platforms, and techniques. https://raw.githubusercontent.com/IS-ENES3/IS-ENES-Website/main/pdf_documents/IS-ENES3_D4.5.pdf

Kindermann, S., Stockhause, M., Nuzzo, A., **Juckes, M.**, Hassel, D., Levavasseur, G., Spinuso, A., 2023. Final KPI and TA report for ENES CDI data services. https://raw.githubusercontent.com/IS-ENES3/IS-ENES-Website/main/pdf_documents/IS-ENES3_D7.6.pdf

Kershaw, P. J., Parker, R., Quaife, T., Pinnington, E., Beavis, P., Building a Digital Twin Earth Precursor: The Climate Impact Explorer (Presentation), ESA Living Planet Symposium, 26 May 2022

Kershaw, P. J., Stephens, A., Smith, R., Evans, Pepler, S., Parton, G. A., Wilson, A., Rahman, M., Massey, N., Building cross-domain environmental catalogues for big data archives with STAC (Poster), ESA Living Planet Symposium, 26 May 2022

Kershaw, P. J., Nassisi, P., ESGF: Main achievements for CMIP6, main evolutions and towards CMIP7 (Presentation), IS-ENES3 Final General Assembly, January 2023.

Kershaw, P. J., IS-ENES3 Project Deliverable D5.3, Architecture Design Plans, November 2022 https://raw.githubusercontent.com/IS-ENES3/IS-ENES-Website/main/pdf_documents/IS-ENES3_D5.3.pdf

Kershaw, P. J., Earth System Grid Federation: A Globally Distributed Software Infrastructure to Address The Challenges of Data Management, Access and Analysis for High Volume Climate Model Data, Convection-Permitting Modelling Workshop (Presentation), Buenos Aires, 7-9 September 2022, https://docs.google.com/presentation/d/1iBM3KQMdy8WYgHB-mPgUkrKoe7eBmd-x/edit#slide=id.p1

Lawrence, B., Kunkel, J., Cole, J., Lister, G., Pedro, L., Wilson, S., McKinstry, A., **Massey, N**., & Meurdesoif, Y. (2023). Data Handling Software Documentation and Roadmap - Deliverable D4.3. Zenodo. https://doi.org/10.5281/zenodo.7599330

Magagna, B., Moncoiffé, G., Devaraju, A., Stoica, M., Schindler, S. and **Pamment, A**: InteroperAble Descriptions of Observable Property Terminologies (I-ADOPT) WG Outputs and Recommendations, RDA Endorsement 3rd May 2022. DOI:10.15497/RDA00071.

Pamment, A., Bartholomew, S., Castelão, G., Cofiño, A., Davis, E., Hassell, D., Heaps, A., Sevadjian, J., Climate and Forecast (CF) community training event, September 2022. https://cfconventions.org/Training/2022-Training-Workshop.html

Parton G. A., Brooks B., **Stephens A., Garland W. E.**, The UK's NCAS Data Project: establishing transparent observational data workflows from field to user, Feb 2023. https://doi.org/10.5194/egusphere-egu23-16288

Parton, G. A., Levelling up data discovery at a UK domain repository, Research Data Alliance 19th Plenary (RDA19), Seoul, June 2022. Zenodo. https://doi.org/10.5281/zenodo.6685739

Ayres, B., Lehtsalu, L., **Parton, G.**, Száldobágyi, Á., Warren, E., Whyte, A., & Zimmer, N. (2022). RDA Professionalising Data Stewardship - Current Models of Data Stewardship: Survey Report (1.0). Zenodo. https://doi.org/10.15497/RDA00075

Ayres, B., Lehtsalu, L., Kuchma, I., **Parton, G.**, Száldobágyi, Á., Warren, E., Whyte, A., & Zimmer, N. (2022). RDA Data Stewardship Organisational Models Survey 2021 Output Dataset [Data set]. Zenodo. https://doi.org/10.5281/zenodo.6665306

David, R., Specht, A., O'Brien, M., Wyborn, L., Drummond, C., Edmunds, R., Filippone, C., Machicao, J., Miyairi, N., **Parton, G.**, Pignatari Drucker, D., Stall, S. & Zimmer, N., (2022). Multilingual Data Challenges in Professionalizing Data Stewardship worldwide (V1.1). RDA 19th Plenary Meeting, Part Of International Data Week, 20–23 June 2022 (RDA Plenary 19th), Seoul, South Korea. Zenodo. https://doi.org/10.5281/zenodo.6588167

Siddorn, J., Blair, G. S., Boot, D., Buck, J. J. H., Kingdon, A., Kloker, A., Kokkinaki, A., Moncoiffé, G., Blyth, E., Fry, M., Heaven, R., Lewis, E., Marchant, B., Napier, B., **Pascoe, C.**, Passmore, J., **Pepler, S.**, **Townsend, P.**, & Watkins, J. (2022). An Information Management Framework for Environmental Digital Twins (IMFe). Zenodo. https://doi.org/10.5281/zenodo.7004351

Huard, D., Pirani, A., Chen, R., Gutiérrez, J. M., **Juckes, M.**, Krey, V., Spinuso, A., & Stockhause, M. (2022). IPCC Data and Code Licensing Guidelines (1.1). Zenodo. https://doi.org/10.5281/zenodo.7431834

Pirani, A., Alegria, A., Al Khourdajie, A., Gunawan, W., Gutiérrez, J. M., Holsman, K., Huard, D., **Juckes, M.**, Kawamiya, M., Klutse, N., Krey, V., Matthews, R., Milward, A., **Pascoe, C.**, van der Shrier, G., Spinuso, A., Stockhause, M., & Xing X., (2022). The implementation of FAIR data principles in the IPCC AR6 assessment process. Zenodo. https://doi.org/10.5281/zenodo.6504469

Pritchard, M., JASMIN's data transfer zone (Presentation) JISC online workshop for research network engineers, February 2023.

Stephens, A., Kershaw, P. J., Plenary Presentation: ESA Earth Observation Exploitation Platform Common Architecture (Presentation), International ESGF Hybrid Conference Toulouse/Oak Ridge, January 2023.

Lambert, S., **Stephens, A.**, & Kayumbi, G. (2023). Net Zero DRI Case Studies Report. Zenodo. https://doi.org/10.5281/zenodo.7828052

Townsend, P., Schien, D., Hays, J., Bane, M., Turner, A., Basden, A., **Stephens, A.**, Lambert, S., Smith, L., Friday, A., Boulton, J., McCarroll, N., Ahrens, F., McGuire, L., & Richardson, H. (2023, February 6). Net Zero DRI Project Partners Webinars - slides for all talks. Zenodo. https://doi.org/10.5281/zenodo.7612560

Training Events

All training events run or supported by CEDA are shown below - for further information check out our website. Overall CEDA provided direct training for over 300 people this year!

Introduction to Scientific Computing course

Taught virtually in 2022-23 (Open registration)

Module 1 - Introduction to the bash shell (Unix/Linux), 17th February 2023 - 12 attendees

Module 2 - Introduction to Git and GitHub (Software version control), 20th February 2023 - 11 attendees

Module 3 - Introduction to Python Programming, 21st-23rd February 2023 - 9 attendees

Module 4 - Python: Working with Data, 28th February - 2nd March 2023 - 12 attendees

JASMIN Beginners' Workshops

(1-day workshop for new JASMIN users)

Leeds, 5th April 2022, 17 attendees from SENSE Centre for Doctoral Training

Virtual, 12th October 2022, 25 attendees, including 14 from AI4ER Centre for Doctoral Training

Virtual, 26th January 2023, 26 attendees (open registration)

Leeds, 28th March 2023, 13 attendees from SENSE Centre for Doctoral Training

CPDN workshop event for early career scientists in Brazil between 12th-23rd Sep 2022 (daily afternoon sessions). Organised by the University of Oxford, 20 attendees used training accounts and computer resources

JASMIN webinars

1-2 hour virtual events, focussing on a single topic (open registration)

Hydro-JULES Summer School, 11-15th July 2022: Jasmin Virtual Field Trip (13th July 2022), 13 attendees

Hydro-JULES Winter School, 16-20th January 2023: Jasmin Virtual Field Trip (18th January 2023), 20 attendees

JASMIN's new GPU cluster, ORCHID, 31st January 2023, over 100 attendees

CF Training

A 3-hour workshop for the CF metadata community (open registration)

Virtual, 8th September 2022, 25 international attendees

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

JASMIN

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

UKRI STFC and NERC

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