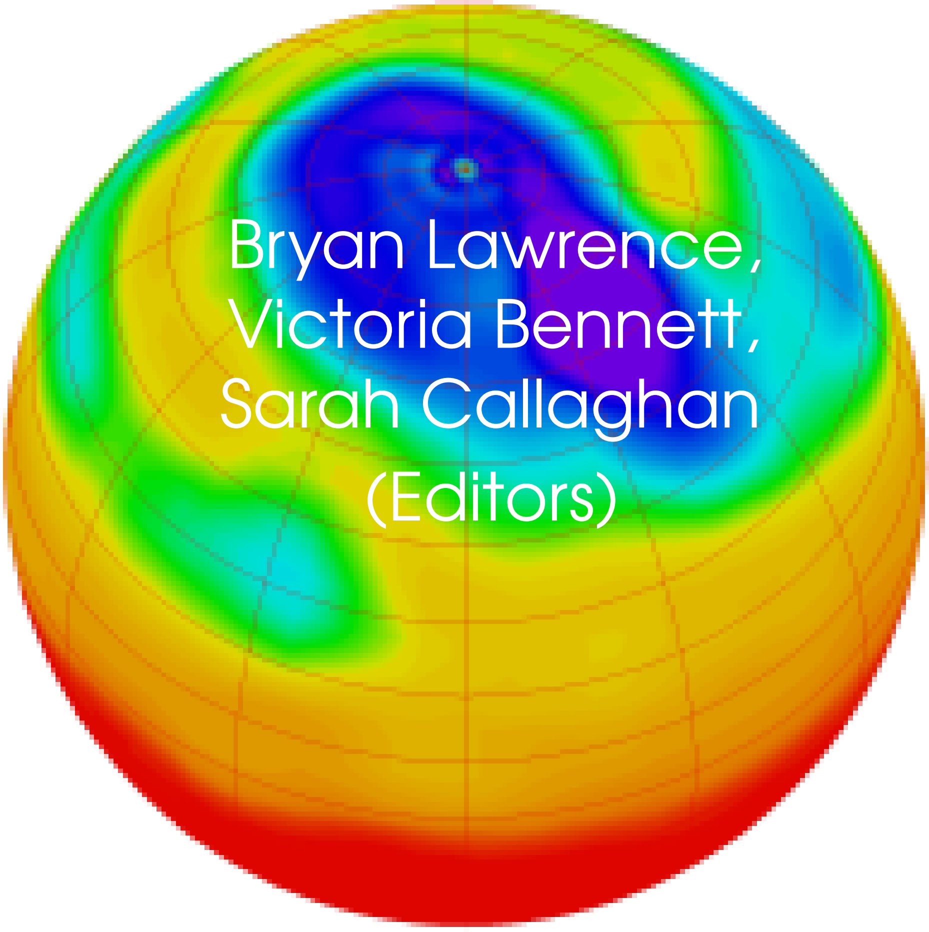
Science & Technology
Facilities Council

Centre for Environmental Data Archival (CEDA) Annual Report 2015

(Apr-2014 to Mar-2015)



Bryan Lawrence,
Victoria Bennett,
Sarah Callaghan
(Editors)

British Atmospheric
Data Centre
NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE
NATURAL ENVIRONMENT RESEARCH COUNCIL

NERC SCIENCE OF THE ENVIRONMENT

National Centre for
Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

Introduction from the Director

The mission of the Centre for Environmental Archival (CEDA) is to deliver long term curation of scientifically important environmental data at the same time as facilitating the use of data by the environmental science community.

CEDA was established by the amalgamation of the activities of two of the Natural Environment Research Council (NERC) designated data centres: the British Atmospheric Data Centre, and the NERC Earth Observation Data Centre, and consolidated annual reports have been produced since 2009. This annual report presents key statistics for the year past (2014-2015) as well as a series of snapshots of activity, expressed as short highlights and short reports. Key data centre metrics are also provided.

This year was characterised by the major upgrade in JASMIN capability (JASMIN being the data intensive supercomputer which provides the fabric upon which CEDA and the CEDA services are delivered). As in the advent of JASMIN itself, in early 2012, this has had a major impact both on what CEDA can do, and the services that can be offered to the community — and the uptake in the community is clear to see in the statistics presented here. The consequences for CEDA itself are still being worked through, but it is clear that there is a significant expansion in the role of CEDA staff whose expertise must now embrace support for "big data" tools and algorithms by archive and other users on JASMIN. Notwithstanding this new role, as in previous years, CEDA staff are involved in nearly all the major atmospheric science programmes underway in the UK, in many earth observation programmes, and in a wide range of informatics activities.

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

Bryan Lawrence

July 31st, 2015

Contents

Part I

Summary

The original CEDA mission is 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. As in previous years, CEDA enhanced this role through the development and maintenance of tools and services to aid data preservation, curation, discovery and visualisation — adding value for the world-wide user community.

In recent years this role has further expanded to include support for the JASMIN data analysis environment, so this section of the annual report has been reorganised to present not only summaries of CEDA usage, but also of JASMIN.

1. Usage of CEDA data

CEDA delivers a number of data centres for different customers, but the primary strategic activity is the delivery of the British Atmospheric Data Centre for the National Centre for Atmospheric Science and the NERC Earth Observation Data Centre for the National Centre for Earth Observation along with the UK Solar System Data Centre (UKSSDC) jointly for NERC and STFC.

Smaller data centre activities include managing aspects of the IPCC Data Distribution Centre for the IPCC under contract to the UK Department of Energy and Climate Change, the UK Energy Research Data Centre (EDC) and the data centre for the WCRP programme Stratospheric Processes and their Role in Climate (SPARC).

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

Annual CEDA Usage: April 2014 to March 2015	
Total number of users	10,214
Total data downloaded	444 TiB
Total number of accesses	11,831,481
Total days activity	41,803

Table 1.1: Summary figures for usage by CEDA consumers during the reporting year. Here we define an "access" as a file download, whether to disk, or into a browser — not just a website hit.

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

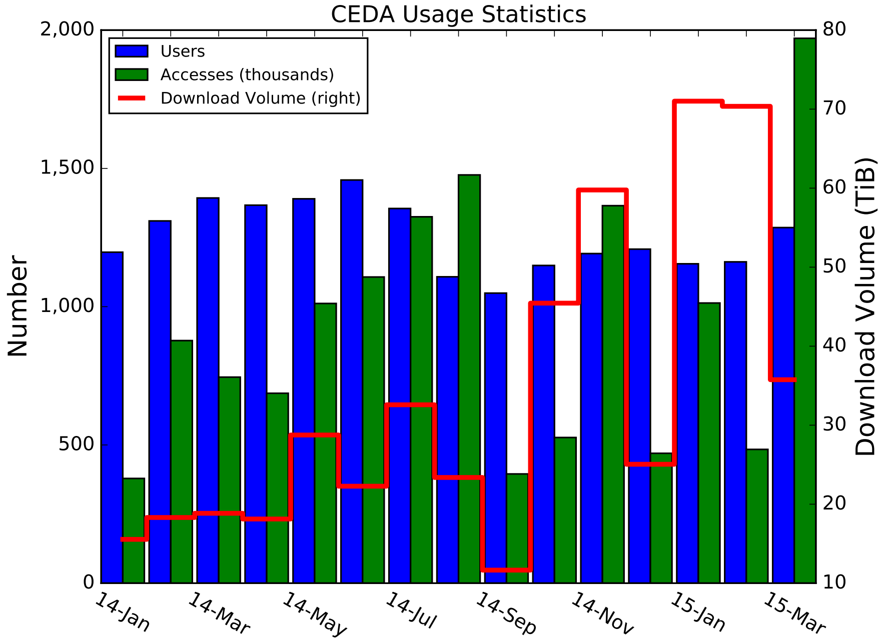


Figure 1.1: Breakdown of CEDA usage by month. The number of users and access (in thousands, so for example there were 378,000 accesses in January 2014) share the left hand axis and the download volumes use the right hand axis.

2. Usage of JASMIN

JASMIN is a custom integrated data intensive super-computer providing four main functions: storage, batch computing, managed infrastructure compute and a private cloud to provide infrastructure as a service (the JASMIN "un-managed cloud").

2.1 JASMIN Storage

Most of the storage growth in 2014/2015 came from expansion by users in "Group Work Space" storage. Note that without the JASMIN phase2 upgrade in spring 2014, the system would have been full by the end of 2014 (JASMIN phase1 had a maximum of 5 PB of fast disk).

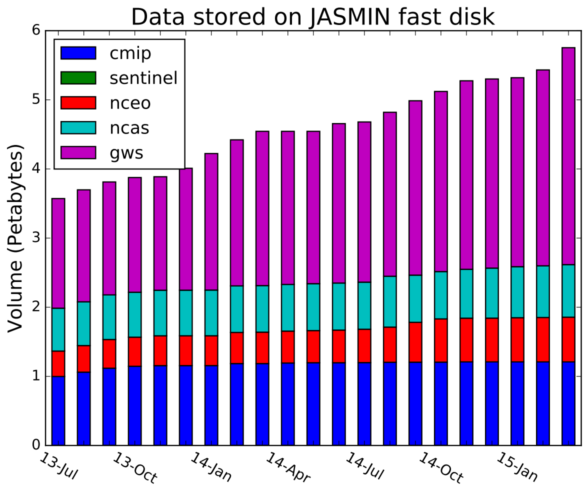


Figure 2.1: Usage of JASMIN storage broken down into major categories of stored data: (1) CMIP data; (2) data from the Sentinel satellites (none in 2014/15); (3) other earth observation data held in the CEDA archive (NCEO); (4) other atmospheric and climate science held in the CEDA archive (NCAS), and; (5) the generic user storage or Group Workspaces, GWS. (Data volumes from the UKSSDC, the EDC, and SPARC are negligible at this scale and not included.)

Within this period (April 2014 to March 2015), the CEDA archive grew in volume by about 0.3 PB, a growth of order 10%, but it grew by around 30% in terms of the number of files, primarily due the ingestion of UKSSDC data into the CEDA system.

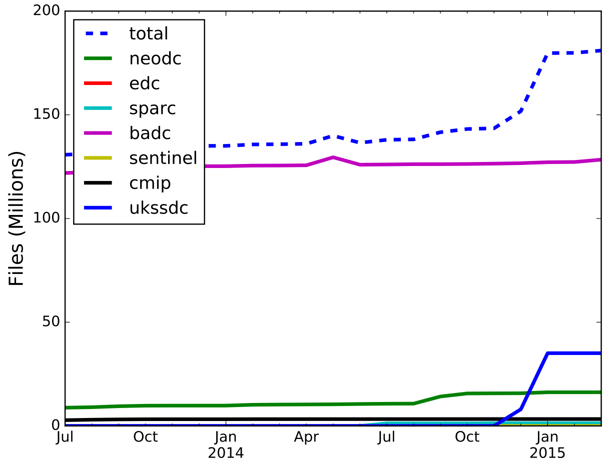


Figure 2.2: Files curated by CEDA as a function of time by data centre (EDC, NEODC, BADC, SPARC, UKSSDC) and major dataset (CMIP, Sentinel). File numbers are dominated by the observational data held by the BADC and UKSSDC (the large increase in early 2015 in the UKSSDC numbers are due to the ingestion of EISCAT data into the CEDA archive).

2.2 LOTUS — JASMIN batch computing

At the beginning of the reporting period the Lotus batch cluster was heavily utilised, with cpu-utilisation reaching 50% and data movement peaking at 3 PB/month. Unlike traditional HPC systems, the priority for Lotus is to be available for data processing with low wait times — not to be running at 80%+ utilisation. In mid-2014 the cluster was significantly upgraded. Within a year not only had data movement doubled, but utilisation was back to similar figures.

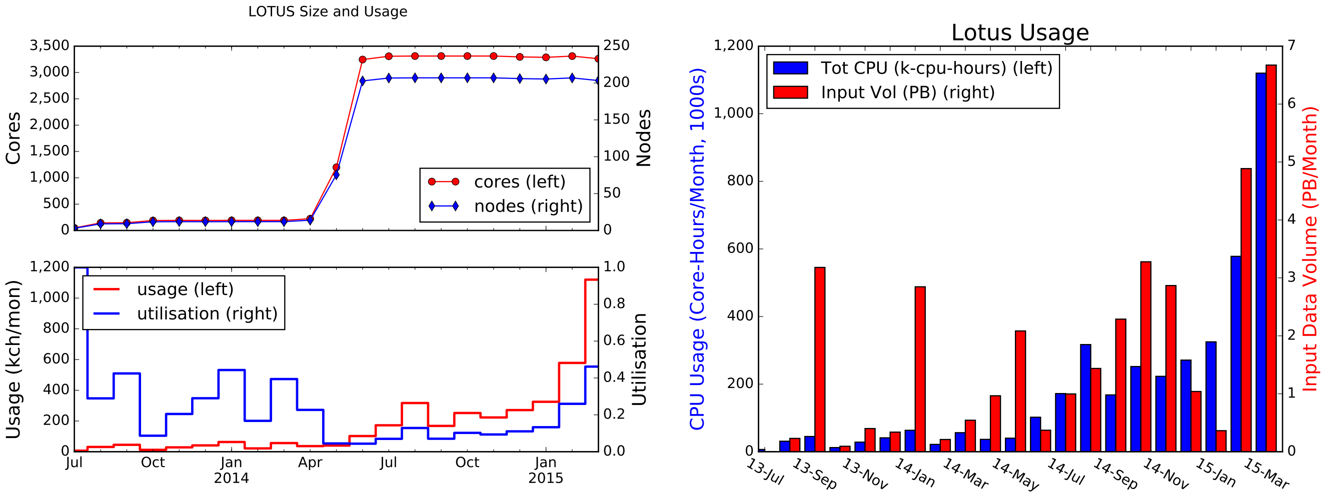


Figure 2.3: Lotus batch cluster size and usage. Left hand panel shows the size of the cluster and cpu hours and utilisation since mid-2013. Right hand panel shows the correlation between cpu usage and data movement into and out of the cluster.

2.3 JASMIN Managed Infrastructure

In the managed infrastructure we measure CPU utilisation only in terms of cores available and core-hours; given most of the cores are virtualised, normal utilisation figures in terms of cpu-hours divided by core-hours-available would be misleading. Data movement in the managed infrastructure is broken down into

1. The data movement into the science and ceda machines, which gives us a measure of how the managed infrastructure is being used for data manipulation (and which can be compared with Lotus), and
2. Data movement in and out of the JASMIN analysis environment via the user transfer machines, and the CEDA download services — and movement to and from tape. These give us a measure of how much data would otherwise have been handled in the wider community (and likely duplicated many times).

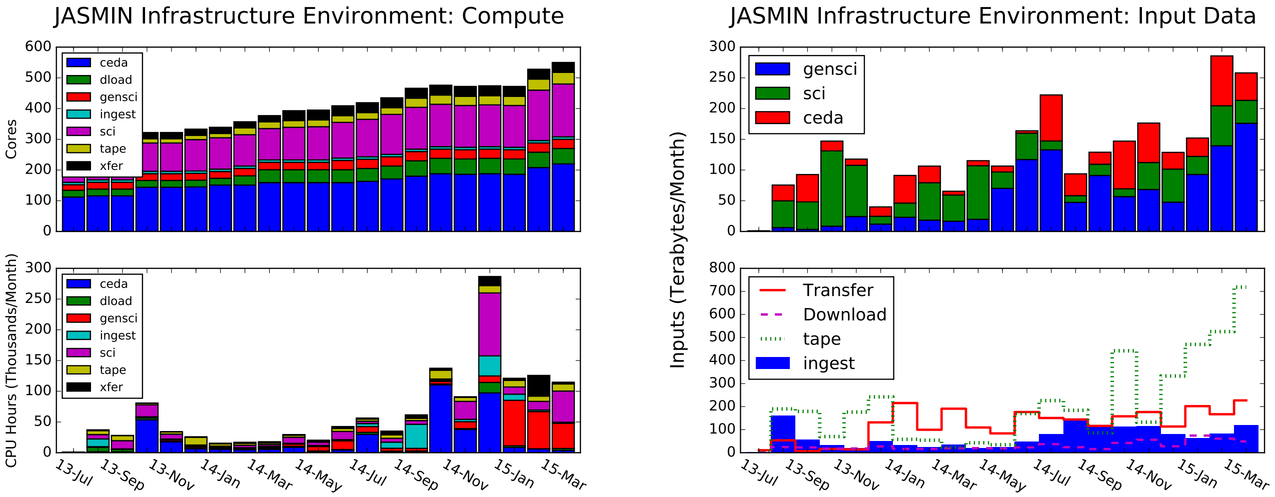


Figure 2.4: Managed Cloud and Service Machine usage. Usage in this service environment is broken into seven categories: (a) ceda — the bulk of the machines used in data management; (b) dload — machines (such as the ftp server) accessed by users for download services; (c) gensci — the generic science machines used by most JASMIN users; (d) ingest — the CEDA gateway machines for data going into the archive; (e) sci — bespoke science machines for specific communities; (f) tape — machines for handling tape traffic; and (g) xfer — machines customised for bulk data input/output by JASMIN users. The top left panel shows the number of cores (most of which were virtualised) deployed in the service environment. The bottom left panel shows the CPU usage generated by those cores. The top right panel shows the data input in the service environment while the bottom right panel shows the data passing through the edge machines (transfer in and out, download, to the tape machines, and ingested into the archive). (Note that the tape movements include multiple transfers for each copy to tape, so these are not the aggregate volume actually written to tape).

There was a near doubling in the number of cores deployed in the managed infrastructure during 2014-15, most of which were virtual machines and came from growth in customised machines for CEDA and for specific science activities (such as to support near real time data handling, the Met Office, etc). CPU utilisation in these machines is variable, but at times was comparable with Lotus, suggesting that for simple computing tasks users valued customised computing as much as raw power on vanilla machines —however, for data handling, the compute nodes in the service environment were moving more than ten times less data than Lotus.

Data transfers in and out dwarfed those of the conventional CEDA archive downloads.

3. Collaborations

CEDA continues to support international climate modelling community through its interactions with the large global collaboration to deliver an “Earth System Grid Federation” under the auspices of the Global Organisation for Earth System Science Portals (GO-ESSP). Other major international collaborations include participation in the infrastructure projects (originally IS-ENES and now IS-ENES2) to support the European Network for Earth system Simulation (ENES).

3.1 Major Collaborations

In 2014/2015, significant national and international collaborations have been continued and/or begun. On the national scale, CEDA itself reflects a collaboration between the earth observation community, the atmospheric sciences community (via NCEO and NCAS) and the space weather community.

Additionally, CEDA is:

1. Working closely with the other NERC Environmental Data Centres, as part of the NERC Data Operations Group.
2. Operating and evolving the Earth System Grid Federation in partnership with the US Programme for Climate Model Diagnosis and Intercomparison and a range of global partners in support of the sixth Coupled Model Intercomparison Project (CMIP6).
3. A leading partner in many major European projects including IS-ENES 2 (developing an InfraStructure for a European Network for Earth system Simulation), CHARMe (Characterization of metadata to enable high-quality climate applications and services) and CLIPC (Climate Information Portal for Copernicus).
4. Working with the wider UK atmospheric science and earth observation communities, via a range of projects, with NCAS and other NERC funding.
5. Working with the European Space Agency on projects such as the ESA Climate Change Initiative (CCI) portal.
6. Working with commercial and academic partners and the Satellite Applications Catapult, on the facility for Climate and Environmental Monitoring from Space (CEMS), to support both academic research and opportunities for commercial applications and downstream services from EO and Climate data.

4. Funding

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

4.1 Annual total funding

	2008-2009	2009-10	2010-11	2011-12	2012-13	2013-14	2014-15
NCAS income	970	866	906	883	935	829	829
NCEO income	378	389	450	419	445	392	390
Other NERC income	788	481	341	527	287	272	600
Other income	461	710	1144	1099	1283	1486	1394
Total income	2597	2446	2841	2928	2950	2979	3213

Table 4.1: Overall funding for CEDA for financial years 2008 — 2009 to 2014 — 2015 (in £k)

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

4.2 Externally funded projects for the year 2014-2015

Name	Description	Topic Funder	Dates	Value
CHARMe	Characterization of metadata to enable high quality climate applications and services	FP7	Jan-2013 to Dec-2014	£272k
Agricultural Nat GHC Inventory	Agricultural Greenhouse Gas Inventory Research Platform	DEFRA	Oct-2010 to Mar-2015	£120k
CLIPC	Climate Information Portal for Copernicus	FP7	Dec-2013 to Dec-2016	£676k
DECC Climate Archive Support	Providing support for the use of climate data, including the IPCC data centre	DECC	Aug-2012 to Mar-2015	£699k
ESA LTDP support	ESA Long-Term Data Preservation	ESTEC	Nov-2012 to Jul-2014	£15k
ESA Optirad	OPTIRAD (OPTImisation environment for joint retrieval of multi-sensor RADiances) aims to advance the state of the art in EO data assimilation in land surface processes.	ESA	Apr-2014 -to Jan-2016	£22k
EUFAR-2	The European Facility for Airborne Research (EUFAR) aims at coordinating the operations of the European fleet of instrumented aircraft in the field of environmental research in the atmospheric, marine, terrestrial and Earth sciences.	FP7	Feb 2014 to Jan 2018	£174k
IS ENES 2	InfraStructure for the European Network for Earth System Modelling - Phase 2 (IS-ENES II)	FP7	Apr-2013 to Feb-2017	£551k
MONSooN Overflow 2	Providing support for MONSooN users on JASMIN	Met Office	Apr-2014 to Mar-2017	£199k
ExArch	Developing a software management infrastructure which will scale to the multi-exabyte archives of climate data which are likely to be crucial to major policy decisions in by the end of the decade.	NERC	Apr-2011 to Dec-2014	£289k
OpenAIREplus	2nd Generation of Open Access Infrastructure for Research in Europe - linking peer-reviewed literature to associated data `https://www.openaire.eu/`	FP7	Dec 2011 to Dec 2014	£162k
Propagation Elements	for broad band satellite systems	ESTEC	Oct-2013 to Sep-2015	£89k
SeaDataNet2	CEDA is contributing expertise with INSPIRE and ISO/OGC standards to enable the SeaDataNet infrastructure to become compliant and interoperable with wider initiatives.	EC	Apr-2012 to Sep-2015	£41k
SPECS Data Archive	Data archive for the SPECS FW7 seasonal prediction project	ICDCDC [1]	Aug-2013 to Oct-2016	£121k
H2020 Fiduceo	Fidelity and uncertainty in climate data records from Earth Observations `http://www.fiduceo.eu/`	EC	Feb-2015 to Mar-2015	£102.5k
H2020 BACI	Towards A Biosphere Atmosphere Change Index `http://baci-h2020.eu/index.php/`	EC	Jan-2015 to Oct-2018	£38.5k
H2020 Eustace	EU Surface Temperature for All Corners of Earth	EC	Jan-/2015 to Jun-2018	£142.9k
Acclimatise	Climate Consultancy to generate a high-quality set of climate statistics from the CMIP5 and CORDEX projects	Acclimatise/ Shell	Jan-2015 to Mar-2015	£46.3k

Table 4.2: Externally funded projects for 2014-2015 (non-core NERC)

Part II

Highlights and Major Activities

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

5. Highlights

5.1 Twenty years of data management in the British Atmospheric Data Centre

Sam Pepler, Sarah Callaghan

Sam Pepler is the CEDA Curation Manager. He joined the BADC in 1996.

The BADC has now hit early adulthood

The British Atmospheric Data Centre (BADC) was formed in 1994 to assist researchers in locating, accessing and interpreting atmospheric data. It also provided a location for the long-term storage of data produced by NERC projects. The facility has grown and now manages around 2PB of data including climate models, satellite measurements, weather radar and atmospheric chemistry results.

Figure 5.1: Past and present BADC staff celebrating 20 years of the facility

What are the plans for the future?

Our user base continues to grow not just in number but also in diversity. Originally the user base was 100% from the UK academia, currently 40% of users are international and 30% are from non-university organisations. We expect this trend to continue. The storage infrastructure will continue to evolve. Our current systems use our own cloud based services and a scalable parallel file system, but we are already looking at the next generation of technology, object stores, and seeing how they can be adapted to hold data for the next 20 years.

Growing number of users and impact

It is difficult to share larger volumes of data, partly because of the changing nature of the storage technology, but also because of the effort needed to make data discoverable and usable. The data centre allows us to do this at scale, not only handling the large volumes, but thornier problems like licensing.

In 1995 we held 60 GB and delivered data to 100 users, by 2014 the archive was 2 PB and the number of downloading users is around 4000 annually.

Data management is increasingly a subject for everyone. We have had a strong advocacy role supporting open data policies, data standards and the use of data citation.

The BADC is primarily supported by the Natural Environment Research Council

5.2 Improving climate risk decisions related to offshore assets

Ag Stephens

Ag Stephens (Head of Partnerships at CEDA) is an expert in data management and software development.

Climate statistics for risk screening

STFC CEDA was funded by a climate consultancy to generate a high-quality set of climate statistics from the CMIP5 (5th Coupled Model Intercomparison Project) and CORDEX (CO-ordinated Regional Downscaling EXperiment) projects. Baseline and future change climate statistics, such as maxima, minima and percentiles, were generated from raw multi-model output across a range of scenarios.

Supporting business

The primary customer is a large petro-chemical company that is undertaking risk analysis on the impacts of future climate change on its major assets.

Terabytes of simulations were processed on the JASMIN platform to generate a set of spatially consistent and quality-controlled statistics. CEDA provides the data via its Web Processing Service (WPS) interface allowing other tools to dynamically interrogate the data set. The major tool using this service for this project is an an externally developed web-portal serving the client's scientists and engineers. They can select spatial sites and receive an ensemble of statistics from multiple models through the bespoke web-portal or directly from the CEDA WPS in the form of summary spreadsheets.

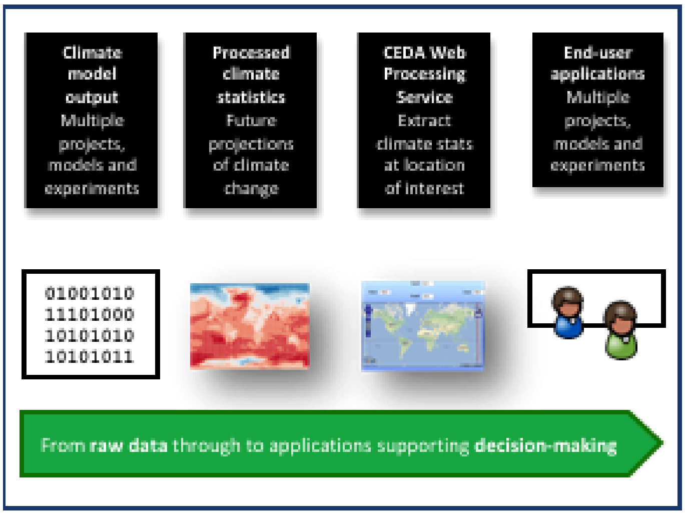

Climate model output
Multiple projects, models and experiments
Processed climate statistics
Future projections of climate change
CEDA Web Processing Service
Extract climate stats at location of interest
End-user applications
Multiple projects, models and experiments
01001010
11101000
10101010
10101011
From raw data through to applications supporting decision-making

Figure 5.2: The arrow indicates the flow of data processing from raw data, through processed statistics, web-delivery tools and end-user applications.

Helping research through open access

As an extension to this work CEDA intends to publish these climate statistics to its catalogue along with a simple web-interface that allows selection of the data for a given location. The underlying data, and the web-tool will be available to academic and commercial users. Additionally, further variables are currently being considered as an extension to the data set. These will provide significantly more information about sea level rise in future scenarios.

This research was funded by Acclimatise Group Limited

5.3 Major upgrade to JASMIN data analysis infrastructure

Matt Pritchard

Matt Pritchard is the JASMIN manager, responsible for the finances, services, and resource allocation.

Launch event puts JASMIN in the spotlight

The 2nd June 2014 saw the JASMIN launch event, attended by key stakeholders to celebrate a major upgrade to NERC's unique data analysis infrastructure. With vast storage capacity, increased computing power and an immensely capable core network, JASMIN is now able to provide a high-performance environment for the entire NERC community to manipulate and store environmental data.

Part supercomputer, part data store, part private cloud

JASMIN is a globally unique compute environment: a cross between a supercomputer and a private cloud coupled to a large array of high-performance disk. JASMIN provides four major capabilities: the storage, a batch cluster (LOTUS), a virtualised infrastructure zone, and a private cloud. Unlike supercomputers which are used to run simulations and generate data, JASMIN is optimised for the bringing together and analysis of existing data and the sharing of results — allowing users to "bring their computations to the data". Climate research scientist, Matthew Mizielinski said, *"The data gymnastics I've been able to do using JASMIN would not have been practical, or indeed possible, anywhere else. The JASMIN platform has become the pivotal tool for the work of the High Resolution Climate Modelling (HRCM) team, and we look forward to testing the capabilities of the upgrade"*.

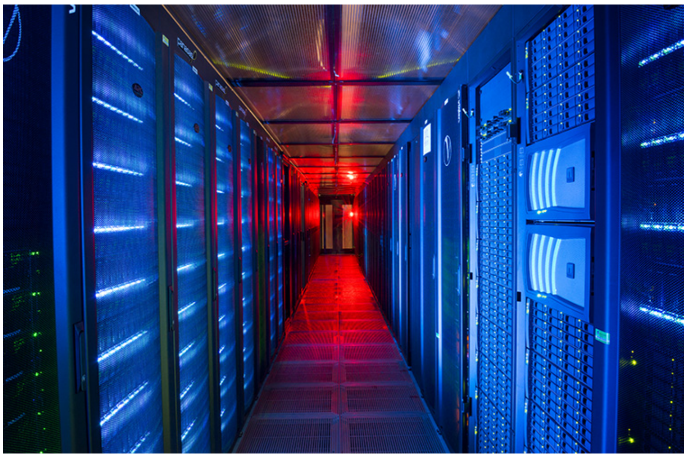


Figure 5.3: The recent upgrade to JASMIN provides a massive 16 Petabytes of storage, 4000 cores of computing power and a super-high-speed network at its core to enable efficient manipulation and analysis of data. The JASMIN infrastructure is operated for CEDA by STFC's Scientific Computing Department.

Next Steps

JASMIN already provides high-performance data analysis space for over 100 scientific projects, using over 10 Petabytes of storage capacity, with the LOTUS processing cluster in constant use. Projects can also be provided with resources within a private cloud environment to meet their needs, either within the managed high-performance environment alongside the CEDA archive, or within an area of NERC network space where they have the freedom to construct their own workflows and services.

JASMIN is expected to grow as the place to be for big environmental science!

JASMIN Phases 2 and 3 were funded by capital investment from the Natural Environment Research Council.

5.4 User service updates in a live archive

Graham Parton

Graham Parton is a senior data scientist, supporting the archival, discovery, access to and understanding of environmental data

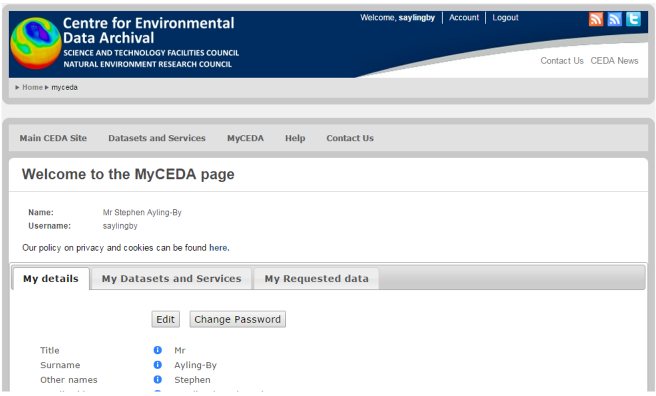
Updating the CEDA systems

CEDA manages many petabytes of constantly on-line environmental data, accessed every year by a global community of around 6000 registered users. While the primary task is data access, delivery requires being able to support data access control for both public and restricted data and reporting on user community demographics (funders are keen to know many things about the users: who, where, and why etc). Allowing thousands of users to view and control their personal information and data access applications are key aspects of data management within CEDA. Any change to systems underpinning these crucial operational systems has to be done with minimal disruption. Rolling out an updated system for user registration, account management and access application system in a live environment is no mean feat, but CEDA achieved this in 2014 with minimal impact to users!

The new MyCEDA service

CEDA has been operating world-leading environmental data centres for over 20 years. Whilst technology moves on, user demand never ceases. Making sure scientists can keep accessing key data via the CEDA infrastructure is mission critical for CEDA, but such services needs to scale with growing user community demands and changing technologies.

Following months of development work CEDA managed a full scale roll out of its new user management system, "myCEDA", in 2014. The myCEDA service was fully integrated into the CEDA's services and websites during a seamless transition for users.



Welcome to the MyCEDA page

Name: Mr Stephen Ayling-By
Username: saylingby

Our policy on privacy and cookies can be found here.

Title Mr
Surname Ayling-By
Other names Stephen

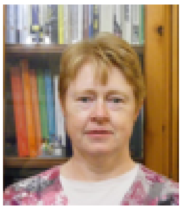
Figure 5.4: Screenshot of the new MyCEDA web page.

Plans for the future

CEDA will continue to develop other services, integrating these to provide smoother, more sustainable and scalable data centre services. Our biggest planned service update is integrating our new data centre catalogue, ensuring the UK academic environmental community can meet its EU legislative requirements for discovering geo-spatial data. This new catalogue will also be seamlessly integrated into the family of CEDA sites and services as part of an on-going drive to improve our user services.

CEDA systems are supported by contributions from many projects

5.5 CEDA Training Courses



Alison Pamment is a data scientist specialising in the climate forecast metadata conventions and the development of training courses

Alison Pamment[1], Sam Pepler[1], Ag Stephens[1], Sarah Callaghan[1], Esther Conway[1], Eduardo Damasio Da Costa[1], Steve Donegan[1], Wendy Garland[1], Anabelle Guillory[1], James Groves[2], Andy Heaps[3], Alan Iwi[1], Graham Parton[1], Charlotte Pascoe[1], Stephen Pascoe[1], Matt Pritchard[1], Charles Roberts[3], Louise Whitehouse[2]

[1] *CEDA,* [2] *NCAS Operations Group,* [3] *NCAS CMS*

Computing skills for environmental scientists

CEDA offers training in the skills needed for effective data management and analysis. Whether a scientist is handling “big data” such as that produced by climate models or large numbers of small data files resulting from instrumental observations, appropriate IT skills are essential. CEDA staff have developed two training courses: “Introduction to scientific computing” and “Introduction to JASMIN”.

Sharing our own experience of handling data

Our training is presented by a team made up of CEDA data scientists and computational experts drawn from the wider NCAS. In addition to presenting the formal elements of the courses, such as an introduction to programming in Python, the team are well placed to pass on their own “real world” experience of data handling, management, and analysis. The “Introduction to scientific computing” course is designed for the needs of PhD students and early career researchers in atmospheric and environmental science while the JASMIN workshop is intended for any scientist who is new to using the system and its services.

Figure 5.5: CEDA provides training in essential IT skills for data management and analysis. One of our courses is an introduction for new users of the JASMIN system.

Future Training Courses

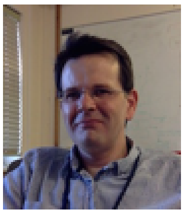
“Introduction to JASMIN” workshops are run periodically, often as a half day “roadshow” event. Recent examples are workshops run at the Met Office and the universities of Leeds and Reading. CEDA plan to run “Introduction to Scientific Computing” courses as an annual event. The next course will take place at the University of Leeds, April 11-15 2016. Up to twenty-five students can be accommodated on each five day course and registration is via the NCAS website.

More details are available on the the CEDA, JASMIN, and NCAS websites at `http://www.ceda.ac.uk/training/`, `http://www.jasmin.ac.uk/how-to-use-jasmin/jasmin-training/` and `https://www.ncas.ac.uk/index.php/en/nerc-ppsda-schools`

The 2013-2014 training courses were funded by NCAS and NERC.

5.6 CHARMe - Linking annotations to Data

Philip Kershaw



Philip Kershaw is the CEDA technical manager responsible for developing CEDA information and computing systems.

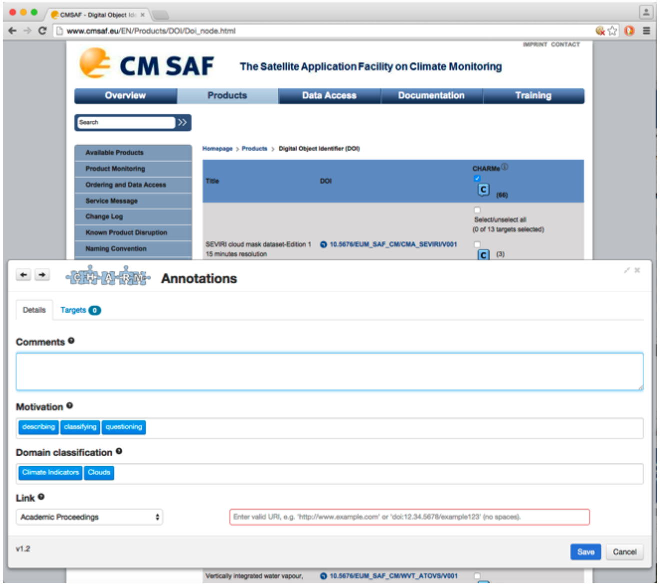
Linking climate data to extra information to better enable use

Climate data has an increasingly important role as policy makers from government and industry sectors seek to address and mitigate the effects of climate change. This data is both diverse and complex making it difficult for potential users outside the research community to effectively exploit it. Each user community requires different classes of information, for example the method by which the dataset was produced, the robustness of the algorithms used, and the terms of use of the data.

Much previous work has been done on producing this extra information for climate data, but it may originate from many sources besides the original provider. There is as yet no consistent mechanism to collate this information and link it to the datasets themselves. The CHARMe (**CHAR**acterization of **Me**tadata to enable high-quality climate applications and services) project was conceived to address this need: it provides a means to crowd-source information — *Commentary metadata* — from the user community by providing a system to annotate datasets and link it with other relevant sources.

How does CHARMe work?

The CHARMe system provides a plugin application which can easily be integrated into third party websites where data is hosted. The screenshot on the right illustrates this with project partner DWD's (German Weather Centre) site. Users can click on any given data and provide annotation information. This could, for example, allow users to link data to a relevant scientific paper. Once recorded the information is stored in a single central node hosted for the project.



Title	DOI	CHARMe
		☑ C (66)
		□ Select/unselect all (0 of 13 targets selected)
SEVIRI cloud mask dataset-Edition 1 15 minutes resolution	10.5676/EUM_SAF_CM/CMA_SEVIRI/V001	□ C (3)
Vertically integrated water vapour,	10.5676/EUM_SAF_CM/WVT_ATOVS/V001	□

Comments

Motivation: describing classifying questioning

Domain classification: Climate Indicators Clouds

Link: Academic Proceedings

Enter valid URI, e.g. 'http://www.example.com' or 'doi:12.34.5678/example123' (no spaces).

Figure 5.6: Annotating data with CHARMe on the website of the DWD Satellite Application Facility on Climate Monitoring .

This enables other users and applications to discover it and re-use. CHARMe exploits the Linked Data technologies advocated by web pioneer Tim Berners-Lee, as successfully applied in many venues, including with the NERC Vocabulary Service. Using this standards-based approach ensures that Commentary information collected can be readily used by other applications on the web and so more widely disseminated.

CHARMe will be integrated into normal CEDA operations in the next year or so, providing direct benefit in CEDA operations from informatics research.

This research was funded by the European Union Framework 7 programme

6. Short Reports

6.1 CEDA Metadata Services

Steve Donegan

CEDA provides access to many petabytes of Atmospheric, Model and Earth Observation data to the UK academic community and beyond. It is vitally important that CEDA maintains a metadata catalogue of its data holdings that record important aspects of the archived data, such as location, measurement period, science keywords, data originator as well as important peripheral information as ownership and access constraints. This metadata catalogue is used to link searches made by CEDA users to the data in the archive.

During 2014/2015 CEDA migrated to an updated version of its internal metadata catalogue based on the MOLES model (Metadata Object Links in Environmental Science). A unique aspect of the MOLES metadata model and catalogue is that it identifies different aspects of the data held by CEDA and links them together. These aspects include information on the instrument, the platform on which it is deployed, and the project. This observation data is associated with and ultimately links these objects to a parent dataset, or observation collection. So for example, it is easy to identify all data obtained using a single airborne instrument across different archive locations, or all datasets associated with a single platform such as a satellite, airplane or climate model. There is a searchable web interface on the CEDA homepage (`www.ceda.ac.uk`) to the MOLES catalogue with results pages clearly displaying links to the download and the archive access applications.

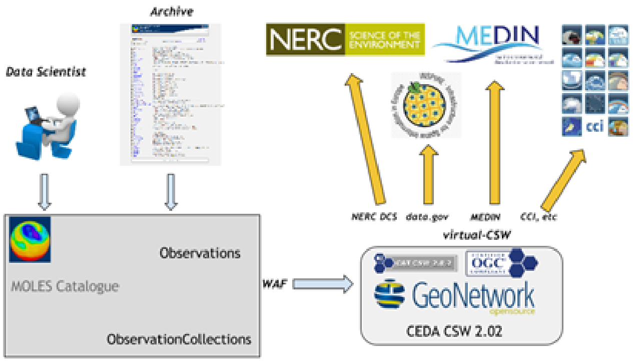


Figure 6.1: How CEDA publishes its metadata. Metadata is generated from MOLES content where it is harvested and placed in an Open Geospatial Consortium (OGC) Catalogue Service for the Web (CSW) from where clients can access the metadata in UK Gemini 2.2 format.

CEDA publishes metadata based on the MOLES catalogue for consumption by a number of web portals and systems that provide data discovery services to a much wider global community. In this way CEDA meets its obligations to NCAS, NCEO and NERC as well as legal obligations such as the EU INSPIRE directive. CEDA is also increasingly involved with international projects that require the ability to find and access data using standardised metadata.

In order to conform to EU INSPIRE legislation, the UK national portal (`www.data.gov.uk`) specified a metadata profile based on the ISO19115 schema for geospatial data and services. This profile, the UK Gemini 2.2 metadata specification, is used as the metadata format for publishing to various services such as the NERC Data Catalogue Service (DCS) (`data-search.nerc.ac.uk`), the MEDIN Discovery Portal (`portal.oceannet.org`) as well as data.gov.uk and various other projects such as the CCI (ESA Climate Change Initiative) portal.

The CEDA MOLES catalogue content publishes metadata via a Web Accessible Folder (WAF) located on a dedicated catalogue server. The UK Gemini format records are transformed dynamically on request to the MOLES WAF, and so will always reflect the current catalogue content at the time of request. CEDA uses an open source Geonetworks OGC Catalogue Server for the Web (CSW) 2.0.2 (`geonetwork-opensource.org`) to periodically harvest the UK Gemini records from the MOLES WAF and make them available to the wider community via published CSW endpoints. The CSW is a specification that makes metadata content searchable to client servers and returns information on available datasets and services at a range of content levels.

6.2 More data from sea to sky

Graham Parton

The Centre for Environmental Data Analysis (CEDA) has been archiving environmental data for over two decades, with a steadily growing number of complete, ongoing and new dataset collections. 2014-15 was no exception with 175 new datasets within 23 collections added, whilst many more continued to be updated within the archives. This brought the total number of datasets to nearly 2400 in 309 collections — a vast and unique collection!

CEDA exists to facilitate the long-term preservation of the UK's atmospheric and earth observation data resources. The archives cover one of the most diverse collections encompassing data from ocean and atmospheric measurements and models including airborne and satellite data. Actively curating these data is an important responsibility to ensure long-term use of these data. During 2014-15 the key collections added to the CEDA archives included:

- Airborne data from EUFAR and OFCAP projects
- Chemistry model output from the MAAM, SPARC and ACCACIA programmes
- Climate datasets from the Climate Reseach Unit and the SPECS modelling project
- Satellite data from AATSR, CALIOP, MODIS and MISR instruments
- Rescue of aerial imagery and land surface data from the landmap.ac.uk service following its closure in 2014.

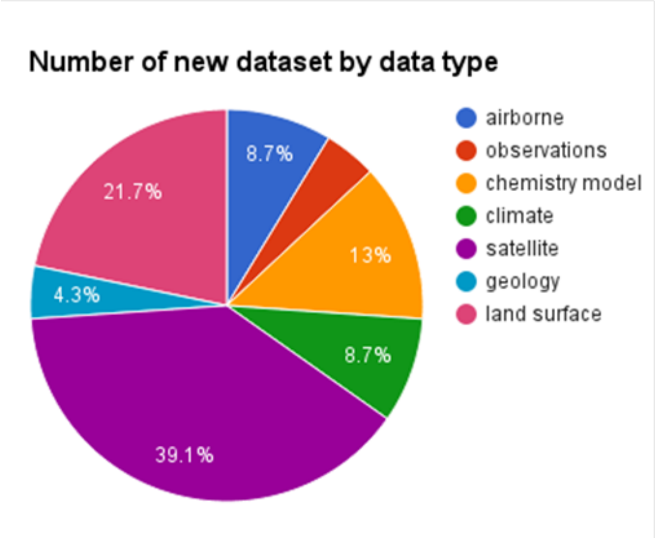


Figure 6.2: Percentage split of 175 datasets published by CEDA during 2014-15 by data types.

Curating these data, though, isn't just a case of adding them to the archive. CEDA staff ensure that these data remain of long-term value by ensuring the data are acquired in well-supported formats and have accompanying documentation to help describe them. An important aspect of this is also to catalogue these data within the CEDA data catalogue, which also ensures that these data become discoverable, especially as these records also appear in places such as data.gov.uk and the NERC Data Catalogue Service.

Also in 2014 the UK Solar System Data Centre undertook a project to digitise its holdings of solar heliosphere images on glass plates and film. These are now being catalogued and will eventually be released to the community.

At the same time, CEDA staff continue to enhance the existing archive holdings. Around 415 datasets continue to be updated on various timescales from every 5 minutes to a few times a year. Meanwhile CEDA staff have been updating supporting material for all the datasets in the CEDA data catalogue, enhancing its detail, accuracy and record connectivity.

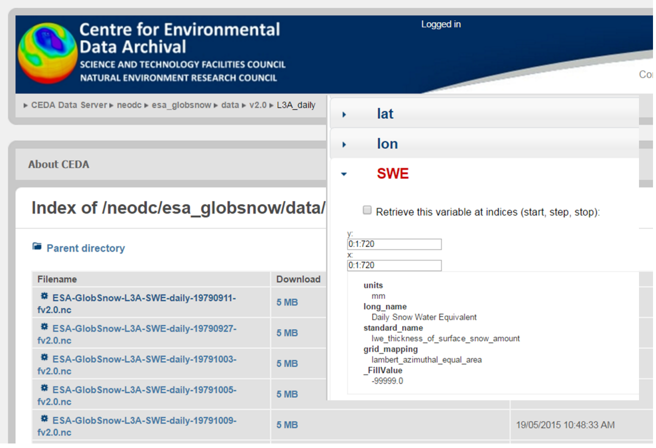
Overall, the ongoing data curation activities for both new and existing holdings within CEDA archives are ensuring that CEDA continues to delivery a well structured, highly relevant and respected service to all users, supporting research from sea to sky.

6.3 Launching CEDA OPeNDAP: data subsets from remote networks

Ag Stephens, Andrew Harwood and Phil Kershaw

CEDA manages a multi-petabyte archive of data in a variety of different formats, sizes and structures. Through the JASMIN big-data platform we allow many scientists to log in and run processing code directly on the data files. However, many end-users would prefer to access data remotely at their own institution ? which could be anywhere in the world. The sheer volume prohibits the downloading of entire data sets so other solutions must be sought to reduce the transfer burden. CEDA has recently launched its OPeNDAP service to allow end-users to subset data files through a web-browser or directly from their processing code.

OPeNDAP is a framework and protocol that enables data to be accessed via remote clients as if they were being read locally. It is built on top of the netCDF data format; providing an interactive connection to the file across the network that allows the end-user to both interrogate metadata and to extract subsets from the data arrays. CEDA has employed the PyDAP technology to deliver an OPeNDAPservice. PyDAP connects a browsable web-view of the data directories and files to a service that is available to any OPeNDAP-aware client in the world. Simple Python code is shown that will connect to the ESA GlobSnow data set, extract relevant metadata and subset the data array. This is a powerful approach that we expect to become much more prevalent in future scientific workflows. The ability to read data from remote archives during the initialisation of a simulation or in a multi-model comparison calculation greatly reduces the onus for the scientist to manage his/her own data transfers and local archival. The CEDA PyDAP service is available at: `http://dap.ceda.ac.uk/`


Centre for Environmental Data Archival
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL
Logged in
Co
▸ CEDA Data Server ▸ neodc ▸ esa_globsnow ▸ data ▸ v2.0 ▸ L3A_daily
About CEDA
▸ lat
▸ lon
▾ SWE
Parent directory

Index of /neodc/esa_globsnow/data/

Filename	Download	
ESA-GlobSnow-L3A-SWE-daily-19790911-fv2.0.nc	5 MB	
ESA-GlobSnow-L3A-SWE-daily-19790927-fv2.0.nc	5 MB	
ESA-GlobSnow-L3A-SWE-daily-19791003-fv2.0.nc	5 MB	
ESA-GlobSnow-L3A-SWE-daily-19791005-fv2.0.nc	5 MB	
ESA-GlobSnow-L3A-SWE-daily-19791009-fv2.0.nc	5 MB	19/05/2015 10:48:33 AM

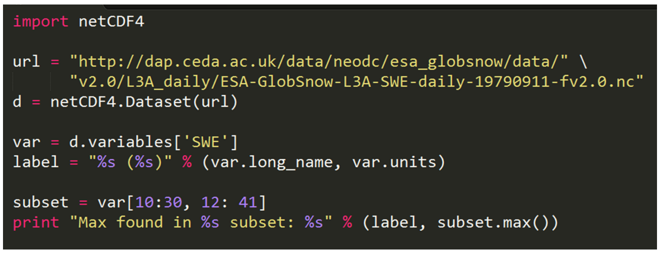
□ Retrieve this variable at indices (start, step, stop):

y: 0:1:720
x: 0:1:720

units
mm
long_name
Daily Snow Water Equivalent
standard_name
lwe_thickness_of_surface_snow_amount
grid_mapping
lambert_azimuthal_equal_area
_FillValue
-99999.0

Figure 6.3: PyDAP web-interface showing download or subset options for GlobSnow netCDF files. The inset shows how a netCDF file can be interrogated and a subset extracted via the browser.

CEDA provides other data access services such as the data “browse” service and the Web Processing Service (WPS), which provides a common interface and structure for interacting with processing code. A single web-application provides a secured user interface to multiple “processes” delivering a range of different functions. We recently deployed the CF-netCDF Checker tool within the WPS. This allows scientists to upload data file to check for compliance with Climate and Forecasts (CF) Metadata Conventions for NetCDF. Other processes include a general “subsetter”, a number of CMIP5-specific manipulation tools and wrappers for the Climate Data Operators (CDO) toolkit. At CEDA we continue to develop these tools and we are keen to engage with end-users to help us ensure that they are optimised for usability by the scientific community.



```
import netCDF4

url = "http://dap.ceda.ac.uk/data/neodc/esa_globsnow/data/" \
      "v2.0/L3A_daily/ESA-GlobSnow-L3A-SWE-daily-19790911-fv2.0.nc"
d = netCDF4.Dataset(url)

var = d.variables['SWE']
label = "%s (%s)" % (var.long_name, var.units)

subset = var[10:30, 12: 41]
print "Max found in %s subset: %s" % (label, subset.max())
```

Figure 6.4: Python code to interact to read a remote netCDF file from the CEDA PyDAP service.

6.4 Mirroring Sentinel data for UK researchers

Victoria Bennett, Steve Donegan, Sam Pepler

The European Sentinel series of satellites, developed by the European Space Agency (ESA), consists of a suite of missions, carrying different instruments, for land, ocean and atmospheric monitoring. Each Sentinel mission is based on a constellation of two satellites, carrying a range of technologies, such as radar and multi-spectral imaging instruments. CEDA has a key role in making data from the Sentinel satellites available to UK researchers.

The first of the Sentinel satellites, Sentinel-1A was launched in April 2014, followed by Sentinel-2A in July 2015. The data product archive volumes are unprecedentedly large: around 2 terabytes per day from Sentinel-1A, with similar expected soon from Sentinel-2A. The data are made available for all users to download from the ESA’s scientific data hub, but for global or long time-series scientific processing and analysis, the data transfer speed and user’s own local storage and processing resources can be insufficient to meet the needs.

In the UK, CEDA provides infrastructure to support the analysis of such data. CEDA will provide both a mirror archive of Sentinel data, and an environment to exploit that data alongside other datasets. Sentinel-1A Level 1 products (Single Look Complex, SLC, and Ground Range Detected, GRD) from March 2015 onwards (tens of TB) are already available in the archive and the data volumes will continue to rise considerably as more data are acquired.

Figure 6.5: Artist’s impression of Sentinel-1 satellite. Copyright ESA/ATG medialab

The data are stored on, and made accessible to the UK science community via, the JASMIN super data cluster and the academic CEMS(Climate Environment and Monitoring from Space) facility — itself hosted on JASMIN. JASMIN incorporates over 16 PB of disk, co-located with tape and computing facilities for data analysis via batch, hosted and cloud computing. Recent data are stored on-line for direct access to users;

older data will be moved to a near-line tape archive, reinstating it for users on demand. It is expected that most UK science users will access, process and analyse the data in the JASMIN-CEMS hosted environment avoiding the need to download and store data on their local machines. Sentinel 2 and Sentinel 3 data will follow soon.

CEDA are also part of the UK Collaborative Ground Segment for Copernicus, collaborating with the Satellite Applications Catapult and Airbus-DS in Farnborough to ensure Sentinel data access for UK users in academia, government and the private sector.

The UK academic community has already achieved impressive science results using Sentinel 1 data. Applications of the data so far include earthquake mapping, monitoring of landcover changes in cloud-covered regions, global forest biomass mapping to constrain and validate climate models, observations of ice loss in polar regions and the detection of deforestation in the tropics.

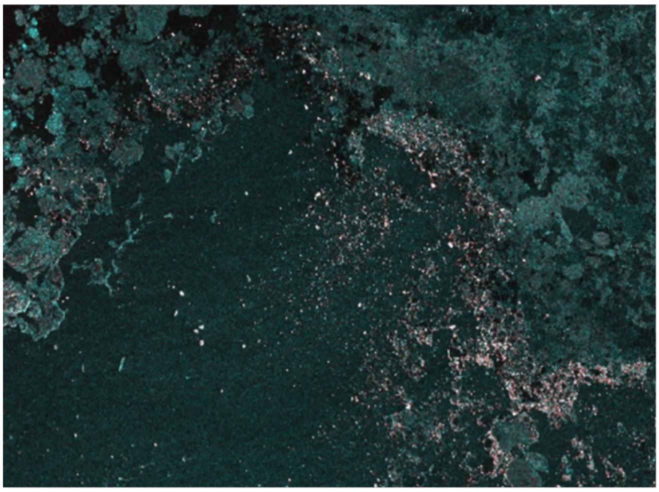


Figure 6.6: Sentinel-1A image of Jacobshavn Glacier, Greenland, September 2014. Copyright ESA

6.5 The Community Intercomparison Suite (CIS)

Victoria Bennett, Phil Kershaw

The Community Intercomparison Suite (CIS) is tool initially developed for JASMIN, kicked off in 2013, to develop an automated suite of tools for scientific analysis of heterogeneous datasets. One of its strengths is that it provides a bridge supporting access to both model and observation based data.

The CIS is a collaboration between CEDA and the Department of Atmospheric Oceanic and Planetary Physics (AOPP) at the University of Oxford. CIS aims to simplify a wide range of tedious tasks in dataset intercomparison (ingest of a range of gridded and ungridded model data and observations, reduction, co-location, and analysis) to a set of simple commands.

Visual representation and comparison of geoscientific datasets presents a huge challenge due to the large variety of file formats and spatio-temporal sampling of data (whether observations or simulations). CIS attempts to greatly simplify these tasks for users by offering an intelligent but simple command line tool for visualisation and colocation of diverse datasets. In addition, CIS can subset and aggregate large datasets into smaller more manageable datasets. The CIS philosophy is to remove as much as possible the need for specialist knowledge by the user of the structure of a dataset.

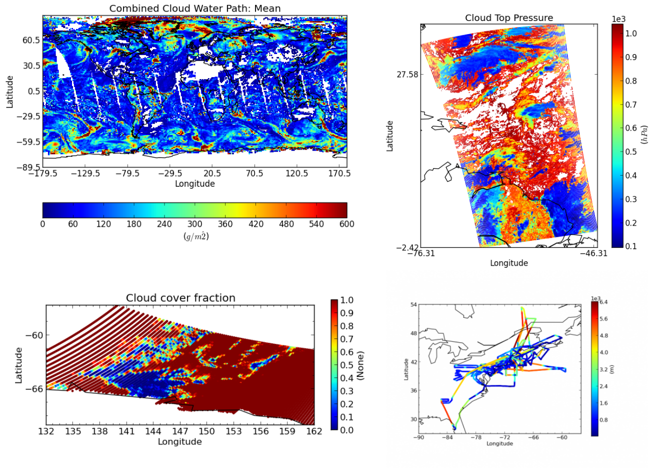


Figure 6.7: Example plots from Community Intercomparison Suite

The key elements of CIS are support for:

- Reading CF-netCDF format data: both model data and observations
- Data reduction: sub-setting and averaging
- Co-location: to a common spatio-temporal domain
- Statistics: from co-located data
- Plotting: of reduced/co-located data
- Writing CF-netCDF format data
- Workflow documentation

CIS can deal with both gridded and ungridded datasets of 2, 3 or 4 spatio-temporal dimensions. It can handle different spatial coordinates (e.g. longitude or distance, altitude or pressure level). CIS supports HDF, netCDF and ASCII file formats. The suite is written in Python with all dependencies publicly available open source packages (e.g. the IRIS package from the MetOffice). A web-based developer hub includes a manual and simple examples.

The CIS is implemented as a consistent command line interface to underlying tools based on science use cases. CIS can be run on any system that supports Python and has recently been integrated into the Anaconda package management system making it easy to deploy. It is deployed on JASMIN where the presence of large scientific datasets is already facilitating early users' atmospheric and environmental science research. The CIS website includes instructions on how to download and use the software, and presents numerous examples of how to read and plot individual and combinations of datasets. Details, instructions and examples are available at `www.cistools.net`.

6.6 CEDA's Data Catalogue meets International Standards

Graham Parton, Ag Stephens

Over the past few decades, huge volumes of environmental data have been collected by field campaigns, ongoing observation networks, satellites, ships and aircraft as well ever more complex torrents of output from weather forecast and climate change models.

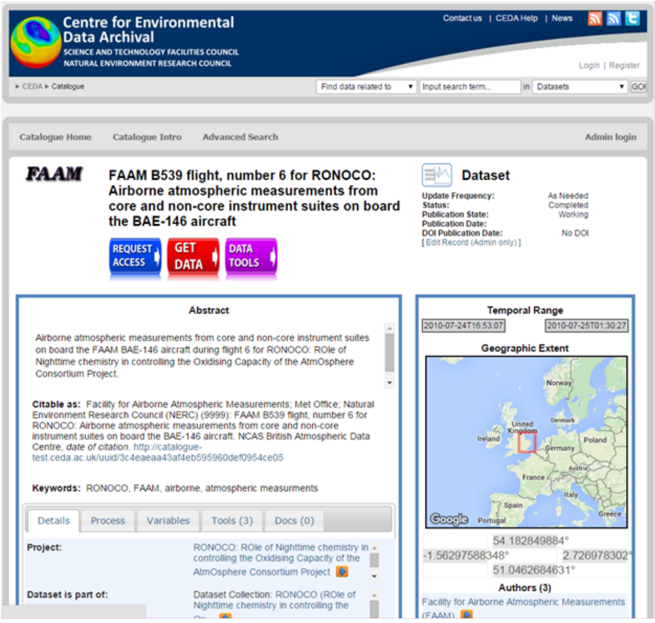
CEDA has been at the forefront of curating such data in its archives and, crucially, helping users discover, access and understand these data through its data catalogue. The catalogue also links records that describe the background "who", "where", "when", "why" and "how" of these data. These context records can be common to many datasets, meaning a highly connected catalogue allows users to browse the catalogue, discovering other really useful related data.

To build this catalogue CEDA staff followed a model, called MOLES — "Metadata Objects Linking Environmental Sciences". MOLES was first developed by CEDA around 10 years ago, with the aim of allowing the sharing of catalogues between all the NERC data centres. Since then, the research and archive communities have moved on, and the wealth of public funded geospatial data has been recognised, with the EU passing the INSPIRE legislation requiring such data to be catalogued using international standards.

Unfortunately, MOLES didn't follow these international standards and more work was needed. Eventually a new version of the MOLES model was produced which not only followed these standards but kept the background and context records too. Implementing the new CEDA catalogue following the updated MOLES model wasn't a straight forward task. Though the early catalogue model was very similar to the new model, there were significant differences. Plus all the old information in nearly 6000, hand crafted records needed to be at least maintained or even improved upon. Eventually in September 2014 CEDA managed the switch and went live with its new catalogue, successfully migrating all the old records and carrying out a huge tidy-up operation at the same time. The catalogue content has never looked so good!

Following the switch we've been improving the user experience of the catalogue thanks to their feedback and we're busy updating and creating new content too. Under the hood, we're developing connections with our other systems to make sure the content remains as fresh as possible and helping our staff manage the thousands of records it now holds - all this while the catalogue and archives continue their relentless growth. Finally, we're exporting more and more of our records to other catalogues and portals such as the NERC Data Catalogue Service and data.gov.uk, making these valuable assets even more discoverable by the wider communities.

Thanks to the new international standards driven catalogue at CEDA, users can discover and browse to find data, not only via CEDA sites, but also in other websites too. The task isn't over, though, as the new catalogue interface continues to improve, old content is revised and new content is added. We are now set to grow our new catalogue in response to all our users' and archive needs.



FAAM B539 flight, number 6 for RONOCO: Airborne atmospheric measurements from core and non-core instrument suites on board the BAE-146 aircraft

Dataset

Update Frequency:	As Needed
Status:	Completed
Publication State:	Working
Publication Date:	
DOI Publication Date:	No DOI

[Edit Record (Admin only)]

Abstract

Airborne atmospheric measurements from core and non-core instrument suites on board the FAAM BAE-146 aircraft during flight 6 for RONOCO: ROle of Nighttime chemistry in controlling the Oxidising Capacity of the AtmOsphere Consortium Project.

Citable as: Facility for Airborne Atmospheric Measurements; Met Office; Natural Environment Research Council (NERC) (9999): FAAM B539 flight, number 6 for RONOCO: Airborne atmospheric measurements from core and non-core instrument suites on board the BAE-146 aircraft. NCAS British Atmospheric Data Centre, *date of citation*. http://catalogue-test.ceda.ac.uk/uuid/3c4eaeaa43af4eb595960def0954ce05

Keywords: RONOCO, FAAM, airborne, atmospheric measurments

Project: RONOCO: ROle of Nighttime chemistry in controlling the Oxidising Capacity of the AtmOsphere Consortium Project

Dataset is part of: Dataset Collection: RONOCO (ROle of Nighttime chemistry in controlling the

Temporal Range

2010-07-24T16:53:07 2010-07-25T01:30:27

Geographic Extent

54.182849884°
-1.56297588348° 2.726978302°
51.0462684631°

Authors (3)

Facility for Airborne Atmospheric Measurements (FAAM)

Figure 6.8: A typical dataset record in CEDA catalogue, showing when, where, how, why, by whom and what the dataset is and how to access it.

Part III

Metrics and Publications

7. Additional Data Centre Metrics

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

Note that a considerable amount of use of CEDA data use is by users on JASMIN, who would not be measured in most of these statistics because the data is directly available on the file system.

7.1 Access related metrics

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

Area	Q1	Q2	Q3	Q4
UK	2534 (61%)	2589 (61%)	2633 (62%)	2686 (64%)
Europe	494 (12%)	499(12%)	485 (11%)	457 (11%)
Rest of the world	1024 (25%)	1055 (25%)	1041 (25%)	1015 (24%)
Unknown	79 (2%)	80 (2%)	79 (2%)	77 (2%)

Table 7.1: Users by area

Institute Type	Q1	Q2	Q3	Q4
University	2934 (71%)	2998 (71%)	2994 (71%)	3022 (71%)
Government	694 (17%)	696 (17%)	700 (17%)	678 (16%)
NERC	160(4%)	166 (4%)	179 (4%)	185 (4%)
Other	277 (7%)	280 (7%)	279 (7%)	267 (6%)
Commercial	42 (1%)	41 (1%)	43 (1%)	41 (1%)
School	35 (1%)	38 (1%)	39 (1%)	38 (1%)

Table 7.2: Users by institute type

7.2 Data Holdings

	Q1	Q2	Q3	Q4
collections	BADC 240 NEODC 33	CEDA 299	CEDA 301	CEDA 303
datasets		CEDA 2853	CEDA 2872	CEDA 2921

Table 7.3: Number of “dataset” collections and “datasets” identified by CEDA and displayed via CEDA catalogue. Note the transition from separate catalogues to combined catalogues during the year.

Data Centre	Q1	Q2	Q3	Q4
NEODC	26	40	40	40
BADC	242	247	247	245
UKSSDC	11	19	19	18

Table 7.4: Number of “dataset” discovery records held in the NERC data discovery service.

7.3 Help Desk Responsiveness

	Q1	Q2	Q3	Q4
Received	388	447	427	612
% 1-day Response	96	92	96	96
% Closed within 3-days	84	81	83	78
Closed	362	467	454	616

Table 7.5: Help desk queries received and closed by quarter, including the one day response and three-day closure rates. These queries cover all aspects of data support *except* resource application and approval correspondence.

	Q1	Q2	Q3	Q4
Received	860	852	842	894
% 1-day Response	98	97	98	96
% Closed within 3-days	88	86	86	84
Closed	838	897	861	931

Table 7.6: Help desk queries specifically about access authorisation for restricted CEDA datasets and services received and closed by quarter, including the one day response and three-day closure rates.

8. Publications and Presentations

- Andre, J-C, G. Aloisio, J. Biercamp, R. Budich, S. Joussaume, B. Lawrence and S. Valcke. (2014) High Performance Computing for Climate Modelling. BAMS 10.1175/BAMS-D-13-00098.1
- Blower, J.D., R. Alegre, V. Bennett, D.J. Clifford, P.J. Kershaw, B.N. Lawrence, J.P. Lewis, K. Marsh, M. Nagni, A. O'Neill, R.A. Phipps. (2014) Understanding Climate Data through Commentary Metadata: the CHARMe project. Communications in Computer and Information Science, 416, pp 28-39, 10.1007/978-3-319-08425-1_4
- Callaghan, S., (2015) Data without Peer: Examples of Data Peer Review in the Earth Sciences, D-Lib Magazine January/February 2015, Volume 21, Number ?, 10.1045/january2015-callaghan
- Callaghan, S. (2014) Preserving the integrity of the scientific record: data citation and linking, Learned Publishing, Volume 27, Number 5, 1 September 2014, pp. 15-24(10)
- Callaghan, S., J. Tedds, R. Lawrence, F. Murphy, T. Roberts, W. Wilcox (2014) Cross-Linking Between Journal Publications and Data Repositories: A Selection of Examples, International Journal of Digital Curation, Vol. 9, No. 1, pp. 164-175, 10.2218/ijdc.v9i1.310
- Callaghan, S., J. Tedds, J. Kunze, V. Khodiyar, R. Lawrence, M. S. Mayernik, F. Murphy, T. Roberts, A. Whyte (2014) Guidelines on Recommending Data Repositories as Partners in Publishing Research Data, International Journal of Digital Curation, Vol. 9, No. 1, pp. 152-163, 10.2218/ijdc.v9i1.309
- Mayernik, M.S., S. Callaghan, R. Leigh, J. Tedds, S. Worley, (2015) Peer Review of Datasets: When, Why, and How, Bulletin of the American Meteorological Society 96 (2), 191-201
- Mizielinski, M. S., Roberts, M. J., Vidale, P. L., Schiemann, R., Demory, M.-E., Strachan, J., Edwards, T., Stephens, A., Lawrence, B. N., Pritchard, M., Chiu, P., Iwi, A., Churchill, J., del Cano Novales, C., Kettleborough, J., Roseblade, W., Selwood, P., Foster, M., Glover, M., and Malcolm, A. (2014) High-resolution global climate modelling: the UPSCALE project, a large-simulation campaign, Geosci. Model Dev., 7, 1629-1640, 10.5194/gmd-7-1629-2014.
- Moine, M-P, C. Pascoe, A. Alias, V. Balaji, P. Bentley, G. Devine, R.W. Ford, E. Guilyardi, B.N. Lawrence, S.Valcke (2014) Development and Exploitation of a Controlled Vocabulary in support of Climate Modelling. Geosci. Model Dev., 7, 479-493, doi:10.5194/gmd-7-479-2014.
- Parton, G.A., S. Donegan, S. Pascoe, A. Stephens, S. Ventouras, B.N. Lawrence (2015). MOLES3: Implementing an ISO standards driven data catalogue. International Journal of Digital Curation, 2015, Vol. 10, No. 1, pp. 249-259. 10.2218/ijdc.v10i1.365
- Pepler, S., S. Callaghan, (2015) Twenty Years of Data Management in the British Atmospheric Data Centre, International Journal of Digital Curation, Vol. 10, No. 2, pp. 23-32 10.2218/ijdc.v10i2.379 Conference proceedings
- Bennett, V., P. Kershaw, M. Pritchard, J. Churchill, C. Del Cano Novales, M. Juckes, S. Pascoe, S.

Pepler, A. Stephens, B. Lawrence, J.- P. Muller, S. Kharbouche, B. Latter, J. Styles. (2014) EO science from big EO data on the JASMIN-CEMS infrastructure. Conference on Big Data from Space (BiDS '14) 10.2788/1823

- Bennett, V., P. Kershaw, M. Pritchard, J. Churchill, C. Del Cano Novales, M. Juckes, S. Pascoe, A. Stephens, B. Lawrence, S. Pepler, (2014) EO Science from Big EO Data on the JASMIN-CEMS Infrastructure, ESA BiDS'14 Conference, ESRIN, Frascati, Nov 2014
- Blower J., B. Lawrence, P. Kershaw, M. Nagni, (2014) Commentary metadata for Climate Science: collecting, linking and sharing user feedback on climate datasets. Geophysical Research Abstracts, Vol 16, EGU 2014-14574-1
- Callaghan, S., T. Carpenter, J.E. Kratz, (2015) Walk softly and carry a large carrot: how to give credit for academic work. In: FORCE2015, 12-13 January 2015, Oxford, UK.
- Clifford, D., J. Blower, R. Alegre, R. Phipps, V. Bennett, P. Kershaw, (2014) Annotating Climate Data with Commentary: the CHARMe Project. Conference on Big Data from Space (BiDS '14) 10.2788/1823
- Clifford, D., J. Blower, R. Phipps, R. Alegre and P. Kershaw, (2014) Annotating climate data with commentary: the CHARMe project, ESA BiDS'14 Conference, ESRIN, Frascati, Nov 2014
- Gomez-Dans, J., P. Lewis, N. Pounder, J. Styles, P. Kershaw, T. Kaminski and P. Van Bodegom, (2014) OPTImisation environment for joint retrieval of multisensor RADiances (OPTIRAD), ESA BiDS'14 Conference, ESRIN, Frascati, Nov 2014
- Henry, A., A. Wood, I. Mustafee, R. Alegre, J. Blower, P. Kershaw, M. Nagni, P. Harwood, R. Phipps, (2014) CHARMe: Earth Observation Metadata and the Semantic Web, 5th International Astronautical Congress, Toronto, Canada, Sept 2014
- Juckes, M., R. Swart, P. Thysse, W. Som de Cerff, A. Groot, V. Bennett, L. Costa, J. Lückenkötter, S. Callaghan, (2015) A Climate Information Portal for Copernicus: a central portal for European climate services, EGU General Assembly Conference Abstracts,Volume 17, Pages 11111
- Kershaw, P., B. Lawrence, J. Churchill, M. Pritchard (2014) The JASMIN Cloud: specialised and hybrid to meet the needs of the Environmental Sciences Community. Geophysical Research Abstracts Vol. 16, EGU2014-10820-3.
- Kershaw, P., R. Ananthakrishnan, (2014), ESGF IdEA — Identity, Entitlement and Access Management, ESGF Face to Face Meeting, Livermore CA, Dec 2014
- Kershaw, P., J. Churchill, B. Lawrence, (2014) The JASMIN Cloud: specialised and hybrid to meet the needs of the Environmental Sciences Community, abstract and presentation, EGU April 2014, Vienna
- Kershaw, P., J. Churchill, S. Pascoe, M. Pritchard, B. Lawrence, (2015) JASMIN Cloud, ESGF Face to Face Meeting, Livermore CA, Dec 2014
- Kershaw, Philip, Bryan Lawrence, Jose Gomez-Dans, and John Holt, (2015) Cloud hosting of the IPython Notebook to Provide Collaborative Research Environments for Big Data Analysis, EGU2015, Vienna, Apr 2015
- Lawrence, B., V. Bennett, J. Churchill, M. Juckes, P. Kershaw, S. Pepler, M. Pritchard, A. Stephens, (2015) Beating the tyranny of scale with a private cloud configured for Big Data, EGU2015, Vienna, Apr 2015
- Murphy, F., S. Callaghan, J. Tedds, V. Khodiyar, J. Kunze, R. Lawrence, M. Mayernik, T. Roberts, A. Whyte, (2014) Guidelines, Policies and Procedures for Data Publication from the PREPARDE project, EGU General Assembly 2014, held 27 April - 2 May, 2014 in Vienna, Austria
- Osprey, A., Riley, G. D., Manjunathaiah, M., and Lawrence, B. N. (2014). The development of a data-driven application benchmarking approach to performance modelling. In 2014 International Conference on High Performance Computing Simulation (HPCS) (pp. 715-723). 10.1109/HPCSim.2014.6903760
- Nagni, M., J. Blower, B. Lawrence, P. Kershaw, (2014) CHARMe Commentary metadata for Climate Science: collecting, linking and sharing user feedback on climate datasets, abstract and presentation, EGU April 2014, Vienna

- Pascoe, S., A. Iwi, P. Kershaw, A. Stephens, B. Lawrence, (2014) JASMIN Analysis Platform ? bridging the gap between traditional climate data practices and data-centric analysis paradigms. Geophysical Research Abstracts Vol. 16, EGU2014-10630
- Schutgens, N., P. Stier, P. Kershaw, S. Pascoe, (2015) Comparing apples and oranges: the Community Intercomparison Suite, EGU2015, Vienna, Apr 2015
- Lawrence, B. (2015) Taking Compute to the Data. Position paper for BDEC Barcelona '15.Pamment, A. (2014). Scientific Computing Training for NERC Researchers, poster, presented at NCAS Science Meeting, Marriott City Hotel, Bristol, July 17, 18 2014.
- Parton, G. A., S. Callaghan, (2014) Environmental Data Archival: Practices and Benefits. In: Problems in Information Sciences: Scholarly E-Publishing summer school., 04 July 2014, Weston Room, Maughan Library, Kings College London.
- Pritchard M., J. Churchill, (2015) "JASMIN: Petascale storage and terabit networking for environmental science" Presentation at JANET Networkshop 43, University of Exeter 31/3/2015 `https://networkshop.ja.net/events/networkshop43/programme/7689/1415`