

JASMIN Science Case (2016)



Bryan Lawrence

V1.1: Last Modified: 9th November, 2016; Approved 16th November 2016
(Science Case approved does not mean the business case has been approved and finance allocated.)

Executive Summary

JASMIN exists to provide the UK environmental sciences the compute facility they need to deliver cost-effective world class science and impact from the exploitation of data. A £17M investment is needed for the next generation of JASMIN, to maintain the UK's scientific and competitive edge, facilitating the exploitation of world-class environmental science to meet the global societal challenges of the future. Such an investment would build on international leadership and would support:

- The merging of extremely large environmental data sets with the latest earth system models: building downstream growth in space-based environmental services; and underpinning international collaborations.
- The transformation of data into information products and services; JASMIN provides the foundation for the UK environmental information ecosystem: investment will provide greater access to knowledge for a range of users.
- Enabling researchers to better support government usage of environmental hazard data resulting in large-scale societal benefit e.g. development of earthquake monitoring systems.
- The next generation of earth observation and environmental simulation, including for the next phase of the World Climate Research Programme (WCRP) global model intercomparison project (CMIP6).

The updated JASMIN will deliver cost-effective, world-class environmental science, exploiting data for societal benefit. It will be a cutting-edge novel computational environment, ensuring highly-skilled people are retained in the UK from systems engineers to environmental data users, from data scientists and analysts to mathematicians.

Contents

1	Introduction	3
2	Benefits of JASMIN Investment	4
2.1	Economic and Policy Impact	4
3	Why and what is JASMIN?	7
3.1	Science Context	7
3.2	Underlying Trends	8
3.3	The JASMIN Value Proposition	9
3.4	JASMIN Services	10
3.5	JASMIN in the UK ecosystem	11
4	Scientific Use of JASMIN	12
4.1	List of Science Exemplars	12
5	Metrics of JASMIN usage	21
5.1	Storage usage	21
5.2	Compute usage	23
6	JASMIN Architecture: Now and in the Future	26
6.1	JASMIN in late 2016	26
6.2	Influences on future Architecture	27
6.3	Key technical requirements	30
7	JASMIN Investment Plan	33
7.1	Management and Governance	34
7.2	Recurrent Support	34
7.3	User Engagement and Support	35
7.4	Next steps	35

1 Introduction

The Natural Environment Research Council (NERC) requires a shared compute facility to enable the exploitation of the data it produces as part of delivering world class impactful science. The need for such a shared facility was first identified by a Joint Weather and Climate Research Programme workshop in 2010, at the same time as it was becoming apparent that the earth observation community also needed a suitable facility for shared product development. The starting size and scope of a suitable facility were established from underlying science requirements, with initial funds provided via a government capital

investment in 2011. The first phase of JASMIN was deployed in March 2012 with several phases following, but in 2016 it is time to assess progress, and make the case for replacement and extension.

This document provides a description of the JASMIN facility, and links aspirations for equipment replacement and expansion to requirements arising from environmental science. It begins with a section stating the key benefits expected from an investment in JASMIN, reviews key elements of the results achieved thus far, before proceeding to a summary of the key underlying drivers for where JASMIN is today and where it needs to go. A summary of the current architecture and usage is presented before discussing

The UK Space and Innovation Growth Strategy (updated in 2015) has set a target to grow the UK share of the world's space economy to 10% by 2030 — an estimated £40 billion per annum of space-enabled turnover and the creation of 100,000 new jobs. Part of the original investment in JASMIN came from the UK Space Agency (UKSA) in response.

One objective within that plan was to support the community to implement a Climate Services Centre for Europe in the UK. The Centre of Environmental Data Analysis (CEDA) has been influential in advancing this agenda — in partnership with the Satellite Applications Catapult.

CEDA is hosting the European Space Agency's Climate Change Initiative (CCI) archive and portal on JASMIN and leading on the provision of climate projections to the European Commission's Copernicus Climate Change Service (C3S) — again exploiting JASMIN. The "Climate Data from Space" zone on JASMIN, co-funded by the UKSA and the NERC National Centre for Earth Observation, is enabling both UK industry and academia to contribute to these important programmes as well as facilitate the use of their products by the wider community.

Box 1: Enabling Space Innovation and Growth — a key objective for JASMIN



specific plans for the future. Exemplars of the existing scientific use of JASMIN are presented in "scientific use case" boxes throughout, using text provided by the groups involved.

decadal climate datasets (e.g. land and lake temperatures, albedo, greenhouses gases etc.), and comparing and contrasting environmental models and data.

2 Benefits of JASMIN Investment

JASMIN infrastructure underpins all NERC science discipline areas — new investment will expand the scope and interdisciplinarity of that work and align with NERC's high-level objectives. Examples include:

- Benefiting from natural resources: JASMIN has been, and will be used further, to understand wave and wind risk to oil and gas extraction in the North Sea and to exploit earth observation data to understand forestry patterns on a global scale.
- Building resilience to natural hazards: JASMIN will help improve high-resolution simulations of flooding and land-use, and provide understanding of the deformation of land surface in earthquakes.
- Managing environmental change: JASMIN will also help support the next generation of climate projections, developing essential

2.1 Economic and Policy Impact

JASMIN has previously benefited from "Eight Great Technologies" investment and would continue to deliver under the themes of Big Data and Energy Efficient Computing and Satellites and Commercial Applications of Space. Ongoing investment in JASMIN will ensure it continues to form a key part of the national e-Infrastructure — providing an essential step in a workflow that begins with satellites and traditional high-performance computing, requires dedicated data analysis facilities and curated data archives, and results in innovative information products. Such investments in JASMIN would also contribute to the Government's target to fill the digital skills gap, identified in the Big Data Dilemma¹ and Digital Skills Crisis² reports.

¹<http://www.publications.parliament.uk/pa/cm201516/cmselect/cmsctech/468/468.pdf>

²<http://www.publications.parliament.uk/pa/cm201617/cmselect/cmsctech/270/270.pdf>

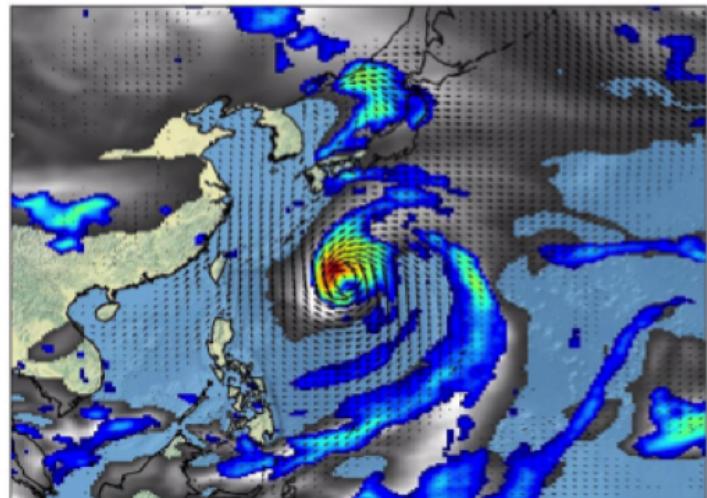
HRCM — Simulating the building blocks of climate

The High Resolution Climate Modelling (HRCM) is a collaboration between the Hadley Centre (UK Met Office) and the Climate Directorate within NCAS. High resolution is often necessary to simulate important processes in the atmosphere. For example, hurricanes evolve and grow from local sea and air conditions. Simulating hurricane climate (where, how strong, and how often) depends on being able to resolve small spatial scales over long periods of time.

Members of the programme utilise a range of high performance computing platforms to simulate current and future climate (e.g. see box 12). In 2013, the group exploited the HERMIT supercomputer in Stuttgart, Germany, via a single allocation of time from PRACE (in the UPSCALE project, still the largest single high performance computing project supported in Europe). One year of simulations produced 400 terabytes of data which was moved to and analysed on JASMIN. (In data output, equivalent to a quarter of the total output archived in the Earth System Grid Federation by the entire global modelling community as part of the fifth Climate Model Intercomparison Project, CMIP5.)

These data are still stored on JASMIN, although most is now on tape with “only” 80TB online in the UPSCALE group workspace, and will contribute to scientific and industrial impact for many years to come. JASMIN was the only European facility that could receive the whole UPSCALE data set at the required data rate (up to 4-5TB/day) and provide access to all project participants.

One part of this work has been the routine tracking of tropical cyclones – on JASMIN 50 years of global 25km data can now be processed in one day with just 50 jobs. Another has been the analysis of “eddy vectors” where total processing time has been reduced from 3 months to merely 24 hours with 1600 batch jobs. These examples demonstrate the **seismic difference in analysis time** that the JASMIN/LOTUS combination has achieved for data processing in “data heavy” research areas, such as climate modelling.



Scientific Use Case 1: Simulating key climate processes such as hurricanes

Contact Prof P.L. Vidale (NCAS, University of Reading) or visit <http://hrcm.ceda.ac.uk>

Enhanced JASMIN capability would directly enable the UK to achieve the following:

- Manage a national curated archive of globally sourced environmental modelling products — placing the UK at the very center of international research and providing UK researchers with immediate access to the most comprehensive datasets.
- Centralise the management and provision of large-scale data across all the NERC disciplines, enabling more and improved cross-disciplinary science.
- With the UKSA, increase the downstream

impact of NERC research into markets identified in the Space Innovation and Growth Strategy thus contributing to the economic growth in a key industry sector and positioning the UK at the leading edge of earth observation data exploitation.

- Support access to large-scale datasets by scientists eligible for Official Development Assistance, both as part of joint research programmes, and in response to emergencies. This will help the UK deliver the Global Challenges Research Fund (GCRF) by supporting cutting-edge research that will aid

JASMIN: Upskilling and supporting the community

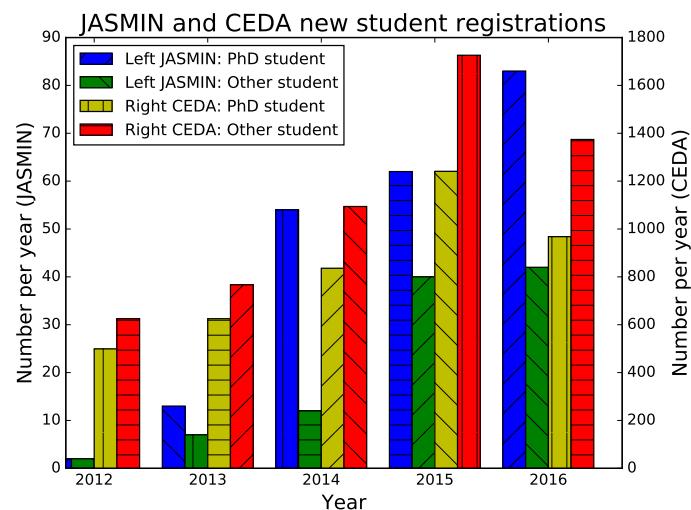
Both the 2010 and the 2012 refresh of the NERC assessment of the most wanted postgraduate and professional skills needed in the environment sector identified modelling, data management, and multi-disciplinarity as three of the most wanted skills. Industry regularly identifies big data skills, and data analytics training as crucial to the future of the UK's digital economy.

JASMIN provides a platform to address all these skills, from supporting modelling, sharing and data management at scale, to the development and deployment of complex diagnostic algorithms.

Over a hundred (mostly doctoral) students register to use JASMIN every year (data for 2016 to October), with around two thousand more registering each year to exploit data via the CEDA download services. Online, workshop and conference training are provided by the CEDA and JASMIN teams.

The next generation of JASMIN will also support a wider variety of computational environments to further enhance opportunities for skills development, including that necessary to exploit cloud computing in general.

Box 2: As well as post-doctoral and other research scientists, a steady stream of graduate and post-graduate students are being educated to use big data technology, while thousands more are supported in their data requirements by CEDA!



developing countries e.g. simulating hurricane climate (where, how strong, and how often).

- More effective use of the national HPC facility (ARCHER), particularly when undertaking large scale environmental simulations at higher spatial and temporal resolutions.

Without further investment the following major international opportunities would be impacted:

- The UK and Europe have invested heavily in satellite programmes from the EC funded ESA Sentinel programme; without improved capability and capacity, the UK academic community will not be able to fully exploit all the mission data; reducing the impact of the initial economic investment in the ESA sentinel programme.
- CMIP6 (box 11) is beginning in late 2016 — this is expected to produce tens of petabytes of data at modelling centres across the globe.

Without investment in JASMIN the UK will not be well positioned to exploit the CMIP6 data expected both to underpin future IPCC assessment reports, future climate advice to government and the burgeoning climate service sector.

Other high level risks with not investing:

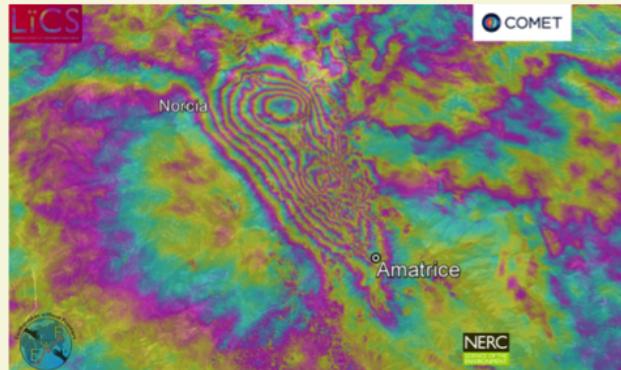
- Lack of competitiveness of UK climate and environmental space services industry given inability of primary research to deliver fundamental datasets and understanding.
- Insufficient return on investment on the primary national HPC capability (currently ARCHER); the transition to higher spatial and temporal resolutions means that large scale simulation on national platforms depends on the presence of suitable storage and analysis facility such as JASMIN.
- Decreased ability to produce scientists with big-data and high-performance computing skills.

Customised Data Intensive Computing for Fault Analysis

Earth observation data is voluminous (just one radar on the Sentinel 1a satellite generates petabytes of data per year — a petabyte being roughly equivalent to 2000 years of MP3 music).

As well as storing the Sentinel data, machines in the JASMIN private cloud can be customised to provide topic specific software environments. One such customisation supports earthquake monitoring. COMET scientists are automating the production of deformation maps using data from Sentinel-1, allowing them to measure surface deformation in multiple images with millimetre accuracy; this big data project would be unfeasible without the unique computational and data storage capability of JASMIN.

As well as responding to sudden events such as earthquakes and volcanic eruptions, the data will be used for long-term monitoring of earthquake faults and volcanoes. The data will also be used to investigate other sources of ground movement, such as subsidence caused by resource extraction (e.g. fracking).



Scientific Use Case 2: Analysis of Faults from Space

Contact Prof Tim Wright (COMET, University of Leeds or visit <http://comet.nerc.ac.uk>

- Lack of capacity to support overseas development assistance projects that depend on direct data access and exploitation e.g. the UK's central role in developing prediction systems for the Indian monsoon.

environmental process or assembly of processes, and comparison of those simulations with observations — and the observations might range from instruments deployed in the field, on platforms such as ships or aircraft, or in space.

The more complex the assembly of processes, or the larger the scale of interest, the greater the volume (and variety) of data required. In addition, over time the influence of faster cheaper computing has also led to higher volumes of data from both sensors and simulations, as both access higher spatial and temporal resolutions.

As a consequence, much important science now involves very large amounts of data in the production, analysis and evaluation, particularly when working at the global or high-resolution regional scale. For the communities involved it is no longer possible to do world leading science with the amounts of data which can be affordably stored and managed locally — and the number of such communities is growing rapidly. These communities are also developing analysis algorithms which require customised computing environments alongside the data.

For other communities, and for those who need information products rather than raw data, it

3 Why and what is JASMIN?

The “Joint Analysis System” was conceived of to deliver a coordinated national response to significant trends in science which had led to the need for a new type of data intensive computing facility. In this section we describe those trends, explain the JASMIN value proposition in the context of those trends, and briefly describe the key services delivered by JASMIN.

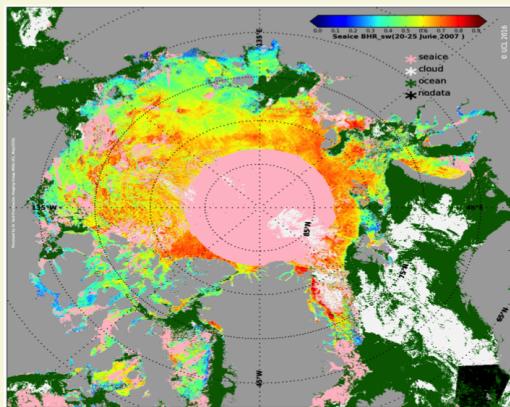
3.1 Science Context

The primary drivers for JASMIN come from the cross-disciplinary environmental simulation and earth observation communities. The main goal of these activities is to document the current state of the environment, predict the near term future, and project longer-term futures under a variety of possible scenarios. In most cases the science involves some sort of simulation of an

Quality Assured Earth Observation Data

Satellites are providing a continuous global measurement of the land surface and atmospheric chemistry; revolutionising the way scientists assess climate change and air quality.

New essential climate variables are being generated on JASMIN as part of the European funded QA4ECV project. This involves processing hundreds of terabytes (TB) of data to generate a 35 year record of quality assured global land surface products from European and US EO instruments.



The time taken to process these satellite datasets has been dramatically improved (by 2 orders of magnitude); one example showed completion in just 3 days — 81 times faster than on a local machine. Many other processing tasks show similar benefits e.g. ≈ 100 times faster processing.

A dedicated project workspace (≈ 700 TB) on JASMIN has been funded through the UK NERC Big Data programme. The project has also benefited from a high-performance transfer server, which has allowed input datasets (MODIS, MISR) to be transferred from US data providers onto JASMIN very efficiently, at rates up to 28 TB/day.

Without JASMIN, this project would have been very difficult to pursue.

Scientific Use Case 3: Land Surface Products generated from Earth Observation

Contact: Prof Jan-Peter Muller (University College London) or visit <http://www.qa4ecv.eu/>

is often the case that the products they need are subsets in space or time of the primary large datasets, or that they need a bespoke workflow to generate the product they need. For them, it is important that it is possible to do the data reduction where the primary data exists, rather than again, attempting to download and process locally.

Even where volume is not the problem, it can be the case that expert management of a variety of datasets in a co-located environment can lead to significant benefits for users — whether they are actually bringing their computation to the data, or downloading subsets for remote working.

Impactful science also requires larger collaborative teams, and sometimes industrial partners who can exploit data to produce information products of commercial or policy relevance. Such teams and partnerships need to be able to work on common data, even if they develop different interfaces and tools (such as websites which differentiate between downstream academic and commercial use to provide very different products and information).

3.2 Underlying Trends

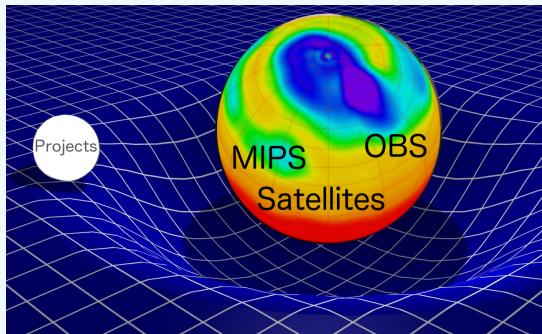
This environmental science context can be characterised by four main trends:

1. Growth in the use of direct numerical simulation (accompanied by significant increases in simulated data),
2. Growth in the volume of remotely sensed earth observation data, especially from space,
3. Increases in the complexity of the algorithms used to exploit data, and
4. Increases in the sizes of teams working on data analysis.

JASMIN was originally conceived of in response to the last of these: the need for scientists from different institutions to share data and analysis techniques, but that requirement coincided with an additional growing requirement from the NERC data management community:

5. The need to manage and preserve increasing volumes of data products from the NERC community, and facilitate community access to growing amounts of high volume data produced elsewhere.

The Data Commons — Enabling the bringing of computation to data



The value proposition:

- Provide a state-of-the art storage and computational environment
- Provide and populate a managed (curated) data environment with key datasets (the “archive”).
- Provide **flexible** methods of exploiting the computational environment.
- Encourage and facilitate the bringing of data and/or computation alongside/to the archive.

... in doing so, exploit the “data gravity” associated with managed data so that users want to bring their projects to the the JASMIN environment because it enables them to either do something they couldn’t do before, or to do things fast enough to change the way they approach their science.

Box 3: The JASMIN value proposition: a data commons with enough compute to use the data. To enable doing something that couldn’t be done before, or make it much more efficient!

None of these trends are abating, with data growth continuing to accelerate, the advent of data science as a discipline introducing new algorithmic approaches, and teams growing in both numbers and breadth of background. Data management issues are growing in response. There are also three new trends:

6. The re-emergence and acceleration of heterogeneity in compute and storage hardware, even as
7. More compute and storage work is migrating to “the cloud”.
8. The advent of the “internet of things” and of “sensor webs”.

These drivers and their influence on JASMIN configuration, now and in the future, are discussed more below.

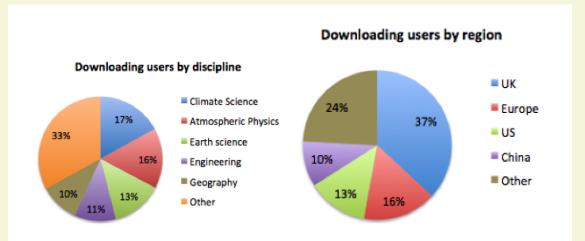
3.3 The JASMIN Value Proposition

Many of the trends outlined in the previous section can provide insurmountable hurdles to individual scientists: too much data to handle, insufficient storage with the compute, difficult to share data and workflows. They are not even easily addressed at the institutional level — there may not be a large enough critical mass of scientists with similar problems to justify a dedicated facility and a shared cross-disciplinary facility wouldn’t necessarily solve the problem either.

The solution involves providing a suitable national facility which provides three key properties: it has adequate storage to deal with shared large-scale long-term persistent data, it has adequate storage to deal with the transient data needed and produced within individual projects, and it has appropriate (adequate and flexible) compute alongside the data so that moving the data is not required to use it.

Curation and Facilitation at CEDA — a core JASMIN tenant

The screenshot shows the homepage of the Centre for Environmental Data Archival. At the top, there's a header with the logo, name, and contact information. Below the header, a navigation bar includes links for Home, Data Centres, Services, Projects, For Academics, For Business, Training, Help, and Contact Us. A search bar is also present. The main content area is titled 'Data Centres' and lists several data centres: British Atmospheric Data Centre, The British Atmospheric Data Centre (NERC Earth Observation Data Centre), The UK Solar System Data Centre, and The Intergovernmental Panel on Climate Change (IPCC) Data Distribution Centre. Each entry includes a small logo and a brief description.



The Centre for Environmental Data Analysis supports a range of data centres. In 2016 it curates nearly 3PB of data and supports over 30,000 users. In 2015 4,000 registered users and 6,000 anonymous users downloaded a total of 600 TB of data in 14 million files.

In 2011, the CEDA computing system needed a complete refresh. This was accommodated during 2012 by migrating the data onto JASMIN, making CEDA the first, and biggest JASMIN tenant. CEDA provides a key component of the data gravity which provides the JASMIN value proposition (box 3).

Scientific Use Case 4: The Centre for Environmental Data Analysis: curating data at scale.

Contact: Dr Sam Pepler (NCAS and CEDA) or visit <http://ceda.ac.uk>

In this context, the JASMIN value proposition (box 3) has evolved to deliver these properties, in particular by providing the three canonical cloud services:

1. Infrastructure as a Service (IaaS): the provision of virtualised hardware which users can configure with their own operating systems, software, and user management.
2. Platform as a Service (PaaS): the provision of complete computing systems which have been pre-configured with software and are ready to use, and
3. Software as a Service (SaaS): the provision of scalable software services providing applications users need without exposing the underlying computing systems (e.g. download services from the curated archive).

In doing so, to enable:

1. Scientists to be able to do things they otherwise couldn't do at all,
2. Scientists to be able to do things fast enough that it completely changes the nature of their workflow. If something can be done in days or hours that used to take months to years, innovation is enhanced, as is throughput and efficiency.
3. The development, management and dissemi-

nation of information products utilising the same platform as the underlying scientific data — minimising the friction between science and impact.

3.4 JASMIN Services

JASMIN itself is hosted in the Scientific Computing Department of the Science and Technology Facilities Council (STFC, in the Rutherford Appleton Laboratory) and managed by the joint NERC/STFC Centre for Environmental Data Analysis (CEDA) on behalf of NERC.

The primary use case for JASMIN is to enable the storage and exploitation of high volume data by groups, by improving their ability to work together and in doing so minimise the expensive and inefficient replication of data across academia. To do this, JASMIN is configured to support users who fall into “consortia” themselves organised into “tenancies”. The JASMIN consortia cover the major disciplines of NERC science as well as an extra consortium representing CEDA itself. The tenancies are self-organised groups who have defined requirements in terms of storage or compute or both. Tenants are allocated resources within an overall envelope of resources allocated to a consortium.



ARCHER: The national “Tier-1” high performance computing platform in Edinburgh, procured by the Engineering and Physical Sciences Research Council (EPSRC) on behalf of EPSRC and NERC.



MONSOON: The shared Met Office and NERC “Tier-2” development platform in Exeter, purchased by the Met Office on behalf of the Joint Weather and Climate Research Programme.



University “Tier-2” machines; such as the POLARIS N8 machine based at the University of Leeds.



Remote data archives, from the Copernicus ground segment, NASA, JAXA etc, are accessible via national and international networks.

Box 4: Key hardware components of UK environmental e-infrastructure. JASMIN is connected to ARCHER, MONSOON and the University of Leeds with both normal JANET backbone links and dedicated lightpath network links. European and global data archives are accessible via GEANT.

CEDA itself is a tenant, providing and managing an archive (using JASMIN storage and compute) to all the other tenants. All tenants (including CEDA) are allocated resources from or on one or more of:

1. Group Work Spaces (GWS) — finite volumes of disk on the fast storage accompanied by provision of access to tape storage via the “Elastic Tape” service³.
2. Access to “Generic Platform Compute” — login and access to machines configured for generic scientific analysis and data transfer.
3. Hosted Platform Compute — bespoke machines deployed on their behalf in the “JASMIN Managed Cloud”.
4. Infrastructure Compute — access to the JASMIN private cloud portal and a quota of machines upon which to configure their own resources.

5. The Lotus Batch Cluster — a traditionally configured batch cluster with a range of processors and memory configurations.

The JASMIN system has 15 petabytes (PB) of usable high performance disk which is used to support the CEDA Archive and the Group Work Spaces, supported by over 30 PB of tape capacity and \approx 0.5 PB of bulk storage in the cloud. Approximately 5,000 cores are deployed in the Lotus Batch cluster and the hypervisors which support the platform and infrastructure compute. More details of the system are in section 6.1.

3.5 JASMIN in the UK ecosystem

JASMIN is only one part of the computing system underpinning environmental science; other key UK components listed in Box 4 include the two major HPC platforms ARCHER and MONSOON, as well as other UK HPC platforms such as those in universities. From a data source perspective, equally important are the data distribution centres associated with the Copernicus

³CEDA makes use of an alternative tape service for the archive: Storage-D, which is more suitable for use by the archive maintainers.

ground segment, such as the ESA collaborative data hub (<https://colhub.copernicus.eu/>).

The UK community also make occasional use of the European Partnership for Advanced Computing (PRACE) resources as well as next generation test systems provided by the STFC Hartree Centre. There is limited, but growing use of the public cloud.

In addition to hardware, another key component of the ecosystem is software support, which for the environmental sciences is mainly provided by the NCAS Computational Modelling Services, with additional projects supported by the ARCHER Distributed Software Engineering programme. Software support necessarily includes that for the workflow associated with data production on national HPC platforms and migration to JASMIN.

It is worth noting that by providing a central data analysis environment for the UK environmental sciences, the community avoids not only up to N copies of data for $N (> 3)$ HPC centres, it avoids up to $N \times (N - 1)$ data flows which would otherwise be necessary to compare data from, and at, those N different sites. At petascale, this is a potential saving which runs into the hundreds of thousands of pounds!

4 Scientific Use of JASMIN

We can consider the usage of JASMIN from two perspectives: how the equipment is utilised, and how that impacts on the science. In section 5 we present the use of JASMIN in terms of user numbers and how they exploit the various characteristics of the system such as storage and compute. Here we concentrate on the scientific outcomes. In doing so, we have selected exemplars of usage chosen to represent the breadth, maturity of science, and scale of volume, variety and community. We have examples from graduate student science to large international programmes, from NERC research grants, to third party contracts (e.g. from ESA).

Most of these examples were provided by the JASMIN user community, and we have presented each as a use-case box with the contact details of the information providers. Not all examples are presented in this section, some are distributed

elsewhere in the document: the following is a list of all the scientific use cases presented, and their location in the document.

4.1 List of Science Exemplars

	Use Case	Page
1	Simulating key climate processes such such as hurricanes	5
2	Analysis of Faults from Space	7
3	Land Surface Products generated from Earth Observation	8
4	The Centre for Environmental Data Analysis: curating data at scale.	10
5	Measuring greenhouse gases from space	13
6	Remote Sensing from Space of Smoke Emissions	14
7	Estimating wildfire gas emissions from earth observation	14
8	High Resolution Ocean Modelling	15
9	Comparing simulations of ozone with observations	15
10	Examining extreme weather events in future climate	16
11	Using the NAME dispersion model to predict research flight paths . .	16
12	Modelling and Observing Oxidant Chemistry over the Indian Sub Continent	17
13	Understanding the connection between climate change and marine coastal zones	17
14	The influence of small ocean scales on weather forecasting	18
15	Sharing and exploiting coupled ocean-atmosphere simulations . .	18
16	Measuring changes in deep ocean heat content	19
17	Forestry Thematic Platform	19
18	Deploying a Data Portal for the Climate Change Initiative	20
19	MAJIC: Managing Access to JULES in the cloud	20
20	Using Global Sea Surface Temperature Measurements	22
21	Near real-time volcanic plumes	32

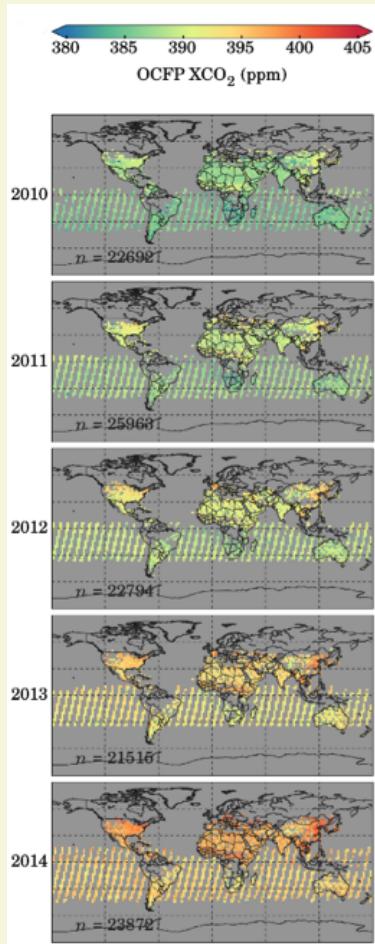
Reprocessing satellite missions on JASMIN

Carbon dioxide (CO₂) and methane (CH₄) are the two most important anthropogenic greenhouse gases (GHGs). The capability to accurately and frequently measure them on a global scale is vital to understanding their impact upon the climate system. The requirement for large spatial coverage and regular sampling means that satellite observations are a key component of such a monitoring system. These observations, combined with atmospheric transport modelling, help to improve our knowledge of the sources and sinks of gases and ultimately improving future climate predictions.

A complex algorithm is required to infer the amount of GHGs in a satellite remote sensing measurement. For new satellites, such as the NASA Orbiting Carbon Observatory-2, this number can be in excess of 30 million measurements each month (24 measurements every second) - necessitating vast processing capabilities with tens to hundreds of terabytes of storage.

JASMIN enabled the ESA GHG-CCI project to reprocess the entire GOSAT mission (2009-present) with the improved CO₂ algorithm within several weeks, as opposed to the many months that it would take locally.

The intensive number of measurements and data volume provide a huge challenge for individual institutions. Future missions, such as Sentinel 5P, are expected to see the number of measurements and data volume preclude any local processing; thus a national capability is vital in continuing to produce important climate datasets.



Scientific Use Case 5: Measuring greenhouse gases from space

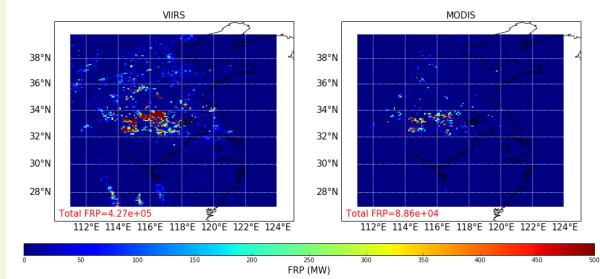
Contact: Robert Parker, Peter Somkuti, Hartmut Boesch (NCEO, University of Leicester) or visit
<http://www.leos.le.ac.uk/GHG/>

- 22 The OPTIRAD “software-as-a-service” ipython-notebook architecture. 33

Emissions from Smoke Pollution

Agricultural or landscape burning is carried out to clear the land for future uses, however the associated emissions of smoke pollution can be extremely harmful. For example, widespread burning of crop residues are suspected of contributing significantly to China's air quality problems, including in the mega-cities of Shanghai and Beijing. Such emissions can be observed from space.

New data from the VIIRS and Himawari-8 Japanese satellites are being used to estimate fire activity, exploiting their high spatial resolutions (existing datasets are known to under-represent the number and extent of fires in agricultural regions). These new data are tens of TBs in volume. JASMIN is essential to processing these large datasets, using new algorithms, successfully built, tested and optimised to process the data.



Initial work shows that existing inventories may be underestimated by around an order of magnitude. This work is helping to support the UK's official development assistance (ODA) which aims to assist science and innovation partnerships for economic development in overseas countries. The next steps will involve a complete smoke emission inventory, which will be used as a key input to improve future air quality models; these could then be used by policy makers to appropriately target how air quality could be improved with legislation.

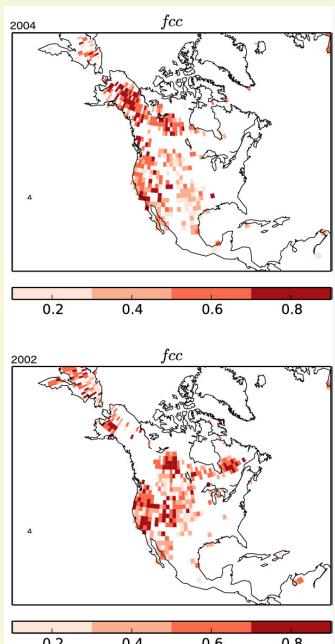
Scientific Use Case 6: Remote Sensing from Space of Smoke Emissions

Contact: Tianran Zhang, Weidong Xu, Jiangping He and Martin Wooster (NCEO, Kings College London)

Deriving the impact of fire on vegetation from EO data

Fire is ubiquitous in the Earth System. There is a pressing need to assess fire and its role in the carbon cycle, as well as its impact on economics and biodiversity.

UCL scientists have developed methods to quantify the impact of fire on vegetation. These metrics are important as they allow us to see whether over the satellite record, fire impact has changed. In order to produce a decadal dataset of fire impacts, a large amount of input data needs to be processed.



This work is normally constrained by both storage space and processing capabilities: JASMIN is a fantastic tool on both counts: not only is there adequate storage, but accessing large datasets is also very efficient — allowing complicated processing on typical tasks to be done in a few weeks, instead of months to years. Without JASMIN this intensive processing would involve large amounts of time and effort, but now scientists can focus their time on addressing the science rather than processing logistics.

These datasets will be used to improve greenhouse gas emission estimates from wildfires, in the past, and in the future.

Scientific Use Case 7: Estimating wildfire gas emissions from earth observation

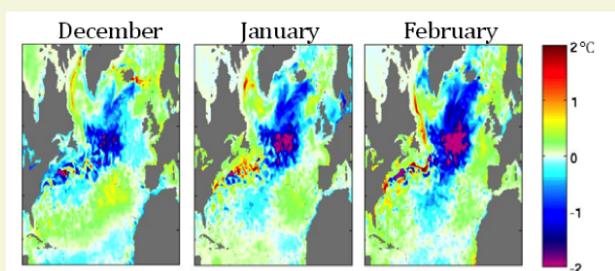
Contact: Jose Gomez-Dans, Prof P Lewis and J Brennan (University College London)

Oceanic Influences on the prediction of UK winter climate

Skilful prediction of winter climate over the UK and Europe is of major societal benefit as even modest skill at predicting likely winter temperatures, precipitation and winds a season in advance would allow advance contingency planning in diverse areas such as; health, transport infrastructure, power generation and flood prevention/mitigation.

There is potential to greatly improve the accuracy and reliability of our forecasts; however we need to resolve oceanic features down to scales of 25 km or less, thus complex ocean models are necessary.

CHARISMA is a Met Office – National Oceanography Centre (NOC) collaboration exploring the use of very high resolution ocean models in seasonal and decadal forecasting and ultimately in climate change projections. These models produce vast amounts of data and are run either on the NERC national supercomputing platform, ARCHER, or the Met Office supercomputer.



JASMIN is absolutely essential for the success of the CHARISMA project. 200 forecast simulations have been run with high-resolution ocean and atmosphere components. The fast links between the Met Office and Archer supercomputers to JASMIN, its high storage capacity and the advanced analysis procedures it provides are the key features which made this study possible.

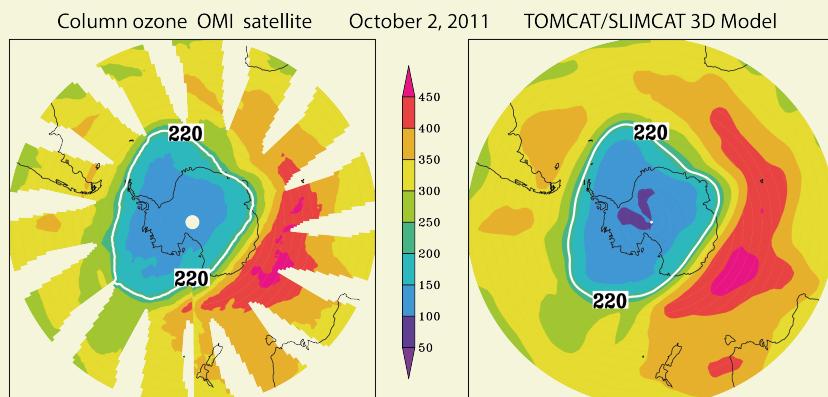
Scientific Use Case 8: High Resolution Ocean Modelling

Contact: Dr Bablu Sinha (NOC, Southampton)

Antarctic Ozone Hole: Comparing Simulations with Observations

Ozone strongly absorbs ultraviolet radiation which is harmful to living organisms. Ozone depletion within the stratosphere is therefore a serious global threat to humans, animals and plants.

Comparisons between ozone simulations and satellites observations (such as those shown in the image) have allowed scientists to gain a good understanding of stratospheric ozone depletion and how it will evolve in the future. This information is important for policy makers who need to evaluate the success of the Montreal Protocol.



JASMIN is a good platform for this research because it provides flexible compute nodes suitable for a relatively inexpensive parallel 3-D model such as TOMCAT alongside the large amount of disk space required for (1) the meteorological analyses used to force the model, (2) the daily model output ,and (3) satellite comparisons from the archive.

Scientific Use Case 9: Comparing simulations of ozone with observations

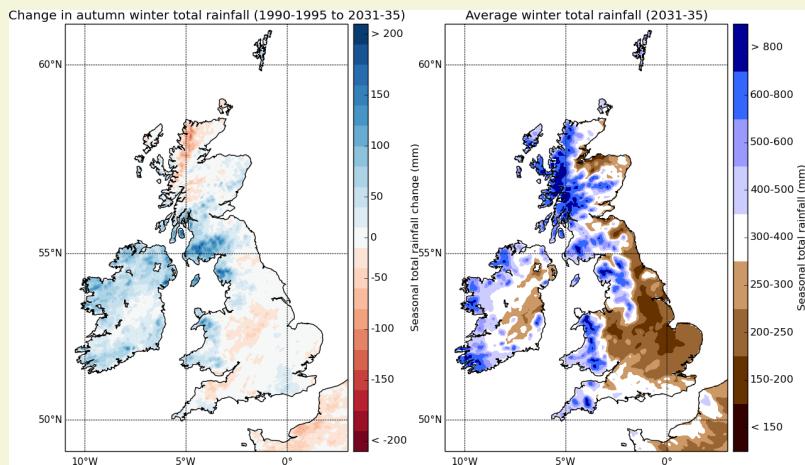
Contact: Prof Martin Chipperfield (NCEO and University of Leeds, Dr Wuhu Feng (NCAS), Chris Wilson and Richard Pope (NCEO) or visit <http://www.see.leeds.ac.uk/tomcat>

Analysing the output of high resolution weather runs

Extreme weather events are becoming increasingly important in our warming climate. The WISER project specifically aims to address how these extreme events affects UK weather over the next two decades, with resolutions of near 20 km globally down to 3km locally.

Most of the computations were completed on Archer (each model month takes 30 hours with 4,000 cores). However, the output data still needs to be analysed, requiring fast disk access for each half decadal run. JASMIN is ideally suited and invaluable for this task.

The figures show the UK winter rainfall accumulations for (2021-2035) inclusive, and the change from the (1990-1995) values.



These simulations are extremely important to policy makers and scientists for understanding and predicting high impact events, such as flooding.

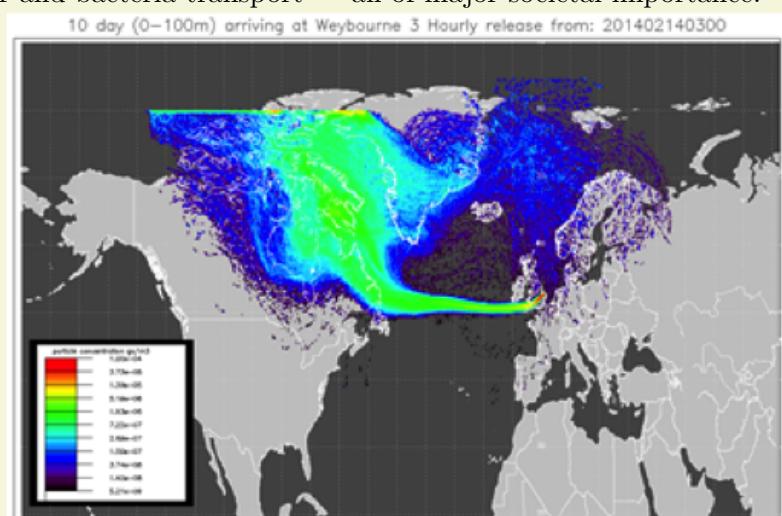
Scientific Use Case 10: Examining extreme weather events in future climate

Contact: Dr Alan Gadian and Dr Ralph Burton (NCAS, University of Leeds) or visit
<http://homepages.see.leeds.ac.uk/~lecag/wiser>

Modelling atmospheric dispersion events on JASMIN

The dispersion of air during nuclear accidents and volcanic eruptions is complicated and requires complex models to simulate. The Met Office NAME model can be used to simulate such events, and also to understand smoke, pollen, odour and bacteria transport — all of major societal importance.

The NAME-on-JASMIN service was opened to users in 2014, who can now exploit $\approx 20\text{TB}$ of input weather-model data — allowing NAME runs at various resolutions and locations. The ease of access has allowed users with little modelling experience to use and easily adapt a specialised model, sharing best practice, output data and analysis code. HPC experts and novice users have been brought together; greatly benefitting the UK research community



GAUGE (a project looking at UK greenhouse gas emission estimates) used NAME-on-JASMIN with near real-time forecasts to support research aircraft science flights. Using JASMIN allowed the scientists to forecast exactly where the aircraft should fly to get the best measurements.

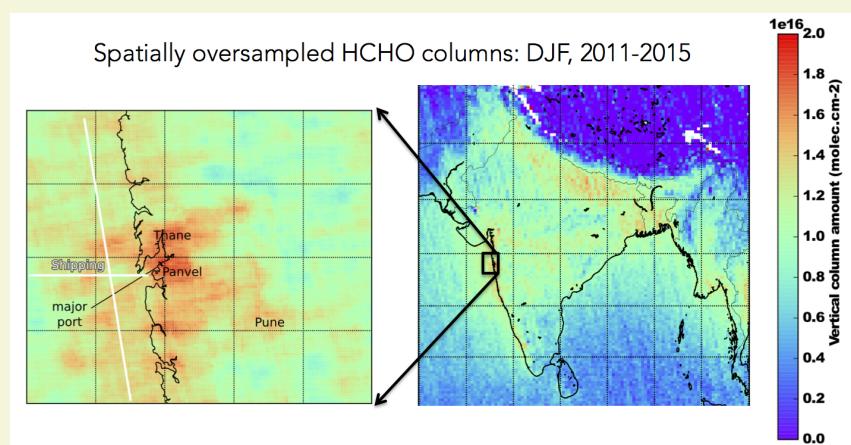
Scientific Use Case 11: Using the NAME dispersion model to predict research flight paths

Contact: Dr Zoe Fleming (NCAS, University of Leicester) or visit
<http://www.metoffice.gov.uk/research/modelling-systems/dispersion-model>

Combining model simulations with observations of air pollution

Understanding the balance between natural and anthropogenic chemistry is vital when observing air quality, especially over countries known for their poor air quality, such as India.

Formaldehyde is a product of volatile organic compounds that are a precursor for tropospheric ozone and organic aerosol, both of which are harmful at elevated concentrations to human health. JASMIN is being used to run the GEOS-Chem model at 25 km over India to interpret satellite observations for formaldehyde from the Ozone Monitoring Instrument (OMI).



The technical challenge is to run the model at the necessary spatial resolution to pick up individual cities over India; the advantage of using JASMIN is that the model output, earth observation and meteorology data are collocated so there is a seamless transition between model analysis and data analysis that is required to achieve these cutting-edge science objectives.

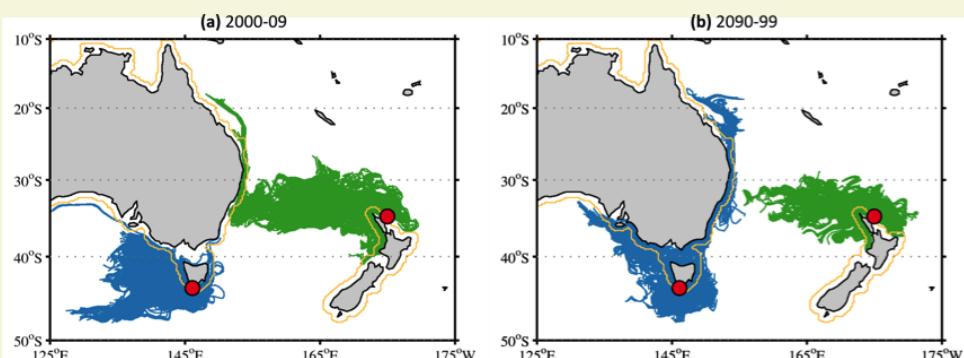
Scientific Use Case 12: Modelling and Observing Oxidant Chemistry over the Indian Sub Continent

Contact: Dr Luke Surl and Prof Paul Palmer (NCEO, University of Edinburgh).

The impact of climate change on coastal zones and local economies

Climate change affects our ecosystems in many ways. One particular focus is on the marine coastal zones that are immensely biodiverse and provide a habitat for a large proportion of global marine life. From a socio-economical perspective, these regions are particularly important for the goods and services they provide (estimated to be worth USD 14 trillion). As well as surface temperature rise, another important climate change impact is changing ocean circulation which threatens to modify the ocean-flow pathways used by many marine species. In the case of Tasmania, the invasion of tropical species like urchins could lead to depletion of the kelp forest and outcompetition of the local species like abalone, oyster and lobster that fuel the local fisheries economy.

High-resolution global ocean models, run under climate scenarios for over a century, are needed to reproduce these pathways and understand how they could change.



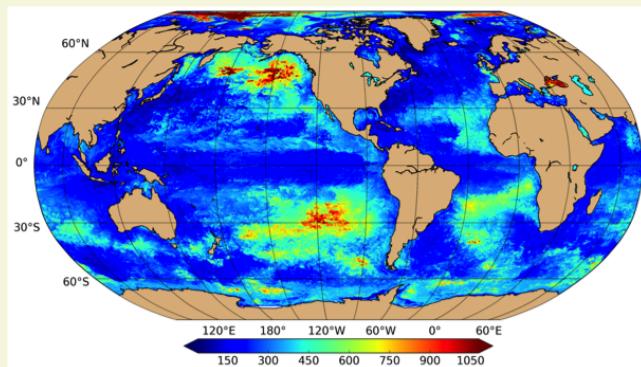
The computational capabilities and access to stored results offered by JASMIN have enabled calculation of future circulation pathway changes and their economic consequences for the entire global shelf.

Scientific Use Case 13: Understanding the connection between climate change and marine coastal zones

Contact: Simon van Gennip (NOC Southampton)

The ‘Weather’ of the Ocean

Mesoscale ocean features such as eddies are present at the surface of the ocean and are sometimes described as “the weather” of the ocean. They are typically turbulent systems, with less than 100 km in spatial scale but with timescales on the order of a month. The ocean changes more slowly than the atmosphere, and it stores a lot of heat - meaning the ocean can provide an important source of forecast skill. However, current ocean model resolutions are too coarse to resolve these small features - thus we don’t know how they may affect forecasts.



To simulate these features, high model resolution is necessary: 8 km at equator and 4 km at 60N, and so national HPC facilities (ARCHER) are required, producing large output volumes. JASMIN is unique in that it is possible to transfer, store and analyse tens of TB of data with similar workflow to that necessary for a few GB on a local machine - no radical rethinking of how to process or analyse the data is necessary. This research would not have been possible without JASMIN, purely due to data volume.

Ultimately this work will contribute to improvements in forecasts, via work with the UK Met Office to develop ocean models for their forecasting systems through the Joint Ocean Modelling Programme.

Scientific Use Case 14: The influence of small ocean scales on weather forecasting

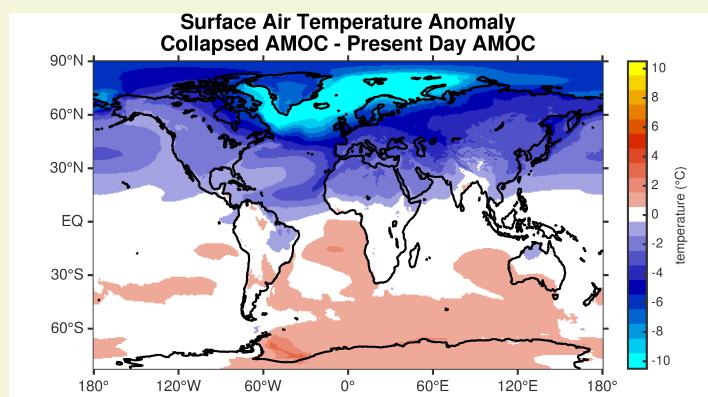
Contact: Dr Adam Blaker (NOC, Southampton)

Abrupt Climate Change — AMOC collapse

The Atlantic Meridional Overturning Circulation (AMOC) brings warm, saline waters from the Atlantic Equatorial region northward, where it cools, becomes denser and sinks. If the AMOC were to collapse regions surrounding the North Atlantic could experience a cooling of up to 10C. The latest IPCC report showed that the AMOC is very likely to weaken over the 21st century.

Recently work to understand the physics, as well as investigate the impacts of a collapsed AMOC in the output of a current generation coupled climate model, has been carried out on JASMIN.

The coupled model used (HadGEM3) has an ocean resolution of 0.25 degrees, which is quite a bit higher than a typical ocean model used in the most recent IPCC report (≈ 1 degree), leading to large volumes of data.



Working with JASMIN has allowed the analysis of large volumes of both ocean and atmosphere data, and supported ongoing activity by people in various places. Such sharing and analysing of this large volume of data has been made a lot easier through the use of JASMIN!

Scientific Use Case 15: Sharing and exploiting coupled ocean-atmosphere simulations

Contact: Dr Jennifer Mecking (University of Southampton) or see Mecking et al. (2016) DOI:10.1007/s00382-016-2975-0

Evaluating observing strategies with simulated measurements

Ocean heat content is a key metric for understanding change in the climate system and for predicting future climate change effects e.g. global and regional sea level rise. Accurate estimates of heat content change in the ocean, below 2000m depth, are vital to constrain full depth heat content. These extreme depths are very difficult to measure; at present the only method for measuring the deep ocean is via deployment of a CTD (an instrument measuring temperature, salinity and density) from a ship.



High resolution model runs are used to evaluate if these CTD measurements alone are sufficient to estimate heat content change accurately. However, high resolution ocean and climate model runs take up a huge amount of space, and each university/research centre has limited availability that is likely under pressure at this time. JASMIN provides a space for UK researchers to store and access the same model runs, reducing duplication of data in different research centres and simplifying across-institute collaboration.

JASMIN provides users with easy access to many more model runs than are available at their own institutes, enabling them to focus more time on science questions.

Scientific Use Case 16: Measuring changes in deep ocean heat content

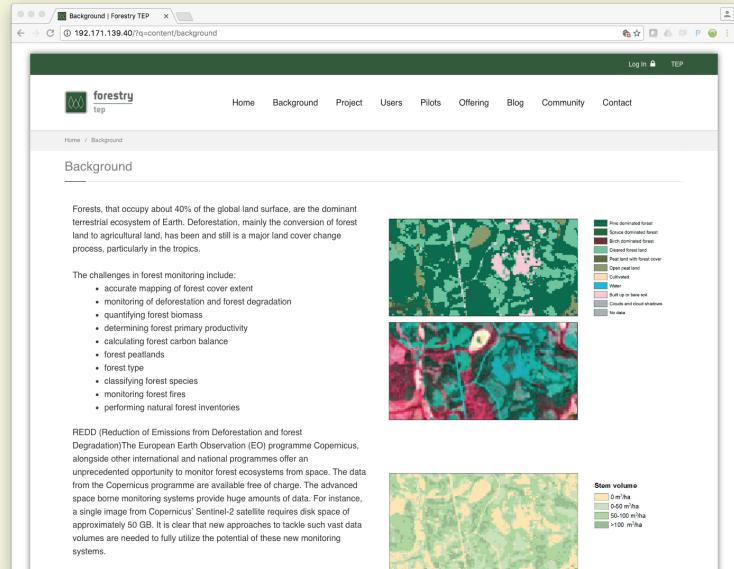
Contact Freya Garry (NOC, Southampton)

Thematic Exploitation Platforms for ESA



Forestry TEP

- A one-stop shop for forestry remote sensing services for the academic and commercial sectors.
- Offers access to pre-processed satellite and ancillary data, computing power, software access and hosting.
- The Forestry TEP is being developed in the JASMIN cloud (where the Polar TEP is also likely to be hosted).



Development by VTT Technical Research Centre & Arbonaut (FIN), CGI IT & STFC (UK), and Spacebel (BEL).

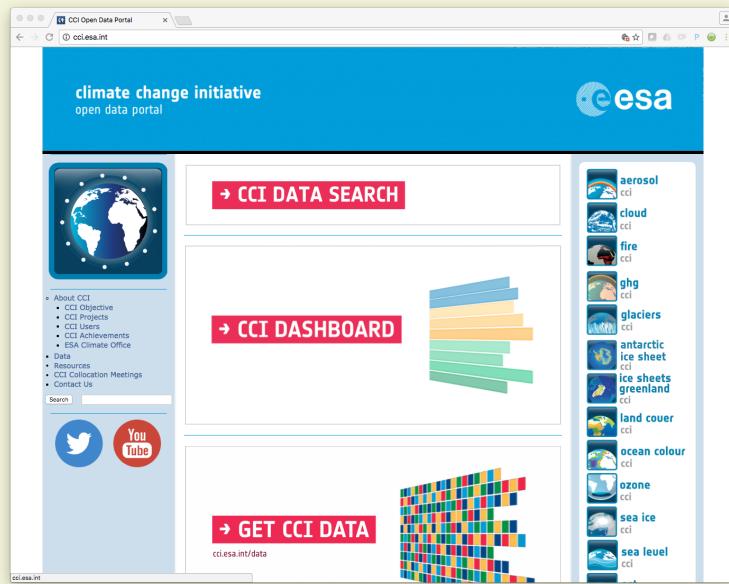
Scientific Use Case 17: Forestry Thematic Platform

Contact: Mr Phil Kershaw (NCEO and CEDA)

CCI Open Data Portal for ESA

The Climate Change Initiative

- Exploiting Europe's EO space assets to generate robust long-term global records of essential climate variables such as greenhouse-gas concentrations, sea-ice extent and thickness, and sea-surface temperature and salinity.
- The CCI Open Data Portal is hosted on the JASMIN cloud and exploits a near complete copy of the CCI datasets held in the CEDA archive.



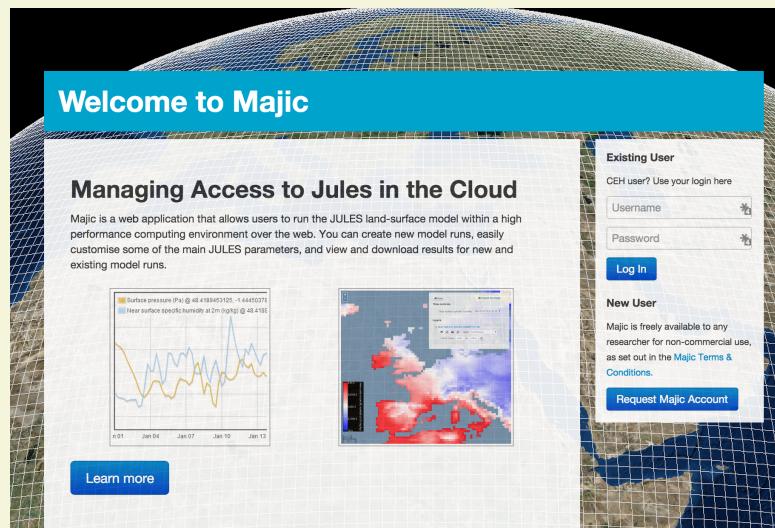
Scientific Use Case 18: Deploying a Data Portal for the Climate Change Initiative

Contact: Dr Victoria Bennett (NCEO and CEDA)

Running JULES on Lotus from a website in the Managed Cloud



- JULES is a community land surface model incorporating processes such as surface energy balance, the hydrological cycle, carbon cycle, dynamic vegetation etc.
- MAJIC provides a web portal running in the un-managed cloud which allows users to configure JULES to run on the JASMIN/LOTUS batch cluster and return results.



Scientific Use Case 19: MAJIC: Managing Access to JULES in the cloud

Contact: Matt Fry, Centre for Ecology and Hydrology

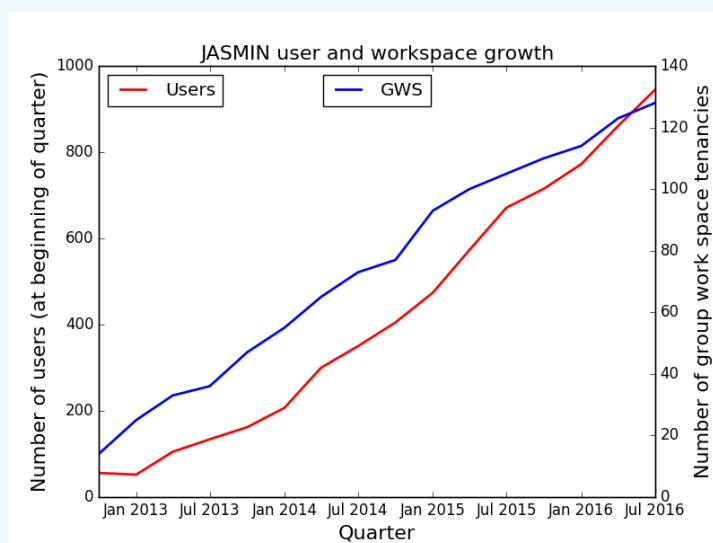
Growth in core JASMIN usage

JASMIN usage falls into a range of categories: thousands use JASMIN indirectly via services hosted on JASMIN (e.g. CEDA itself), and many use JASMIN via the un-managed infrastructure cloud.

The figure presented here shows only the usage of the main platform compute and storage by presenting time-series of both the number of unique users who have login access and the number of tenancies supported in the fast storage (that is, individual projects and communities). More details of the users of those tenancies is presented in box 7.

The initial usage came from the very large projects such as UPSCALE (science case 1), which could not have been delivered without JASMIN, followed rapidly by users with large earth observation datasets who could see the direct benefit of accelerating their workflows, and then by larger swathes of the community who could also see the benefit of direct exploitation of the existing archives.

Box 5: The growth of login users and group workspace tenancies.



5 Metrics of JASMIN usage

The first components of the JASMIN system were turned on in March 2012. The earliest users included the CEDA archive team, who needed both to rescue their data from ageing inappropriate hardware, and simultaneously cope with the advent of the CMIP5 archive, and those in the science community who had to cope with either massive volumes of data or excessive data rates (or both, such as the participants in the UPSCALE project — Use Case 1). User growth has been linear since, and well correlated with the addition of new group work space tenancies (box 5), suggesting a steady increase in communities whose scientific workflow has exceeded what can be provided locally and/or more sharing of data and workflow.

Some understanding of how these users exploit JASMIN, and therefore of future requirements for JASMIN, can be gained by considering the observed growth and usage along three further axes: storage use, core compute use, and use of

the unmanaged “infrastructure compute”. Each of these are considered below.

5.1 Storage usage

Storage use on JASMIN comes from two primary directions: use of the archive, and use of the group work spaces. It can be seen (box 6) that there is near linear growth in both over the lifetime of JASMIN, although group workspace growth is currently exceeding archive growth.

Archive growth in the CEDA data centre (primarily supporting the National Centres of Atmospheric Science and Earth Observation, NCAS and NCEO) come from two primary directions:

1. Data produced by NERC programmes, which are provided to CEDA in accordance with the NERC data policy⁴ which requires NERC funded scientists to make their data openly available and in return commits NERC to manage such data for the long-

⁴<http://www.nerc.ac.uk/research/sites/data/policy/>

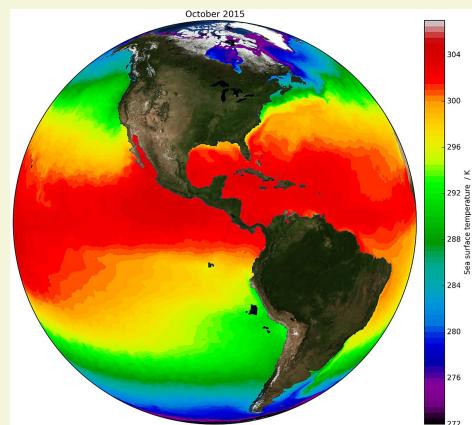
Sea Surface Temperature from Space

Sea surface temperature (SST) provides fundamental information on the global climate system; it is an essential parameter in weather and climate prediction. Earth observation (EO) data collected via satellites provides global observations of SST, this can then be used in the running, validation, and interpretation of high resolution atmospheric and ocean models.

The ESA Climate Change Initiative (CCI) programme utilises $\approx 180\text{TB}$ of raw EO data currently archived at CEDA, producing $\approx 50\text{ TB}$ of high level products. JASMIN allows scientists to generate 30+ years of datasets in just a few days, rather than months or years. It is not practical for JASMIN users to transfer this volume of data elsewhere — there is nowhere else to store it and nowhere else that can process it (with interim products, the SST-CCI alone needs 260TB of group workspace for product development).

The SST section of the Copernicus Climate Change Service (C3S SST) will also make use of JASMIN to obtain near-real time satellite data and generate short delay products useable by ECMWF and others — ensuring no duplication of effort.

New and future satellite missions will involve much larger volumes of data; 10-100s TB each year. There is a clear requirement from the community for both CEDA and JASMIN — archival and processing — with the emphasis being on these resources located together!



Scientific Use Case 20: Using Global Sea Surface Temperature Measurements

Prof. Chris Merchant and Dr Owen Embury (NCEO, University of Reading)

term.

2. Third party data needed within NERC science programmes and where a central cache significantly improves the productivity of NERC science; such data is added to the CEDA archive and also managed long-term (often as part of global preservation activities). Some key drivers to archive growth from this category are described in box 11.

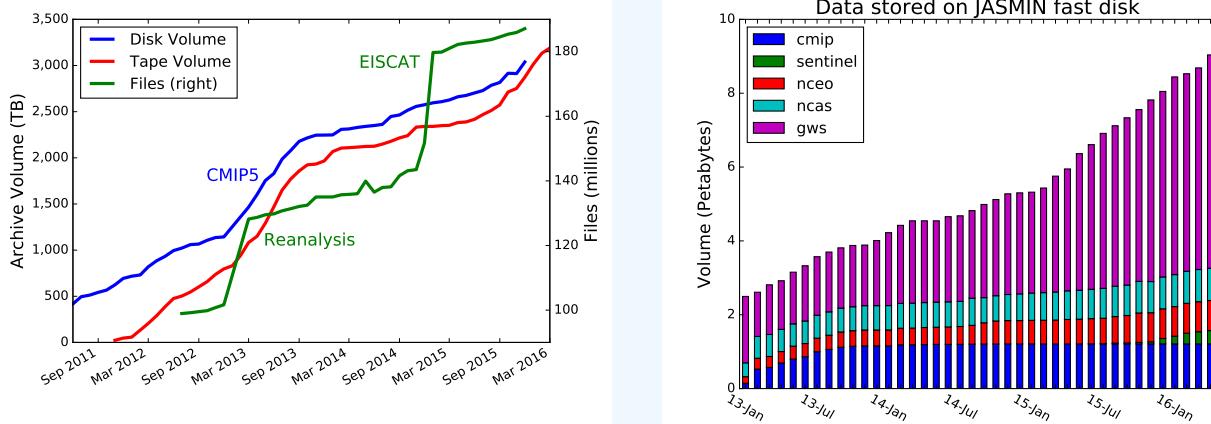
As already seen, the number of group work spaces is growing near linearly, but not all group workspaces are the same. Each group work space is effectively an individual JASMIN tenancy, with a manager who controls who can have access to their group workspace. Some are large, with few users, and some small, with many users: fifty now have allocations of 10 TB or more and more than 5 users (box 7). One of the reasons why group workspace growth exceeds archive growth can be understood by considering the construction of sea surface temperature datasets (Use Case 20) — a 280 TB group workspace is necessary to exploit o(180) TB of archive to produce a few TB of final

products.

As of the end of 2016, nearly all the fast disk is allocated to tenancies, meaning that there is no route to growth within any tenancies, or space to add new tenancies without shrinking existing tenancies. One of the issues that JASMIN faces is that the residence time of data on storage is measured in years to support the long analysis times which follow data production whether it be from observations or simulations (e.g. box 12).

There is considerable capacity available in the “elastic tape” system. Group work space managers are initially allocated the same storage capacity on tape as they are given on disk, and it is expected that they can request more (hence “elastic”, as in the capacity is extensible). However, while the system is functional, and being used the system supporting this provision is not optimal, with issues around metadata management and performance that will need to get resolved in the next phase of JASMIN.

Observed Data Growth on JASMIN



The **left panel** shows the growth in the data stored in the CEDA archive: The blue and red curves show data volumes (of disk and tape respectively), and the green the number of files. The step change in volume associated with ingesting CMIP5 is clear, as are the step changes in file numbers associated with the ingestion of the reanalysis and EISCAT datasets.

The **right panel** shows the growth in the volume of all data stored on JASMIN: The NCEO and NCAS bars show the data stored in the atmospheric and earth observation components of the CEDA archive, neglecting CMIP and Sentinel data (shown separately). It is clear that the major source of data growth is currently coming from users exploiting the group work spaces.

Box 6: Growth in data stored on JASMIN fast disk.

5.2 Compute usage

Compute usage falls in four classes: batch cluster, generic and hosted platform compute and infrastructure compute.

Platform compute and batch cluster usage are shown in box 8. In this analysis we have not distinguished between the virtualised platform compute and the physical platform compute (large memory and data transfer nodes). The observed growth in usage arises both from the growth in JASMIN user numbers and user communities (box 5) and from existing users working out how to make the best use of the JASMIN environment.

In the JASMIN environment the expensive part of the system is the storage, so ideally the compute environment is responsive enough that users do not need to keep their data online in group work spaces for extensive periods. Over the life time of the systems, the longer the wait times on compute, the less efficiently we use the storage. The trend is towards thinking about the compute as free, at least in comparison to the storage cost, which means minimising wait time and not maximising utilisation – exactly the opposite ap-

proach to a traditional HPC system. To that end, the Lotus environment would ideally prioritise response time over utilisation, but by early 2016, Lotus utilisation had reached over 70% (box 8), which is not ideal (even a traditional HPC systems is considered “full” at 80%, higher loads mean the scheduler cannot operate satisfactorily). Although as yet difficult to quantify the optimal rate⁵, a utilisation rate of around 60% would seem to offer a good compromise between storage residence time, Lotus response time, and Lotus utilisation.

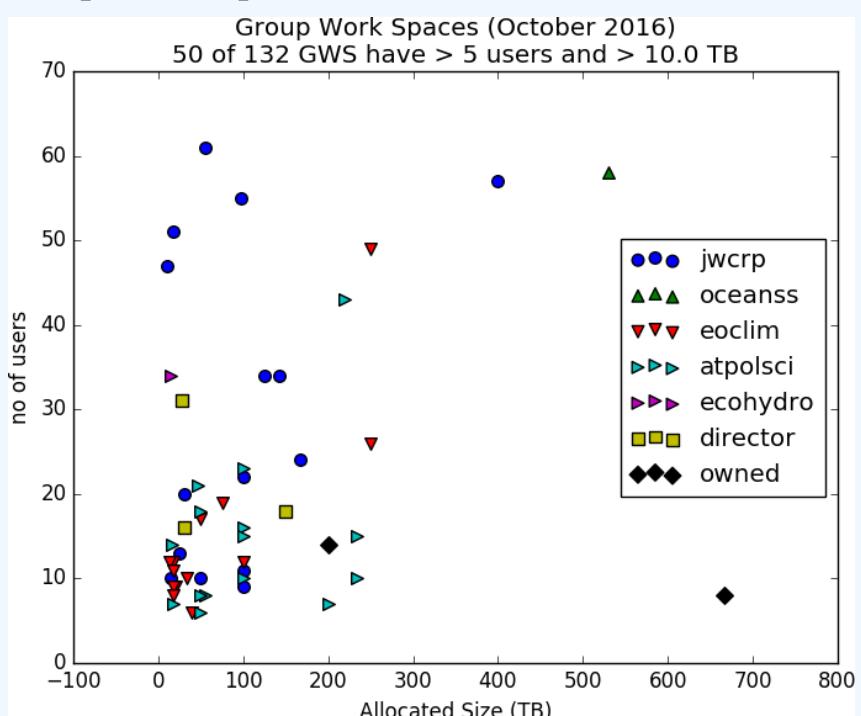
At the advent of JASMIN, cloud computing was still in its infancy, both in terms of uptake in academia (whether in the public cloud or a private cloud), and in terms of available tools for managing and administering a private cloud. Even in 2016, this is still the case, although the use of virtualisation is relatively common. Over the years

⁵The optimal balance for a system like JASMIN is not known, finding it out is a research issue, one that the JASMIN team expect to address in the near future in partnership with an appropriate Computer Science group.

The 50 Largest Group Workspace Tenancies on JASMIN

The largest consortia

atpolsci	Atmospheric and Polar Science
director	CEDA directors allocation (mainly supporting H2020)
ecohydro	Ecology and Hydrology
eoclim	Earth observation and climate services
jwcrp	Joint Weather and Climate Research Programme
oceansss	Oceans and Shelf Seas
owned	Resources owned by third parties within the JASMIN partnership



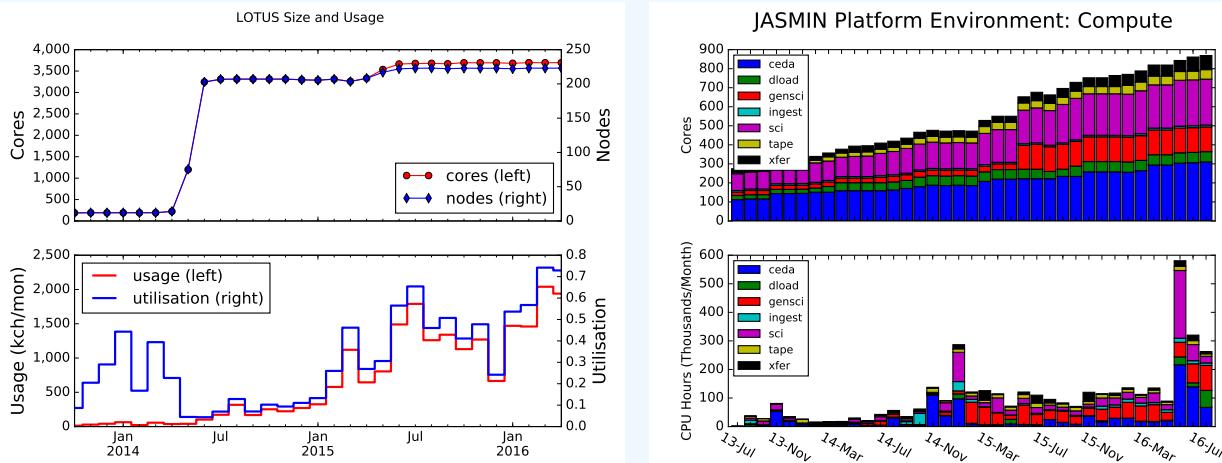
There are 132 group workspace tenancies, supporting 822 users (of the 1059 with JASMIN login access at the time). The largest group workspace is owned by the University College of London. Most of the rest of the GWS (less than 10 TB) have a handful of users, but there are also 7 which have less than 10TB and more than 10 users — even for relatively small data volumes, sharing and co-location is important.

Box 7: Group Workspace Allocations, October 2016

of JASMIN, significant work has gone into the development of a “cloud portal” to ease the user configuration of machines, but this has only recently been rolled out, and statistics of use are not yet available. Nonetheless, we know that despite relatively little uptake of the JASMIN “Infrastructure as a Service” delivery, we have nearly filled the available capacity — and it is clear there is latent demand hindered by technical issues which we would expect to address in the proposed upgrades (section 6.3). However, there have been some interesting use cases, some of which are in production. Six exemplars are presented in box 9 with three presented in more detail as science use cases (17, 18, and 19): these exemplars cover hosted environments under contract to ESA, websites which are managed in the JASMIN cloud, but which deliver services from the JASMIN platform compute and/or Lotus, and

a range of systems supporting customised environments such as providing Desktop-as-a-Service for the genomics community where JASMIN can provide larger-memory systems than are typically available. There is a prototype service offering the use of ipython-notebooks in containers orchestrated in the JASMIN infrastructure cloud. The experience of these early adopters has had significant influence on the JASMIN technical roadmap and planning.

Compute usage

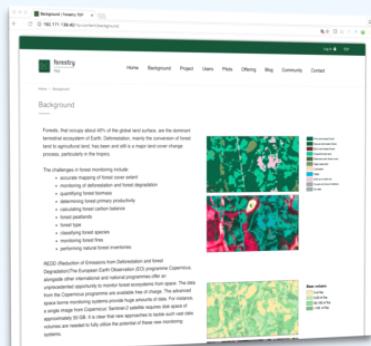


Left panels: The evolution and usage of the Lotus batch cluster: By mid-2016 Lotus usage had reached over two million core hours per month exceeding 70% utilisation — Lotus is targeting a lower rate of utilisation as discussed in the text.

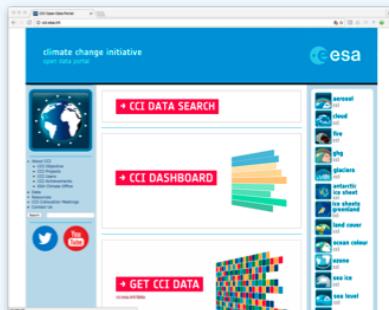
Right panels: Platform Compute: Top right, cores (mostly, but not all, virtualised) to support: CEDA general compute (ceda), CEDA download services (e.g. ftp etc: dload), general science machines (for user login and general compute: gensci), CEDA ingestion services (ingest), hosted platform compute (e.g. websites: sci), tape systems (tape), and efficient transfer in/out of the JASMIN environment (xfer). Bottom right, how much compute work these various categories were doing.

Box 8: The growth in usage and utilisation of the JASMIN core compute environment.

Virtual Environments deployed on JASMIN



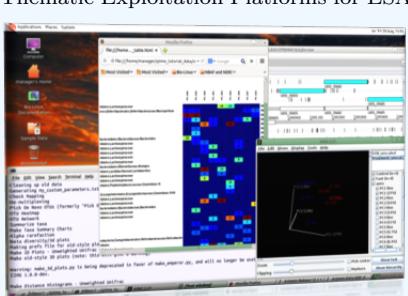
Thematic Exploitation Platforms for ESA



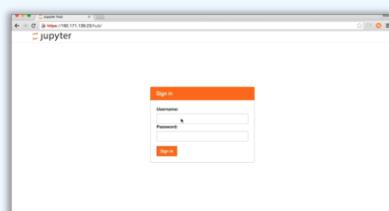
CCI Open Data Portal for ESA



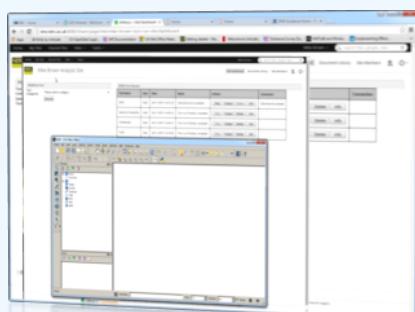
MAJIC interface to JULES model



EOS Cloud — Desktop-as-a-Service for Environmental Genomics



Hosted Ipython Notebooks



NERC Environmental Workbench

Box 9: Exemplar projects deployed in the JASMIN infrastructure zone — “the unmanaged” cloud.

6 JASMIN Architecture: Now and in the Future

In this section we provide a more detailed look at the JASMIN architecture, enough to provide context for the technical decisions which will need to be addressed in the future phases of JASMIN. We then revisit the major scientific drivers influencing JASMIN requirements, again in the context of the technical decisions which will be required. The section concludes with a summary of the architectural requirements for JASMIN evolution.

6.1 JASMIN in late 2016

A high level view of the JASMIN architecture is shown in box 10. The underlying hardware has been procured in three major phases (and a number of minor phases). JASMIN phase 1 was procured in 2011, and delivered in March 2012 with some minor upgrades in the following year. Phases 2 and 3 were procured in 2013 and 2014 and delivered in March 2014 and 2015 respectively. As configured in late 2015, JASMIN consists of:

1. The Lotus Batch Cluster (approximately 3,200 cores, mostly phase 2),
2. The JASMIN Hypervisor Cluster (approximately 800 cores, mostly phase 2), which are deployed to support
 - a) The Managed Infrastructure: a set of systems which deliver a PaaS hosted compute environment where users can both bring their own data, and directly exploit the archive data. This managed environment is used to provide the CEDA archive services.
 - b) The Infrastructure Cloud Service (IaaS): a cloud environment where NERC and STFC partners can bring, install, manage and utilise their own computing environments. SaaS to access data held in the CEDA archive data across fast local links.

Users in both environments can exploit SaaS deployed on other parts of the infrastructure to gain access to data held in the archive, and users in the infrastructure cloud can deploy their own services to third parties outside of

JASMIN.

3. Fast disk storage (\approx 15PB usable in two storage pools, one associated with phase 1 and one with phases 2 and 3), and
4. Bulk disk storage (\approx 1 PB, phase 2), and
5. Interfaces to, and capacity in, a massive tape store, consisting of
 - a) Tape media, capable of storing 30 PB of data held in Oracle StorageTek tape libraries, fronted by:
 - b) The “Storage-D” software and hardware cache interface used by CEDA for archiving data, and
 - c) The “Elastic-Tape” software and hardware cache interface used by JASMIN users for offline storage to supplement disk resources.

6. An internal fast network linked both with traditional firewalls to the wider world and via a “Science-DMZ” for high-performance data transfer.

Both the batch and hypervisor clusters include a range of hardware with high memory machines to support memory intensive computing and lower memory machines to support less memory intensive computing. It is possible to move hardware from the batch cluster to the hypervisor cluster, although this is a manual task. The balance of cores allocated between Lotus and the hypervisor cluster reflect usage over the last eighteen months.

It can be seen that key to the JASMIN fabric are the cloud interfaces, and currently there are three significant components for which significant investments have been made:

1. The basic virtualisation layer is provided by VMware for which commercial licenses have been obtained.
2. A bespoke portal for managing JASMIN cloud resources has been developed which is currently layered on top of VMware but which has been designed to support alternative underlying virtualisation technologies, and
3. The PaaS offering is based on the JASMIN Analysis Platform, a customised version of the Linux operating system.

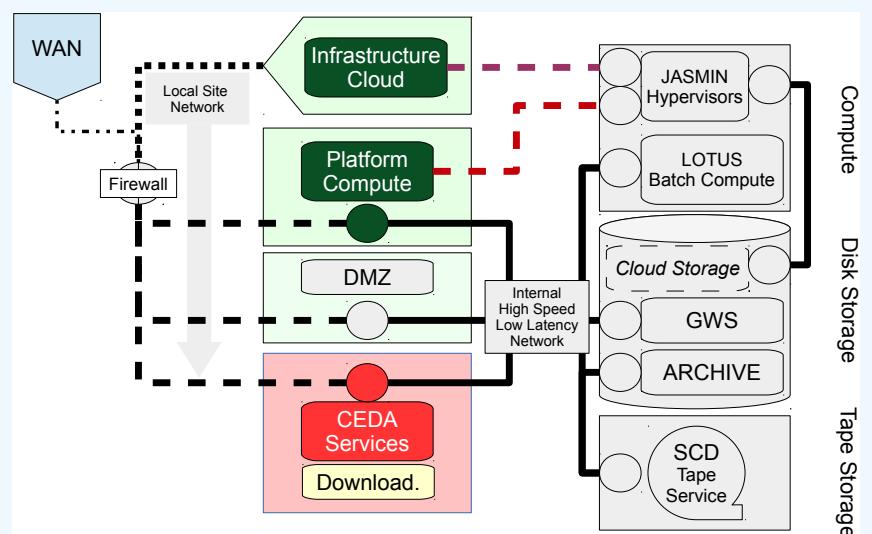
All these components, hardware and software, have defined life cycles. The timelines associated with hardware replacement are mainly controlled

High level view of JASMIN in 2016

A high level view of JASMIN services was presented in section 3.4. These services are delivered in a physical infrastructure which consists of compute, data storage, and tape storage.

Compute: Two classes of hypervisor, one for managed compute, and one for infrastructure cloud. The Lotus batch cluster, including physical machines to support dedicated services (e.g. the high performance network DMZ).

Storage: The Group workspaces and archive share parallel fast disk (on Panasas hardware), and a block file service via iSCSI (on Dell hardware).



Tape Storage: Tape access is provided via Oracle tape library fronted by the CASTOR tape software, itself fronted by two bespoke STFC tape software services (Elastic Tape and Storage-D).

All the JASMIN components share a high speed low latency internal network, but the infrastructure compute talks to the rest of the services via the front-door firewall for security reasons.

Box 10: Relationship between services and components in JASMIN as of 2016

by maintenance cycles, some components become too expensive to maintain and must be replaced after a few years. Timelines associated with software need to be responsive to supporting both the way users want to work, and the hardware available. Meeting these timelines is a crucial part of the future procurement plan.

6.2 Influences on future Architecture

The driving theme for JASMIN is supporting the joint analysis (computing) of data (storage). The system and data storage need to address three key timescales

1. decadal to centennial (to support the long term curation requirements of CEDA),
2. many years (to support large programmes like the high resolution climate programme, see box 12, and
3. short term, to support high performance data analysis.

Taking these three timescales in turn, we note that until 2016, CEDA had managed two decades of curating the entire archive on disk, with only

backup copies on tape. However, projections for future storage demand from climate and earth observation are likely to exceed both disk affordability, and the ability to fit the disk in the existing machine room. In the next two years we expect the data rate from just two projects (CMIP6, climate; and the Sentinels, earth observation, see box 11) to require more storage than is feasible on disk alone. These are not surprises, being consistent with the underlying trends discussed in section 3.2, and the advent of sensor webs will only compound the scale of the problem!

It has already been noted that despite these large numbers, over a long period, the group workspace growth supporting the data analysis community is growing faster than the archive, and this is driven by the nature of the workflow, the growth in users, and the growth in the number of tenancies. It is also important to consider that for real analysis workflows, the data needs to be available, and the system fast enough, for both exploratory data analysis, and "final product production". While the latter may be amenable to timescales of up to weeks, the former really needs

Major Drivers of Future Data Growth on JASMIN



The Coupled Model Intercomparison Projects are integral within WCRP. CMIP6 will begin in 2017, with PB of data from dozens of modelling groups worldwide being shared — much via JASMIN, including the European contribution to hiresMIP which alone is expected to exceed 2 PB.

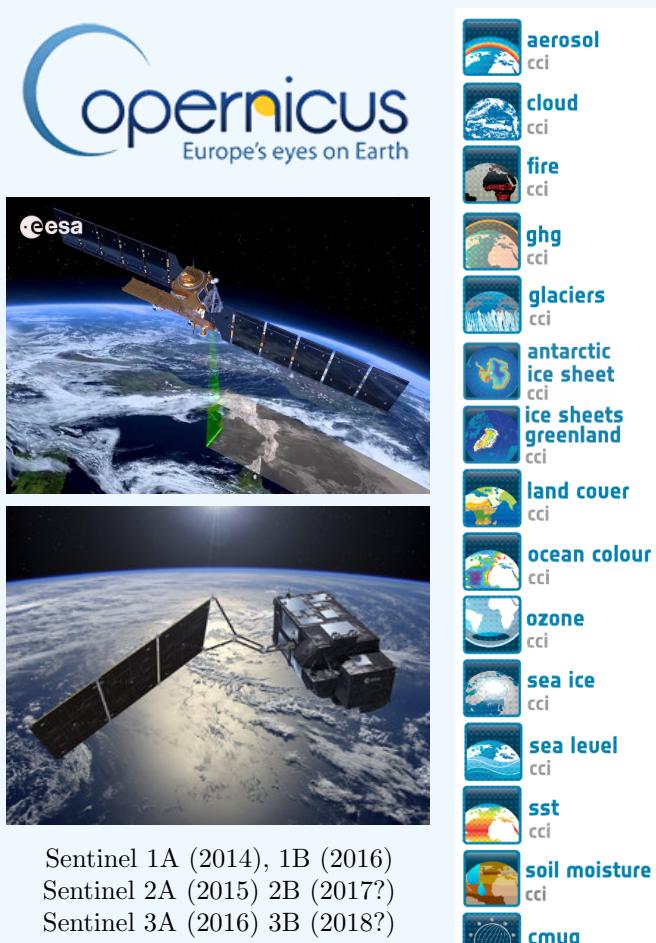
A key role for JASMIN is enabling the intercomparison of models with observational data from ground based, airborne, and satellite remote sensing. JASMIN is unique in having large co-located simulation and observation archives.

Box 11: Data from earth observation and simulation campaigns will dominate archive growth in the next few years. Although CMIP6 data rates and replicated data volumes are not yet known, CEDA is planning on up to 10 PB of CMIP6 data on spinning disk, as well as all the ESA Climate Change Initiative data, and a multi-petabyte cache of Sentinel data (with the rest on tape). Current experience (Box 6) suggests archive growth will be dwarfed by group work space growth.

to be completed within hours so that codes can be developed, executed, and interpreted on human timescales. The issue for modern science is that those “exploratory codes” may be handling datasets which are tens of terabytes in scale (for example, to look at high frequency waves in one year of a relatively modern high resolution atmospheric model requires processing a 9.8TB array, just to look at one four-dimensional field). Such data sizes put physical constraints on the size of, and bandwidth needed between, the various

components of the system.

To some extent of course, scientific demand of computing is infinite. Thus far JASMIN has followed the philosophy that either costs should be covered, or if some part of NERC has paid for some part of the science programme (either directly or the programme involves NERC funded staff), then JASMIN should attempt to support that programme given that the project will have been reviewed elsewhere, and double-jeopardy in review is to be avoided. However, it is clear that



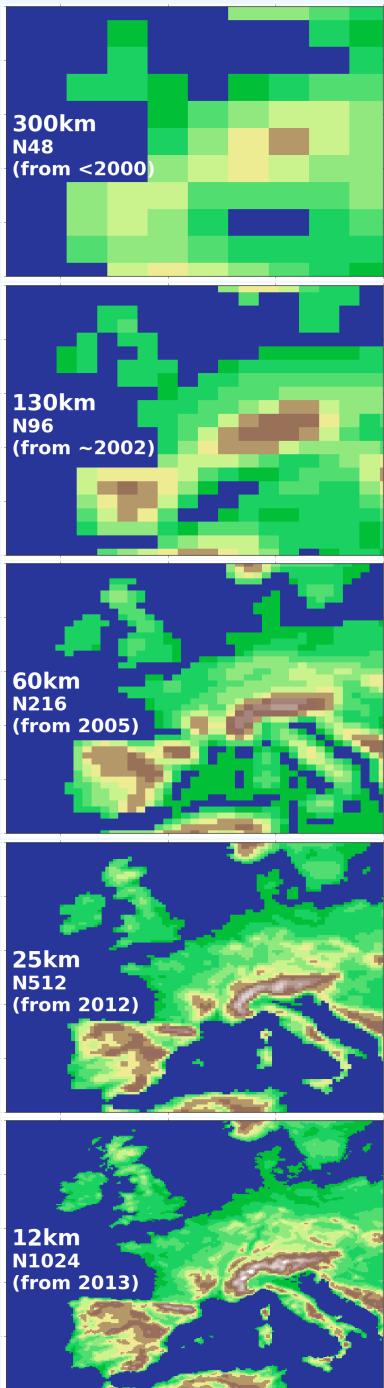
Sentinel 1A (2014), 1B (2016)

Sentinel 2A (2015) 2B (2017?)

Sentinel 3A (2016) 3B (2018?)

Data rate: 0(6) PB/year

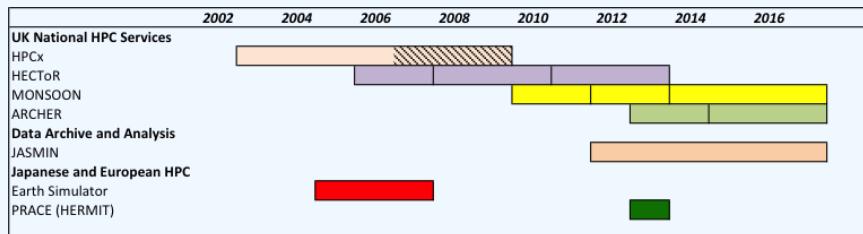
High Resolution Global Climate Modelling — in it for the long game!



Five generations of the Unified Model showing the increase in resolution (in a portion of the global domain) over two decades of development. The lower resolution models are still being developed for use in complex models with more processes and/or bigger ensembles (more runs), while the higher resolution models push the capability of the largest supercomputers, even to do development and/or single runs.

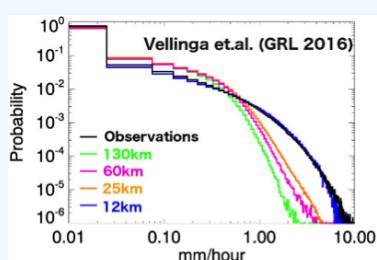
Many simulation programmes are best thought of as long-term campaigns taking decades from inception to impactful results: much like the design, building and deploying of earth observation satellites.

One such programme is the UK High Resolution Global Climate Modelling programme (a partnership between the NERC National Centre for Atmospheric Science, NCAS, and the Met Office). Over the last fifteen years it has exploited all the national HPC platforms (HPCx, HECToR and ARCHER), JASMIN, Met Office computing (both the shared NERC/Met Office MONSooN platform, and the main Met Office computer), as well as the Japanese Earth Simulator and a European PRACE supercomputer (HERMIT).

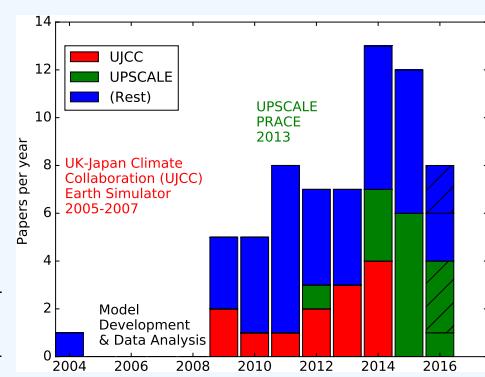


HPC platforms contributing to the High Resolution Climate Programme since 2002

Each of these systems has contributed to different aspects of the programme; from code development and optimisation (the highest resolution models are initially developed on leading HPC and later in life can be maintained on smaller systems), to routine science runs and then, at intervals, massive campaigns such as UJCC which exploited the then largest computer in the world (the Japanese Earth Simulator) and later, UPSCALE. UPSCALE used the equivalent of a quarter of the entire UK academic HPC capability for an entire year, and is the largest computing project ever supported by the European Partnership for Advanced Research Computing (PRACE) — but it could not have been completed without the years of earlier work on national HPC platforms or the use of JASMIN for the data storage and analysis!



One the many benefits of increased global resolution can be seen in the convergence of simulated rainfall over the Sahel to observations when resolution reaches 12km.



The Sahel paper was published in 2016 using data produced during, and since, the 2013 UPSCALE campaign. These lead times (years between simulation and exploitation) are not untypical as can be seen by the histogram of publication dates following model campaigns.

Box 12: High Resolution Climate Modelling: very long time-scales utilising multiple HPC platforms. Prior to JASMIN, data migration and management over the necessary timescales has been problematic!

at some point further demand management may be necessary, and this will be regularly considered as part of both the technical and governance environments.

As well as the scientific trends, section 3.2 also identified the re-emergence of heterogeneous computing and the advent of the cloud as technical trends to consider. These two trends push in opposing directions, and in the context of JASMIN we can see them as pushing simultaneously towards three orthogonal directions: a “hyperconverged” architecture (with storage and compute in the same nodes), or dedicated data intensive hardware (with complex layers of memory, burst-buffers, or other I/O accelerators such as the Cray DataWarpTM), or handing the problem over to the public cloud. These trends are also operating in the presence of a political imperative to join up activities across the UK research council e-infrastructure. Putting these all together, it can be seen that JASMIN evolution is complicated by three key factors, leading to three possible futures. The key factors are:

1. Evolving opportunities for large scale infrastructure collaboration;
2. The rapid development of new commercial software, hardware and service technologies, and
3. The growth in demand from the science base.

These lead to the three possible futures for JASMIN evolution:

1. Continue to evolve the existing JASMIN infrastructure (with internal heterogeneity),
2. Integrate JASMIN into a shared (cross disciplinary RCUK) national facility, or
3. Migrate JASMIN functionality into a public cloud.

In the next few years it is not considered technically feasible to integrate JASMIN into a shared national facility, primarily because there are quite different requirements of input/output performance required by the JASMIN community than the other parts of the community. However, sharing a machine room and network fabric with the other tenants of the STFC SCD machine room (in particular the UK involvement in particle physics science including the Large Hadron Collider), continues to be desirable.

With current and foreseeable pricing for both networks and data storage, and with the current

configuration of public cloud storage offerings⁶ (in particular the lack of affordable POSIX disk interfaces), it is considered that it is neither technically nor financially feasible to consider using a public cloud in the next few years.

Accordingly, the remainder of this document considers a plan which revolves around delivering an evolution of the existing JASMIN infrastructure over the next five years. However, both technology and cloud offerings will change over time, so the question as to how to deliver “JASMIN-like” functionality for the environmental sciences should be regularly revisited, and the plan to be presented will include in the final year another look at this issue — in time to start to migrate the JASMIN infrastructure towards another possible future if such is required.

6.3 Key technical requirements

The most important requirement is to replace existing storage before it leaves maintenance! The replacement of storage will occur in two important contexts, one technical, and one arising from the scientific demand:

1. The maturing of object-based disk storage hardware as an alternative to parallel file system disk for CEDA archive data, and
2. The likelihood that the storage demand from the science is likely to exceed the disk storage that is available under any feasible financial support for a disk-based storage scenario.

With these factors in mind, a key part of the future is obviously to try and buy more and cheaper disk, but it must also include a higher profile for the use of tape in both archive and scientific workflows. However, if the tape system introduces too much latency much of the scientific advantages of JASMIN would be neutralised, hence another key future requirement will include the delivery of a new “smart caching” system to first allow the CEDA archive, and then the other PaaS and IaaS users, access to high-throughput high-volume disk

⁶There are also issues around permanence in the public cloud to consider for the curated long-term archive — what happens to data if RCUK is too slow to pay its bills, or can’t pay for a year or two because of some internal issues? While these latter issues may not be insurmountable, it would take significant time and money to work through the legal constraints necessary for appropriate contracts.

cache backed by tape. This will involve the migration from a more traditional “backup” role for tape to a more sophisticated system more akin to the MASS and MARS systems used by the Met Office and ECMWF, although the JASMIN situation will be complicated by the wider variety of use cases and greater heterogeneity of data.

For the disk replacement per-se, there are effectively three choices: continue on the existing strategy of buying high-performance parallel file-systems, or move to a high performance object store, or to a combination of both. Moving to any form of object store has major implications for how users work with the data, and cannot be undertaken without a comprehensive understanding of the possible software solutions. Nonetheless, there are compelling reasons to consider at least some of the system exploiting object stores, such as the ability to more dynamically add storage as it is needed, as well as lower costs per unit stored, and, potentially higher density storage (less physical space used).

JASMIN compute was relatively lightly stressed during phase 1, but the direction of travel was clear, hence the significant incremental compute upgrades in JASMIN thus far. It can be expected that as the storage and user base grows, more compute will be necessary (and older compute will need to be retired and replaced). The difficult questions then will be “how much is needed?” and “in what configuration”?

In terms of how much compute, the aim should be to balance the Lotus utilisation with the storage, adding Lotus compute capacity as necessary to keep utilisation below about 60%. If that cannot be achieved within budget, then further demand management could be applied via the queuing system.

In terms of hardware compute configuration, it is unlikely that CPU heterogeneity will be required in the near future, but that option will be kept open. JASMIN already supports some compute heterogeneity in that there are a range of memory configurations in both the platform compute and batch cluster environments (from 128GB to 2TB in the physical machines, with total flexibility inside the hypervisor capability up to 480 GB in the virtual environment). The most likely avenue of further heterogeneity will be added local storage on some nodes (possibly

including extended tiering of memory) in order to support algorithms exploiting such locality of storage. To that end, a small part of a future upgrade to the compute should be configured to provide an experimental zone for such work, possibly in the infrastructure cloud.

One major requirement of future upgrades will be to support more flexibility both in how the compute is configured between the batch cluster, the platform compute zone, and the infrastructure compute cloud; and within the infrastructure cloud (to allow software defined clusters etc). This will have ramifications for the physical network environment, the virtualisation fabric used, and the software environments (including the capability of the cloud portal). Two key technical issues which have been hindering cloud take-up will also need to be addressed in software procurement: the lack of secure access to the archive from the cloud infrastructure and the lack of an appropriate object store interface.

The major existing components of system software to be considered include the basic virtualisation fabric, the standard environment for platform as a service, and the software infrastructure needed to manage JASMIN itself. In JASMIN thus far the commercial VMware environment has been used for the primary virtualisation system mainly because in 2011 when JASMIN was being commissioned it was the only available option, despite considerable expense. However, in 2016, VMware is no longer the only alternative for the cloud infrastructure component, and is currently limited in key functionality, and so it will probably be necessary to migrate a considerable amount of the work to a new basic virtualisation environment.

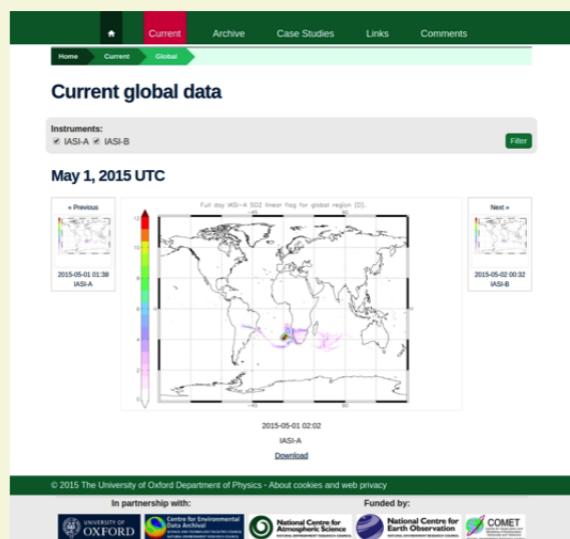
Whatever the basic functionality providing JASMIN virtualisation, cloud tenants need a portal for configuring virtual machines within their tenancy. Unfortunately experience has shown that the VMware portal products are not suitable for deployment without considerable support for the users, and of course they are bespoke — which means that any experience built up in the community would be lost with the migration to another environment. To that end, ongoing investment will be needed in the existing portal for configuring machines inside a tenancy. The software is constructed so that details of the un-

Hosted processing of near real time data for volcano alerts

Near real time observations of volcanic plumes of ash and SO₂ are becoming increasingly important, especially for air travel. Timely data can be used to alert volcanic ash advisory centres to a new eruption or volcanic plume, so as to warn aircraft, and also to monitor the progression of the plume.

A volcanic SO₂ monitoring website has been launched on JASMIN which displays near real time data from both IASI satellite instruments within 3 hours of measurement. A number of larger volcanic eruptions have been observed on the website including; Calbuco, Chile (April 2015), Wolf Island, Galápagos Islands (May 2015, as shown here), and Popocatpetl, Mexico (January 2016).

JASMIN has provided access to the IASI dataset in near real time and the concomitant computer resources to analyse the data within the required time. This unique relationship between data archive and data processing facilities has been, and will be, invaluable for observing and understanding the plumes from eruptions.



Scientific Use Case 21: Near real-time volcanic plumes

Contact: Dr Elisa Carboni (NCEO and University of Oxford) or visit <http://www.nrt-atmos.cems.rl.ac.uk>

derlying virtualisation can be hidden from the users.

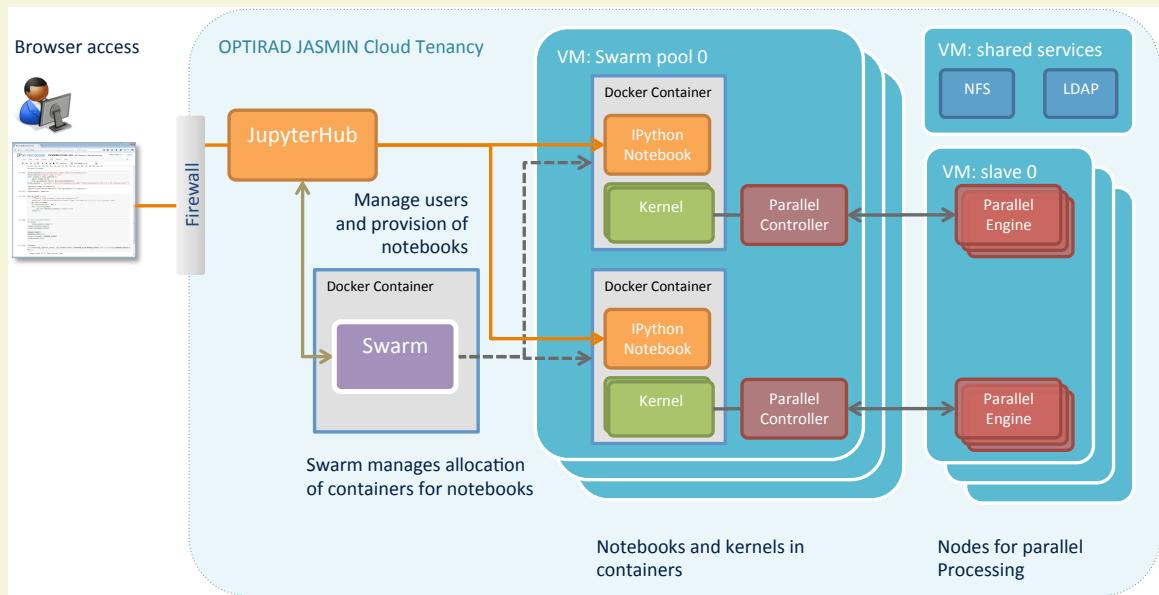
One of the earliest aspirations of the JASMIN cloud infrastructure was to support “cloud bursting” that is for users to be able to offload virtual tasks onto the public cloud to provide compute capacity beyond that available within JASMIN. Early work to provide that capacity foundered on issues around the lack of standards across public cloud providers, and sensible mechanisms for understanding cost implications of such workloads. However, the aspiration still remains — ideally workflows could be developed which “reduce data on JASMIN” before further computational analysis in the public cloud! This functionality will need to be on the roadmap within the portal development described above, allowing users to construct workflows which do data-reduction on JASMIN, followed by scalable compute in the public cloud.

Another ongoing requirement is to ensure that the JASMIN software infrastructure, particularly in the platform environment, is fit for purpose. JASMIN upgrade plans will need to continue to include support for variants of the “JASMIN Anal-

ysis Platform” software environment (whether as now, via packages and conda, or containers or some other mechanism to be agreed as part of ongoing technical evaluation). Such support may need to include key parallel data analysis packages where such support is not available via other methods. There is considerable community interest in helping set requirements for, and collaborating on the development of, these tools (especially from the Met Office).

Dynamic allocation of machines in a workflow is one of the main use-cases for which cloud computing currently out-performs traditional batch computing environments (where the “pilot job” method of resource orchestration is known to be relatively inefficient where the load is varying). An example of such a workflow is shown in Use Case 22, which also demonstrates how containers - lightweight virtual hosts which can run inside other systems - can be used to parcel up specific software environments which can be replicated as needed within workflows. “Containerisation” as this has process has become known, is likely to become integral to most environmental analysis workflows, but the challenge includes dynamically

Next Generation support for containers and parallelisation



The ipython notebook environment allows users to share combined notebooks containing shared code, notes and visualisations. These notebooks can themselves execute complicated parallel workflows. The architecture shown here allows a tenancy in the infrastructure cloud to manage its own users (in the LDAP database), and support them instantiating ipython-notebooks, each in their own containers which are distributed and orchestrated across a cluster of virtual machines. Where the notebooks deploy parallel code, the parallelisation is across another set of virtual machines hosting the parallel processing engine. Data access is currently via NFS to shared storage in the tenancy, or OPENDAP to the CEDA archive, and the amount of compute provided is under the control of the tenancy (within limits of the tenancy allocation on JASMIN).

Scientific Use Case 22: The OPTIRAD “software-as-a-service” ipython-notebook architecture.

Contact: Jos Gmez-Dans (NCEO, UCL) or Phil Kershaw (NCEO, CEDA)

allocating the machines on which the containers are deployed (in this case, the notebook and parallel pools) as demand fluctuates. This means systems will need to be developed based around technologies to manage containers, to manage hosts for containers and to manage service scaling. While all these systems exist, integrating them into the JASMIN environment so that they are easily available to users may be non-trivial, especially if it is necessary to transition some or all of the virtualisation away from VMware on the same timescale.

7 JASMIN Investment Plan

JASMIN requires £17M capital over the next five years — phased as three £5M phases intervened with two £1M phases. This phasing is due to the unique nature of JASMIN. To provide the

best possible data-intensive environment, it is necessary to deliver a custom integrated compute system (that is, rather than buy a supercomputer, JASMIN requires the purchase of a range of components which are integrated together by the STFC team). It is also necessary to avoid buying too much storage before it is necessary. The large/small phasing of investment represents a tick/tock development profile: with the gradual integration of extra storage as it is required (tick), followed by customised reconfiguration and the addition of compute and tape (tock) as required by the workload. (A side effect of the JASMIN initiative is that it provides a world-class example of innovative system engineering!)

The first tick phase of investment would be aimed at a) replacing obsolete hardware from the initial JASMIN investment and b) providing sufficient capacity for primary Sentinel and

Year	2016/17	2017/18	2018/19	2019/20	2020/21
Capital (£K)	5,000	1,000	5,000	1,000	5,000
Recurrent (NC only) (£K)	652	662	672	682	692

Table 1: Overall finance plan assuming phases are one financial year apart, and including the NERC national capability component of the recurrent budget.

All indicative figures until the first full design meeting, technical sign-off, and project board approval. Phasing designed to ensure both a flexible and iterated solution and avoiding buying storage before it is needed. Component costs based on extrapolation from costs of JASMIN phases 1-3 with requirements based on reasonable expectations of future science programmes. Phasing based on past experience of growth and integration issues.		Phase a	Phase b	Phase c	Phase d	Phase e
Storage Systems Hardware	fast disk and object storage, dedicated tape library and cache system					
Compute Systems Hardware	batch system and hypervisors for virtualisation					
Infrastructure Hardware	support systems for cloud (storage etc), switches, networks, server-room adaptation					
Cloud Software	portal and templates					
Infrastructure Software	workflow, tape and object store support					
Application Software	customised operating system, analysis s/w					
Integration & Contingency	Integration of systems, documentation, contingency					
	Totals	5000	1000	5000	1000	5000

Table 2: Overall capital breakdown by phase and category

CMIP data, subsequent investments would support further replacement/expansion to support the ongoing data and analysis growth. Technical evolution would continue throughout.

(on project planning and management), ISO9001 (quality management), the Cabinet Office major projects assurance toolkit, and the OGC managing successful programmes guidance.

7.1 Management and Governance

Overall project governance will be provided by the CEDA board representing the major stakeholders: the Natural Environment Research Council, NERC, and the National Centres of Atmospheric Science and Earth Observation (NERC, NCAS, and NCEO respectively) as well as the delivery partners (NCAS and the Science and Technology Facilities Council, STFC).

JASMIN project leadership is provided by NCAS, with a project manager and technical design provided by STFC. The system design, procurement, and integration will be carried out by the same team who successfully delivered earlier phases of JASMIN (under the then Department of Business, Innovation and Skills with Cabinet Office oversight). Project planning will exploit the STFC project management framework — which includes elements of PRINCE2, BS6079

7.2 Recurrent Support

JASMIN operates as a partnership, with recurrent support coming from five key routes:

1. Core support from the NCAS national capability budget,
2. Core support from the NCEO national capability budget,
3. Core support from the NERC HPC national capability budget,
4. STFC support for power and networks, and
5. Income from project funding, where appropriate.

The first three of these are shown in the recurrent budget in table 1. It is expected that only inflation adjusted increases in core support will be needed to accomplish the upgrade plans — any increases in recurrent costs will be recovered via project funding and/or recurrent funding from new JASMIN partners. Details of the partner-

ship and financial model will be provided in the proposed operations plan (section 7.4).

7.3 User Engagement and Support

A key part of the delivery of JASMIN is, and will be, support for the user community. Such support is provided in three ways:

1. Primary systems support is delivered by the STFC Scientific Computing Department's JASMIN team, who provide the compute fabric, and provide support to the operators of the platform and infrastructure compute.
2. Secondary support for use of the platform compute is provided by the CEDA JASMIN team, who manage resources, run training, and provide advice on the use of Lotus, Group Work Spaces and the JASMIN analysis platform.
3. Tertiary systems and application support, particularly for the end users of the Infrastructure Cloud, is provided in the community by the JASMIN partners themselves.

As the JASMIN community expands beyond the core NCAS and NCEO delivery (where direct support is directly funded within the JASMIN partnership), new partners will need to fund their own systems and application support.

The user community also provides direct input to the evolution of JASMIN (hardware and software) formally via the CEDA board, and informally at the annual JASMIN user conference, training sessions and other activities. Ongoing research into the nature of the algorithms used will also inform the system evolution.

7.4 Next steps

The next key phase is to establish an operations and business plan, outlining the relationship between capital investment, recurrent expenditure, and the return on investment. As noted in section 3.3 (and box 3) the JASMIN value proposition involves providing a data commons: with a centralised environment, JASMIN will reduce expensive duplication of data in the academic community and expensive dedicated computing, even as scientific workflows will be accelerated leading to faster and more pervasive impact.

The advantage of the data commons approach has been recognised by the community, with

some university groups investing directly in JASMIN themselves — alongside the core RCUK and UKSA investment. With the new investment described here, other investors are likely from the space, earth resources and climate service sectors, all aiming to exploit the data gravity associated with an existing managed archive with co-located cloud and HPC. Their investments will need to be included in the operations plan.

At the same time, the JASMIN architecture team will be progressing the technical architecture, ready for an architecture review by the end of calendar year 2016. The initial plan (table 2) was developed within the regular JASMIN operations meetings. This plan covers

- **Hardware:** fast disk, tape systems, the Lotus batch cluster, hypervisors, the first use of object stores, a migration away from VMware for the main cloud systems, and upgrades to the network and cloud disk.
- **Software:** cloud portal and templates, workflow and tape and object store environment support, the platform compute software environment and specialised software.
- **Integration and Contingency:** Network design, contractor support for testing and cloud systems development and integration, hardware and software contingency.

The most important technology changes anticipated will be support for containerised workflows and analysis workflows which start and end on tape with data migration onto customised hardware as part of “the job”.

Given the pace of change in technology, and the existing hardware profile, this plan reflects a clear technical vision and relatively concrete plans for the next two years, with scope for considerable change on longer timescales (to get the best scientific benefit from technology changes). It will be updated annually in the context of the operations and business plan.

Acknowledgements

This document received significant improvement from input and careful reading by Victoria Bennett, Jonathan Churchill, Philip Kershaw, and Ag Stephens. The science cases were assembled and co-edited by Poppy Townsend.