Institution: University of Reading		
Unit of Assessment: 11 – Computer Science and Informatics		
Title of case study: New computer systems for exploiting big environmental data for worldwide usability, model and policy development		
Period when the underpinning research was undertaken: Between 2011 and 2019		
Details of staff conducting the underpinning research from the submitting unit:		
Name(s):	**Role(s) (e.g. job title):**	**Period(s) employed by submitting HEI:**
B. N. Lawrence	Professor of Weather and Climate Computing	September 2011 to Present
Period when the claimed impact occurred: Between 1 August 2013 and 31 July 2020		
Is this case study continued from a case study submitted in 2014? No		

1. Summary of the impact

Cross-disciplinary environmental simulation and earth observation communities require prohibitively large amounts of heterogeneous data, beyond the level sensible to replicate and manage at most institutions. Research at Reading has addressed these issues by developing technologies to handle millions of files and petabytes of data. These have made it possible for the UK to be world leaders in delivering large-scale environmental data analytics. The resulting technologies underpin one of the world's largest multi-petabyte online environmental data archive, CEDA (the Centre for Environmental Data Analysis) hosted on a unique computing facility, JASMIN – a world-leading computational facility that delivers a petascale analytical environment (a 'data commons'). Experience with JASMIN resulted in: the development of new computer systems at the UK Met Office; improved commercial exploitation of satellite data; and the use of the global Earth System Grid Federation to support the Intergovernmental Panel on Climate Change (IPCC). All these activities led, and lead, directly to climate science outcomes of relevance to our society.

2. Underpinning research

In the early 2000s, the computing and data centre environment consisted of data centre silos decoupled from serious computing. Research carried out under the auspices of the UK e-science programme started to address the problems which resulted, such as the plethora of data types arising from different communities, the location of data, delivering appropriate compute platforms, and handling provenance in the face of millions of files and, eventually, petabytes of data. This work provided the foundations for the developments below.

Data services which need to be interoperable between communities and countries require common understandings of the data: both in terms of inherent meaning and methods of digital encoding. The climate forecast (CF) conventions (https://cfconventions.org) have been developed over two decades to address these problems. The University of Reading team's major research contribution was to provide a formal understanding of how these conventions could be coded in an interoperable way via a data model and software implementation [R3.1]. The data model and tools allow software engineers to validate their software against expected outcomes and supports precision in the CF specification.

Early experience with grids led to the Reading researchers' recognition that working towards a single solution was not the right approach, and that there were two separate use cases which could be better resolved with the provision of separate, but complementary, solutions. These two use cases, termed 'common communities, distributed data', and 'disparate communities with some common data', required independent solutions, the development of which form the body of this case.

Common communities, distributed data:
This problem was dealt with first, by co-developing (in a larger international consortium) the Earth System Grid Federation (ESGF), which provides distributed management of climate data to support discovery and download services [R3.2]. The ESGF depends on harvesting metadata from data formatted according to an extension of the CF conventions, distributing that metadata, and then using that harvested metadata in multiple portals providing data discovery via faceted browse. This includes user authentication and authorisation, allowing data download from data wherever it is held - whether locally or further afield. The underlying use case for the ESGF was

to support coupled, model-intercomparison projects (CMIPs) – which involve modelling groups worldwide that are carrying out specific numerical simulations and producing specific data for intercomparison. The CMIP “Data Reference Request’ [R3.3] and the ‘Data Reference Syntax’ (online) were key inputs to the process, providing scientists with guidance on what was required and how to format it for the ESGF (and further intercomparison). With European Commission support [G3.1], an ontology for providing data provenance for numerical simulations, [R3.4] was developed. This work provides the first comprehensive data model for complex numerical simulation workflows in environmental science, and the first ontological descriptions of numerical experiments prior to their execution. Additional work at the University of Reading exploiting a decade of investment using NERC national capability funding [G3.2] and targeting understanding and supporting the climate forecast conventions for the NetCDF data format, culminated in a formal data model for ‘CF compliant archive metadata’ along with a comprehensive Python implementation [R3.1]. All three activities (data reference syntax, simulation ontology and CF data model with python implementation) exploited fundamental concepts in data modelling, applied in new applications.

Disparate communities with (some) common data:
The CEDA holds over 13PB of environmental data, organised into over 2,000 datasets in 300 dataset collections. There are hundreds of parameters, including aircraft campaigns, satellites, radars, automatic weather stations, climate models, and more. The research challenge is making these data discoverable, accessible, and usable alongside other user-centric sources of data. Reading research to address these issues included improved cataloguing systems [R3.5] and architecting and implementing JASMIN [R3.6]. JASMIN is a unique datacentric computing platform which provides a range of storage technologies and customisable computing.

[R3.5] exploited a taxonomy and concepts developed during the e-Science era to create a new cataloguing system, capable of providing more than just traditional data discovery by providing a browse-based method of moving between data descriptors. The Reading research explored several implementations using a range of technologies before setting on the database structures described in [R3.5] – which are still in use today and underpin all the various views of more than 50,000 different metadata artefacts. With PB distributed in hundreds of millions of files, the catalogue system is integral to the use of JASMIN, as no available file system supports performant browsing at this scale.

The new JASMIN architecture solves the ‘disparate communities with common data’ problem by providing software, platform, infrastructure, and data as a service for a range of communities. JASMIN was shaped by e-science experience, and a sustained programme of research [G3.1], [G3.2], [G3.3]. It was the first large computing system designed primarily for petascale data analysis that included a curated petascale archive alongside petascale compute resources. Over the decade since JASMIN was first commissioned, it has progressively included more cloud computing capability and new storage technologies (twice deploying the world’s largest pools of new storage technologies, based on proof-of-concept research carried out by the JASMIN team). The ‘Software-as-a-service’ work developed many different methods and models for allowing users to manipulate data ‘server-side’ (e.g. [G3.4]), the most recent of which is the European Space Agency (ESA) Open Data Portal [R3.7] which utilises the JASMIN cloud.

3. References to the research

The research resulted from sustained national capability and external competitive funding; all except [R3.5] were published in peer-reviewed journals ([R3.5] appears in a peer-reviewed journal, but in a non-reviewed section). The research meets and exceeds 2* quality level definitions through defining and implementing new processes, techniques and methodologies for dealing with and managing large-scale environmental data and introducing fundamental new ideas for describing simulation requirements and properties.

R3.1 Hassell, D., Gregory, J., Blower, J., **Lawrence, B. N.**, & Taylor, K. E. (2017). ‘A data model of the Climate and Forecast metadata conventions (CF-1.6) with a software implementation (cf-python v2.1)’. *Geoscientific Model Development*, **10**(12), 4619–4646. DOI: https://doi.org/10.5194/gmd-10-4619-2017

R3.2 Balaji, V., Taylor, K. E., Juckes, M., **Lawrence, B. N.**, Durack, P. J., Lautenschlager, M., Blanton, C., Cinquini, L., Denvil, S., Elkington, M., Guglielmo, F., Guilyardi, E., Hassell, D., Kharin, S., Kindermann, S., Nikonov, S., Radhakrishnan, A., Stockhause, M., Weigel, T., & Williams, D. (2018). 'Requirements for a global data infrastructure in support of CMIP6'. *Geoscientific Model Development*, **11**, 3659–3680. DOI: https://doi.org/10.5194/gmd-11-3659-2018

R3.3 Juckes, M., Taylor, K. E., Durack, P. J., **Lawrence, B.**, Mizielinski, M. S., Pamment, A., Peterschmitt, J.-Y., Rixen, M., & Sénési, S. (2020). 'The CMIP6 Data Request (DREQ, version 01.00.31)'. *Geoscientific Model Development*, **13**(1), 201–224. DOI: https://doi.org/10.5194/gmd-13-201-2020

R3.4 Pascoe, C., **Lawrence, B. N.**, Guilyardi, E., Juckes, M., & Taylor, K. E. (2020). 'Documenting numerical experiments in support of the Coupled Model Intercomparison Project Phase 6 (CMIP6)'. *Geoscientific Model Development*, **13**(5), 2149–2167. DOI: https://doi.org/10.5194/gmd-13-2149-2020

R3.5 Parton, G. A., Donegan, S., Pascoe, S., Stephens, A., Ventouras, S., & **Lawrence, B. N.** (2015). 'MOLES3: Implementing an ISO standards driven data catalogue'. *International Journal of Digital Curation*. **10**(1). DOI: https://doi.org/10.2218/ijdc.v10i1.365

R3.6 **Lawrence, B.N.**, Bennett, V. L., Churchill, J., Juckes, M., Kershaw, P., Pascoe, S., Pepler, S., Pritchard, M., & Stephens, A. (2013). 'Storing and manipulating environmental big data with JASMIN'. *2013 IEEE International Conference on Big Data*, 68–75. DOI: https://doi.org/10.1109/BigData.2013.6691556

R3.7 Kershaw, P., Halsall, K., **Lawrence, B. N.**, Bennett, V., Donegan, S., Iwi, A., Juckes, M., Pechorro, E., Petrie, R., Singleton, J., Stephens, A., Waterfall, A., Wilson, A., & Wood, A. (2020). 'Developing an Open Data Portal for the ESA Climate Change Initiative'. *Data Science Journal*, **19**, 16. DOI: https://doi.org/10.5334/dsj-2020-016

Key Projects/Grants (all with Lawrence as the principal investigator):

G3.1 Infrastructure for the European Network for Earth System Modelling (ENES) series of grants from the European Commission: Metafor (between 2008 and2011), IS-ENES2 (2013-2017), IS-ENES3 (2019-2022).

- Supported the ontology work and now supports the Climate Forecast conventions (CF).

G3.2 NCAS Computational Model Services (Annual NERC National Capability Contract via the University of Leeds, between 2012 and 2018; from 2019 subsumed into the overall NCAS contract).

- Delivered the support for JASMIN R&D (JASMIN operations funded at STFC). Supported the CF data model work prior to EC Funding.

G3.3 European Centre of Excellence in Weather and Climate Computing (European Commission): ESiWACE (between 2015 and 2019), ESiWACE2 (2019-2022).

- Supported work on new storage systems and data access software.

G3.4 Copernicus Programme: C3S-MAGIC (between 2016 and 2017).

- Supported work on 'climate diagnostics' that can be run 'server-side' as part of a climate toolbox (in the Copernicus Climate Store).

G3.5 Met Office Contracts: 'HPC QA for the Met Office Supercomputing Procurement' (between 2014 and 2015, Contract H514900). 'HPC Support' (2013-2019, Contract H5169100)

- Supported advice to the Met Office on their 'JASMIN-like' procurement (SPICE), and research on parallel data manipulation on JASMIN.

4. Details of the impact

Environmental science is heavily dependent on big data and has always been dependent on bleeding edge data handling technologies. The research described here has directly enabled science outcomes which would have otherwise been difficult, or even impossible, to achieve given data volumes and disparate communities. These outcomes exploited three separate activities: (i) the use of the JASMIN data commons, (ii) the use of the Earth System Grid Federation, and (iii) the re-use of metadata techniques and tools by third parties; all of which arose from the research summarised here.

JASMIN data commons usage: JASMIN is used by a range of environmental scientists who wish to either share data, or directly exploit the massive volume of data held in the archives of

CEDA. CEDA itself is a key tenant of the JASMIN system, alongside a number of other organisations and individuals. The architecture and GBP20m implementation of the JASMIN was a primary research outcome – the operational implementation is now supported by annual budget in excess of GBP 1m via contributions from the National Centre for Atmospheric Science (NCAS), National Centre for Earth Observation (NCEO), the Natural Environment Research Council (NERC), and the UK Space Agency.

The JASMIN data commons depend on the data gravity asserted by the petascale CEDA archive: users are incentivised to bring their data alongside the archive and bring other users (and their data) for collaborations. In excess of 1,500 users exploit JASMIN directly, with approximately 25,000 additional users exploiting CEDA services hosted on JASMIN (Statistics, [S5.1]). The 1,500 direct users are organised into more than 250 distinct groups (tenancies) sharing data and or compute resources for data analysis – users can, and do, belong to multiple groups. JASMIN is now an integral part of UK and global science, yielding a wide variety of outcomes. Representative key JASMIN usage examples are listed below (but see also [S5.2]):

(i) The ESA climate change portal (http://cci.esa.int/data) is deployed in the JASMIN cloud, making use of metadata and data distribution technologies which have evolved from the original DataGrid work [R3.7].

(ii) JASMIN is providing the underpinning services for UKSA contingency planning for retaining access to EC Copernicus data post-Brexit and for the commercial activities of the Science and Technology Facilities Council's RALSpace and others [S5.3].

(iii) The production of unique rainfall estimates and insurance products for over three million African farmers are now predicated on climate services deployed on JASMIN [S5.4].

(iv) The dominant mode of analysis for UK CMIP6 work feeding into the sixth assessment report of the IPCC is the use of JASMIN, [S5.3] and [S5.5].

(v) The most recent UK State of Nature Report, a key government indicator for guiding environmental policy, covering how biodiversity has changed from 1970 to 2015 analysed 12,000 species using over 34 million records on JASMIN [S5.6].

All these outcomes were predicated on the co-location of curated data, user data, and specialised computing which arose from the University of Reading research into 'disparate communities with common data'.

Earth System Grid Federation (ESGF) and policy change: The ESGF provides the mechanisms for global management and distribution of data products produced by environmental groups worldwide – the 'common community distributed data' problem. In particular, it was designed to support the fifth and sixth global coupled model intercomparison projects – CMIP5 and CMIP6 – which themselves were timed to support the fifth and sixth assessment reports of the IPCC. The IPCC is a UN entity created to provide policymakers with regular scientific assessments on climate change, its implications and potential future risks, as well as to put forward adaptation and mitigation options. Those responsible for the fourth assessment report of the IPCC shared in the 2007 Nobel Peace Prize.

As of July 2020, the ESGF hosts 28 different projects with 21 petabytes of data. These are from 7.3 million datasets, distributed across 15 data nodes, which provide data downloads to 140 countries [S5.7]. The UK ESGF data nodes are deployed by CEDA and hosted on JASMIN, one cataloguing UK data for third-party download, and one hosting third-party data which has been replicated to JASMIN to avoid multiple petascale downloads to the UK. The research described here was integral to the initial architecture and deployment in CMIP5 [S5.3].

Much of the ESGF data is simulation data, used directly by academics, climate change consultants, and indirectly by those making commercial and policy decisions. All can find it difficult to understand the difference between different simulations. The Reading research on simulation provenance [R3.4] provided technologies to support the creation, extraction, and comparison of simulation documentation directly addressing this issue. As a consequence, the Reading researchers developed new systems and all modelling groups have been mandated [R3.2] to use them to construct simulation documentation for the sixth global model intercomparison project (CMIP6). (The mandate is described in [R3.3] and further evidenced in [S5.5]).

Reuse: Two examples are presented: (1) how the climate forecast (CF) metadata conventions [R3.1] are integral to weather forecasting and earth observation [S5.8]; and (2) how the JASMIN concept was copied and implemented by the Met Office to provide them with internal capability (SPICE) which matched the JASMIN provision to the wider UK community [S5.9].

The importance of the CF conventions to applications in weather and climate is well known and accepted, but prior to the advent of Reading's data model, they were not very interoperable with other systems. With this data model in place, the World Meteorological Office is now exploring the development of new regulations to encourage such interoperability. Whether or not these are put in place, the use of CF and the data model is now changing information transfer in global weather forecasting (on top of its integral role in climate model intercomparison). CF is now also becoming integral to earth observation, allowing major satellite data providers to lower costs and make data available to more users. The Reading team's data model has been a key part in making data products accessible to more of the users of Europe's largest provider of satellite based meteorological data, EUMETSAT [S5.8].

JASMIN was the first environmental ‘‘super data computer’’, showing the benefit of centralising high-volume data with dedicated high-performance analysis compute. The existence of JASMIN has enabled international research collaborations like PRIMAVERA (an EU Horizon 2020 research project involving the University of Reading and 19 European partners with the aim of developing a new generation of advanced global climate models for the benefit of governments, business and society) [S5.10]. What is more, JASMIN has sped up research workflows to the extent that some tasks which previously would have taken years, now take days. This is allowing some people to have, and test, as many ideas in a year, as they could have done in their entire career with traditional systems. The impact on research science was significant enough that the Met Office designed and procured their own ‘‘mini-JASMIN’’ called SPICE (Scientific Processing and Intensive Compute Environment) – contracting Lawrence to provide advice and quality assurance [G3.5]. SPICE now underpins Met Office research across both weather and climate [S5.9].

Summary: Large-scale environmental science depends on vast volumes (petabytes) of data, typically in millions of files. The research carried out at Reading on techniques for managing and manipulating data at this scale (on vocabularies and tools, the earth system grid federation, and the design and implementation of JASMIN - a computational facility that delivers a ground-breaking petascale analytical environment) has been essential to the work of multi-national agencies (for example, EUMETSAT). It has underpinned the research that led to delivery of the UK climate impact projections, the Paris Agreement of the UN Framework for Climate Change [S5.5], and the research that will underpin the next assessment of the IPCC. The next decade will see new challenges in moving from petabytes to exabytes, necessitating further developments in tools and information systems.

5. Sources to corroborate the impact

S5.1 (i) CEDA/JASMIN statistics via annual reports at https://www.ceda.ac.uk/about/highlights/ and (ii) https://manage.jasmin.ac.uk/projects/
S5.2 JASMIN Science Case http://cedadocs.ceda.ac.uk/1350/1/JASMIN_Science_Case.pdf
S5.3 Testimonial from Director of RAL Space, July 2020
S5.4 Testimonial from Director of TAMSAT, September 2020
S5.5 World Climate Research Programme (WCRP), Testimonial from Head of Understanding Climate Change, Met Office, July 2020
S5.6 Using JASMIN for the largest ever UK wildlife assessment https://www.ceda.ac.uk/blog/using-jasmin-for-the-largest-ever-uk-wildlife-assessment/
S5.7 ESGF statistics http://esgf-ui.cmcc.it/esgf-dashboard-ui/.
S5.8 Testimonial from EUMETSTAT, June 2020
S5.9 Testimonial from Met Office re SPICE, July 2020
S5.10 Testimonial from Royal Netherlands Meteorological Institute re PRIMAVERA, July 2020