CEMS: A NEW INFRASTRUCTURE FOR EO AND CLIMATE SCIENCE

Victoria L. Bennett [(1)], Philip Kershaw [(1)], Geoff Busswell [(2)], Richard Hilton [(3)], Alan O'Neill [(4)]

[(1)] Centre for Environmental Data Archival, RAL Space, STFC, Harwell Oxford, Didcot, OX11 0QX, UK,
Email: victoria.bennett@stfc.ac.uk, philip.kershaw@stfc.ac.uk
[(2)] CGI, The Office Park, Springfield Drive, Leatherhead, Surrey KT22 7LP, UK,
Email: geoff.busswell@cgi.com
[(3)] Satellite Applications Catapult, Electron Building, Fermi Avenue, Harwell Oxford, Didcot, OX11 0QR, UK,
Email: richard.hilton@sa.catapult.org.uk
[(4)] NERC National Centre for Earth Observation, Dept of Meteorology, The University of Reading,
Earley Gate Bldg 58, Reading RG6 6BB, UK,
Email:alan.oneill@nceo.ac.uk

ABSTRACT

CEMS, the facility for Climate and Environmental Monitoring from Space, has been created as a collaboration between UK academic and industrial partners at Harwell, Oxfordshire, UK, offering Climate and Earth Observation (EO) data and services. Since going operational in September 2012, CEMS has been supporting a range of research and commercial users. Applications include production of climate-quality long-term global datasets, processing satellite observations, and development of novel algorithms and products combining EO with other environmental datasets.

This paper briefly describes the CEMS infrastructure, present some example uses with initial indications of benefits of the CEMS environment, and outline plans for future evolution.

1. INTRODUCTION

CEMS, the facility for Climate and Environmental Monitoring from Space, supports research in the climate and environmental science community, and provides commercial sector opportunities for new business, based on exploitation of EO data and development of downstream services.

Data volumes used in climate and environmental science and applications are growing rapidly, both from new generation satellite instruments and large-scale climate modelling activities. Users wish to analyse and process these datasets, and combine or intercompare observations and simulations. It is becoming increasingly impractical for multiple organisations to retrieve these data over networks, and store them. EO data volumes are already of order 10's TB for single instruments, with many applications requiring input data from several missions.

The production of Essential Climate Variables (ECVs), e.g. through the ESA Climate Change Initiative (CCI), requires efficient access to multiple input datasets (typically several satellite data streams and ancillary meteorological data), multi-year processing and subsequent validation, analysis and intercomparison activities, as well as data dissemination to new user communities.

Another significant driver is the Copernicus programme: with the imminent launch of Sentinel satellites providing data volumes around 25 times larger than Envisat, an associated increase in complexity of data handling and processing, and the potential to exploit the data in diverse downstream services, the need for community data infrastructures is clear.

Three key features for CEMS were identified in a requirements study in 2011 funded by TSB, the UK Technology Strategy Board [1]:
1) a flexible hardware infrastructure, configurable to support both academic research and business opportunities for the commercial sector
2) access to large-volume EO and climate datasets, co-located with high performance computing
3) expertise and tools for provision of data quality and integrity information, giving users confidence and transparency in data provenance, services and products.

CEMS was developed and deployed in 2011/2012 as a partnership between UK academic and commercial organisations, originally through the International Space Innovation Centre (ISIC) collaboration. ISIC has since been superseded by the Satellite Applications Catapult[1], established in 2013 as a centre for the development and commercial exploitation of space and satellite-based products, services and applications.

The CEMS facility is now jointly operated by the Satellite Applications Catapult (primarily serving the commercial sector), and the UK Science and

[1] https://sa.catapult.org.uk/

Proc. 'ESA Living Planet Symposium 2013', Edinburgh, UK
9–13 September 2013 (ESA SP-722, December 2013)

Technology Facilities Council (STFC) Centre for Environmental Data Archival (CEDA)[2] (primarily serving the academic research sector, and in particular NCEO, the UK National Centre for Earth Observation).

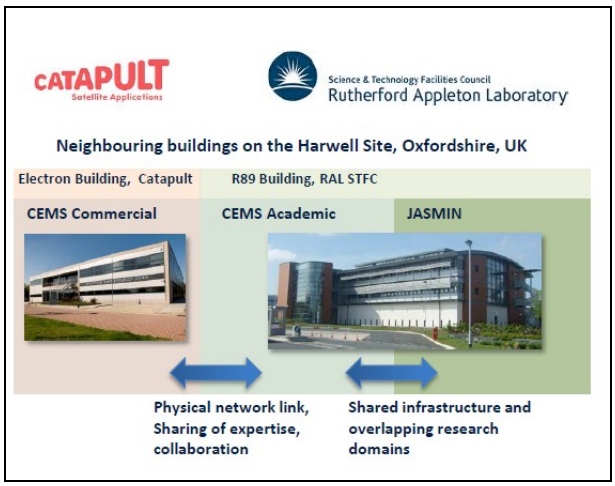

Figure 1. CEMS infrastructure and organisation across the Satellite Applications Catapult and STFC-CEDA at the Rutherford Appleton Laboratory (RAL)

As shown in figure 1, the "academic" CEMS infrastructure is fully integrated with the JASMIN super-data-cluster [2], a high-performance, high-volume data analysis environment for the UK environmental science community. The "commercial" CEMS infrastructure is co-located with a range of facilities which are designed to support organisations in generating ideas, applications and technologies for products and research derived from satellite data. A dedicated fast network link connects the academic and commercial sides, enabling data and services to be shared, and staff collaboration ensures sharing of expertise.

2. TECHNICAL DETAILS

To address the increasing need for on-line storage and flexible access to large and complex datasets, the CEMS system is configured for high volume storage, alongside compute capacity to support processing and analysis next to the data. CEMS deploys approximately 2.5PB of fast, low-latency disk storage and 450 computing cores for local computation.

The storage technology is implemented with a Panasas® system, which provides a global file system with fast performance, eliminating input/output bottlenecks between the processing nodes and storage hardware. Compute resources are managed by making use of virtualisation technology and a cloud-based service model (VMware vCloud). This enables a variety of working environments to be created and supported for different applications across different user communities.

In addition, fast network links connect CEMS and JASMIN to other relevant data stores and infrastructures in the UK and Europe, including the UK Met Office and the UK-PAC/PAF in Farnborough (UK ESA Multimission Processing and Archiving Centre).

JASMIN also includes a small batch compute system, "Lotus HPC" and is connected to the STFC SCARF HPC facility (4000 cores for STFC users).

The CEMS storage includes a dedicated allocation for holding "community datasets" (currently ~500TB, including data from a number of ESA and Eumetsat missions) which authorised users can access, process and analyse on the system. Further space is made available to users as temporary workspaces, to hold user input data as well as intermediate and output products.

3. EARLY PROJECTS

In its first year of operation, CEMS has been supporting a range of research and commercial users.

Projects are underway which span both public and private sector led initiatives, and concern areas such as climate change, carbon markets and renewable energy. Applications already include production of climate-quality long-term global datasets, processing satellite observations, and development of novel algorithms and products combining EO with other environmental datasets.

In the Satellite Applications Catapult, industry users including SMEs (Small and Medium-sized Enterprises) and larger organisations are hosting and managing datasets, and developing products and tools for bespoke applications with impact for decision makers. Examples include:

- Forest carbon stocks monitoring (Rezatec): using CEMS to collate, manage and host EO and associated data which will enable customers to access quality assured carbon values
- Weather and climate analysis for agricultural production (WeatherSafe): combining satellite derived information and data gathered directly from farms to provide practical and targeted suggestions for farmers
- Quality indicators for EO data products (CGI): developing procedures to improve the confidence of end users in the uptake of EO data within a carbon auditing service
- Products to support habitat monitoring and assessment (GeoSeren): Bringing together relevant information from a range of sources and combine them to produce tailored mapping products

In addition, CEMS is being used to trial the use of an

[2] http://www.ceda.ac.uk/

"exploitation platform" environment for EO ground segment applications, and new capabilities are being built, including satellite scheduling and metadata management.

Early academic users of CEMS include research groups from UK universities and the NCEO, who are using CEMS as a hosted processing environment, capitalising on the flexible environment to initially trial, then scale up, large-scale EO data processing tasks with direct access to input and output data stores [4].

Example academic CEMS projects are:

- Trial processing of AATSR for Climate (ARC) Sea Surface Temperature Data (University of Edinburgh, University of Reading): evaluation of the CEMS processing environment for retrievals of sea surface temperature. Input datasets directly available on CEMS.
- Land Surface Temperature processing from AATSR data (University of Leicester): parallelisation of monthly data processing for faster whole-mission reprocessing. Input datasets directory available on CEMS.
- Processing and hosting GlobAlbedo data products (MSSL, UCL): large (100 TB) workspace for high-volume land surface data products, and processing capability to generate new products. Product dissemination to users.
- ATSR1+2 Full Mission Reprocessing (RAL): efficient processing on CEMS and Lotus, with input data directly available on CEMS and output data made available via CEMS archive
- Cloud Essential Climate Variables processing (RAL), data processing using CEMS and SCARF compute, large workspace (120 TB) to store and disseminate output products, input datasets available on CEMS.

Projects are able to use different approaches to process their data, but typically users have started to develop and test their code on one Virtual Machine (VM), then either scale out to a larger number of cloned VMs, or add more cores to existing VMs, or shift their processing onto the Lotus compute cluster.

Users have reported considerably shortened processing times due to the fast input/output between compute and storage, but also the ability to parallelise processing tasks across the available compute, typically completing tasks 100 times faster than on their previous analysis environment [3]. This level of speed-up completely changes the way scientific research can be carried out with the data, and the limiting factor in the process is now scientists' time to analyse and evaluate the results rather than the processing itself. In addition, the available datasets and scalable computational resource, including memory, storage and processing power, mean that new scientific approaches can be explored.

4. FORWARD LOOK

Over the next two years, significant investments in hardware expansion are planned, both in STFC-CEDA infrastructure, and in the Satellite Applications Catapult.

In STFC-CEDA, it is planned to scale up the hardware with approximately an additional 3000 cores, 5 PB disk (across JASMIN and CEMS) and 10 PB tape storage. In addition the range of services will be expanded through an internal private cloud to enable more flexible access to the data archive and compute resources. New capability will also be added to federate with other cloud providers. Users will be able to *cloud-burst*, supplementing compute and storage resources available on the CEMS private cloud with resources from public clouds as needed.

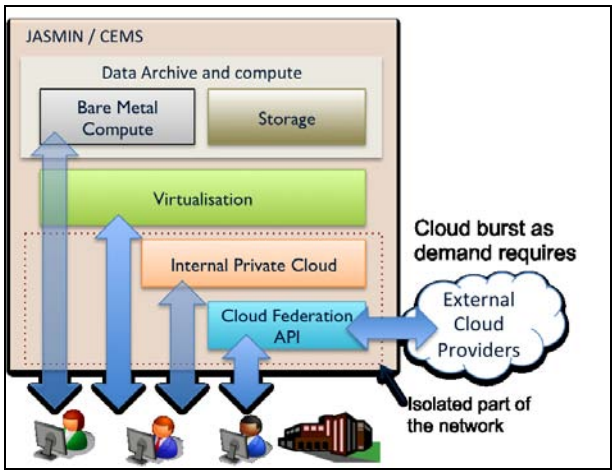

Figure 2. Schematic showing current and future CEMS service models, ranging from bare metal compute with direct but restricted access to the data archive, to a private cloud environment with greater abstraction from the data archive but also greater autonomy for management of user environments

In the Satellite Applications Catapult, further investment has recently been made to enhance the underpinning computing facilities in terms of redundancy and availability, plus providing a more focussed range of technical facilities spanning both private cloud functions and High Performance Computing (HPC). Support to SME and start-up companies is continuing and the ability to provide these organisations with significant computing resources under novel economic arrangements should start to produce real growth in the sector. Lastly, significant work is currently underway to attract commercial data providers and software/service vendors to utilise CEMS to tap into and grow the climate and environmental monitoring markets both externally, but also via the use of the CEMS platform itself as a marketplace. The legal and business practices to support this are being developed with the aim that the user experience can be

as simple as possible whilst ensuring that the CEMS system is enriched as far as possible.

5. SUMMARY

We have described the CEMS infrastructure and organisation, and given some examples of its usage in its first full year of operation. Early operations have shown that there is considerable demand for such a facility, and that there are clear benefits to a community shared data and compute infrastructure.

While industry users are beginning to develop applications and services, and science users are capitalising on the hosted processing capability, there is scope to develop both aspects, continuing to enable innovative use of Earth Observation Data from Space for Climate and Environmental applications.

6. REFERENCES

1. Logica, NCEO, Astrium-GEO, RAL Space, 2011, A User Requirements Analysis on a Facility for Climate and Environmental Monitoring from Space (CEMS), proposal to the Technology Strategy Board, Issue v1.0

2. Lawrence, B.N. , V. Bennett, J. Churchill, M. Juckes, P. Kershaw, P. Oliver, M. Pritchard, and A. Stephens, "The JASMIN super-data-cluster," *ArXiv e-prints*, Apr. 2012.

3. Lawrence, B.N. , V.L. Bennett, J. Churchill, M. Juckes, P. Kershaw, S. Pascoe, S. Pepler, M. Pritchard and A. Stephens, "Storing and manipulating environmental big data with JASMIN", submitted and accepted for 2013 IEEE International Conference on Big Data (*BigData 2013*), Santa Clara, CA, USA, Oct 6-9, 2013.

4. Kershaw, P.J., A. Smith, J. Churchill, R. Brugge, G. Corlett, D. Ghent, C. Merchant, "First Results from CEMS (A)ATSR Hosted Processing Pilot Projects", *Big Data From Space*, ESA-ESRIN, Frascati, Italy, June 2013. http://www.congrexprojects.com/docs/default-source/13c10_docs/abstractbook.pdf?sfvrsn=2