Geophysical Research Abstracts
Vol. 21, EGU2019-18192, 2019
EGU General Assembly 2019

Delivering resilient access to global climate projections data for the Copernicus Climate Data Store using a distributed data infrastructure and hybrid cloud model

Philip Kershaw (1), Matt Pryor (2), Alan Iwi (2), Ag Stephens (2), Martin Juckes (2), Ruth Petrie (2), Stephan Kindermann (3), Carsten Ehbrecht (3), Sébastien Denvil (4), Sébastien Gardoll (4), and Bryan Lawrence (5)

(1) NCEO/Centre for Environmental Data Analysis, STFC Rutherford Appleton Laboratory, Didcot, United Kingdom, (2) NCAS/Centre for Environmental Data Analysis, STFC Rutherford Appleton Laboratory, Didcot, United Kingdom, (3) DKRZ, Hamburg, Germany, (4) CNRS / IPSL, Paris, France, (5) NCAS / University of Reading, Reading, United Kingdom

We describe the development of a resilient, distributed data infrastructure created to serve global climate projections data to the Copernicus Climate Change Service (C3S) Climate Data Store (CDS). The CDS provides a single public point of access for freely available climate-related observations, re-analyses and model data. This is achieved using a distributed architecture, with individual participating data providers hosting and serving data through to the CDS via an agreed set of web service interfaces. The CP4CDS project was established to provide access to a quality controlled subset of CMIP5 model data and is led by CEDA working together with partners DKRZ and IPSL.

From the outset, an architecture was proposed leveraging the existing Earth System Grid Federation (ESGF), a globally distributed software infrastructure operated to provide access to CMIP data and other earth sciences data. However, the project requirements specified a high level of resilience for the data services (98% uptime) which no single site could meet as part of its standard SLA. A novel architecture was developed to meet this requirement based on a distributed deployment model and taking advantage of public cloud capabilities. Each of the three partner site provides an identical mirror of the data services. These are combined using load-balancing to deliver a single point of access thus providing an aggregate uptime which meets requirements. Amazon Web Services' Route 53 was selected in order to provide DNS-based load balancing. Cloud computing technologies have also had an important role in the hosting of the services themselves with CEDA's services hosted on-premise on the JASMIN data analysis facility's community cloud. In addition, container and container orchestration technologies have have many attractive features for building a highly available system. Over the last year, the ESGF collaboration have worked together to port the ESGF software stack to Docker and Kubernetes. This has facilitated a deployment of part of the system to Google Cloud.

Public cloud hosting provides the means to meet the requirements for availability and uptime with a single deployment and potentially dispenses with the need for a distributed architecture. However, this must be weighed up against the high relative cost for the hosting of large volumes of model data (100s of TBs) when compared with on-premise hosting. In the short to medium term, a likely scenario in the wider ESGF community is the continued use of on-premise hosting for large datasets together with increased adoption of public cloud for the hosting of lighter weight services such a data search where there are much smaller storage requirements.

The project has successfully delivered an operational system then, based on a hybrid model of a distributed data infrastructure together with public cloud. The same approach will be applied for a project to deliver regional model data CORDEX4CDS led by IPSL and will also influence the evolution of the architecture for ESGF.