Institution: University of Exeter		
Unit of Assessment: UoA 11 Computer Science and Informatics		
Title of case study: Pangeo – transforming geoscience research		
Period when the underpinning research was undertaken: 2018 - 2020		
Details of staff conducting the underpinning research from the submitting unit:		
Name(s): Alberto Arribas Herranz Naill H Robinson	**Role(s) (e.g. job title):** Professor Senior Lecturer	**Period(s) employed by submitting HEI:** 2018 – present 2019 – present
Period when the claimed impact occurred: 2018 - present		
Is this case study continued from a case study submitted in 2014? N		

1. Summary of the impact

The ability of geoscience and other sectors and fields of research to generate data has outstripped the infrastructure required for research and downstream data analysis. To tackle this issue, scientists from the University of Exeter have developed Pangeo, a ground-breaking new approach employing an ecosystem of interoperable tools and architectures, enabling users to work with multiple large datasets using nothing more than a laptop.

Pangeo has had a transformative impact on the geoscience sector, radically improving the quality, speed, scope, scale and reproducibility of research in this discipline and has delivered the following impacts:

- It has become the primary access point for environmental datasets in Microsoft Azure, utilised in their ‘AI for Earth’ and ‘Planetary Computer’ initiatives.
- Used by 10,000+ researchers in academia and across public and private sector organisations, including the Alan Turing Institute, to produce innovative tools and large-scale geoscience research.
- Allowing researchers to better understand the relationship between Covid-19 and environmental factors.

2. Underpinning research

Context

A major issue in geoscience research (but also true in other scientific disciplines such as astrophysics and particle physics) is that the analysis tools are no longer able to cope with the high data volumes involved; this slows down and complicates the process of scientific research in these fields. To address this problem, Arribas and Robinson led research at the University of Exeter, building on their previous work at the Met Office, to develop Pangeo in 2018, an innovative ecosystem of interoperable tools and architectures, sufficiently flexible for a range of different applications and use cases.

Design principles

To overcome the key challenges the team employed a novel combination of architectural design principles, outlined below **[3.1; 3.2].**

- **“Separation of concerns”** – coupling together a series of unitary components which do one (and only one) thing well – this modularity makes it practical to adapt the individual components as long as they still fulfil the same purpose, thus maximizing the potential for Pangeo to be used in a broad range of applications.
- **“Specializing late”** – as few components as possible are affected by domain specific design decisions to ensure maximum adaptability to different use cases Keeping components of the systems as generic as possible “future proofs” them,

allowing the thin layer of domain specific functionality built on top of them to be adapted to new use cases. In Pangeo, analysts interact with objects which represent earth science data (Xarray or Iris) and these systems invoke more generic operations in other components, such as NumPy and Dask.

- **"Co-locating compute and data"** – Pangeo makes use of JupyterLab, a web-based interactive development environment which enables scientists to send commands and queries via the internet to compute resources co-located with data.
- **"Parallel computing"** – this is not yet widespread in scientific analysis mainly due to the limitations of existing analysis software. However, in Pangeo, Dask is used to distribute calculations across the many nodes of an HPC or cloud-based cluster, or across the available cores on a single computer.
- **"Elastic scaling"** – compute nodes are rapidly allocated and deallocated in response to user activity – Pangeo has been deployed in elastic scaling mode in both HPC and cloud environments. By using Dask the user does not have to explicitly parallelize their work, meaning they can execute the same scientific analysis as previously but get their answer faster.
- **"Lazy evaluation"** – doing the minimal amount of analysis possible by maintaining canonical, base-level datasets alongside "lazy" derived datasets, which access and process the canonical dataset on demand. This was previously impractical as processing times were too long for on-demand data products; however, it is possible with elastic scaling and parallel computing (see above).
- **Lowering the cognitive burden for accessing large datasets** – a range of approaches have been employed, including exploring new ways of handling data such as Parquet and Zarr, which are optimized for the storage of very large datasets across many individual files. A complementary approach is to use a data-broker layer, such as Intake, which mediates between the data files and the composite dataset object used by scientists in their analysis.
- **Adopting open standards and building on common interfaces to avoid vendor lock-in**. Pangeo has made use of Kubernetes, an open source, cloud-native system which mediates between running processes and infrastructure. By using it, the Pangeo project has run data platforms on AWS, Microsoft Azure, Google Compute Cloud, and Alibaba.

Application

The flexibility of the systems architecture **[3.3]** developed using the principles described above, means that Pangeo deployments can be created by implementing an automated recipe which defines the entire interconnected system, a technique known as "Infrastructure as Code" (IaC). The key advantages are that it is:

i) possible to integrate data archive and cloud computing seamlessly; and
ii) possible to apportion costs to different stakeholders – as third parties can mirror portions of the platform on their own cloud computing account.

Therefore, the resulting platform makes it possible, for the first time, for an individual researcher to:

- do sophisticated mathematical operations over petabyte-scale datasets without having to invest, or gain access to, tens of millions of pounds worth of infrastructure and software licenses beforehand;
- combine multiple complex datasets hosted in different clouds/archives in a seamless way; and
- ensure the traceability and reproducibility of their results, automatically promoting scientific best practice.

In addition, as the researcher is no longer required to download large datasets to an onsite resource, nor constrained by storage capacity, much more rapid and in-depth research is possible. Researchers no longer need to select short running tasks over long ones due to infrastructure, nor are they required to reduce data points due to storage capacity.

3. References to the research

3.1 **Robinson NH**, Hamman J and Abernathy R (2020). Seven Principles for Effective Scientific Big-Data Systems. arXiv:1908.03356v2

3.2 Abernathy R, **Robinson NH** *et al.* (November 2020) Cloud-Native Repositories for Big Scientific Data. IEEE Computing in Science and Engineering. *Preprint on Authorea* DOI: 10.22541/au.160443768.88917719/v1.

3.3 Technical Architecture (2018), Pangeo website: http://pangeo.io/architecture.html.

4. Details of the impact

The development of Pangeo has had a transformative impact on geoscience research globally. It has led to a major international network focused on utilising and developing Pangeo environments, enabling numerous geoscience and research projects, and the development of tools based on big data which would have previously been unfeasible. Now part of a major collaboration at Microsoft it is a key element in Microsoft's 'AI for Earth' and 'Planetary Computer' initiatives, and provides the primary entry point to access environmental datasets in Microsoft Azure.

Facilitating large scale research tackling environmental challenges through Microsoft Azure

Recognising the transformative opportunities offered by Pangeo, Microsoft has developed a strong partnership with the UoE team to provide Pangeo tools via Microsoft Azure **[5.1]**. The importance of this collaboration has been highlighted by the AI for Earth Programme Director: "*Prof Arribas and his team have been instrumental in bringing Pangeo tools to the Azure cloud, and thus have dramatically accelerated our ability to support the wider computational needs of the environmental science and earth science communities*" **[5.1]**.

Microsoft's AI for Earth initiative was established as a global movement to accelerate technology development for environmental sustainability. Pangeo has formed a key element within it **[5.1]**. As the Programme Director explains, "*one major challenge we have faced in the past is difficulty that most organizations and individual researchers struggle with accessing, combining, and interrogating huge geoscience datasets. Pangeo tools abstract away the complexities of managing thousands of compute nodes, so environmental scientists can focus on science, not infrastructure*." **[5.1]**

The AI for Earth Programme is currently supporting hundreds of organisations, high-impact startups and cutting-edge research projects in over 81 countries, to facilitate research, insight and impact on environmental challenges. One example is the OOICloud Project **[5.2]** which is working to make data from the Ocean Observatories Initiative (OOI) widely available through the cloud and accessible through a Pangeo interface for research. The collaboration with Azure has significantly extended the reach of Pangeo, providing a single access point for Azure users to link with and interrogate multiple geoscience datasets.

The Pangeo tools are also key to another Microsoft initiative, the Planetary Computer, announced by Microsoft President Brad Smith in April 2020. The Planetary Computer, explains Programme Director, is part of the company's commitment to become an "*ecosystem-positive company and support [its] grantees and customers in their ecosystem-positive goals, none of which can be achieved without global-scale analysis*". They go on to say "*The Planetary Computer will essentially constitute a data catalog coupled with a Pangeo installation, facilitating large-scale analysis of remote sensing data, ecosystem health information, and climate data. The centrality of Pangeo in this project highlights the importance of our collaboration with Prof. Arribas to our corporate environmental sustainability commitments*." **[5.1]**.

Uptake by the Geoscience community and creation of new geoscience research tools
Estimated to have over 10000 users **[5.3]**, Pangeo packages have received over 4 million downloads to date **[5.3]**. Pangeo deployments are facilitating diverse and high impact geoscience research which previously would not have been feasible, due to time or infrastructure constraints. Over 50 self-reported publications have been produced by researchers utilizing Pangeo, covering topics such as ice thickness of glaciers and the relationship between changes in climate and glacier length, ocean mixing linked to climate variability, and predictability in modelling the global carbon cycle **[5.4]**.

Additionally, Pangeo has catalysed a new geoscience community of practice, whereby an international network of over 500 people have come together to utilise, promote, and advance Pangeo's scalable cloud infrastructure **[5.3]**, the community have used the platform to develop tools and software for key areas of geoscience research. For example, Pangeo based software packages formed the basis for OceanSpy **[5.5]**, a python package to facilitate ocean model data analysis and visualization. To date OceanSpy has had over 5000 downloads and is now part of the larger NSF-funded Poseidon project **[5.5]**, which aims to make Peta- and Exascale ocean simulations widely available to the public for the first time. Other geoscience tools developed from the Pangeo packages include the Open Global Glacier Model **[5.5]** and Psyplot interactive visualization framework **[5.5]**.

Transforming big data research capability at the Alan Turing Institute (ATI)

As a leading research institution for data science and artificial intelligence globally, the ATI quickly recognised Pangeo's potential to improve its capability to facilitate big data research. Following Prof Arribas's appointment as a Turing Fellow, he has been working with the ATI research engineering group to explore the applications of Pangeo on the various research projects being undertaken by the institute. This has allowed the ATI to support novel, diverse projects utilizing big data geoscience methodologies, helping to address challenges such as:

- Monitoring of iceberg calving in Antarctica
- Incorporating estimates of pest and disease risk into estimations of the impacts of climate change on agriculture globally; and
- Improving the resolution of environmental models through the combination of satellite data with surface sensors and physics-based model simulations **[5.6]**.

As a result of Prof. Arribas's advocacy and expertise the Alan Turing Institute has also been able to leverage a $300,000 gift from Microsoft to extend the infrastructure to an additional cloud provider **[5.6]**. The Programme Lead for Tools, Practices and Systems at the Alan Turing Institute said:

"*The Pangeo architecture has allowed projects to progress at a greater speed, with a reduced requirement for physical infrastructure and incorporating far more data than we would have been able to facilitate previously. This has led to the Turing being able to offer a more a rapid, high quality research capability, which results in downstream benefits to the sectors and industries related to our research*." **[5.6]**.

Understanding Covid-19 and environmental factors

Pangeo has also offered opportunities for researchers to respond to topical issues, such as COVID-19. The Met Office COVID platform, a partnership between Pangeo, the Met Office, Microsoft Azure, the Environmental Futures and Big Data Impact Lab and the EDRF, is providing a resource for researchers to explore the relationships between COVID-19 and environmental factors. To provide customisable compute environments for researchers, since April 2020 an instance of Pangeo was created which has provided easy access to data. This has enabled researchers to easily access data and undertake studies which they could not otherwise have done, supporting national responses to the pandemic **[5.7]**.

The DELVE (Data Evaluation and Learning for Viral Epidemics) group used the platform to complement its work informing the UKs strategic response to the pandemic. Since August 2020, the data was used to generate population weighted daily weather averages to investigate links between weather and COVID transmission. This was fed into the DELVE Global COVID-19 Dataset. The platform also fed into two reports submitted to SAGE (Scientific Advisory Group for Emergencies) in November 2020 **[5.8],** where the Pangeo environment was highlighted an exemplar of good data readiness practise. These reports have helped inform government responses to pandemics, through the adoption of new data governance practises, allowing rapid-turnaround, data driven answers to policy questions **[5.8]**.

The platform was also used to further funders' research funding approaches. For example, in the NERC Covid-19 Digital Sprint **[5.9]**, which set a number of challenges around the environmental impacts and consequences of COVID-19, as well as in investigations being carried out by the UCL Centre for Longitudinal Studies to model pre- and post-lockdown compliance, and its effects on Covid-19 mortality **[5.9].** In 2020, the Met Office COVID platform has been accessed by a variety of national and international organisations and higher education institutions to support their modelling endeavours. The organisations include: Department for Business, Energy and Industrial Strategy (BEIS), Google DeepMind, the Bank of England, University of Oxford and the London School of Hygiene and Tropical Medicine **[5.10]**.

5. Sources to corroborate the impact

5.1 Support letter from Microsoft AI for Earth Programme Director. Available as PDF.

5.2 Microsoft blog discussing OOICloud project. Available at: https://web.archive.org/web/20210118121326/https://www.microsoft.com/en-us/ai/ai-for-earth-ooicloud

5.3 Evidence of Pangeo user numbers

5.4 Self-reported publications utilising Pangeo. Available at: https://web.archive.org/web/20201001031913/https://pangeo.io/publications.html

5.5 Examples of tools developed using Pangeo architecture

5.6 Support letter from Alan Turing Institute Programme Lead for Tools, Practices and System. Available as PDF.

5.7 Theo McCaie (9th April 2020) “Met Office and partners offer data and compute platform for COVID-19 researchers” *Medium*. Available at: https://web.archive.org/web/20210118121700/https://medium.com/informatics-lab/met-office-and-partners-offer-data-and-compute-platform-for-covid-19-researchers-83848ac55f5f

5.8 DELVE Reports on Data Readiness and Organisational Data Maturity

5.9 Evidence of the applications of Covid-19 Hub

5.10 Evidence of organisational access to Covid-19 hub