European Commission

Proceedings of the 2023 conference on

Big Data from Space (BiDS'23)

From Foresight to Impact

6-9 November 2023
Austrian Center, Vienna

Edited by P. Soille, S. Lumnitz, and S. Albani

SatCen

European Commission

esa

Joint Research Centre

PANGEO@EOSC: ENABLING OPEN, REPRODUCIBLE, AND SCALABLE BIG DATA GEOSCIENCE FOR EVERYONE

Anne Fouilloux[1], *Tina Erica Odaka*[2], *Alejandro Coca-Castro*[3], *Sebastian Luna-Valero*[4], *Pier Lorenzo Marasco*[5], *Guillaume Eynard-Bontemps*[6]

[1]Simula Research Laboratory, Oslo, Norway
[2]LOPS UMR 6523, CNRS-IFREMER-IRD-Univ.Brest-IUEM, Brest, France
[3]The Alan Turing Institute, London, UK
[4]EGI Foundation, Amsterdam, Netherlands
[5]SEIDOR, Italy
[6]CNES, Toulouse, France

ABSTRACT

The exponential growth of data in the field of geoscience presents both great opportunities and challenges. Extracting insights from vast volumes of Earth observation data requires sophisticated analysis techniques. When these techniques are open, reproducible, and scalable they accelerate reusability. This paper explores Pangeo@EOSC, an innovative platform uniting the power of the European Open Science Cloud (EOSC) and the Pangeo ecosystem to enable open, reproducible, and scalable big data geoscience for researchers and practitioners worldwide. We discuss key components, highlighting benefits for efficient data analysis and collaboration. To increase take-up, Pangeo@EOSC offers a Pangeo Training Infrastructure as a Service (PTIaaS), which facilitates the provision of Pangeo deployments to individuals interested in delivering training to researchers and students. This service enables seamless access to real-world datasets and analysis capabilities within the Pangeo environment. We conclude by discussing future prospects and potential impact of Pangeo@EOSC to advance the field of big data geoscience and promote a culture of open and collaborative research.

Index Terms— community, scalable, reproducible, FAIR, EOSC

1. INTRODUCTION

The era of big data has revolutionized the field of geoscience, with Earth observation satellites generating vast amounts of data on a daily basis. However, the sheer volume and complexity of this data pose challenges for researchers and scientists who aim to extract meaningful insights. In this section, we provide an overview of the challenges and opportunities in big data geoscience and introduce the concept of Pangeo@EOSC as a solution.

2. TOWARDS A EUROPEAN FEDERATED AND OPEN SCIENCE CLOUD

The benefits of Open Science (OS) and FAIR principles - Findable, Accessible, Interoperable and Reusable - are increasingly valued by academia, although what OS and FAIR entail in practice are largely misunderstood. Also, even if researchers manage to grasp OS and FAIR principles, they are often hit by technical difficulties, typically due to a lack of appropriate infrastructure, services and support. These are precisely gaps that the European Open Science Cloud (EOSC) was meant to fill in. EOSC is the main initiative in Europe for providing a federated and open multi-disciplinary environment where European researchers, innovators, companies and citizens can share, publish, find and re-use data, tools and services for research, innovation and educational purposes. One of the goals of the EOSC is to co-design with communities the tools and services that are useful for their daily research activities, facilitate collaboration, and promote the wider adoption of Open Science practices.

3. PANGEO@EOSC

In this section, we explain how the core technologies of the Pangeo ecosystem, such as Jupyter notebooks, Dask, and Xarray, which form the backbone of Pangeo's capabilities were integrated within EOSC. Additionally, we examine the data storage and sharing mechanisms provided by EOSC and how Pangeo leverages these resources to enhance accessibility and data-driven research.

The Pangeo community, consisting of scientists and developers worldwide, is dedicated to creating user-friendly and community-driven platforms for big data geoscience. While various services based on Jupyter Notebooks were already available, there was no publicly accessible Pangeo deployments providing swift access to substantial amounts of data

doi:10.2760/46796

and compute resources within the EOSC environment. Most of the existing cloud-based Pangeo deployments were centered in the US, leaving European members of the Pangeo community without a shared platform to exchange knowledge and experiences. To address this gap, Pangeo collaborated with two EOSC projects, namely EGI-ACE and C-SCALE. This partnership aimed to demonstrate the deployment and utilization of Pangeo on EOSC ([1]), highlighting the benefits it offers to the European research community.

Together with EGI, the Pangeo Europe Community deployed a DaskHub comprising a Dask Gateway and JupyterHub, backed by a Kubernetes cluster on EOSC using the infrastructure of the EGI Federation. The Pangeo EOSC JupyterHub deployment leverages the EGI Check-in for user registration, providing authenticated and authorized access to the Pangeo JupyterHub portal and the underlying distributed compute infrastructure. It also utilizes the EGI Cloud Compute and cloud-based EGI Online Storage to distribute computational tasks across a scalable compute platform and store intermediate results generated by user jobs.

To enable future deployments of Pangeo enabled JupyterHub on EOSC (Pangeo@EOSC) a specific recipe for the user-friendly Infrastructure Manager (IM) Dashboard [1] has been implemented [2]. This flexibility enables future Pangeo deployments across a broad range of private and public cloud providers, including EGI Federated Cloud, OpenNebula, OpenStack, AWS, GCP and more. Moreover, our developed recipe and IM enable the seamless deployment of the Pangeo@EOSC platform across diverse infrastructures, which not only enhances scalability, allowing you to transition from a local playground to a versatile platform, but also serves as a powerful safeguard against the potential risks of vendor lock-in. At present, computing and storage resources are provided by CESNET as part of the EGI-ACE and C-SCALE projects. These resources not only provide ample support for both teaching and research but also exceed the typical computing power users have for free. These deployments offer substantial advantages by enabling training participants to tackle realistic, large, and complex data analysis problems that directly align with their research. Participants learn how to efficiently access and analyze extensive online datasets using tools such as Xarray, Dask, and others. Moreover, this collaborative environment allows attendees to seek assistance, collaborate with fellow researchers and Research Software Engineers, and adopt Open Science practices. This approach eliminates the burden of building their own infrastructure and having to install all the necessary packages and libraries by themselves.

[1] https://im.egi.eu/im-dashboard/login
[2] https://github.com/pangeo-data/pangeo-eosc

4. COMMUNITY OF PRACTICE WITH THE EDS BOOK

With a continually growing community, over 100 researchers have already undergone training on Pangeo@EOSC deployments in the first year (see Section 5). This expansion sparked discussions on the importance of establishing clear practices for writing and publishing FAIR software that can be readily reused and built upon for future research endeavors.

To address this need, Pangeo joined forces with the Environmental Data Science Book (EDS Book, https://edsbook.org), an inclusive pan-European resource driven by the community. Hosted on GitHub and powered by Jupyter Book, the EDS Book provides practical guidelines and templates to assist researchers in transforming their research outputs into well-curated, interactive, shareable, and reproducible executable notebooks. In addition a collaborative and transparent reviewing process, supported by GitHub-related technologies, ensures the quality of FAIR notebooks. Through the utilization of EDS Book, Jupyter Notebooks created and published by researchers become highly modular and reusable, fostering knowledge exchange and enhancing the overall research experience.

5. PANGEO TRAINING INFRASTRUCTURE AS A SERVICE (PTIAAS)

”Pangeo Training Infrastructure as a Service (PTIaaS)” ([2]) is a concept and approach developed by Pangeo to provide a dedicated infrastructure for conducting training sessions and workshops on the Pangeo ecosystem. It aims to offer a scalable and reproducible environment where participants can learn and practice using Pangeo tools and technologies straight away, without having to think about technicalities.

The PTIaaS model involves deploying a Pangeo JupyterHub instance specifically tailored for training purposes. This infrastructure is designed to accommodate a large number of users simultaneously and provide them with access to computational resources and data storage capabilities. The infrastructure can be provisioned on cloud platforms or dedicated servers, depending on the requirements and preferences of the training organizers.

By adopting the PTIaaS approach, training providers can ensure a consistent and controlled environment for participants to learn and explore the functionalities of Pangeo. The infrastructure can be customized with different language kernels (Python, R, Julia), pre-installed software packages (like Pytorch and Tensorflow for machine learning), sample datasets, and bespoke teaching materials to facilitate a smooth learning experience. Additionally, the PTIaaS model allows for easy scaling of resources (”elastic” model) based on the number of participants and their computational needs.

The PTIaaS infrastructure is free for anyone willing to deliver a training event based on Pangeo and it can be ordered

doi:10.2760/46796

through an online form [3]. It is typically made available for a specific training period (up to several months), ensuring that participants have uninterrupted access during the training sessions. It enables hands-on exercises, collaborative work, and interactive learning through Jupyter notebooks, allowing participants to gain practical experience with Pangeo tools and workflows.

Overall, the PTIaaS approach enhances the effectiveness of Pangeo training initiatives by providing a dedicated and optimized infrastructure that supports the learning objectives of participants. This simplifies the setup and management of the training environments, enabling trainers to focus on delivering high-quality content and promoting the adoption of Pangeo among the geoscience community. To showcase PTIaaS in action, let's explore specific training events where it was effectively used.

5.1. FOSS4G workshop

The newly deployed infrastructure was highly successful in facilitating user onboarding during the FOSS4G conference via a dedicated course. Developed collaboratively and made openly available under a CC-BY-4 license, this training program witnessed active participation. In addition to instructors and helpers, it served over thirty users simultaneously. The FOSS4G Pangeo 101 workshop took place on August 23, 2022 and lasted 4 hours, including breaks. It leveraged the capabilities of the Pangeo infrastructure and sparked engaging discussions, generating growing interest among participants in utilizing this infrastructure for their data analysis endeavors. Participants gained comprehensive insights into the Pangeo ecosystem and acquired crucial skills for conducting open, reproducible, and scalable Earth science on the European Open Science Cloud infrastructure. The workshop employed fully reproducible Jupyter Notebooks, meticulously prepared to guide attendees in various critical tasks. These included accessing both local and remote data sources, loading and analyzing data using the potent Xarray library, visualizing data with the interactive Hvplot visualization package, and understanding techniques for scaling computations with Dask. Throughout the workshop, participants worked with Sentinel-3 NDVI (Normalized Difference Vegetation Index) Analysis Ready Data, sourced from the Copernicus Global Land Service. This provided them with hands-on experience using real-world geospatial datasets.

5.2. eScience course

An annual eScience course on linking observations with modeling in Climate Science underwent a significant transformation with the integration of the EOSC infrastructure. The 6th edition of the course, held at Tjärnö Marine Laboratory in Sweden, attracted students from various countries. The course was open to master and PhD students, with a majority of participants being master students. A JupyterHub, similar to the one used in the FOSS4G workshop but with an increased amount of memory (48GB for each frontend), was provided to students, mentors, and instructors. Dedicated assistants guided the students in conducting small studies on specific topics related to climate science. Here also up to thirty users were simultaneously working, and the infrastructure perfectly held this workload. Topics covered diverse areas such as Arctic ocean biology, aerosol dynamics, and ozone changes. The course fostered active engagement and lively discussions. Students benefited from effective training materials and utilized GitHub repositories for collaboration and for sharing their work: a GitHub organization (https://github.com/orgs/eScience-course/repositories) was created and each group had its own repository (seven public repositories under the MIT license). Students stored their Jupyter notebooks and reports in these repositories, following a template provided by the Pangeo community for the course [4]. The integration of GitHub into the JupyterHub was highly appreciated. Each student presented their Jupyter notebooks, and a hybrid meeting was held to present their findings. For more information about the course and the training material (available under the CC-BY-4 license) visit https://pangeo-data.github.io/escience-2022/intro.html.

5.3. CLIVAR/IASC CMIP6 Arctic Bootcamp

The idea of the Arctic processes in CMIP6 (the 6th Coupled Model Inter-comparison Project) CLIVAR Bootcamp [5] was different from a "classical" training workshop: the goal was to have Early Career Researchers focus on "grand challenges" and to try to move the science forward using state-of-the-art tools, and get some practical experience within their small groups as well as through the guidance of scientists mentors. The Pangeo community and EGI-ACE provided the technical infrastructure (Pangeo@EOSC) and technical support (best software practices for writing shareable data analysis codes and data). The training material was adapted from the previous workshop and is also open source (CC-BY-4 licence). It is available at:
https://pangeo-data.github.io/clivar-2022/intro.html. Early Career Researchers and their mentors learned how to work with CMIP6 data from many models in one go, and via the cloud computing infrastructure delivered by the Pangeo Community and EGI-ACE, and the rest was spent doing actual science, i.e. apply the tools to answer the research question participants had chosen. Attendees were split into groups: each group worked on different scientific

[3] https://forms.gle/GGoHAZX1rPr52fmP8

[4] https://github.com/eScience-course/escience2022-template

[5] https://www.clivar.org/events/arctic-processes-cmip6-bootcamp

doi:10.2760/46796

questions and used different datasets. A Github organisation was created [6] to host the work of each group: access rights (private/public) were left to each group. Five groups chose to have private repositories but added an MIT licence to be ready for future publications) and two groups created public repositories.

This CLIVAR bootcamp was also the first time s3-like object storage was introduced to read and write files on Swift object storage. A total of 840 virtual CPUs, 3.2 TB memory and 20 TB of storage were provided to attendees through a Pangeo JupyterHub and Dask clusters.

5.4. Reproducibility Challenge

The 2023 Climate Informatics Reproducibility Challenge [7] offered an inclusive opportunity for individuals worldwide to showcase their expertise and collaborate in teams (seven teams comprising three people each) to reproduce scientific papers from the Environmental Data Science Journal[8], part of Cambridge University Press and Assessment. Participants were encouraged to use the templates and guidelines provided by EDS Book. The challenge aimed to foster transparency and openness in climate informatics research. Among the initial teams formed, the event witnessed the submission of three Jupyter notebooks, meeting rigorous standards of quality and reproducibility. These outstanding notebooks are published in EDS Book, further contributing to the collective knowledge and understanding in the field of environmental data science.

6. CONCLUSION AND FUTURE WORK

The collaboration between Pangeo and EOSC is a success: Pangeo delivered highly efficient and scalable Pangeo JupyterHubs, which proved to be instrumental in onboarding over 100 users from more than 10 countries. The feedback received from the users has been overwhelmingly positive. Although some attendees faced challenges with EGI Check-in during the FOSS4G conference, primarily due to the time gap between registration and resource access, proactive measures were taken for subsequent events. Attendees were encouraged to register for EOSC services at least one week (preferably two weeks) prior to the workshop or training to address any potential issues in advance. Additionally, the documentation to freely access the EOSC Pangeo JupyterHub and connecting with EGI Check-in has been enhanced to provide clearer instructions [9].

In addition to delivering online events, the Pangeo community maintains an ongoing collaboration with EGI to improve Pangeo deployment and facilitate Open Science practices. For instance, work on the deployment of a Binder instance with a Dask gateway is under way. This collaborative effort aims to provide a unified approach to spatial data analysis that is independent of data and infrastructure providers.

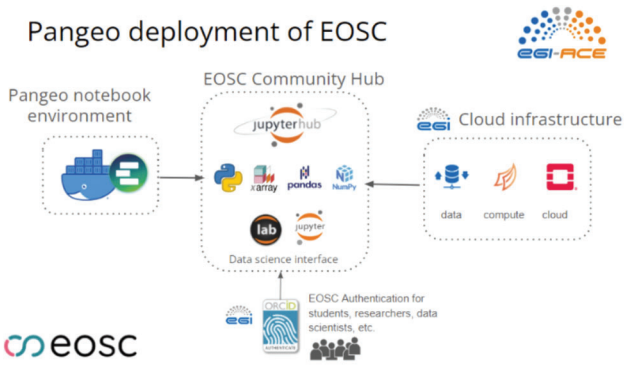

Fig. 1. Diagram illustrating the Pangeo deployment on EOSC: every user can authenticate using EGI Check-in, a unified way to login to any EOSC services. For instance, users can login with their ORCID Identifier or Linkedin or Github account, etc.

7. ACKNOWLEDGMENT

This work benefited from services and resources provided by the EGI-ACE Project (receiving funding from the European Union's Horizon 2020 research and innovation under Grant Agreement no. 101017567), with the dedicated support of CESNET-MCC.

REFERENCES

[1] Guillaume Eynard-Bontemps, Jean Iaquinta, Sebastian Luna-Valero, Miguel Caballer, Frederic Paul, Anne Fouilloux, Benjamin Ragan-Kelley, Pier Lorenzo Marasco, and Tina Odaka. Pangeo@EOSC: deployment of PANGEO ecosystem on the european open science cloud. May 2023. doi: 10.5194/egusphere-egu23-9095.

[2] Anne Fouilloux, Pier Lorenzo Marasco, Tina Odaka, Ruth Mottram, Paul Zieger, Michael Schulz, Alejandro Coca-Castro, Jean Iaquinta, and Guillaume Eynard Bontemps. Pangeo framework for training: experience with FOSS4g, the CLIVAR bootcamp and the eScience course. May 2023. doi: 10.5194/egusphere-egu23-8756.

[6] https://github.com/orgs/clivar-bootcamp2022/repositories

[7] https://eds-book.github.io/reproducibility-challenge-2023/intro.html

[8] https://www.cambridge.org/core/journals/environmental-data-science

[9] https://pangeo-data.github.io/pangeo-eosc/