Open-source federated learning across multi cloud environment

Anonymous Authors[1]

Abstract

Hundreds of Petabyte of data is generated daily from Satellite, Weather and Earth models. Machine-Learning (ML) and Artificial Intelligence (AI) promises great advances in understanding the impact of climate change on Earth and discovering climate adaptation and mitigation solutions. The main challenge in advancing AI for Climate is rooted in the vast logical and physical distribution of climate-relevant data and information. The data is usually hosted in a openly accessible cloud environment. However, assembling an AI-ready datacube requires accessing individual platforms and applying spatial and temporal filters. We describe a distributed open-source data platform that aggregates data through a Spatio-Temporal Asset Catalog (STAC) for quick data discovery and to download only the data of interest. Various datasets are harmonized using the openEO framework that assembles the AI datacubes efficiently, enabling a speed-up in their generation.

1. Introduction

Geospatial-temporal data such as weather, satellite, maps, land use, and land cover can reveal the local versus global impact of climate change. To better quantify the changes, local observations as detected by satellite observations can be combined with weather and climate data that has a larger predictive time horizon and associated uncertainties. To operationalize the predictions, there is a need to assemble datacubes that can easily used by AI models. Climate-relevant data is not only complex, often with a geospatial context, but is also very big (in most cases), making it impractical or even impossible to move. We define this challenge as "data gravity". In addition, most climate-relevant information is "dark", or not readily discoverable, meaning that one needs to inspect (*i.e.*, download) the information before one knows whether a particular data set contains the required information. There is still a lack of general geospatial-temporal modeling tools and frameworks which would enable modularization of common geospatial-temporal model tasks, such as exploration, training, testing, and deployment, not only within a single compute environment but across different instances.

The challenge of the vast logical and physical distribution of climate data was the main driver for us to develop AnonymousTool, which is composed of services that enable federated discovery and modeling across a network of nodes, where each node has its own compute resources with some climate-relevant information. Through visibility into climate-relevant information, the federated discovery services accelerate the finding of required data drastically to better support development of climate adaptation and mitigation solutions. It will not only un-darken climate information but also drastically reduce data gravity challenges as one can work with "targeted" and smaller data sets. The modeling services are not only key to reduce data gravity challenges, for example by providing distributed geospatial insights summarization or federated workflows for geospatial model training and scaling, but also accelerate model development in general. These services support modern AI workflows and common requirements of climate change model development, for example, driving models with dynamic climate data and static geospatial data, the need to calibrate and validate models, to scale them efficiently in time and space, and to provide not only predictions but quantification of uncertainties in those predictions.

2. Data Discovery

2.1. Core technologies for distributed learning

The main goal of the discovery technologies is to make geospatial-temporal climate-relevant data and information discoverable. The discovery process often reduces data volume significantly thereby reducing data gravity challenges (Lu & Hamann, 2021).

Key to the discovery of the climate-relevant information lies in the efficient spatial and temporal harmonization of the data because most discovery tasks invovle data with multi-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

ple, differing spatial and temporal resolutions. Towards that end, AnonymousTool leverages two key innovations. First, AnonymousTool includes the concept of a common spatial-temporal coordinate and reference system accompanied with a set of nested resolution layers, which creates efficient spatial and temporal joins between different data sets. Second, AnonymousTool supports a geospatial-temporal key-value representation of the climate data information. These two innovations enable not only a "data cube" for raster data but also for combined vector and raster data, which accelerates downstream discovery and modeling tasks significantly. AnonymousTool also includes a multi-modal data curation and ingestion service, which supports hundreds of different geospatial-temporal data representations and formats.

2.2. Core data-level discovery operators

Powered by a unique geospatial-temporal indexing scheme, these functions allow rapid pixel level search. For example, finding all locations and timestamps with extreme rain events in rasterized weather data. Because of the nested resolution layers, such search can be extended by very fast spatial and temporal joins and filtering, which is a common task in the discovery process. Another unique feature are overview layers or indexes (Ives & Taylor, 2008; Quoc et al., 2018) where statistical metrics of a matrix of pixels are stored. These indexes not only reduce the time complexity of search and discovery tasks by *O(log n)*, but also enable smart sampling and optimized data retrieval, access, and rapid visualization during data exploration.

2.3. On-demand featurization

A key differentiator of AnonymousTool is on-demand featurization services and functions, which is key to coping with data gravity challenges. AnonymousTool allows users to define featurization functions using arbitrary math, where only the features are serviced to the end users for further downstream modeling tasks. For example, with AnonymousTool a user can discover extremes of weather data, determine quantiles and standard deviations, or run regressions and decision trees all as a service, without the need for accessing or even downloading the raw data, which can provide significant acceleration in the overall workflow.

3. Modeling Technologies

The main purpose of the modeling technologies is the acceleration of the development of geospatial-temporal AI models regardless of the compute environment (*e.g.*, single instance, cloud, hybrid, on-prem, or across different instances). These modeling technologies are critical in overcoming data gravity challenges and accelerating the development of high-impact climate applications in general (Edwards et al., 2022).

3.1. Core modeling services

The modeling services are based on two key concepts: modules and workflows. Modules are blocks of code, which can include actions like invoking discovery services, running simulation or AI models, processing model outputs, etc. These modules are the atomic blocks for composing workflows. Workflows are the pipelines of these modules that pass data between them to achieve a specific task. Such workflows can be serial, but can also include branching and even iteration. The modeling services include pre-built models (*e.g.*, climate impact models such as pluvial floods), modules (*e.g.*, downscaling via Weather Generators) and workflow templates (*e.g.*, for uncertainty quantification). They enable model version control, metadata management, deployment, and orchestration as well as lifecycle management. We envision in the future that foundation models will be served for downstream tasks. In such a scenario, a foundation model is trained to learn a representation of a large data set, which then can be used for specific downstream tasks by users, for example for extreme weather forecasting or land use detection.

3.2. Federated modeling services

The key result of the concept of modules and workflows is the modularization of workflows, which is not only essential for efficient model development in a given compute environment but critical for distributing tasks. We envision extending AnonymousTool's modeling services to enable distributed and federated workflow orchestration. This is particularly important as climate impact applications often involve multiple "staggered" models, where one model feeds data into the next. The decisions of what compute environments are best suited for each model, how to minimize data movement, and what features can be computed where and when within the workflow, is envisioned to be managed by AnonymousTool's federated modeling services. (Zillman, 2009; Prein et al., 2015)

4. Data Harmonization

4.1. STAC-Enabled Workflow for Creating AI-Ready GHG Emission Data

Preparing geospatial data for AI applications is often labor-intensive and fragmented. Analysts typically spend significant time aggregating and pre-processing data, only to find that a small portion is suitable for AI pipelines. This issue becomes even more pronounced when aligning multiple modalities - such as satellite imagery, inventory records, weather data, or sensor outputs - across both space and time.

To address these challenges, we propose a workflow that

relies on SpatioTemporal Asset Catalog (STAC) [1] for organizing and accessing GHG emission data from diverse sources. Each source is treated as a distinct modality, and the goal is to transform them into harmonized, AI-ready datacubes through a sequence of structured processing steps.

Figure 1 shows an overview of the AI-ready data cube workflow and its registration back into STAC within our openEO (Mohr et al., 2025) implementation. openEO is an Application Programming Interface (API) specification focused on Earth observation data processing, which also includes a data processing specification

The workflow begins with reading metadata from STAC, such as bounding boxes and observation timestamps. Using this information, the first step matches and aligns modalities spatially and temporally - a critical requirement for GHG applications where time synchronization is essential. Next, the pipeline filters these candidate pairs based on spatial overlap, data quality, or other criteria. For example, pairs lacking sufficient overlap for patch extraction are discarded. Additional filtering conditions can be applied as needed. After filtering, the actual data for each modality is retrieved and processed. This includes spatial-temporal operations like imputation, reconstruction, and aggregation. For example, sensors with multi-band imagery may require resolution harmonization and band-specific filtering. Some modalities, such as Sentinel-5P CH4, exhibit significant data gaps. Here, AI-based gap-filling methods can outperform traditional approaches by incorporating auxiliary features.

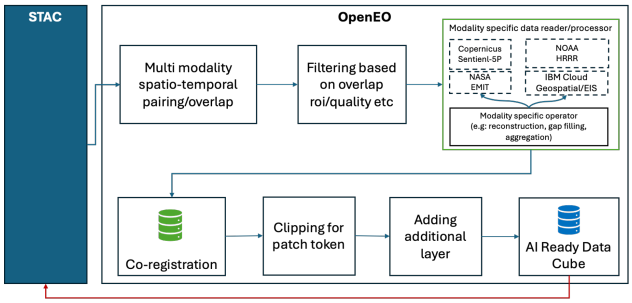

Figure 1. Overview of the AI-ready data cube workflow and its registration back into STAC within the openEO framework

A subsequent step involves registering all modalities to a common grid and projection system using openEO. Once aligned, the data is divided into spatial patches (*e.g.*, 224×224) to create AI-ready tokens. As shown in Figure 1, there's also an optional `Adding additional layer` module that supports label data integration for supervised learning use cases. This can also be used to incorporate categorical information - such as Land Use and Land Cover - to facilitate downstream tasks like class-balanced sampling or stratified splitting of the AI-ready dataset into training, validation, and testing subsets.

[1] https://stacspec.org/en

4.2. Data Management Overview

The goal of our data management strategy is to unify datasets from multiple repositories to support the development of AI models focused on methane emissions.

As outlined above, we leverage STAC and openEO as key components of this system. STAC serves as the central access point for all data, while data processing is handled through a set of well-defined pipelines in openEO. These pipelines operate in three modes:

- Direct Import: Datasets are imported into Cloud Object Storage (COS) and registered in STAC. This approach is used when a specific datacube is known to be needed in advance and it is desirable to host the data internally.
- External Reference: Publicly available datasets are referenced directly and registered in STAC without being hosted locally. This approach is applied when the data is already publicly available in a usable format. In this case, we register the dataset in STAC without hosting it, thereby avoiding redundant storage while still allowing access to this data and use in the openEO pipelines.
- Lazy Onboarding: Data is imported and registered in STAC only when explicitly requested. This approach is used to minimize storage and avoid registering unnecessary data. Here, data is only imported and registered when a request is made, which introduces a delay due to on-demand processing.

These pipelines transform datasets from remote sources into formats suitable for our openEO implementation, such as NetCDF and Cloud-Optimized GeoTIFF (COG).

4.3. Processing remote sensing imagery for CH4 plume detection

In this section, let's briefly discuss a few examples of data manipulation. For instance, consider the choices to be made when manipulating satellite imagery data, such as the hyperspectral images of methane plumes from point source emissions like oil and gas infrastructure from NASA's Earth Surface Mineral Dust Source Investigation (EMIT) mission. These images are freely available from link in either radiance (corresponding to L1B processing level) (Green, 2022a) or L2A for surface reflectance (Green, 2022b) . For a given acquisition and processing level, each file is in NetCDF format, containing 285 spectral bands. The images typically have dimensions of 1,280×1,242 pixels, covering

approximately 76.8×74.5 km on the Earth's surface with a pixel size of 60 meters. Each NetCDF file contains all bands, along with ancillary information, such as viewing geometry angles, which can be used for georeferencing.

Registering such data in STAC and being able to open it using openEO is attractive, but users should consider a few practical points. For example, imagine a scenario where patches of these images are sampled to train an AI model to detect plume patterns in unseen locations. In this case, sampling patches directly from the row/column raster images (*i.e.*, without geometric projection) may arguably be preferable to preserving the original pixel spectral signature, which could carry a very weak methane signal. However, if the patches are georeferenced and aligned to a target grid of different scale, further distortions may be introduced, which could impact the detectability by an AI model.

Now consider an alternative case where the user wants to employ a traditional matched filter on hyperspectral imagery, a proven technique to estimate methane concentration (in ppm.m) at the pixel level, through comparison with a target signature estimate from physical modeling (Foote et al., 2020). A sophisticated open-source implementation of the matched filter is the mag1c algorithm[2]. We have adapted this code to work with the original EMIT NetCDF as input due to the ancillary information it contains, and to consider the full image (not patches) to model the statistics of background methane-free pixels. In this scenario, users must consider a few options: adapting the matched filter code to work with image patches, such as those that could be make accessible from openEO, or running the matched filter on the original NetCDF directly and storing the output estimated methane image layer result into STAC. Both options have advantages and disadvantages that require careful consideration, particularly in relation to i) data modification (*e.g.*, georeferencing the raw data and then running the code, or running the code on the raw data and then georeferencing its result), ii) code adaptation efforts, and the iii) convenience of sampling large datasets using openEO for training AI models.

5. Conclusion and Future Work

Climate-relevant data is difficult to discover and to process given that it is massive, distributed, and heterogeneous. The main contribution of this paper is AnonymousTool, a set of geospatial services and tooling that support building AI models for climate. It enhances data discovery by using nested resolution layers integrated with STAC. It reduces data gravity by creating a federation of openEO and STAC instances deployed on different computing clusters. Furthermore, AnonymousTool provides a flexible approach to data storage: data can be pulled on demand using the lazy data onboard approach, externally referenced, or directly imported to improve performance.

One avenue for future work is to conduct a rigorous evaluation of each tool in order to compare against state-of-the-art and state-of-the-practice approaches.

[2] https://github.com/markusfoote/mag1c

References

Edwards, B., Fraccaro, P., Stoyanov, N., Bore, N., Kuehnert, J., Weldemariam, K., and Jones, A. Cimf: Climate impact modelling framework. In *ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, 2022.

Foote, M. D., Dennison, P. E., Thorpe, A. K., Thompson, D. R., Jongaramrungruang, S., Frankenberg, C., and Joshi, S. C. Fast and accurate retrieval of methane concentration from imaging spectrometer data using sparsity prior. *IEEE Trans. on Geoscience and Remote Sensing*, 58(9): 6480–6492, 2020.

Green, R. EMIT L1B At-Sensor Calibrated Radiance and Geolocation Data 60 m V001, 2022a.

Green, R. EMIT L2A Estimated Surface Reflectance and Uncertainty and Masks 60 m V001, 2022b.

Ives, Z. G. and Taylor, N. E. Sideways information passing for push-style query processing. In *IEEE Intl. Conf. on Data Engineering*, pp. 774–783, 2008.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proc. of the Intl. Conf. on Machine Learning*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lu, S. and Hamann, H. F. IBM PAIRS: Scalable big geospatial-temporal data and analytics as-a-service. In *Handb. of Big Geospatial Data*, pp. 3–34. Springer, 2021.

Mohr, M., Pebesma, E., Dries, J., Lippens, S., Janssen, B., Thiex, D., Milcinski, G., Schumacher, B., Briese, C., Claus, M., Jacob, A., Sacramento, P., and Griffiths, P. Federated and reusable processing of Earth observation data. *Scientific Data*, 12(1), 2025.

Prein, A. F., Langhans, W., Fosser, G., Ferrone, A., Ban, N., Goergen, K., Keller, M., Tölle, M., Gutjahr, O., Feser, F., et al. A review on regional convection-permitting climate modeling: Demonstrations, prospects, and challenges. *Reviews of geophysics*, 53(2):323–361, 2015.

Quoc, D. L., Akkus, I. E., Bhatotia, P., Blanas, S., Chen, R., Fetzer, C., and Strufe, T. Approxjoin: Approximate distributed joins. In *Proc. of the ACM Symposium on Cloud Computing*, pp. 426–438, 2018.

Zillman, J. W. A history of climate activities. *World Meteorological Organization Bulletin*, 58(3):141, 2009.