Terragon: A Unified Framework for Earth Observation Data Cube Generation

Adrian Höhl [1,2¶], **Paul Höhn** [1,3], **and Xiao Xiang Zhu** [1,2]

1 Chair of Data Science in Earth Observation, Technical University of Munich **2** Munich Center for Machine Learning **3** Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR) ¶ Corresponding author

DOI: 10.21105/joss.08857

Software

- Review
- Repository
- Archive

Editor: Jack Atkinson

Reviewers:

- @usethedata
- @AtmaMani

Submitted: 28 April 2025
Published: 29 September 2025

Summary

Terragon (Earth(Poly)gon) is a Python package facilitating access to remote sensing and Earth observation data from multiple sources. Its goal is to unify the process of downloading data in a simple and efficient manner. While existing tools focus on specific satellites or data providers, Terragon offers a more flexible solution. The package offers a consistent way to search, filter, and download data from various data sources. It utilizes a polygon format to define the region of interest and creates a spatio-temporal data cube (Mahecha et al., 2020) of rasterized data in the Xarray Dataset format, as illustrated in Figure 1. Additionally, it ensures the alignment of projections and resolutions, organizing the data according to the selected resolution and coordinate reference system.

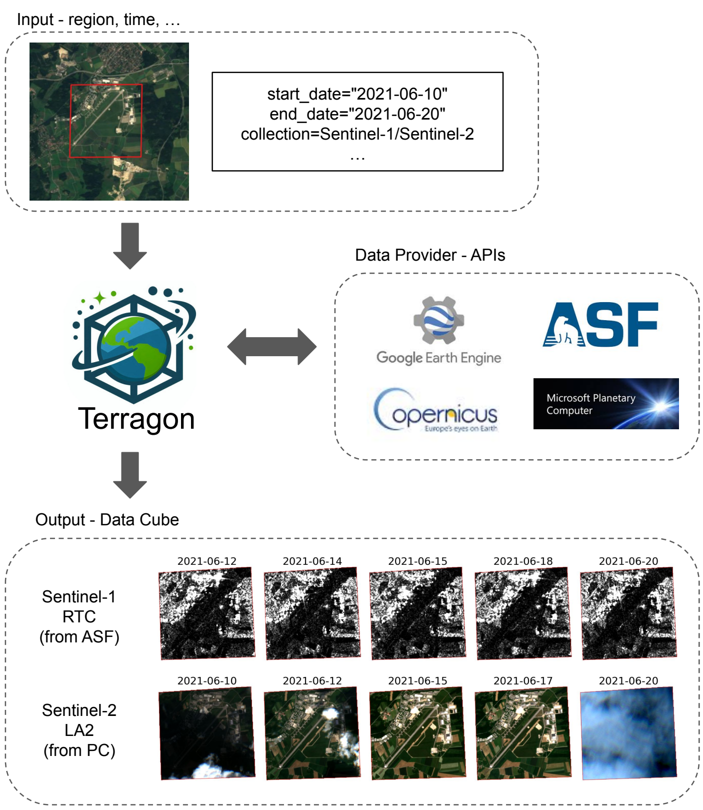

Figure 1: Terragon Workflow

Currently, several common data providers are supported, including Google Earth Engine (Gorelick et al., 2017), Planetary Computer (Microsoft Open Source et al., 2022), Copernicus Data Space Ecosystem (Copernicus, 2024) and Alaska Satellite Facility (Alaska Satellite Facility, 2025). The goal is to further develop and maintain this tool by incorporating additional data providers and implementing processing techniques, such as mosaicing and resampling, based on community needs. The software was leveraged to prepare two large-scale datasets Sen12Landslides (Höhn et al., 2025) and CropClimateX (Höhl et al., 2025). Sen12Landslides contains 75,000 landslide annotations and has over 12,000 patches from Sentinel 1 and 2. CropClimateX contains 15,500 small data cube spanning 1527 counties in the USA, it spans multiple sensors, weather and extreme events, soil and terrain features. Further projects are currently in preparation.

Overall, Terragon reduces the resources required for the time-consuming process of accessing and downloading data from various APIs, condensing them into a consistent, reusable and cost-efficient framework.

Statement of need

The number of active Earth-observation satellites is increasing (Union of Concerned Scientists, 2025), and platforms like Sentinel-2 are generating vast amounts of data. This development has significantly impacted remote sensing analysis by providing access to large-scale datasets, both free and commercial. Simultaneously, there is an increasing demand for analysis-ready data to develop data-driven methods, such as machine learning models. However, the process of gathering this data remains time-consuming. Existing platforms are fragmented, often providing only specific satellites, formats, and custom APIs. This leads to repeated work and poor compatibility across platforms, highlighting the need for more convenient access and a unified, straightforward framework.

Although many useful tools have already been developed to help people access and process Earth observation data (Chudley & Howat, 2024; Haan et al., 2023; Hogenson et al., 2020; Montero et al., 2024; Taconet & Moiroux, 2024), these tools often support only one satellite, product, or API, or are built for specific programming languages like R. This limits their broader applicability, especially for researchers looking to combine data from multiple sources or work within a Python-based scientific stack, where Python has become the standard for machine learning research. Although there is some functional overlap with cubo (Montero et al., 2024), our design diverges substantially. Cubo builds fixed-size data cubes around a centroid point using Universal Transverse Mercator (UTM) zones, while Terragon supports arbitrary polygons as input and preserves the source coordinate reference system, avoiding pixel misalignment and overlaps. We aim to have a cube that closely mimics the shape of the polygon, making it a far more flexible approach. Cubo's design does not support this paradigm and would require rewriting its core architecture.

Terragon addresses these challenges, allowing users to focus on their applications rather than the technical setup. This simplifies scaling to different platforms and eliminates the need for repeated setups for each data provider. Terragon offers a straightforward and consistent interface for downloading data from various APIs. It produces clean, analysis-ready data cubes that integrate seamlessly into machine learning workflows and other applications.

Target Audience

Our target audience includes anyone who needs to download Earth observation data, particularly researchers and users with limited financial resources who cannot afford the costs associated with commercial providers or cloud services. Additionally, Terragon is designed for users with limited technical knowledge who prefer not to spend excessive time navigating different APIs to download remote sensing data.

Documentation

Documentation of the package is available at readthedocs.

Acknowledgements

The work of A. Höhl was supported by the project ML4Earth by the German Federal Ministry for Economic Affairs and Climate Action under grant number 50EE2201C. P. Höhn was funded by HELMHOLTZ IMAGING, a platform of the Helmholtz Information & Data Science Incubator under grant number: ZT-1-PF-4-028. The work of X.X. Zhu is also supported by the Munich Center for Machine Learning.

References

Alaska Satellite Facility. (2025). *ASF Home*. Alaska Satellite Facility. https://asf.alaska.edu/

Chudley, T. R., & Howat, I. M. (2024). pDEMtools: Conveniently search, download, and process ArcticDEM and REMA products. *Journal of Open Source Software*, *9*(102), 7149. https://doi.org/10.21105/joss.07149

Copernicus. (2024). *Copernicus Data Space Ecosystem | Europe's eyes on Earth*. https://dataspace.copernicus.eu/.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, *202*, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Haan, S., Harianto, J., Butterworth, N., & Bishop, T. (2023). Geodata-harvester: A python package to jumpstart geospatial data extraction and analysis. *Journal of Open Source Software*, *8*(89), 5205. https://doi.org/10.21105/joss.05205

Hogenson, K., Kristenson, H., Kennedy, J., Johnston, A., Rine, J., Logan, T., Zhu, J., Williams, F., Herrmann, J., Smale, J., & Meyer, F. (2020). *Hybrid pluggable processing pipeline (HyP3): A cloud-native infrastructure for generic processing of SAR data*. Zenodo. https://doi.org/10.5281/zenodo.4646138

Höhl, A., Ofori-Ampofo, S., Fernández-Torres, M.-Á., Kuzu, R. S., & Zhu, X. X. (2025). CropClimateX: A large-scale, multitask, multisensory dataset for crop monitoring under climate extremes. *Hugging Face*. https://doi.org/10.57967/hf/5047

Höhn, P., Heidler, K., Behling, R., & Zhu, X. X. (2025). *Sen12Landslides - a spatio-temporal dataset for satellite-based landslide detection*. Hugging Face. https://doi.org/10.57967/hf/5883

Mahecha, M. D., Gans, F., Brandt, G., Christiansen, R., Cornell, S. E., Fomferra, N., Kraemer, G., Peters, J., Bodesheim, P., Camps-Valls, G., Donges, J. F., Dorigo, W., Estupinan-Suarez, L. M., Gutierrez-Velez, V. H., Gutwin, M., Jung, M., Londoño, M. C., Miralles, D. G., Papastefanou, P., & Reichstein, M. (2020). Earth system data cubes unravel global multivariate dynamics. *Earth System Dynamics*, *11*(1), 201–234. https://doi.org/10.5194/esd-11-201-2020

Microsoft Open Source, McFarland, M., Emanuele, R., Morris, D., & Augspurger, T. (2022). *Microsoft/PlanetaryComputer: October 2022* (Version 2022.10.28). Zenodo. https://doi.org/10.5281/zenodo.7261897

Montero, D., Aybar, C., Ji, C., Kraemer, G., Söchting, M., Teber, K., & Mahecha, M. D. (2024). *On-demand earth system data cubes*. https://doi.org/10.48550/ARXIV.2404.13105

Taconet, P., & Moiroux, N. (2024). Modisfast: An r package for fast and efficient access to MODIS, VIIRS and GPM earth observation data. *Journal of Open Source Software*, *9*(103), 7343. https://doi.org/10.21105/joss.07343

Union of Concerned Scientists. (2025). *Satellite Database*. https://www.ucsusa.org/resources/satellite-database