

---

# M3LEO - Supplementary Material

---

**Matt Allen**  
University of Cambridge, UK  
mja78@cam.ac.uk

**Francisco Dorr**  
Independent, Argentina  
fran.dorr@gmail.com

**Joseph A. Gallego-Mejia**  
Drexel University, USA  
jagallegom@unal.edu.co

**Laura Martínez-Ferrer**  
Universitat de València, Spain  
laura.martinez-ferrer@uv.es

**Anna Jungbluth**  
European Space Agency, Climate Office, UK  
anna.jungbluth@esa.int

**Freddie Kalaitzis**  
University of Oxford, UK  
freddie.kalaitzis@cs.ox.ac.uk

**Raúl Ramos-Pollán**  
Universidad de Antioquia, Colombia  
raul.ramos@udea.edu.co

## 1 S.1 Access

### 2 S.1.1 Code

3 Code for our framework is available at [github.com/spaceml-org/M3LEO](https://github.com/spaceml-org/M3LEO) and is published under  
4 the Creative Commons BY-SA 4.0 license.

### 5 S.1.2 Data

6 The **M3LEO-miniset** is available at [huggingface.co/M3LEO-miniset](https://huggingface.co/M3LEO-miniset) and the full version of  
7 **M3LEO** at [huggingface.co/M3LEO](https://huggingface.co/M3LEO).

8 **Metadata** Croissant metadata for each AOI can be found at the links given in Table S.1. Croissant  
9 metadata is automatically generated and may briefly be unavailable following dataset updates.

10 **License** Data is made available under the Creative Commons BY-SA 4.0 license. The authors bear  
11 all responsibility in case of violating the rights under which the component datasets were originally  
12 published.

13 **Reading** The repository linked in Section S.1.1 facilitates the reading of the **M3LEO** dataset and  
14 contains documented examples under the notebooks folder.

15 Table S.1: **Links to croissant metadata** for each AOI in **M3LEO**. Croissant metadata may be  
briefly unavailable following data updates due to automatic generation.

AOI	Full	Miniset
CONUS	M3LEO/conus/croissant*	miniset/conus/croissant
Europe	M3LEO/europe/croissant*	miniset/europe/croissant
China	M3LEO/china/croissant*	miniset/china/croissant
S. America	M3LEO/southamerica/croissant*	miniset/southamerica/croissant
Middle East	M3LEO/middleeast/croissant*	miniset/middleeast/croissant
PAKIN	M3LEO/pakin/croissant*	miniset/pakin/croissant

\* Some components of the full-size dataset are undergoing reformatting to support automatic croissant metadata generation. Croissant metadata will be in place at these links after completion (prior to publication).

## 16 S.2 Documentation/Intended Uses

17 We provide a datasheet for **M3LEO** following [1] in Appendix SA.1.

## 18 References

19 [1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach,  
20 Hal Daumé III, and Kate Crawford. Datasheets for datasets, 2021.

### Motivation

**For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to enable Machine Learning (ML) research on Earth Observation (EO) data at large scales, with a particular view to including interferometric SAR datatypes in addition to more commonly used Sentinel-2 and Sentinel-1 amplitude data.

**Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

This dataset was created during the Frontier Development Lab Europe (<https://fdleurope.org>) 2023 sprint, a public / private partnership between the European Space Agency (ESA), Trillium Technologies, the University of Oxford and leaders in commercial AI supported by Google Cloud and Nvidia.

**Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.

This work was primarily supported by Frontier Development Lab Europe, a public / private partnership between the European Space Agency (ESA), Trillium Technologies, the University of Oxford and leaders in commercial AI supported by Google Cloud and Nvidia. L.M-F. was supported by the European Research Council (ERC) Synergy Grant “Understanding and Modelling the Earth System with Machine Learning (USMILE)” under the Horizon 2020 research and innovation programme (Grant agreement No. 855187). M. J. A. was supported by the UKRI Centre for Doctoral Training in Application of Artificial Intelligence to the study of Environmental Risks [EP/S022961/1].

**Any other comments?**

N/A

### Composition

**What do the instances that comprise the dataset represent (e.g., documents, pho-**

**tos, people, countries)?** Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Each instance comprises a scalar field (image) describing data from a  $4.48 \times 4.48$ km patch, either input data (Optical imagery, SAR, DEMs) or from a labelled dataset (Vegetation, land cover use, built surface).

**How many instances are there in total (of each type, if appropriate)?**

A total of 10, 105, 477 data chips are currently generated, with 4, 195, 320 corresponding to labelled data and 5, 910, 157 for input datasets

**Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**

If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

Coverage of component datasets is limited to match the coverage of GUNW, hence all datasets except GUNW are subsets of the original products, which may not be representative of global coverage. This decision was made to ensure all tiles had aligned interferometric data, in addition to reducing storage requirements.

**What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features?** In either case, please provide a description.

S1GRD and S2RGB data consists of data chips (images) tiled from products as-available from the European Space Agency. SRTM consists of chips tiled from the NASA Shuttle Radar Topographic Mission data available via Google Earth Engine. GSSIC and GUNW are derived from paired Sentinel-1 acquisitions by interferometry, performed by the Alaska Satellite Facility. AGB, MODISVEG, GHSBUILTS contain per-pixel regression labels for each task. ESAWC comprises labels for 11-class semantic segmentation.

**Is there a label or target associated with each instance?** If so, please provide a description.

Each geographic tile is associated with 4 labelled datasets - AGB, MODISVEG, GHSBUILTS and ESAWC. See previous answer for descriptions.

**Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Some channels may be intermittently missing from S1GRD data - for example, ascending acquisitions are more common than descending. This is due to the particular orbit of the satellites. Not all date-pair-season combinations are available for every tile of GSSIC and GUNW, as acquisitions at the required time points cannot be guaranteed by the orbit of the satellites.

**Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** If so, please describe how these relationships are made explicit.

Relationships between data chips are implied by the use of identical geographic tiling across all component datasets.

**Are there recommended data splits (e.g., training, development/validation, testing)?** If so, please provide a description of these splits, explaining the rationale behind them.

We provide a split based on repeated geographic banding, described in the appendices to the main text in detail. We choose to use repeated geographic banding rather than random selection or contiguous train/validation/test sets to balance data leakage and train-test distribution shift respectively. We encourage users to create data splits appropriate for their own applications.

**Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.

A very small number of chips (low 100s) per AOI appear to have some channels with a constant value (e.g. all 0) but do not indicate that this data is missing in the metadata. We have not verified whether this is certainly an error.

**Is the dataset self-contained, or does it link to or otherwise rely on external**

**resources (e.g., websites, tweets, other datasets)?** If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self contained.

**Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)?** If so, please provide a description.

No

**Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** If so, please describe why.

No

**Does the dataset relate to people?** If not, you may skip the remaining questions in this section.

No

**Does the dataset identify any subpopulations (e.g., by age, gender)?** If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

N/A

**Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** If so, please describe how.

N/A

**Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; bio-**

**metric or genetic data; forms of government identification, such as social security numbers; criminal history)?** If so, please provide a description.

N/A

**Any other comments?**

N/A

### Collection Process

**How was the data associated with each instance acquired?** Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

We tiled each dataset from existing products described within the main text. Tile bounds are geographically identical across products.

**What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?** How were these mechanisms or procedures validated?

Datasets available via Google Earth Engine (S2RGB, S1GRD, SRTM, ESAWC, MODISVEG) were tiled using [geetiles](https://github.com/rramosp/geetiles)<sup>1</sup>. The remaining datasets (GSSIC, GUNW, AGB, GHSBUILTS) were tiled using [sartiles](https://github.com/rramosp/sartiles)<sup>2</sup>.

**If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

Data was sampled to match the coverage of GUNW interferograms.

**Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

Data was processed by researchers funded by the grants previously outlined.

<sup>1</sup><https://github.com/rramosp/geetiles>

<sup>2</sup><https://github.com/rramosp/sartiles>

**Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?** If not, please describe the timeframe in which the data associated with the instances was created.

Data was downloaded/tiled in Summer 2023. The data from component datasets is from 2020 with the exceptions of SRTM which was measured in 2000 and MODISVEG which was measured yearly 2016-2020.

**Were any ethical review processes conducted (e.g., by an institutional review board)?** If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No

**Does the dataset relate to people?** If not, you may skip the remaining questions in this section.

No

**Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

N/A

**Were the individuals in question notified about the data collection?** If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

**Did the individuals in question consent to the collection and use of their data?**

If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

**If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future**

**or for certain uses?** If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

**Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?** If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

**Any other comments?**

N/A

### Preprocessing/cleaning/labeling

**Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Data was tiled from the original products, which are linked within the main text.

**Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** If so, please provide a link or other access point to the “raw” data.

No, as this data is already publicly available.

**Is the software used to preprocess/clean/label the instances available?** If so, please provide a link or other access point.

Yes - we provide two tools, geetiles and sartiles, for doing this.

**Any other comments?**

N/A

### Uses

<sup>3</sup><https://arxiv.org/abs/2310.00826>

<sup>4</sup><https://arxiv.org/abs/2310.00119>

<sup>5</sup><https://arxiv.org/abs/2310.03513>

<sup>6</sup><https://arxiv.org/abs/2310.02048>

**Has the dataset been used for any tasks already?** If so, please provide a description.

We provided baselines for the ESAWC, AGB and GHSBUILTS tasks in the main text. This data has been used in previous work on self-supervised learning applied to SAR data<sup>3456</sup>.

**Is there a repository that links to any or all papers or systems that use the dataset?** If so, please provide a link or other access point.

Not currently. The M3LEO framework repository [github.com/spaceml-org/M3LEO](https://github.com/spaceml-org/M3LEO) may be updated with this information in the future.

**What (other) tasks could the dataset be used for?**

Any labels available via Google Earth Engine or in GeoTIFF/NetCDF format can be integrated with the data we provide, although users will need to supply their own compute for tiling.

**Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

Due to the large volume of chips, we have not checked the fidelity of the data for each chip by hand. Some component labelled datasets are based on tertiary remote sensing data and are not ground-verified (i.e no extensive in-person surveys have been conducted to verify labels).

**Are there tasks for which the dataset should not be used?** If so, please provide a description.

We cannot guarantee the accuracy of any of our data. Users should take extreme caution if using this data for highly sensitive/high impact real world applications, or otherwise supply their own ground-verified data (which can be integrated with our framework/data).

**Any other comments?**

N/A

N/A

### Any other comments?

N/A

### Distribution

**Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** If so, please provide a description.

The data is open-access.

**How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?**

The public-facing datasets are distributed via HuggingFace<sup>78</sup>.

**When will the dataset be distributed?**

All data advertised within the text will be available on or prior to October 30<sup>th</sup> 2024. Data for the **M3LEO-miniset** is already in-place excepting some GUNW date-pairs, described in the appendices to the main article.

**Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

Data is distributed under the CC BY-SA 4.0 license.

**Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

N/A

**Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

<sup>7</sup>[huggingface.co/M3LEO](https://huggingface.co/M3LEO)

<sup>8</sup>[huggingface.co/M3LEO-miniset](https://huggingface.co/M3LEO-miniset)

### Maintenance

**Who will be supporting/hosting/maintaining the dataset?**

Public-facing data is hosted by HuggingFace and maintained by the researchers. A private repository is maintained via the Google Cloud Platform.

**How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

Email addresses are provided at the beginning of this supplementary material.

**Is there an erratum?** If so, please provide a link or other access point.

There is not currently an erratum. Future errata will be listed in the repositories linked previously.

**Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

New component datasets may be distributed via the HuggingFace repositories.

**If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?** If so, please describe these limits and explain how they will be enforced.

N/A

**Will older versions of the dataset continue to be supported/hosted/maintained?** If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Updates will be communicated via the HuggingFace repositories. We do not plan to maintain old versions of the dataset.

**If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We provide geotiles and sartiles for users who wish to augment the dataset with their own data. We do not plan to provide distribution for user contributions.

**Any other comments?**

N/A