Article

Not peer-reviewed version

Global Overview of Usable Landsat and Sentinel-2 Data for 1982–2023

Katarzyna Lewińska * , Stefan Ernst , David Frantz , Ulf Leser , Patrick Hostert

Posted Date: 21 February 2024

doi: 10.20944/preprints202402.1169.v1

Keywords: Data availability; aggregation; long-term analyses; terrestrial, vegetation, time series

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2024 doi:10.20944/preprints202402.1169.v1

Article

Global Overview of Usable Landsat and Sentinel-2 Data for 1982–2023

Katarzyna Ewa Lewińska [1,2,*], Stefan Ernst [1], David Frantz [3], Ulf Leser [4] and Patrick Hostert [1,5]

[1] Geography Department, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany
[2] SILVIS Lab, Department of Forest and Wildlife Ecology, University of Wisconsin-Madison, 1630 Linden Drive, Madison, WI 53706, USA
[3] Geoinformatics—Spatial Data Science, Trier University, Behringstraße 21, 54296 Trier, Germany
[4] Department of Mathematics and Computer Science, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany
[5] Integrative Research Institute on Transformations of Human-Environment Systems (IRI THESys), Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany
* Correspondence: lewinska@hu-berlin.de

Abstract: Landsat and Sentinel-2 acquisitions are among the most widely used medium-resolution optical data adopted for terrestrial vegetation applications, such as land cover and land use mapping, vegetation condition and phenology monitoring, and disturbance and change mapping. When combined, both data archives provide over 40 years, and counting, of continuous and consistent observations. Although the spatio-temporal availability of both data archives is well-known at the scene level, information on the actual availability of cloud-, snow-, and shade-free observations at the pixel level is lacking and should be explored individually for each study to correctly parametrize subsequent analyses. However, data exploration is time and resource consuming, thus is rarely performed a-priori. Consequently, the spatio-temporal heterogeneity of usable data is often inadequately accounted for in the analysis design, risking ill-advised selection of algorithms and hypotheses, and thus inferior quality of final results. Here we present precomputed data on the daily 1982-2023 availability of usable Landsat and Sentinel-2 acquisitions across the globe, sampled at a pixel-level in a regular 0.18°-point grid. The dataset comprises separate Landsat- and Sentinel-2-specific data records, and is accompanied by a pixel-specific growing season information to facilitate data exploration. The dataset was derived based on freely available 1982-2023 Landsat surface reflectance (Collection 2) and Sentinel-2 top-of-the-atmosphere reflectance (pre-Collection-1 and Collection-1) scenes from 2015 through 2023, following the methodology developed in the recent study on data availability over Europe [1]. Growing season information was derived based on 2001-2019 time series of the yearly 500 m MODIS land cover dynamics product (MCD12Q2; Collection 6) [1]. As such, the dataset presents a unique overview of the spatio-temporal availability of usable daily Landsat and Sentinel-2 data at the global scale, hence offering much needed a-priori information aiding identification of appropriate methods and challenges for terrestrial vegetation analyses at local to global scale.

Keywords: data availability; aggregation; long-term analyses; terrestrial; vegetation; time series

Specifications Table

Subject	Remote Sensing; Earth-Surface Processes; Big Data Analytics.
Specific subject area	Pixel-level global overview of available of cloud-, snow-, and shade-free Landsat and Sentinel-2 observations for terrestrial vegetation analyses
Data format	Analysed
Type of data	Tabulated data distributed as .csv

Data collection	We based our dataset on satellite data available freely and openly in the public domain. See section on *Data source location*. The used satellite acquisitions were spatially subsampled using a regular 0.18° x 0.18° grid defined in the EPSG:4326 projection and tabulated.
Data source location	Landsat data (Collection 2, doi: 10.5066/P918ROHC [2], doi: 10.5066/P9TU80IG [3], doi: 10.5066/P975CC9B [4]) are freely and openly available in the public domain. We accessed Landsat reflectance Level 2, Tier 1 scenes acquired from 1982 through 2023 through Google Earth Engine in December 2022 – January 2023 and in January-February 2024. Sentinel-2 data (pre-Collection-1 doi: 10.5270/S2_-d8we2fl [5], and Collection-1 doi: 10.5270/S2_-742ikth [6]) are freely and openly available in the public domain. We accessed Sentinel-2 top-of-atmosphere (TOA) reflectance Level-1C scenes acquired between 26 June 2015 and 31 December 2023 through Google Earth Engine in July – November 2023 and in January-February 2024. MODIS land cover dynamics product at 500-m resolution (MCD12Q2; Collection 6 doi: 10.5067/MODIS/MCD12Q2.061) is freely and openly available in the public domain. We access the 2001-2019 time series of data through Google Earth Engine in July 2023.
Data accessibility	Tabular data on 1982-2023 global availability of usable Landsat and Sentinel-2 observations, accompanied by growing season information are publicly available for download in a data repository: Repository name: Dryad Data identification number: doi.org/10.5061/dryad.gb5mkkwxm Direct URL to data: https://doi.org/10.5061/dryad.gb5mkkwxm (will be made publicly available upon acceptance of the paper) Rasterized version of the tabular data on 1982-2023 global data availability based on Landsat and Sentinel-2 archives can be interactively viewed via Google Earth Engine App: https://katarzynaelewinska.users.earthengine.app/view/worlddataaval

Value of the Data

- Understanding data availability is crucial for the appropriate selection and parametrization of algorithms used for terrestrial vegetation analyses. Yet, a-priori data exploration is rarely performed due to its high resource and time requirements. The lack of appropriate understanding of data availability can lead to ill-advised selection of algorithms and poorly framed research hypotheses, and thus inferior quality of the final results. Our dataset provides a ready-to-use, pixel-level global overview of the spatio-temporal availability of cloud-, snow-, and shade-free Landsat and Sentinel-2 observations from 1982 through 2023, allowing for informed decision-making for analyses relying on datasets based on these two data archives.
- The dataset comprises information on the availability of cloud-, snow-, and shade-free Landsat and Sentinel-2 pixels sampled daily for 1984-2023 in a regular 0.18°-point grid at the global scale. Consequently, a user can easily query data availability for their specific area of interest and time window. As such, the dataset facilitates parametrization of time series processing algorithms, selection of optimal length of compositing windows, evaluation of data availability for spectral-temporal metrics, land cover classification, trend analyses, and other analysis specific to terrestrial vegetation.
- The dataset provides separately availability information for the Landsat (1982-2023) and Sentinel-2 (2015-2023) data archives. The corresponding structure of the two

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2024 doi:10.20944/preprints202402.1169.v1

tabulated files comprising the data allows for seamless integration, while catering to users utilizing only one of the data archives. Furthermore, this separation allows for a straightforward assessment of the added value of joint use of Landsat and Sentinel-2 archives after 2015, as compared to Landsat or Sentinel-2 time series alone.

- The pre-calculated overview of usable data provides insight into the quality of formerly derived datasets and results based on Landsat and/or Sentinel-2 time series that lack explicit data-availability quality assessment.
- The accompanying Google Earth Engine App (https://katarzynaelewinska.users.earthengine.app/view/worlddataaval) offers on-the-fly querying of the datasets. Provided functionality allows exploration of the data availability for a selected sensor constellation, using a user-defined length of aggregation period, and allowing to choose an entire calendar year, a vegetation-specific growing season, or other user-defined time periods. As such, the App provides an interface with a basic data query functionality for exploring Landsat and Sentinel-2 data availability that is designed to be used by a wide range of user groups.

Background

While developing analysis workflows based on Landsat and Sentinel-2 time series, we realised that to properly parametrize algorithms for vegetation analysis we need a-priori information on data availability on annual and multi-annual basis. Often, parametrization choices are made based on educated guesses, trial-and-error, or 'expert knowledge' of the availability of satellite data acquisitions. However, many regions are prone to frequent cloud and snow cover, and different observation capacities due to limited download or on-board storage limitations. These factors inflict lower availability of usable data comparing with the theoretically possible availability arising from data acquisition frequency. Specifically, inexperienced users often struggle to find a suitable parameterisation of their analysis workflows. Recognizing the existing information gap, we decided to derive a global 1982-2023 overview of cloud-, snow-, and shade-free Landsat and Sentinel-2 observations for a regular 0.18° x 0.18°-point grid [7]. The dataset provides a readily available overview of the usable data coverage, thus supporting, for example, the informed selection of algorithms and compositing windows, and aiding the parametrization of specific vegetation-focused analyses. The dataset builds upon our previous study on availability of usable Landsat and Sentinel-2 data over Europe [1], now providing global coverage and extending the time series through 2023.

Data Description

The article describes the dataset in the linked repository, which comprises the 1982-2023 global overview of daily availability of cloud-, snow-, and shade-free free Landsat and Sentinel-2 observations [7]. The data were sampled over land at the pixel level in a regular 0.18° x 0.18°-point grid defined in the EPSG:4326 projection and span area between -179.8867°W and 179.5733°E and -59.05167°S and 83.50834°N, totalling 475,150 points. The complete dataset in the linked repository consists of three files comprising pixel-level daily data availability information for i) 1982-2023 Landsat and ii) 2015-2023 Sentinel-2 time series, as well as iii) auxiliary growing season information distributed as a mask in two variants, i.e., for a regular and a leap year (Table 1).

Table 1. Datasets shared through the linked repository [7].

File name	Explanation
GLOBAL_LND_1982-2023_CSO.csv	Daily data availability derived from Landsat 1982-2023 archives
GLOBAL_S2_2015-2022_CSO.csv	Daily data availability derived from Sentinel-2 2015-2023 archives
GLOBAL_GrowingSeason.csv	Growing season information for normal and leap years
README.md	Text file containing basic information on the distributed datasets

Each dataset is distributed in a tabulated format (.csv) and consists of 475,150 observations representing the global sample-point grid. Each observation is characterized by a unique identifier and coordinates (Table 2). The binary information on availability of cloud-, snow-, and shade-free observation (i.e., 1 – valid observation; 0 – no data, or invalid observation) is given for the *GLOBAL_LND_1982-2023_CSO.csv* and *GLOBAL_S2_2015-2023_CSO.csv* files on a daily basis in variables named *L_YYYY_MM_dd* (Table 2). For the Landsat-specific dataset, the valid range of dates is 1982-08-22 through 2023-12-31, while for Sentinel-2-specific dataset the valid range of dates is 23-06-2015 through 2023-12-31. The auxiliary dataset containing information on growing season consist of two sets of variables providing daily growing season masks. The first set of variables is specific to a regular year, whereas the second set characterizes a leap year (Table 2).

The dataset is distributed in .csv format ensuring easy ingestion and facilitating manipulation in scripting languages and data processing software.

Table 2. Overview of variables available in each dataset distributed through the linked repository.

Variable	Explanation
All datasets	
id	Unique identifier
Lat	Latitude [in degrees] (EPSG:4326)
Lon	Longitude [in degrees] (EPSG:4326)
GLOBAL_LND_1982-2023_CSO.csv	
L_YYYY_MM_dd	Data availability (binary information: 1 – valid observation; 0 – no data) for a single day where YYYY indicates year, MM indicates month, and dd indicates day.
GLOBAL_S2_2015-2023_CSO.csv	
L_YYYY_MM_dd	Data availability (binary information: 1 – valid observation; 0 – no data) for a single day where YYYY indicates year, MM indicates month, and dd indicates day.
GLOBAL_GrowingSeason.csv	
Regular_MM_dd	Information on growing season (1 – within the growing season, 0 – outside the growing season) provided daily for a regular year comprising 365 days, where MM indicates month and dd day of a day of interest.
Leap_MM_dd	Information on growing season (1 – within the growing season, 0 – outside the growing season) provided daily for a leap year comprising 366 days, where MM indicates month and dd day of a day of interest.

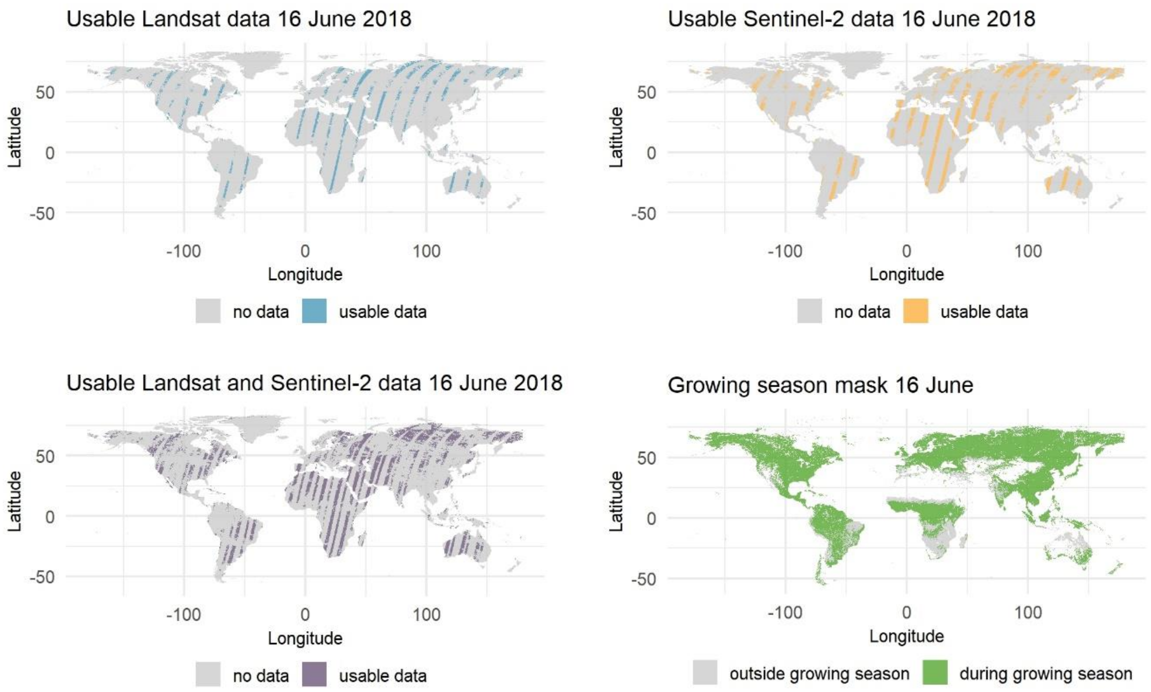

Figure 1. Global availability of usable Landsat and Sentinel-2 data. Example for the 16th June 2018 alongside with the respective growing season mask.

Furthermore, the data are also available through the Google Earth Engine App interface (https://katarzynaelewinska.users.earthengine.app/view/worlddataaval), allowing for on-the-fly interactive query based on a set of predefined criteria.

Experimental Design, Materials and Methods

We based our analyses on freely and openly accessible Landsat and Sentinel-2 data archives available in Google Earth Engine [8]. We used all Landsat surface reflectance Level 2, Tier 1, Collection 2 scenes acquired with the Thematic Mapper (TM) [2], Enhanced Thematic Mapper (ETM+) [3], and Operational Land Imager (OLI) [4] scanners between 22nd August 1982 and 31st December 2023, and Sentinel-2 TOA reflectance Level-1C scenes (pre-Collection-1 [5] and Collection-1 [6]) acquired with the MultiSpectral Instrument (MSI) between 23rd June 2015 and 31st December 2023.

We implemented a conservative pixel-quality screening to identify cloud-, snow-, and shade-free land pixels. For Landsat time series, we relied on the inherent pixel quality bands [9,10] excluding all pixels flagged as cloud, snow or shadow as well as pixels with the fill-in value of 20,000 (scale factor 0.0001; [11]). Furthermore, due to the Landsat 7 orbit drift [12] we excluded all ETM+ scenes acquired after 31st December 2020. Because Sentinel-2 Level-2A quality masks lack the desired scope and accuracy [13,14], we resorted to Level-1C scenes accompanied by the supporting Cloud Probability product. Furthermore, we employed a selection of conditions, including a threshold on Band 10 (SWIR-Cirrus), which is not available at Level-2A. Overall, our Sentinel-2-specific cloud, shadow, and snow screening comprised:

- exclusion of all pixels flagged as clouds and cirrus in the inherent ‘QA60’ cloud mask band;
- exclusion of all pixels with cloud probability >50% as defined in the corresponding Cloud Probability product available for each scene;
- exclusion of cirrus clouds (B10 reflectance >0.01);
- exclusion of clouds based on Cloud Displacement Analysis (CDI<-0.5) [15];

 doi:10.20944/preprints202402.1169.v1

- exclusion of dark pixels (B8 reflectance <0.16) within cloud shadows modelled for each scene with scene-specific sun parameters for the clouds identified in the previous steps. Here we assumed a cloud height of 2,000 m.
- exclusion of pixels within a 40-m buffer (two pixels at 20-m resolution) around each identified cloud and cloud shadow object.
- exclusion of snow pixels identified with a snow mask branch of the Sen2Cor processor [16].

Through applying the data screening, we generated a collection of daily availability records for Landsat and Sentinel-2 data archives. We next subsampled the resulting binary time series with a regular 0.18° x 0.18°-point grid defined in the EPSG:4326 projection, obtaining 475,150 points located over land between -179.8867°W and 179.5733°E and 83.50834°N and -59.05167°S. Owing to the substantial amount of data comprised in the Landsat and Sentinel-2 archives and the computationally demanding process of cloud-, snow-, and shade-screening, we performed the subsampling in batches corresponding to a 4° x 4° regular grid, and consolidated the final data in post-processing.

We derived the pixel-specific growing season information from 2001-2019 time series of the yearly 500-m MODIS land cover dynamics product (MCD12Q2; Collection 6) available in Google Earth Engine. We only used information on the start and the end of a growing season, excluding all pixels with quality below 'best'. When a pixel went through more than one growing cycle per year, we approximated a growing season as the period between the beginning of the first growing cycle and the end of the last growing cycle. To fill in data gaps arising from low quality data and insufficiently pronounced seasonality [17], we used a 5x5 mean moving window filter to ensure better spatial continuity of our growing season datasets. Following [1], we defined the start of the season as the pixel-specific 25th percentile of the 2001-2019 distribution for start of the season dates, and end of the season as the pixel-specific 75th percentile of the 2001-2019 distribution for end of the season dates. Finally, we subsampled the start and end of the season datasets with the same regular 0.18° x 0.18°-point grid defined in the EPSG:4326 projection.

Limitations

Our dataset relies on the cloud, shadow, and snow masking functionality available in Google Earth Engine. While for Landsat we relied on Fmask version 3.3.1 [9,10], for Sentinel-2 we needed to rely on an ensemble of approaches. Although all cloud detection algorithms carry certain level of uncertainty [18], our analyses are conservative and provide a valid generic overview.

Since our dataset is based on a regular 0.18° x 0.18°-point grid, the derived availability may not capture some of the local variability in cloud-, snow-, and shade-free Landsat and Sentinel-2 observations. Nevertheless, the dataset provides a robust overview of the general patterns at landscape to global scale.

The quality of our auxiliary growing season dataset is sometimes restricted due to, at times, low input data quality and insufficiently pronounced seasonality in the original MODIS time series [17].

Author Contributions: Katarzyna Ewa Lewińska: Conceptualization, Methodology, Formal analysis, Investigation, Visualization, Data curation, Writing –original draft, Visualization; **Stefan Ernst**: Data curation, Writing –review & editing; **David Frantz**: Writing –review & editing; **Ulf Leser**: Writing –review & editing, Funding acquisition; **Patrick Hostert**: Supervision, Writing –review & editing, Funding acquisition.

Informed Consent Statement: Not applicable. Our work does not involve any use of human subjects, animal experiments or data collected from social media platforms.

Acknowledgements: Funding: This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 414984028 – SFB 1404 FONDA. This research contributed to the Landsat Science Team 2018–2023.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. K.E. Lewińska, D. Frantz, U. Leser, P. Hostert, Usable Observations over Europe: Evaluation of Compositing Windows for Landsat and Sentinel-2 Time Series, Environmental and Earth Sciences, 2023. https://doi.org/10.20944/preprints202308.2174.v2.
2. Earth Resources Observation And Science (EROS) Center, Collection-2 Landsat 4-5 Thematic Mapper (TM) Level-1 Data Products, (1982). https://doi.org/10.5066/P918ROHC.
3. Earth Resources Observation And Science (EROS) Center, Collection-2 Landsat 7 Enhanced Thematic Mapper Plus (ETM+) Level-1 Data Products, (1999). https://doi.org/10.5066/P9TU80IG.
4. Earth Resources Observation And Science (EROS) Center, Collection-2 Landsat 8-9 OLI (Operational Land Imager) and TIRS (Thermal Infrared Sensor) Level-1 Data Products, (2013). https://doi.org/10.5066/P975CC9B.
5. European Space Agency, Sentinel-2 MSI Level-1C TOA Reflectance, Collection 0, (2021). https://doi.org/10.5270/S2_-d8we2fl.
6. European Space Agency, Sentinel-2 MSI Level-1C TOA Reflectance, (2022). https://doi.org/10.5270/S2_-742ikth.
7. K.E. Lewińska, S. Ernst, D. Frantz, U. Leser, P. Hostert, Global overview of cloud-, snow-, and shade-free Landsat (1982-2023) and Sentinel-2 (2015-2023) data, (2024). https://doi.org/10.5061/dryad.gb5mkkwxm.
8. N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, R. Moore, Google Earth Engine: Planetary-scale geospatial analysis for everyone, Remote Sensing of Environment 202 (2017) 18–27. https://doi.org/10.1016/j.rse.2017.06.031.
9. S. Foga, P.L. Scaramuzza, S. Guo, Z. Zhu, R.D. Dilley, T. Beckmann, G.L. Schmidt, J.L. Dwyer, M. Joseph Hughes, B. Laue, Cloud detection algorithm comparison and validation for operational Landsat data products, Remote Sensing of Environment 194 (2017) 379–390. https://doi.org/10.1016/j.rse.2017.03.026.
10. Z. Zhu, C.E. Woodcock, Object-based cloud and cloud shadow detection in Landsat imagery, Remote Sensing of Environment 118 (2012) 83–94. https://doi.org/10.1016/j.rse.2011.10.028.
11. Y. Zhang, C.E. Woodcock, P. Arévalo, P. Olofsson, X. Tang, R. Stanimirova, E. Bullock, K.R. Tarrio, Z. Zhu, M.A. Friedl, A Global Analysis of the Spatial and Temporal Variability of Usable Landsat Observations at the Pixel Scale, Front. Remote Sens. 3 (2022) 894618. https://doi.org/10.3389/frsen.2022.894618.
12. S. Qiu, Z. Zhu, R. Shang, C.J. Crawford, Can Landsat 7 preserve its science capability with a drifting orbit?, Science of Remote Sensing 4 (2021) 100026. https://doi.org/10.1016/j.srs.2021.100026.
13. L. Baetens, C. Desjardins, O. Hagolle, Validation of Copernicus Sentinel-2 Cloud Masks Obtained from MAJA, Sen2Cor, and FMask Processors Using Reference Cloud Masks Generated with a Supervised Active Learning Procedure, Remote Sensing 11 (2019) 433. https://doi.org/10.3390/rs11040433.
14. R. Coluzzi, V. Imbrenda, M. Lanfredi, T. Simoniello, A first assessment of the Sentinel-2 Level 1-C cloud mask product to support informed surface analyses, Remote Sensing of Environment 217 (2018) 426–443. https://doi.org/10.1016/j.rse.2018.08.009.
15. D. Frantz, E. Haß, A. Uhl, J. Stoffels, J. Hill, Improvement of the Fmask algorithm for Sentinel-2 images: Separating clouds from bright surfaces based on parallax effects, Remote Sensing of Environment 215 (2018) 471–481. https://doi.org/10.1016/j.rse.2018.04.046.
16. M. Main-Knorn, B. Pflug, J. Louis, V. Debaecker, U. Müller-Wilm, F. Gascon, Sen2Cor for Sentinel-2, in: L. Bruzzone, F. Bovolo, J.A. Benediktsson (Eds.), Image and Signal Processing for Remote Sensing XXIII, SPIE, Warsaw, Poland, 2017: p. 3. https://doi.org/10.1117/12.2278218.
17. M. Friedl, G. Josh, D. Sulla-Menashe, MCD12Q2 MODIS/Terra+Aqua Land Cover Dynamics Yearly L3 Global 500m SIN Grid V006, (2019). https://doi.org/10.5067/MODIS/MCD12Q2.006.
18. S. Skakun, J. Wevers, C. Brockmann, G. Doxani, M. Aleksandrov, M. Batič, D. Frantz, F. Gascon, L. Gómez-Chova, O. Hagolle, D. López-Puigdollers, J. Louis, M. Lubej, G. Mateo-García, J. Osman, D. Peressutti, B. Pflug, J. Puc, R. Richter, J.-C. Roger, P. Scaramuzza, E. Vermote, N. Vesel, A. Zupanc, L. Žust, Cloud Mask Intercomparison eXercise (CMIX): An evaluation of cloud masking algorithms for Landsat 8 and Sentinel-2, Remote Sensing of Environment 274 (2022) 112990. https://doi.org/10.1016/j.rse.2022.112990.