https://doi.org/10.5194/egusphere-2026-619
Preprint. Discussion started: 10 February 2026

Scalable Earth Observation Data Cubes for Advanced Analytics of Dynamic Earth Surface Processes: An Open-Source Package for Customized Processing of Sentinel-2 Data on HPCs and Beyond

Baturalp Arisoy[1], Florian Betz[1], Georg Stauch[2], Doris Klein[3], Stefan Dech[3,1], Tobias Ullmann[1]

[1]Earth Observation Research Cluster, Department of Remote Sensing, University of Würzburg, Würzburg, 97074, Germany
[2]Department of Geomorphology, University of Würzburg, Würzburg, 97074, Germany
[3]German Remote Sensing Data Center, German Aerospace Center, Oberpfaffenhofen, 82234, Germany

Correspondence to: Baturalp Arisoy (baturalp.arisoy@uni-wuerzburg.de)

Abstract. Earth Observation archives now encompass petabytes of multispectral imagery, yet transforming these heterogeneous collections into analysis-ready data (ARD) cubes remains a critical bottleneck. We present an open-source Python package that unifies cloud masking, co-registration, and super-resolution into a seamless Xarray-based workflow, tailored specifically to close practical gaps in ARD cube generation. Leveraging scalable high-performance computing (HPC) infrastructure, our framework delivers rapid, reproducible cube construction and incremental updates, enabling users to build or extend large time-series data cubes without reprocessing historical scenes. Besides HPCs, our package is also suitable for local processing of Sentinel-2 data. Our approach integrates (1) s2cloudless, a probabilistic cloud-masking algorithm offering user-defined thresholds to overcome the rigid limitations of the Sentinel-2 Scene Classification Layer (SCL) and STAC metadata; (2) AROSICS, a sliding-window co-registration routine that ensures sub-pixel alignment over complex, dynamic landscapes to produce smoother temporal metrics and more consistent change detection; and (3) SEN2SR, a deep-learning super-resolution model that refines all bands to 2.5 m, revealing fine geomorphic and ecological features previously obscured at native resolutions. Together, these components address three recurring ARD cube gaps in existing Xarray-based toolkits: adaptive cloud filtering, robust time-series alignment, and integrated spatial enhancement within a single, reproducible pipeline. To maximize accessibility and reuse, the package is accompanied by well documented, interactive Python notebooks that guide users through configuration, and end-to-end cube generation. Validated on the German Aerospace Center's terrabyte HPDA clusters, the pipeline runs equally well on local workstation and can be accessed at https://github.com/BaturalpArisoy/stac2cube.

1 Introduction

The rapid expansion of Earth Observation (EO) archives, driven by high-revisit missions such as Sentinel and Landsat, has produced petabyte-scale datasets that strain conventional data management and analysis workflows (Soille et al., 2018). This "Big Earth Data" concept amplifies the familiar Big Data challenges (volume, velocity, variety, veracity, value) with spatial-

https://doi.org/10.5194/egusphere-2026-619
Preprint. Discussion started: 10 February 2026

temporal constraints and application demands for near-real-time, reproducible time-series analysis (Guo et al., 2021; Sudmanns et al., 2020). As public missions continuously accumulate dense, multi-sensor imagery, researchers face two linked problems: (i) the operational burden of ingesting and standardizing huge archives, and (ii) the computational cost of producing analysis-ready, co-registered time series across large areas and long periods.

Earth Observation Data Cubes (EODCs) have emerged as an effective pattern for addressing these issues. An EODC stores EO data as multi-dimensional, gridded arrays, typically indexed by band, time, and spatial coordinates, so users can query and analyze dense time series without repeated preprocessing (Baumann et al., 2019; Dwyer et al., 2018). By delivering scene collections in an analysis-ready format, data cubes reduce friction for large-scale trend analysis, change detection, and machine learning workflows. However, building, maintaining and analyzing EODCs at scale requires substantial compute and careful pre-processing. Long archives and large footprints do not only increase storage but also memory and CPU demand, and sensor- and scene-specific effects (e.g., misalignment, variable resolution, and cloud obstruction) degrade the quality of derived time series unless properly addressed.

Cloud computing and high-performance computing (HPC) provide the scalable infrastructure needed for production-grade EODC workflows. Cloud platforms offer flexible access to storage and distributed computation, while HPC and High-Performance Data Analytics (HPDA) techniques enable parallel processing and the processing capacity required for large-area, long-time analyses (Yang et al., 2017). Despite these capabilities, practical limitations remain. Commercial cloud services often couple dataset access to proprietary APIs or usage limits, and pay-for-tier constraints can restrict the export and bulk processing required for comprehensive ARD generation.

Popular EO platforms illustrate this tradeoff. Google Earth Engine (GEE) provides an extensive catalog and a convenient processing API, but its low-code environment and platform coupling constrain custom pipelines and large exports. Microsoft's Planetary Computer offered Spatio Temporal Asset Catalogs (STAC) based, open catalog access and a more flexible programming environment, but its operational changes in 2024 highlighted the fragility of relying solely on profit-based platforms for long-term research infrastructures. These realities have encouraged an increasing interest in open-source ecosystems built around STAC, Xarray and the Open Data Cube (ODC), which enable self-hosted, transparent EODC production such as 'sits' R package (Simoes et al., 2021) but still leave methodological gaps.

We identify three recurring deficiencies in both open-source and commercial EODC tool chains that limit their utility for challenging, dynamic land surfaces (e.g., braided rivers) and for long-term time-series analysis. First, the level and definition of Analysis-Ready Data (ARD) is often underspecified and overly tailored to a single study area or surface type. Many packages simply expose Sentinel-2 scenes with default Sen2Cor processing for cloud masking, an approach known to perform sub-optimally across heterogeneous land cover, while other higher-quality products (e.g., MAJA) are not delivered as STAC collections and therefore do not integrate cleanly into STAC-driven ingestion pipelines. In practice, this means that users must accept substandard masks or build custom pre-processing outside the cube workflow, reducing reproducibility and scalability. Second, geometric inconsistencies across time are frequently neglected. Consecutive Sentinel-2 scenes are often misaligned at the sub-pixel level; while a handful of studies demonstrate co-registration on isolated image pairs, for example FORCE (Frantz,

https://doi.org/10.5194/egusphere-2026-619
Preprint. Discussion started: 10 February 2026

2019) and GeFolki (Brigot et al., 2016), to the best of our knowledge co-registration has not been systematically applied within EODC production for long time series. Without robust, automated co-registration across all time steps, pixel-level time series inherit geometric noise that corrupts temporal signals, especially in dynamic or heterogeneous areas where small positional errors translate into large radiometric changes.

Third, although deep neural network (DNN) based super-resolution methods have matured and are applied to selected scenes, they have not been implemented at the scale of entire data cubes. Most EODC projects retain the native 10 m Sentinel-2 resolution for multispectral bands, foregoing the potential benefits of bringing all bands to a common, higher spatial fidelity (e.g., 10 m → 2.5 m). As with co-registration, super-resolution is typically demonstrated on limited samples rather than integrated into a reproducible EODC pipeline that processes archives entirely.

In response, this study proposes a complete workflow implemented in an open Python package for generating analysis-ready Sentinel-2 data cubes that explicitly address these three gaps. Concretely, the package implements: (i) robust cloud masking based on s2cloudless but tuned and validated across diverse surface types; (ii) automated co-registration that enforces geometric consistency across all time steps in an EODC; and (iii) scalable cube-level super-resolution to predict higher-resolution representations and harmonize all bands to a common target resolution (default 10 m, with optional upsampling to 2.5 m). The implementation reads STAC catalogues, stores data in Xarray-labelled arrays compatible with the Open Data Cube paradigm and integrates with existing tools while keeping the full pipeline open-source and auditable.

We evaluate the methods on a highly braided, dynamic river system to quantify improvements in temporal consistency, mask quality, and downstream analysis tasks (e.g., index stability and classification). The package was developed and optimized for the "terrabyte" HPC environment (DLR / Leibniz Supercomputing Center) to demonstrate feasibility at scale, but it is fully portable to local machines or other clusters, enabling the entire research community to reproduce the workflow and apply it to other environments beyond the terrabyte user community.

2 Dataset and Study Area

2.1 STAC as Dataset

The open-source EODC structure is designed to retrieve data from an open-access data source. It is evident that a significant number of sources have made requests for Copernicus mission data, with STAC being amongst the most in demand. STAC can be defined as a standardized specification of geospatial data to describe, store and distribute geodata. As STAC is defined as open source, it is written in GeoJSON format, which is both human-readable and machine-readable.

This study compares three STAC catalogues, the terrabyte STAC API (DLR/LRZ), Microsoft's Planetary Computer, and Element84's Earth Search, each containing multiple collections, including standardized Sentinel-2 L2A datasets. In the STAC hierarchy, collections comprise items representing single spatiotemporal assets (e.g., a Sentinel-2 scene), with assets providing spectral bands and metadata such as geometry, EPSG code, platform, cloud cover, orbit state, sun azimuth, and processing baseline, which can also define EODC alignment. While the catalogues share similar structures, key differences include query

https://doi.org/10.5194/egusphere-2026-619
Preprint. Discussion started: 10 February 2026

processing speed and the number of available scenes (Table I): terrabyte offered the most up-to-date data for the German test area but cannot be downloaded without approval; Microsoft's PC was fastest; Earth Search had missing scenes and the slowest performance. Data storage locations also differ, with terrabyte hosted in Germany (Central Europe coverage complete), PC on Azure, and Earth Search on AWS.

For the present study, given that the study area is not located in Central Europe, the Earth Search Sentinel-2 L2A and Sentinel-2 L1C collections will be utilized as a reference. The primary rationale for opting for Earth Search over Microsoft's PC is that Element84 has licensed the catalog as open source, aligning with the motivation of the study.

Table 1: Key characteristics of the selected STAC API services.

Service	Provider	License/ Open-Source	STAC API	Returned Date	Processing Time (s)
terrabyte	DLR	MIT License Copyright (c) 2024 Deutsches Zentrum für Luft- und Raumfahrt e. V / Not open-source deployment	https://stac.terrabyte.lrz.de/public/api/	1,134	24.0
Earth Search	Element84	Apache License 2.0 / Open-source	https://earth-search.aws.element84.com/v1/	1,038	140.5
Planetary Computer	Microsoft	MIT License Copyright (c) Microsoft Corporation. / Not open-source deployment	https://planetarycomputer.microsoft.com/api/stac/v1	1,133	12.2

Note: Returned Date indicates the total number of available Sentinel-2 L2A acquisitions for S2 tile 32UNA (Würzburg, Germany), 2017–2025. Processing times represent the mean of ten repeated STAC API queries measured using the timeit module. All links were last accessed and verified as operational on 16 December 2025.

The Sentinel-2 L1C and Sentinel-2 L2A products are fundamental to the EODC. L1C data provides orthorectified and radiometrically calibrated Top-Of-Atmosphere (TOA) reflectances across 13 spectral bands at 10 m, 20 m, and 60 m spatial resolution, serving as the input for cloud probability estimation; the high radiometric accuracy of L1C ensures reliable discrimination between cloudy and cloud-free pixels. Conversely, L2A products deliver Bottom-Of-Atmosphere (BOA) reflectances subsequent to scene-level atmospheric correction, thereby providing surface reflectance values that are prepared for multi-temporal analysis and time-series decomposition. Both collections of STAC APIs are stored in integer format with a view to saving storage space. However, it is important to note that raw DNs do not accurately represent the actual surface reflectance values. Consequently, scale factors are applied to both missions' DNs to ensure the accuracy of the data. The primary distinction between the two collections is that the scale factor calculation is linear for L2A. However, L1C collection necessitates a dynamic scaling system due to the Sentinel-2 processing baseline. For the majority of STAC APIs, items dated after January 25th, 2022, have a baseline greater than 04.00, with the exception of certain APIs, where the processing baseline of items dated prior to that date is updated with the most current processing baseline (>05.00). Consequently, the scale factoring is performed on the basis of the metadata (attribute) of the returned STAC item.

https://doi.org/10.5194/egusphere-2026-619
Preprint. Discussion started: 10 February 2026

2.2 Study Area

The catchment of the Naryn River (Fig 1a) in Kyrgyzstan was selected as study area due to its highly dynamic nature, which qualifies it as useful case study for demonstrating the capability of the approach implemented. The Naryn River basin, enclosing an area of 52,130 km², exhibits a pronounced altitudinal gradient, ranging from an approximate elevation of 868 m in the lower reaches to a maximum of 5,133 m in the Tian Shan headwaters. In this region, perennial snow and glacier cover persists for a duration of 180–200 days, often exceeding 1 m in depth (Betz et al., 2023; Vydra et al., 2024). The region's climate is strongly continental and semi-arid, with mean annual temperatures of approximately 2.7 °C recorded at an elevation of 2,050 m, exhibiting characteristic features such as cold winters (often below –6 °C) and mild summers (up to 15 °C in July). Total yearly precipitation ranges between 280 and 450 mm, with over 75% of this falling from April to September. From a hydrological perspective, the Naryn River demonstrates a distinctive seasonal discharge pattern, characterized by a surge in monthly flows during the peak summer months of June and July, reaching approximately 235–244 cubic meters per second (m^3/s), attributable to snow and ice melt. Conversely, the river experiences a decline in flow during the winter months, from December to March, with minimum flows ranging from 47 to 92 m^3/s at various gauge sites. At the point at which the Kara Darya River meets it, the average annual discharge of water is approximately 429 m^3/s (the total annual runoff being approximately 13.7 km^3). This supports a series of reservoirs, the most significant of which is the Toktogul Reservoir, which has a capacity of 19.9 km^3. These reservoirs are vital for both the generation of hydroelectric power and the irrigation of land in Kyrgyzstan and the downstream region of Uzbekistan (Xenarios et al., 2018).

The proposed ARD EODC is motivated by the semi-arid climate of the Naryn River catchment, where low precipitation yields more cloud-free Sentinel-2 imagery, providing a strong basis for time series analysis. Its braided river morphology, with unstable channels, gravel islands, and ephemeral vegetation, offers a challenging yet valuable setting to test a customized co-registration algorithm. Furthermore, the small, heterogeneous surface patches make the area ideal for evaluating the study's super-resolution model to extract finer details beyond the native 10 m resolution.

To evaluate the methodologies outlined in the subsequent chapter, a designated segment of the river corridor has been selected for analysis (Fig 1b). This segment, encompassing an area of 59.8 km², is situated between the cities of Naryn City and Ak Tal. The topography of the area is characterized by a variety of land surfaces, centered around the Naryn river. These include a small settlement area (Kulanak), an area of agricultural activities on both sides of the river, a pond, and the foothills of a mountain range with dozens of water streams flowing down to the Naryn river. Furthermore, the riparian zone of the Naryn River is characterized by the presence of gravel bars of varying sizes, and is subject to seasonal greening events on the floodplains and on islands within the active channel. Furthermore, two areas are designated in the region of interest: a dynamic area and a non-dynamic area (Fig 2). Here, "dynamic" refers to geomorphically active terrain in which land-cover classes (e.g., water, bare sediment, vegetation) shift frequently and unpredictably. This applies to the braided river system. By contrast, the agricultural area is treated as "non-dynamic", as its physical structure and land-cover configuration remain comparatively stable despite seasonal spectral variability. The dynamic area covers the active channel of the Naryn's central catchment. This

https://doi.org/10.5194/egusphere-2026-619
Preprint. Discussion started: 10 February 2026

was derived based on Landsat data for over 25 years since 2,000 [10]. The non-dynamic area is mostly covered by agricultural land with distinctly defined boundaries. These two regions are selected to analyze the performance of our cloud masking and co-registration methods on both dynamic and non-dynamic surface characteristics.

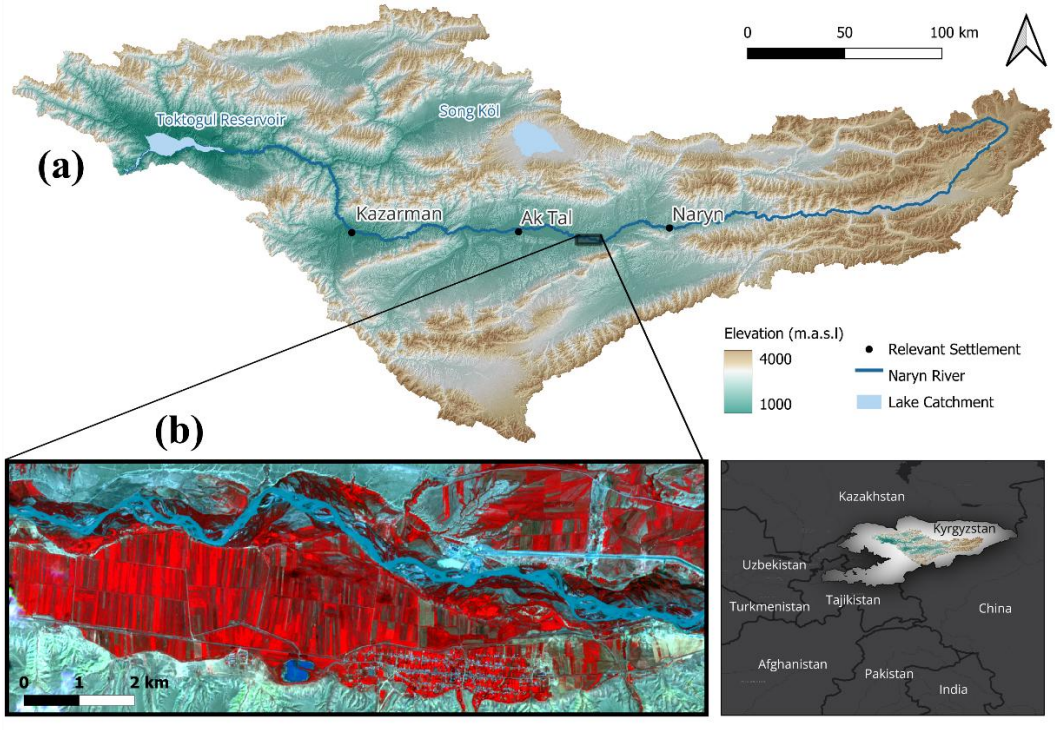


Figure 1: Naryn river catchment (a) and the region of interest (b).

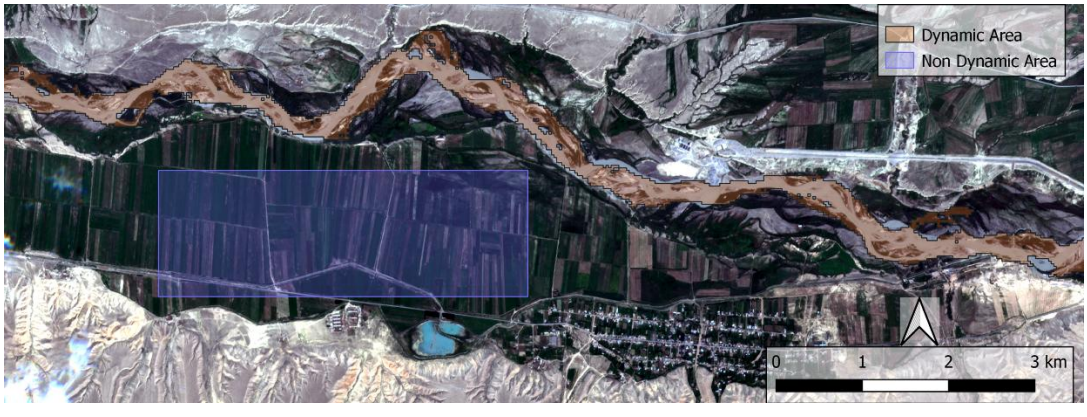


Figure 2: Selected areas within the region of interest based on the dynamism of the surface.

https://doi.org/10.5194/egusphere-2026-619
Preprint. Discussion started: 10 February 2026

3 Methods

3.1 Overview

This study implements a processing workflow for generating analysis-ready Sentinel-2 data cubes from STAC catalogues. The pipeline comprises four main stages: (i) data ingestion and preprocessing from STAC collections, (ii) cloud masking using s2cloudless or Sen2Cor SCL layers, (iii) co-registration with AROSICS to ensure spatial consistency, and (iv) deep-learning-based super-resolution with SEN2SR to enhance spatial detail. An integrated update mechanism allows newly available scenes to be incorporated efficiently without reprocessing the full dataset.

Data from freely available STAC catalogues is accessed via PySTAC and loaded using the Open Data Cube (ODC) approach. Selected items—filtered by bands, time range, cloud coverage, and resolution—are stacked into a four-dimensional array (time, band, latitude, longitude) in UTM coordinates, resampling 20 m bands to 10 m when required. In addition to spatial and dimensional metadata, attributes such as mission name, tile ID, bounding box, and derived indices are stored to enable future updates to the data cube. The software developed for this study is available on GitHub at https://github.com/BaturalpArisoy/stac2cube and includes interactive Jupyter notebooks that demonstrate each processing step and allow users to adapt the workflow to their area of interest, masking strategy, and computational environment.

Our implementation is optimized for the terrabyte High-Performance Data Analytics platform, jointly operated by the German Aerospace Center (DLR) and the Leibniz Supercomputing Center (LRZ). Terrabyte provides fast access to structured Sentinel-2 L1C and L2A STAC catalogues stored on LRZ servers and offers registered users access to CPU and GPU resources. In our Kyrgyzstan case study, a 534-scene, 6-band, 60 km² Sentinel-2 cube (resampled to 10 m) was processed and exported as a ~15 GB NetCDF file in just over five minutes using a single CPU node (40 cores, 512 GB RAM). For users without HPC access, the package supports chunked processing and the same update mechanism to extend a cube's temporal coverage without reprocessing previously ingested data.

Final outputs can be stored either as a single multi-dimensional NetCDF file or as multiple GeoTIFFs, depending on the downstream application. On terrabyte, outputs can be written to the platform's Data Science Storage (DSS) and transferred to local systems via Globus for efficient handling of large datasets.

3.2 Processing Pipeline

The processing of the data cube in this study is performed in four major steps (Fig 3). Initially, the STAC collection of S2-L2A is requested for all the available scenes in the archive. Subsequently, the necessary preprocessing steps are performed, such as scale factoring and the calculation and stacking of selected indices with the raw data along the band dimension. This resulted in the initial data cube structure, which also includes metadata about CRS and further Sentinel-2 parameters. Secondly, the selected cloud masking algorithm, s2cloudless, is applied to the S2-L1C STAC collection of all available scenes. As with the initial data cube processing, the preprocessing steps are performed, with the requirements of S2-L1C data being considered. The result of this process is the generation of an additional data cube, which contains cloud probabilities. The following step

https://doi.org/10.5194/egusphere-2026-619
Preprint. Discussion started: 10 February 2026

in the process is co-registration of S2 scenes of the cloud masked L2A, using different existing co-registration algorithms that are to be tested for the study. Finally, the cloud-free, co-registered data cube is upscaled to 2.5-metre resolution using the SEN2SR algorithm (Aybar et al., 2025). The final ARD data cube provides two major products: firstly, a time-series data cube of all scenes that are meaningful to process (i.e. cloud-free, co-registered, and with higher resolution); and secondly, another dataset that provides selected different temporal aggregations. It is important to note that both results can be updated with newly available scenes without the necessity of processing the previous scenes once again.

As the fundamental steps of data cube generation are already well documented (Baumann et al., 2021; Hoyer and Hamman, 2017), our focus is on a detailed description of the analysis-ready data cube generation, making our processing pipeline distinct from existing methods, cloud masking, co-registration, and super resolution. The details about these processing steps are explained in the subsequent sections.

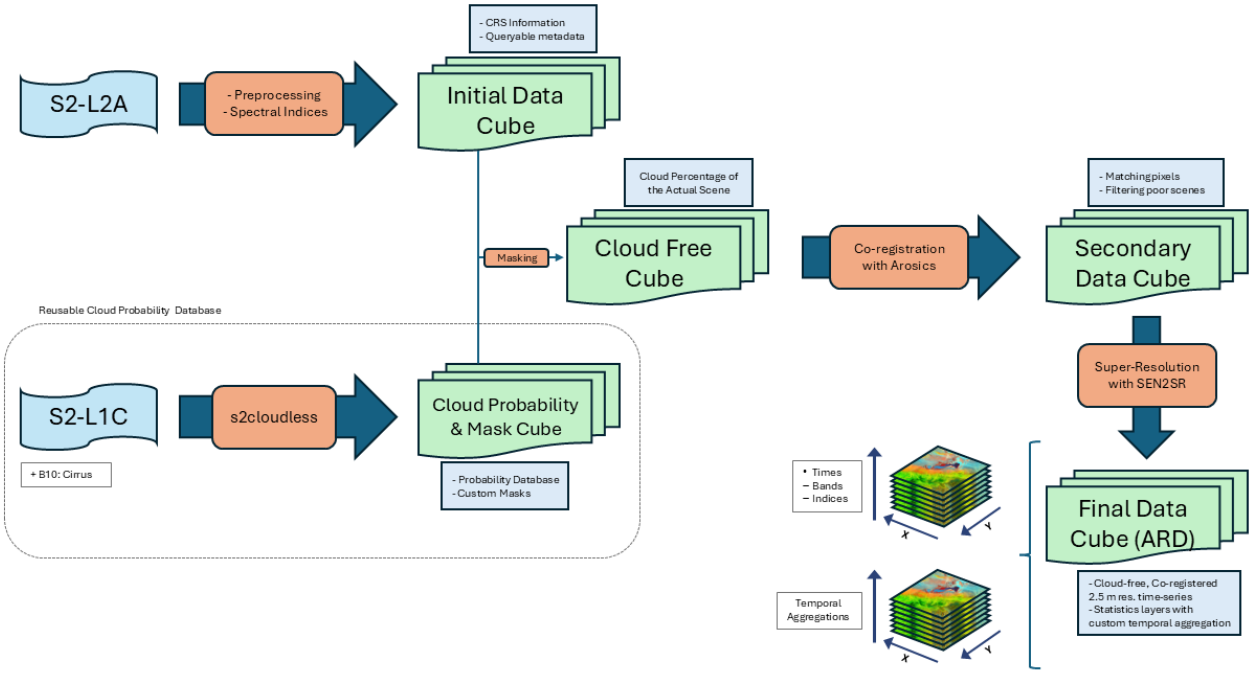


Figure 3: Methodology workflow of the study.

3.2.1 Cloud Masking

Despite cloud masking being fundamental to post-processing, no unifying solution exists. Algorithm performance is surface-dependent, and most methods are trained for specific land covers (Tarrio et al., 2020). We initially considered MAJA, FMask,

https://doi.org/10.5194/egusphere-2026-619
Preprint. Discussion started: 10 February 2026

Sen2Cor and s2cloudless; MAJA and FMask depend on full SAFE tiles and therefore do not fit a STAC-based ingestion, therefore, we here focus on Sen2Cor and s2cloudless.

Sen2Cor provides an L2A Scene Classification Layer (SCL) alongside spectral bands that produce immediate, binary masks and are straightforward to use in STAC-driven cubes. However, SCL classifications show only moderate accuracy: they may miss thin/low clouds and produce false positives over spectrally bright surfaces (e.g., gravels frequently occurring in our dynamic river case study). Crucially, SCLs are binary and offer no user control over mask strength. For this study, we selected only three cloud-related classes - high-probability cloud, medium-probability cloud, and thin cirrus – while cloud shadows and snow/ice are excluded since the method is specifically designed to detect only cloud presence.

S2cloudless, by contrast, yields per-pixel cloud probability maps from a ML model, allowing users to derive multiple binary masks via custom thresholds. This is an important feature for heterogeneous or highly dynamic AOIs where thresholding should be adaptive. S2cloudless is open-source and directly applicable to STAC collections, designed to be applied per scene, which increases computing time. S2cloudless is applied to L1C data as it requires first requires the cirrus band and secondly the ML algorithm has been trained on L1C top of atmosphere data. Thus, this necessitates scene-specific scaling (depending on the processing baseline); we derive appropriate scale factors from STAC metadata. Our implementation resamples 20 m bands to 10 m (bilinear resampling produced more realistic cloud shapes than nearest neighbour), applies s2cloudless per scene, and produces probability maps plus several thresholded binary masks based on the percentage selections by the user. The cloud outputs are stored as a separate four-dimensional cube (time, probability & masks as bands, x, y), save probability bands as integers to reduce storage, and record AOI cloud-cover percentages in cube metadata to enable efficient filtering. Given these trade-offs, s2cloudless provides the flexibility and control needed for robust, STAC-based EODC masking despite higher computational cost. Lastly, the percentage of cloud cover for the selected area of interest is additionally stored in the metadata of the data cube. This facilitates cloud-based filtering of the cube according to the area of interest, as opposed to cloud cover on the level of the entire underlying Sentinel-2 tiles.

3.2.2 Co-registration

Co-registration is essential for Sentinel-2 time-series applications because scenes commonly shift by 1–2 pixels due to geolocation inaccuracies; these shifts corrupt pixel-level time series unless corrected. Co-registration must therefore follow cloud masking, since clouds obscure stable surface features used for matching. For highly dynamic landscapes (e.g., braided rivers) where stable tie-points are scarce, we evaluated available tools and selected the open-source Python package AROSICS (Scheffler et al., 2017). AROSICS supports a fast global mode (single X/Y shift) and a slower local mode (dense tie-point grid); given the predominantly uniform Sentinel-2 shifts and the large number of scenes in an EODC, we adopt global mode for speed and scalability while retaining the option to supply custom matching windows.

We apply a custom, moving matching windows because (i) clouds are present randomly on most satellite images, so a fixed window may land on masked areas, and (ii) scanning multiple windows ensures the algorithm selects the region with the

https://doi.org/10.5194/egusphere-2026-619
Preprint. Discussion started: 10 February 2026

highest tie-point accuracy. This moving-window strategy therefore enables a fully automated system that still finds the most optimal matches across heterogeneous, partially masked scenes.

Building on the moving window logic, AROSICS is enhanced to robustly register masked, heterogeneous time series. The first scene is fixed as the master, and each subsequent scene is treated as a slave. For each slave we place an initial matching window (centered by default), evaluate its tie-point score, and, if it contains excessive no-data, iteratively reduce its size. If reduction fails, the window is shifted automatically across the scene (user-configurable number of windows), each location scored; the highest-scoring window supplies the global X/Y shift. If no window yields a reliable match, the scene is dropped from the cube. After registering all successful slaves to the master, arrays are converted back to Xarray, producing a cloud-filtered, co-registered time series.

3.2.3 Super-resolution

To further enhance the resolution of the data and enable analysis of small-scale features, a deep learning-based super resolution algorithm was implemented. This algorithm enabled the up-sampling of the 20-meter bands of Sentinel-2, particularly the red edge and shortwave infrared bands, to 10-meter resolution. Furthermore, the implementation of super-resolution enables a resolution of 2.5 meters for the final output data product. For this purpose, the "SEN2SR" algorithm is implemented, which was originally developed by (Aybar et al., 2025). The proposed algorithm is founded on a series of neural network models (CNN, Mamba, and SWIN transformer) and is engineered to generate super-resolution outputs of Sentinel-2 imagery that are both spatially and spectrally robust. The training of the models was performed on the SEN2NAIP-synthetic-v2 dataset, an extension of the original SEN2NAIP data (Aybar et al., 2024). An exhaustive evaluation of the SEN2SR algorithm, encompassing downstream tasks, is provided in (Aybar et al., 2025). In the present study, the focus is on a comparison of the generated 2.5 meters Sentinel-2 imagery with high-resolution UAV images acquired on a date close to the Sentinel-2 acquisition. The SEN2SR package offers two primary models: SENSR and SENSRLite. To maintain the simplicity of the processing, the SEN2SRLite_RGBN submodel was selected for testing on the designated study area. The package has been designed to function with Xarray-based data cubes. Therefore, the model is applied to each scene of the time-series.

4 Results

4.1 Cloud Masking

The primary distinction that favors the implementation of SCL masking concerns the application of these methodologies within the data cube framework. The implementation of SCL masking proved to be a relatively uncomplicated process. Meanwhile, a revision of the data cube structure was undertaken to facilitate the application of the s2cloudless algorithm to the scenes. The implementation of s2cloudless within our data cube structure demanded a substantial investment in research and work time. Given that SCLs are already available as STAC assets, this did not affect the processing power or the duration of the masking, resulting in nearly equivalent duration outcomes as the generation of the initial data cube outlined in Chapter 3.1. Consequently,

https://doi.org/10.5194/egusphere-2026-619
Preprint. Discussion started: 10 February 2026

the initial data cube can be generated directly with masked scenes, eliminating the need to retain the original cloud presence scenes. On the other hand, the s2cloudless algorithm must be applied to the initial data cube with clouds. Both the cloud data cube and the cloud-free data cube should be generated as additional files, increasing the size and demanding more storage space. Furthermore, the implementation of s2cloudless in every scene imposes a significant computational burden, primarily due to two factors. Initially, the 10 bands from S2-L1C are selected, and the arrays are subsequently computed. Following this, the implementation of an ML algorithm on each scene is computationally intensive, requiring approximately six hours to generate a cloud cube using the specified HPC configuration. It is noteworthy that increasing the size of the data cube polygon does not have a significant impact on the duration of applying the ML algorithm. Therefore, it is advisable to avoid selecting an area that is too small but rather selecting a larger area to retrieve cloud data. However, it should be noted that expanding the study area also results in an increase in the number of arrays that must be computed with DASK.

In contrast, the following distinctions are found to be in favor of the s2cloudless algorithm. The primary distinction between these two methods is the resultant outcomes. SCLs are classified layers, and once binary cloud composition masks the L2A scenes, the result is cloud-masked images. This approach to cloud masking is not flexible, as it does not allow the user to adjust the strength of the masking algorithm. Therefore, the user is required to use it regardless of whether non-cloud objects are masked. S2cloudless does not provide cloud masks directly; rather, it offers cloud probability maps (Fig 4a). The probability scale ranges from 0 to 100, with each scene's probability being accumulated to form another data cube. The reusability of this data cube is predicated on the capacity of the user to establish a threshold value on the probability map, thereby generating multiple binary cloud masks. It is also possible to stack these masks to generate a cloud data cube. This is since probability maps and binary masks can be aggregated under a given date. Depending on the surface characteristics and the climate of the study area, multiple thresholds can be tested to obtain the best-fitting result. Following this, the initial data cube can be masked with the selected binary cloud mask to generate a second-level, cloud-free data cube.

A comparison of the shapes and qualities of the clouds detected by the two methods reveals another distinctions (Figure 4b). In the example, SCL based masking is not entirely covering single clouds but leave unmasked cloud pixels on the edges, however, performs well on larger cloud clusters by aggregating the entire cloudy area. The performance of s2cloudless masking on the other hand, entirely depends on the threshold value. In the current example, 40% threshold performs the best by covering as many cloud pixels as possible comparing to masks with higher values. When it comes to single clouds, s2cloudless detected clouds exhibit much more realistic shape than SCL, however less performance of capturing the larger cloud cluster, instead exhibits multiple smaller fractions of masks over the cloudy area. One example of solving this issue is to lower the threshold value, meanwhile this might lead potentially false positives.

https://doi.org/10.5194/egusphere-2026-619
Preprint. Discussion started: 10 February 2026

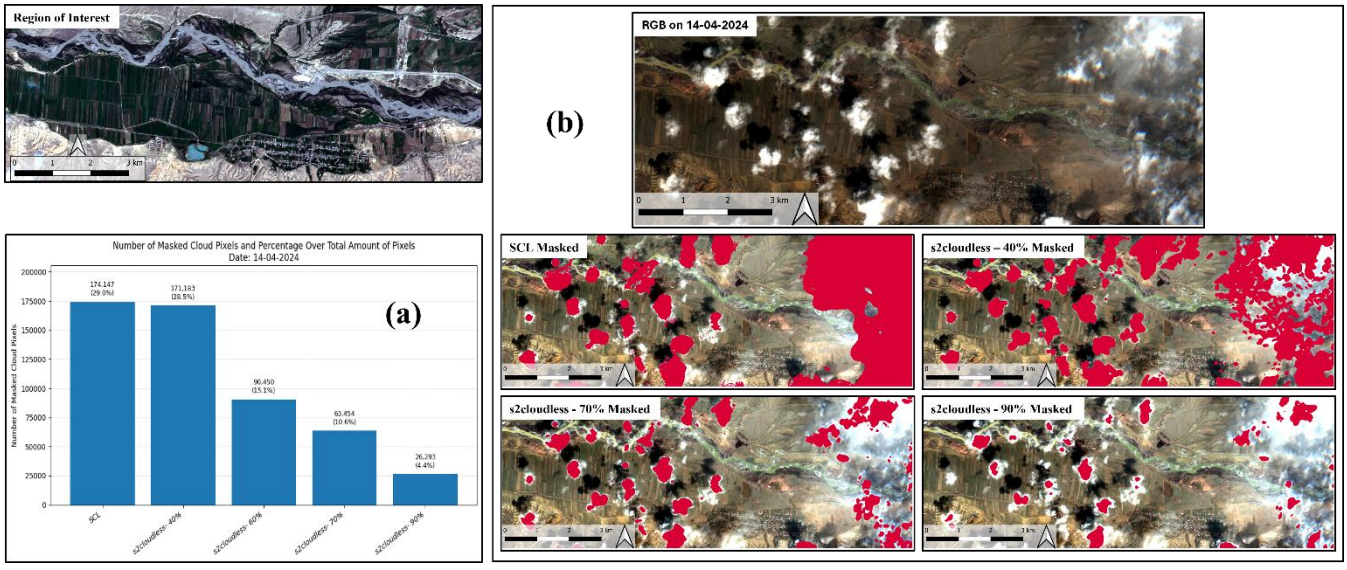


Figure 4: Comparison of the different levels of s2cloudless masks and SCL. (a) Ratio of the masked cloud pixels over the total pixels. (b) Visual comparison of different s2cloudless threshold values and SCL.

In the above example, using SCL based cloud masking might seem still preferable since the result performance is not significantly worse than an appropriate s2cloudless threshold. However, the main advantage of s2cloudless becomes clearer with the detection of false positives. Fig 5 shows two examples from different dates when the satellite scenes are entirely cloud-free. In the first example, SCL based detection masks out pixels along the river corridor which is highly undesirable when working on river systems. 40% thresholded s2cloudless masking performs false positive along the river at minimal and no more false positives at 90%. In the second example, 40% thresholded detection masks out the river channel but SCL does not mask the river but several patches on agricultural surfaces. Meanwhile, increasing the threshold values to 90% solves the false positive detection issue almost entirely by leaving only minimal masked pixels.

In all the above examples, SCL's performance is not consistent and cannot be adjusted. Even though s2cloudless based masks sometimes fail to capture the entire cloud coverage or generate false positive, the user can achieve an optimal performance of resulting the best cloud masking by changing the threshold values. This especially is feasible and efficient thanks to the reusability of the exported cloud probability and masking data cube.

https://doi.org/10.5194/egusphere-2026-619
Preprint. Discussion started: 10 February 2026

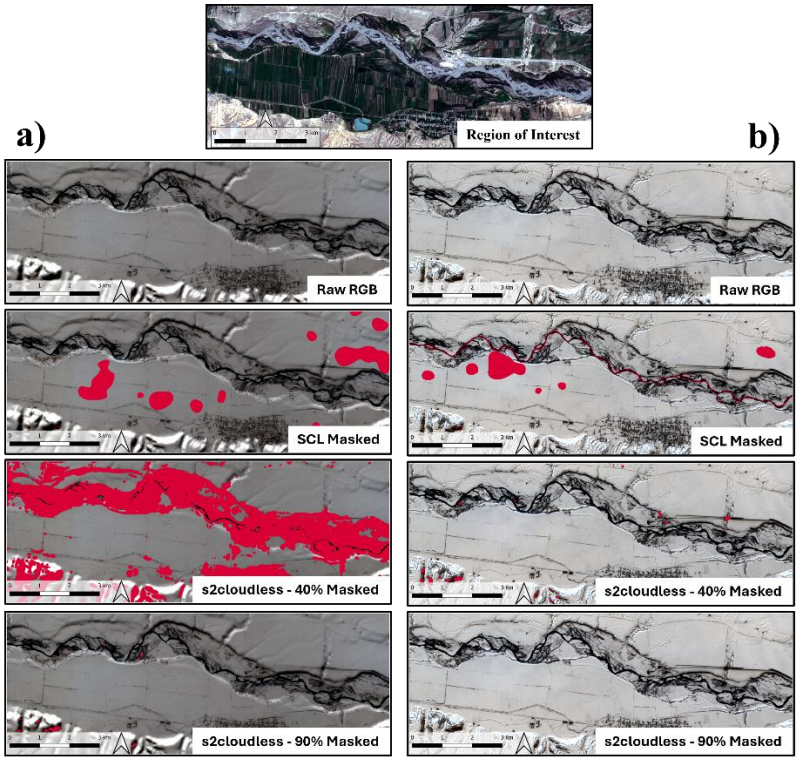


Figure 5: Examples of river channel false positive over several s2cloudless levels and SCL. (a) Cloud free scene on 29.02.2024. (b) Cloud free scene on 24.02.2024.

320 Lastly, the performance of SCL and different thresholded s2loudless masks are compared for both the dynamic and non-dynamic regions of the region of interests (Table II). For this purpose, 1000 random points are sampled for both regions shown in Fig 2, counting the number of masked pixels for SCL, s2cloudless 40%, 60%, 70% and 90% masked time series of the year 2024.

Table 2: Sensitivity of cloud masking approaches to landscape dynamicity: dynamic vs. non-dynamic area comparison.

Masking	Dynamic (%)	Non-dynamic (%)	Absolute Difference
s2cloudless – 40%	40.99	40.44	0.55
s2cloudless – 60%	39.27	39.52	0.25
s2cloudless – 70%	38.37	38.88	0.51
s2cloudless – 90%	35.61	35.78	0.17
SCL	35.28	38.01	2.73

https://doi.org/10.5194/egusphere-2026-619
Preprint. Discussion started: 10 February 2026

The comparison of masked pixel percentages between dynamic and non-dynamic areas reveals notable differences in the behaviour of cloud masking strategies. Across all s2cloudless probability thresholds, the dynamic and non-dynamic regions exhibit only marginal differences, with absolute deviations remaining below 1 percentage point. This indicates that s2cloudless produces a largely spatially consistent mask, independent of underlying land surface change or heterogeneity. In contrast, the SCL-based mask shows a substantially stronger divergence between the two areas (≈2.7 percentage points), implying that its cloud classification is more sensitive to landscape dynamics.

4.2 Co-registration

To evaluate the influence of geometric alignment on vegetation index stability, we compared NDVI time series from the year 2024 using both co-registered and non-registered Sentinel-2 image stacks, restricted to scenes with less than 10% cloud cover. The analysis was conducted separately for a non-dynamic reference area and a highly dynamic braided river floodplain (see Fig 2). In homogeneous landscapes, where NDVI varies smoothly in space, co-registration is expected to have minimal impact. In contrast, dynamic river corridors contain fine-scale mixtures of vegetation, bare sediment, and open water, producing sharp spatial reflectance transitions. In such environments, even small geometric misalignments can cause pixels to shift between surface types, resulting in substantial NDVI discrepancies.

To explicitly assess how spatial heterogeneity modulates this effect, we computed a spatial NDVI gradient index for each time step and grouped sample points into low-, mid-, and high-gradient bins. These bins distinguish homogeneous surfaces from edge-dominated environments, allowing us to quantify how strongly NDVI differences arise from spatial mixing at vegetation–water boundaries.

Fig 6 illustrates the mean absolute NDVI error (MAE) between the co-registered and non-registered image stacks over the full annual time series. The dynamic braided river corridor exhibits consistently elevated MAE values, forming a clear spatial signal along channel bars, floodplain vegetation patches, and water–vegetation boundaries. These are locations where surface composition changes rapidly in space, meaning that even sub-pixel geometric offsets cause NDVI to shift from vegetation to water or sediment, producing large apparent differences.

https://doi.org/10.5194/egusphere-2026-619
Preprint. Discussion started: 10 February 2026

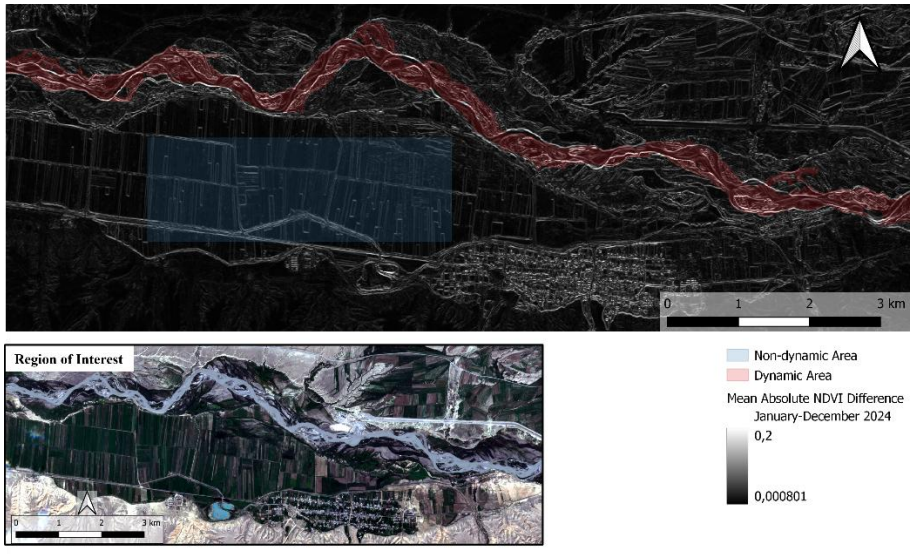


Figure 6: Effect of geometric co-registration on Sentinel-2 NDVI in non-dynamic and dynamic riverine landscapes.

In contrast, the non-dynamic area displays predominantly low MAE values, reflecting stable and spatially homogeneous vegetation cover where co-registration has little influence on NDVI. Localized increases in MAE occur along roads and field edges, where abrupt transitions between vegetation, bare soil, and infrastructure create spatial gradients similar to those in the braided river corridor. Notably, elevated MAE values are also visible in the small settlement of Kulanak, where houses are located on grassland surfaces, illustrating another potential use of our time series co-registration method. These spatial patterns confirm that the magnitude of co-registration effects is not tied to vegetation type, but instead to the degree of spatial heterogeneity and dynamicity.

Fig 7 further shows how co-registration effects vary both seasonally and spatially. In Fig 7a, the non-dynamic area maintains consistently low NDVI differences across the year, indicating that NDVI remains stable without co-registration when surface reflectance patterns are uniform. In contrast, the dynamic braided river area shows substantially higher NDVI differences, particularly during the vegetation season (late spring through early autumn), when vegetation expansion and retreat along bars and channel margins create spatially complex NDVI patterns.

Fig 7b demonstrates that these discrepancies are strongly concentrated in high-gradient regions. While both areas show a modest increase in NDVI difference from low- to mid-gradient classes, only the dynamic area exhibits a pronounced rise in the high-gradient class, corresponding to vegetation–water boundaries and bar-edge transitions. This gradient dependence indicates that co-registration errors arise primarily from sub-pixel spatial mixing rather than from actual temporal changes in vegetation condition.

Together, these results show that co-registration is critical for accurately capturing NDVI behaviour in spatially complex, morphologically active river systems, but provides comparatively little benefit in stable, homogeneous landscapes.

https://doi.org/10.5194/egusphere-2026-619
Preprint. Discussion started: 10 February 2026

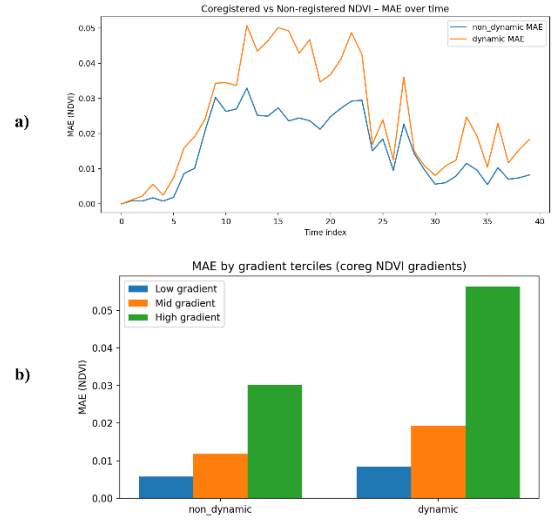


Figure 7: Effect of geometric co-registration on Sentinel-2 NDVI in non-dynamic and dynamic riverine landscapes. (a) Mean absolute NDVI difference (MAE) between co-registered and non-registered NDVI throughout the year. (b) Mean absolute NDVI difference (MAE) conditioned on spatial NDVI gradient.

To illustrate the local impact of geometric misalignment on NDVI dynamics, we extracted the NDVI time series from a single vegetation pixel located on a gravel bar adjacent to open water (Fig 8). This location represents a high-gradient environment, where small spatial shifts can move a pixel footprint across vegetation, bare sediment, and shallow water surfaces. The non-registered time series shows the expected seasonal pattern of NDVI increasing through the summer and declining towards winter. However, the temporal trajectory contains abrupt jumps and drops, including isolated high or low NDVI values that are inconsistent with the surrounding dates. These discontinuities do not reflect ecological change; rather, they arise from sub-pixel spatial shifts caused by the lack of co-registration. Even a 1–2 pixel horizontal offset is sufficient for this pixel to intermittently sample water, shadow, or sparsely vegetated surfaces, producing large artificial fluctuations.

In contrast, the co-registered time series exhibits a smooth and coherent seasonal trajectory, with gradual increases during the vegetation growth period and gradual decreases afterwards. Because co-registration ensures that the same surface location is sampled at every time step, short-term fluctuations are minimized, and the temporal progression reflects actual vegetation dynamics rather than spatial mixing. This effect is particularly noticeable during late summer peaks: for example, on August 28, the co-registered NDVI reaches values above 0.6, while the non-registered series records values below 0.4 for the same nominal pixel. A difference of ≥0.2 at the pixel level is sufficient to bias phenology metrics, temporal composites, or classification thresholds.

https://doi.org/10.5194/egusphere-2026-619
Preprint. Discussion started: 10 February 2026

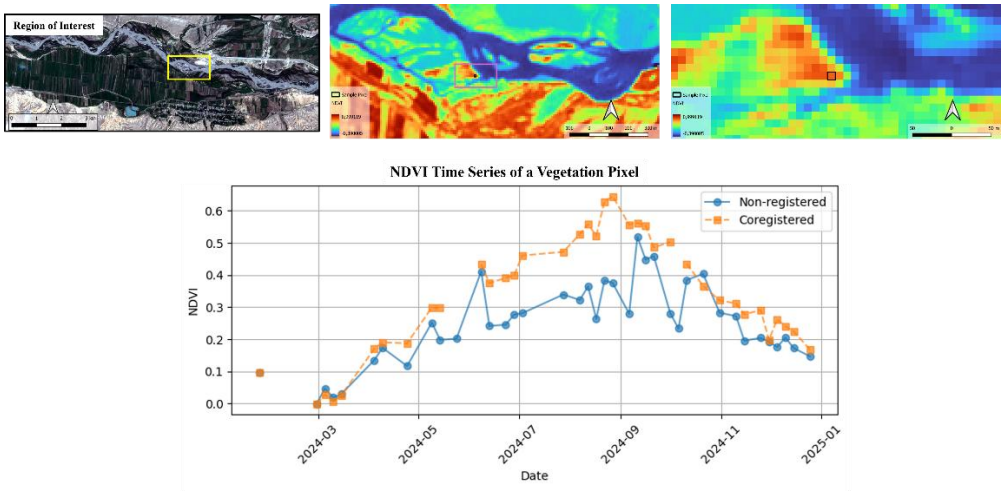


Figure 8: Co-registered and non-registered spectral profile of NDVI timeseries on a selected neighbouring pixel.

This comparison demonstrates that co-registration acts as an implicit correction for spatial aliasing, achieving a result similar to temporal low-pass filtering without smoothing or information loss. By ensuring that each pixel consistently represents the same ground location, co-registered time series yield more reliable estimates of seasonal amplitude, timing of peak greenness, and variability statistics. In settings where vegetation and water co-occur at fine spatial scales, such as braided river systems, co-registration is therefore essential for accurate NDVI-based phenology and trend analysis.

4.3 Super-resolution

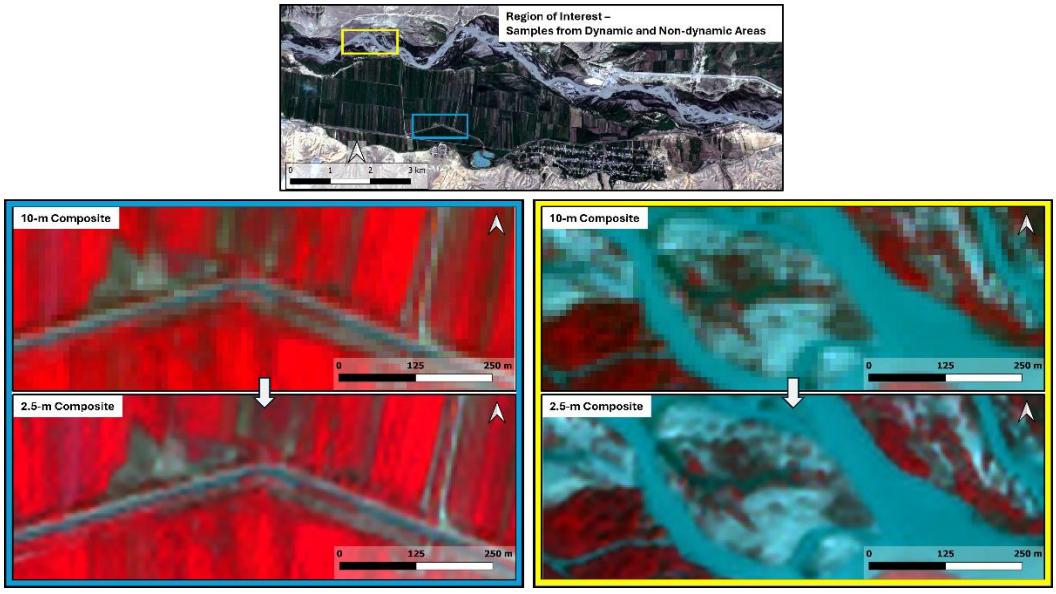
To assess the significance of the super-resolution model, the algorithm was applied to the entire region of interest, and two representative samples were selected: one from a dynamic area and another from a non-dynamic area. Fig 9 illustrates the false-color vegetation composites resulting from the super-resolution process, in which the original 10 m Sentinel-2 data were spatially enhanced to 2.5 m resolution.

The non-dynamic area sample includes multiple agricultural parcels intersected by a rural road. In the original 10 m imagery, the road appears as only one to two mixed pixels, and parcel boundaries are represented by single, spectrally mixed pixels. After super-resolution, the road structure becomes clearly delineated, comprising approximately five pixels, and the parcel boundaries appear smoother and more continuous. Although this spatial enhancement is valuable for infrastructure or cadastral mapping, such areas with uniform surface characteristics do not critically require upscaling for land-surface dynamics analyses.

In contrast, the sample from the dynamic area demonstrates the advantage of the super-resolution model much more prominently. This region contains a mixture of bare gravel, vegetation, and water surfaces. In the original imagery, individual bushes are difficult to distinguish, and narrow water streams are often spectrally mixed with gravel or vegetation pixels. The super-resolved outputs provide smoother and more accurate shoreline delineation, enabling improved river extraction. Vegetation boundaries are more clearly defined, allowing more efficient detection of small vegetation patches. Most notably,

https://doi.org/10.5194/egusphere-2026-619
Preprint. Discussion started: 10 February 2026

415 narrow water channels are now distinctly visible and separable from adjacent gravel and vegetation pixels. This improvement is particularly crucial in highly dynamic braided river systems, where rapid channel shifts may remain undetected in the original Sentinel-2 scenes due to single-pixel representation or spectral confusion.



420 **Figure 9: Examples of super-resolution results from non-dynamic and dynamic areas.**

Nevertheless, Sentinel-2 imagery is still low resolution compared to UAV imagery. Since there is UAV data available from a field campaign in May 2024, we conducted a comparison (Fig 10) between the acquired high-resolution UAV imagery (a), the super-resolved Sentinel-2 image (b), and the native 10 m Sentinel-2 image (c). The UAV mission was carried out in late May 2024, operating the drone DJI Mini 3 Pro at 500 meters. The corresponding Sentinel-2 scene was extracted from the data cube
425 for 24.05.2024, and super-resolved from its native 10 m ground sampling distance to 2.5 m. In the UAV orthomosaic, the separation between water, gravel, and vegetation is very clear, and several smaller gravel islands around the main gravel bar can be identified. By contrast, surface separation in the super-resolved scene is less distinct than in the UAV reference, particularly for the smaller gravel islands and the thin water channels separating them. Nonetheless, the shape of the main gravel island remains well defined and is clearly delineated from the river channel. In the native 10 m Sentinel-2 image, the
430 island shape appears much rougher, and surface transitions are represented by single or spectrally mixed pixels, making the smaller islands very difficult to locate.

Even though the super-resolved image provides a substantial improvement over the native Sentinel-2 scene, the vegetation cover highlights a fundamental limitation of this type of learning-based super-resolution. Individual vegetation units (e.g., bushes) are clearly visible in the 2.5 m UAV image, whereas these units appear aggregated in the super-resolved Sentinel-2
435 product. This is because the super-resolved image is generated from the original 10 m Sentinel-2 data, in which vegetation is already represented as mixed or aggregated pixels. In other words, the super-resolution model does not introduce new surface

https://doi.org/10.5194/egusphere-2026-619
Preprint. Discussion started: 10 February 2026

information but spatially enhances the patterns already present in the base imagery. To further investigate this effect, we included downscaled UAV images at 2.5 m (d) and 10 m (e) using simple nearest-neighbour resampling. At 2.5 m downscaled UAV resolution, individual vegetation patches remain clearly distinguishable, unlike in the 2.5 m super-resolved Sentinel-2 scene. At 10 m downscaled resolution, vegetation–gravel separation is still present, but the exact position of small vegetation units becomes difficult to identify. This comparison indicates that, when UAV data needs to be harmonized with satellite imagery, downscaling UAV imagery to 2.5 m is more optimal than downscaling to 10 m for preserving fine-scale vegetation structure.

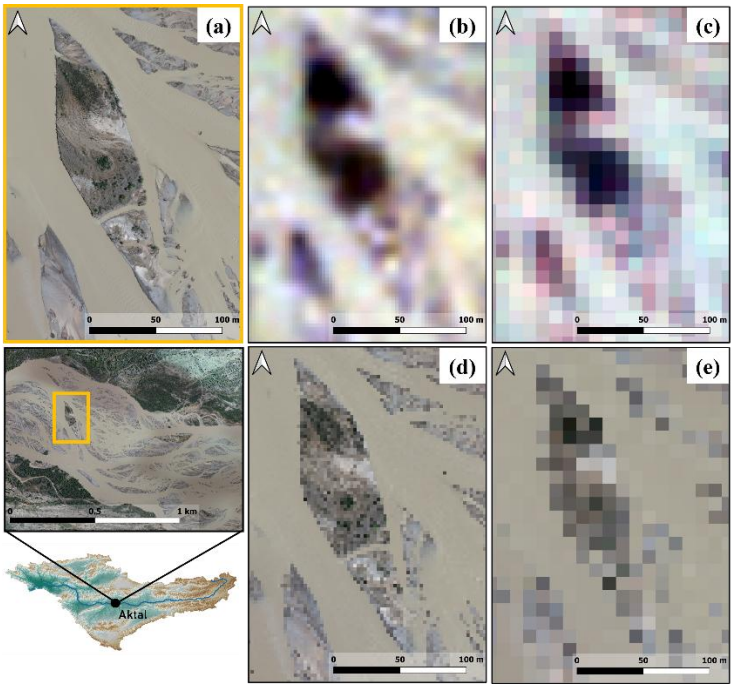


Figure 10: Spatial comparison between UAV images in different resolutions super-resolved Sentinel-2 scene and original Sentinel-2 scene. (a) RGB composite of UAV mission at 0.16 meters. (b) RGB composite of super-resolved Sentinel-2 with SEN2SR at 2.5 meters. (c) RGB composite of Sentinel-2 at 10 meters. (d) RGB composite of UAV mission at 2.5 meters, downscaled with nearest neighbour. (e) RGB composite of UAV mission at 10 meters, downscaled with nearest neighbour.

5 Discussion

The increasing availability of STAC-indexed Sentinel-2 archives has made Xarray-based Earth Observation Data Cubes a widely adopted framework for time-series analysis and environmental monitoring. However, existing open-source data cube workflows typically rely on default Sen2Cor cloud masks, assume geometric alignment across scenes, and retain the native spatial resolution of Sentinel-2 bands. These assumptions are reasonable for homogeneous landscapes but introduce major limitations in spatially heterogeneous and dynamic environments, where small-scale surface transitions amplify the impact of cloud misclassification, spatial misalignment, and mixed pixels. In particular, braided rivers contain sharp reflectance

https://doi.org/10.5194/egusphere-2026-619
Preprint. Discussion started: 10 February 2026

boundaries and temporally shifting surface conditions that make time series highly sensitive to processing decisions. Therefore, there is a growing need for ARD workflows that are flexible, reproducible, and tailored to handle fine-scale spatial heterogeneity.

The approach presented in this study provides such an alternative by building Sentinel-2 data cubes directly from STAC endpoints and implementing three key processing steps that are largely absent from current open-source EODC solutions: (1) probabilistic, threshold-adjustable cloud masking using s2cloudless; (2) scene-to-scene co-registration to enforce pixel-level temporal consistency; and (3) deep-learning-based super-resolution to enhance spatial separability of land–water and land–vegetation boundaries. These components are fully integrated in an Xarray-native workflow, preserving metadata alignment, enabling chunked computation through Dask, and allowing incremental updates of the data cube as new scenes become available. The strength of this approach is therefore not limited to its algorithmic components, but lies in how these components interact to stabilize, clarify, and densify time series in landscapes where traditional cube pipelines fail.

Our comparison of the Sen2Cor Scene Classification Layer (SCL) and s2cloudless cloud probability masks demonstrate two distinct behavioural differences. SCL offers immediate scene-level masks and integrates easily into STAC-based data ingestion, but it is constrained by rigid class definitions and cannot be tuned. This limitation is particularly relevant in spectrally complex environments: we observe consistent false positives along the gravel bars of the Naryn River corridor, reflecting confusion between bright sediment surfaces and certain cloud classes. In contrast, s2cloudless provides a continuous cloud probability field, enabling the user to select an optimal threshold that balances cloud removal against the preservation of valid observations. The threshold-tuning capability is crucial for heterogeneous landscapes because masking performance depends not only on cloud occurrence but also on surface reflectance conditions. The ability to generate multiple masks from the same probability cube allows for retrospective adjustment, masking can be designed post hoc based on surface dynamics, seasonality, or the sensitivity of specific analyses. Furthermore, the comparison of cloud-masked pixel ratios in dynamic and non-dynamic areas shows that s2cloudless is effectively insensitive to surface heterogeneity, while SCL displays systematic over-masking in the dynamic river corridor. This suggests that probabilistic cloud masking is inherently better suited for multi-habitat environments where spatial gradients are sharp and mixed pixels are common.

The results of the co-registration analysis emphasize the importance of geometric consistency for pixel-level time series. In the stable agricultural area, NDVI values change gradually across space, and co-registration has little effect, confirming that homogeneous surfaces are robust to small spatial shifts. However, in the braided river corridor, even a 1–2 pixel displacement introduces substantial NDVI variation because pixel footprints intermittently sample vegetation, gravel, or water. Our gradient-based stratification shows that the largest errors concentrate in high-gradient environments, indicating that co-registration is essential for any setting where small-scale spatial heterogeneity drives reflectance variability. This finding has broader implications for environmental monitoring. Dynamic systems, such as braided rivers, exhibit exactly these fine-scale spatial transitions. In such systems, unregistered time series will systematically overestimate short-term temporal variability while underestimating true ecological trends. Conversely, co-registration reduces artificial noise and enhances the detectability of gradual seasonal and interannual change. The pixel-scale case study (Figure 8) demonstrates this effect clearly: the co-

https://doi.org/10.5194/egusphere-2026-619
Preprint. Discussion started: 10 February 2026

registered NDVI time series follows a coherent seasonal trajectory, whereas the non-registered series shows abrupt, non-ecological fluctuations. Co-registration thus functions as a spatial filter that preserves temporal signal quality without needing temporal smoothing, preserving both seasonal amplitude and timing. From the perspective of analyzing Earth surface processes, the use of satellite time series enables new opportunities to enhance the understanding of dynamic environments such as braided rivers (Betz et al., 2023; Bozzolan et al., 2023). Land surfaces tend to be highly dynamic and heterogeneous, so studies involving time series analysis can benefit significantly from co-registration. The AROSICS algorithm provides a robust solution, even in highly dynamic systems with limited stable areas.

Studies of heterogeneous landscapes can benefit from Sentinel-2 super-resolution, as it provides greater spatial detail, for example clearer delineation of smaller side channels and improved separation of shrub patches. Although the SEN2SR algorithm is considered robust and not prone to creating artefacts (Aybar et al., 2025), the 2.5 m output is still a learning-based estimate that goes beyond the native Sentinel-2 observation. Our comparison with high-resolution UAV imagery confirms that the model sharpens existing spatial patterns (e.g., gravel bars, shorelines, narrow channels) but cannot recover individual vegetation units or very small islands that are already spectrally mixed at 10 m resolution. Users should therefore handle the 2.5 m product with care and evaluate its suitability for their study area, ideally against independent high-resolution data such as UAV orthomosaic. Even when only running inference, the super-resolution step in our Sentinel-2 processing pipeline is computationally expensive, benefits substantially from GPU acceleration due to its Torch dependency, and increases storage requirements because of the higher pixel count. For many applications, a 10 m data cube will therefore remain the recommended default, while the 2.5 m super-resolved product is most useful for specific tasks where spatial heterogeneity is critical, such as in rivers and floodplains with their patchy, dynamic morphology.

A key contribution of this study is that the entire workflow is implemented using STAC-native data access and Xarray/Zarr-based cube construction, meaning that the data cubes remain transparent, reproducible, and system-independent. The workflow can be executed on HPC systems for large regions, but also on local machines where chunked and deferred Dask execution allows scalable out-of-core processing. The update mechanism allows extending the data cube as new Sentinel-2 acquisitions become available, avoiding full reconstruction and dramatically reducing processing cost over long monitoring periods. This is especially important for operational monitoring workflows such as river morphology tracking, floodplain restoration, agricultural crop monitoring, or drought progression assessment.

6 Conclusion

This study presents an open-source workflow for generating analysis-ready Sentinel-2 data cubes that address key limitations of existing Xarray-based cube frameworks. First, we demonstrate that cloud masking based on s2cloudless cloud probability fields provides greater adaptability than the commonly used Sen2Cor SCL, particularly in heterogeneous and dynamic surface environments. The threshold-based masking approach allows users to adjust cloud detection strength and avoid systematic false positives, resulting in more reliable temporal coverage across the region of interest.

https://doi.org/10.5194/egusphere-2026-619
Preprint. Discussion started: 10 February 2026

Second, we show that scene-to-scene co-registration substantially improves the temporal consistency of NDVI time series in the braided river corridor. In spatially homogeneous agricultural areas, co-registration has little effect; however, in morphologically dynamic zones, sub-pixel misalignments were found to introduce abrupt, non-realistic NDVI fluctuations. The co-registered cube preserves smooth seasonal trajectories and allows meaningful interpretation of vegetation dynamics, demonstrating that geometric stability is a prerequisite for time-series analysis in fine-scale, heterogeneous systems.

Third, integrating the SEN2SR super-resolution model into the data cube enhances the separability of land–water and vegetation–sediment boundaries. While the 2.5 m product does not approach the structural detail of UAV imagery, it improves the delineation of geomorphic features beyond the native 10 m resolution, making Sentinel-2 more useful for studying dynamic river landscapes and other patchy surface environments. Together, these results demonstrate that scalable data cube processing workflows must be flexible in cloud masking, explicit in spatial alignment, and optionally enhanced in spatial resolution to support advanced environmental analytics.

In addition to the methodological contributions, the workflow is released as a fully open-source Python package with detailed documentation and user-friendly Jupyter notebooks. These interactive notebooks guide users through data querying, cube construction, cloud masking selection, co-registration, and optional super-resolution, allowing the workflow to be adopted in both research and operational settings. By prioritizing accessibility and transparent processing steps, the package enables reproducible analyses and supports wider application of EODC-based monitoring in dynamic surface environments.

The workflow developed here provides a foundation for reproducible, scalable environmental monitoring using Sentinel-2 time series. Future work should explore the integration of uncertainty metrics for both co-registration and super-resolution outputs, as well as the application of surface-adaptive cloud thresholding strategies guided by spatial gradients. Extending the approach to other dynamic environments, such as deltas, wetlands, and arid river systems, would enable broader evaluation of landscape-dependent processing needs. Furthermore, incorporating machine learning and deep learning tools (e.g., scikit-learn, TorchGeo) directly into the cube will support automated classification, segmentation, and change detection. As data archives continue to expand, the incremental update mechanism will play an increasingly important role in operational monitoring pipelines. Ultimately, this workflow contributes to making advanced surface process analysis more accessible by bridging high-performance computing, open-source methods, and Earth system science.

Code availability

Code is available at Zenodo: DOI: 10.5281/zenodo.18458717. The development repository is hosted on GitHub: https://github.com/BaturalpArisoy/stac2cube.

Author contributions

BA designed the methodology workflow, developed the code, generated the code repository and tested analysis.

https://doi.org/10.5194/egusphere-2026-619
Preprint. Discussion started: 10 February 2026

Competing interests

The author has declared that there are no competing interests.

Disclaimer

Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. The authors bear the ultimate responsibility for providing appropriate place names. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Acknowledgements

This work was supported by the German Aerospace Center (DLR) and benefited from access to the terrabyte high-performance computing infrastructure at the Leibniz Supercomputing Centre (LRZ). We thank the LRZ and DLR support teams for their dedicated technical assistance and sustained maintenance of the platform. We also acknowledge the broader terrabyte user community for constructive discussions and practical feedback that helped streamline the development and execution of the workflow.

Sentinel observations used in this study were provided by the European Union's Copernicus Programme and operated by the European Space Agency (ESA), whose open and consistent Earth Observation data policy and long-term mission continuity are fundamental to reproducible remote-sensing research.

Finally, field work and UAV data acquisition were supported by the German Research Foundation within the FluBig project (Grant No. 522097137).

References

Aybar, C., Montero, D., Contreras, J., Donike, S., Kalaitzis, F., and Gómez-Chova, L.: SEN2NAIP: A large-scale dataset for Sentinel-2 Image Super-Resolution, Sci. Data, 11, 1389, https://doi.org/10.1038/s41597-024-04214-y, 2024.

Aybar, C., Contreras, J., Donike, S., Portalés-Julià, E., Mateo-García, G., and Gómez-Chova, L.: A Radiometrically and Spatially Consistent Super-Resolution Framework for Sentinel-2, https://doi.org/10.2139/ssrn.5247739, 9 May 2025.

Baumann, P., Misev, D., Merticariu, V., and Huu, B. P.: Datacubes: Towards Space/Time Analysis-Ready Data, in: Service-Oriented Mapping, edited by: Döllner, J., Jobst, M., and Schmitz, P., Springer International Publishing, Cham, 269–299, https://doi.org/10.1007/978-3-319-72434-8_14, 2019.

Baumann, P., Misev, D., Merticariu, V., and Huu, B. P.: Array databases: concepts, standards, implementations, J. Big Data, 8, 28, https://doi.org/10.1186/s40537-020-00399-2, 2021.

https://doi.org/10.5194/egusphere-2026-619
Preprint. Discussion started: 10 February 2026

Betz, F., Lauermann, M., and Egger, G.: Biogeomorphology from space: Analyzing the dynamic interactions between hydromorphology and vegetation along the Naryn River in Kyrgyzstan based on dense satellite time series, Remote Sens. Environ., 299, 113890, https://doi.org/10.1016/j.rse.2023.113890, 2023.

Bozzolan, E., Brenna, A., Surian, N., Carbonneau, P., and Bizzi, S.: Quantifying the Impact of Spatiotemporal Resolution on the Interpretation of Fluvial Geomorphic Feature Dynamics From Sentinel 2 Imagery: An Application on a Braided River Reach in Northern Italy, Water Resour. Res., 59, e2023WR034699, https://doi.org/10.1029/2023WR034699, 2023.

Brigot, G., Colin, E., Plyer, A., and Janez, F.: Adaptation and Evaluation of an Optical Flow Method Applied to Coregistration of Forest Remote Sensing Images, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 9, 1–17, https://doi.org/10.1109/JSTARS.2016.2578362, 2016.

Dwyer, J. L., Roy, D. P., Sauer, B., Jenkerson, C. B., Zhang, H. K., and Lymburner, L.: Analysis Ready Data: Enabling Analysis of the Landsat Archive, Remote Sens., 10, 1363, https://doi.org/10.3390/rs10091363, 2018.

Frantz, D.: FORCE—Landsat + Sentinel-2 Analysis Ready Data and Beyond, Remote Sens., 11, 1124, https://doi.org/10.3390/rs11091124, 2019.

Guo, H., Liang, D., Chen, F., and Shirazi, Z.: Innovative approaches to the Sustainable Development Goals using Big Earth Data, Big Earth Data, 5, 263–276, https://doi.org/10.1080/20964471.2021.1939989, 2021.

Hoyer, S. and Hamman, J.: xarray: N-D labeled Arrays and Datasets in Python, J. Open Res. Softw., 5, https://doi.org/10.5334/jors.148, 2017.

Scheffler, D., Hollstein, A., Diedrich, H., Segl, K., Hostert, P., Scheffler, D., Hollstein, A., Diedrich, H., Segl, K., and Hostert, P.: AROSICS: An Automated and Robust Open-Source Image Co-Registration Software for Multi-Sensor Satellite Data, Remote Sens., 9, https://doi.org/10.3390/rs9070676, 2017.

Simoes, R., Camara, G., Queiroz, G., Souza, F., Andrade, P. R., Santos, L., Carvalho, A., and Ferreira, K.: Satellite Image Time Series Analysis for Big Earth Observation Data, Remote Sens., 13, 2428, https://doi.org/10.3390/rs13132428, 2021.

Soille, P., Burger, A., De Marchi, D., Kempeneers, P., Rodriguez, D., Syrris, V., and Vasilev, V.: A versatile data-intensive computing platform for information retrieval from big geospatial data, Future Gener. Comput. Syst., 81, 30–40, https://doi.org/10.1016/j.future.2017.11.007, 2018.

Sudmanns, M., Tiede, D., Lang, S., Bergstedt, H., Trost, G., Augustin, H., Baraldi, A., and Blaschke, T.: Big Earth data: disruptive changes in Earth observation data management and analysis?, Int. J. Digit. Earth, 13, 832–850, https://doi.org/10.1080/17538947.2019.1585976, 2020.

Tarrio, K., Tang, X., Masek, J. G., Claverie, M., Ju, J., Qiu, S., Zhu, Z., and Woodcock, C. E.: Comparison of cloud detection algorithms for Sentinel-2 imagery, Sci. Remote Sens., 2, 100010, https://doi.org/10.1016/j.srs.2020.100010, 2020.

Vydra, C., Dietz, A. J., Roessler, S., Conrad, C., Vydra, C., Dietz, A. J., Roessler, S., and Conrad, C.: The Influence of Snow Cover Variability on the Runoff in Syr Darya Headwater Catchments between 2000 and 2022 Based on the Analysis of Remote Sensing Time Series, Water, 16, https://doi.org/10.3390/w16131902, 2024.

https://doi.org/10.5194/egusphere-2026-619
Preprint. Discussion started: 10 February 2026

Xenarios, S., Shenhav, R., Abdullaev, I., and Mastellari, A.: Current and Future Challenges of Water Security in Central Asia, in: Global Water Security, World Water Council, 117–142, https://doi.org/10.1007/978-981-10-7913-9, 2018.

615 Yang, C., Yu, M., Hu, F., Jiang, Y., and Li, Y.: Utilizing Cloud Computing to address big geospatial data challenges, Comput. Environ. Urban Syst., 61, 120–128, https://doi.org/10.1016/j.compenvurbsys.2016.10.010, 2017.