REVIEW OF MANUSCRIPT

"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, Florian Betz, Georg Stauch, Doris Klein, Stefan Dech, Tobias Ullmann

1 Overview

The paper presents an open-source Python package designed to streamline the generation of Analysis-Ready Data (ARD) cubes from Sentinel-2 imagery. The authors identify three deficiencies in current Xarray-based Earth Observation Data Cube (EODC) toolchains: rigid and inaccurate cloud masking, geometric inconsistencies across time series, and a lack of scalable super-resolution implementation. To address these gaps, the study proposes a fully integrated workflow that utilizes Spatio Temporal Asset Catalogs (STAC) and runs efficiently on High-Performance Computing (HPC) clusters or local environments.

The pipeline relies on three algorithms:

(a) Cloud Masking (s2cloudless): Instead of relying on the standard, binary Scene Classification Layer (SCL) provided by Sen2Cor, the package implements s2cloudless. This provides user-adjustable probabilistic cloud masking, which the authors show is essential for avoiding false positives over spectrally bright surfaces like river gravel bars.
(b) Co-registration (AROSICS): To fix sub-pixel geometric misalignments (typically 1-2 pixels) between Sentinel-2 acquisitions, the authors integrate a sliding-window co-registration routine.
(c) Super-resolution (SEN2SR): The workflow utilises a deep learning model to up-sample native 10m and 20m Sentinel-2 bands to a 2.5m resolution. This is intended to improve the delineation of narrow water channels and patchy vegetation.

2 Points for Improvement

2.1 Data cube definition

There is no definition of "data cube" in the paper. The example provided by the paper uses a small area with specific characteristics.

Different definitions of a data cube exist. Appel and Pebesma (2019) see it as a regular space-time partition with all images sharing the same coordinate projection.

Simoes et al. (2019) extend this to include tiles with different coordinate projections but same spatial and temporal resolutions. The stricter definition requires all tiles to be converted to a single projection, which is hard for large areas. The extended definition allows a data cube to cover large areas.

The authors need to clarify what a "data cube" means for them. The examples suggest they adopt the definition from the "*cubo*" package, which is a limited version of Appel and Pebesma (2019), focusing on small patches in the UTM coordinate system.

2.2 Author's contribution

The authors developed an extension for the "*cubo*" package, including functions to enhance cloud-cover removal, image registration, and produce super-resolution data. However, this does not constitute an "open-source package for customized processing of Sentinel-2," as claimed. Their work lacks functions typical of EODC tools, like raster analysis and machine learning classification. A more accurate description is that they created pre-processing tools for Sentinel-2 data in cube environments.

For this reason, the manuscript title does not accurately reflect the authors' contributions. It is suggested that the title be revised to better reflect what the authors have done. One option is to use a title such as "An Open-Source Package for Customised **Pre-**processing of Sentinel-2 Data for Data Cube Analytics", or something along these lines.

2.3 Relation to previous work

When stating their contributions relative to previous work, researchers need to be careful on how they state previous work. In the paper, the authors state:

> "*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*."

Strong arguments need solid evidence. The terms "deficient" and "limited" suggest these tools can't perform data analysis, but this isn't self-evident. To support this, authors should specify the reviewed "open-source and commercial EODC tool chains" and detail their deficiencies. Fairness in evaluating third-party tools is lacking; examples like Pebesma et al. (2026) and Gomes et al. (2020) illustrate proper evaluation. Despite not being a review, authors should have briefly described the EODC tool chains they consider and pointed out their capacities and limitations.

Considering the current EODC tool chains like *Google Earth Engine*, *openEO*, *sits*, and *Open Data Cube*, calling these tools "deficient" and "limited" is unwarranted. The authors should highlight their work positively, emphasizing the benefits their work offers to data cube users.

2.4 Lack of API package description

When examining similar works on EO cloud computing like Gorelick et al. (2017), Simoes et al. (2019), and Schramm et al. (2021), all describe their packages' APIs, including their functionality, usage, benefits, and limitations. It's important for authors to provide a similar API discussion.

2.5 Limitations of the work

Careful authors inform the reader of their contributions. They frame the work as necessary to avoid "deficiencies" of existing EODC tools, which is not true. Existing tools are widely used without the authors' pre-processing tools. The authors should present their results positively and openly discuss the benefits and limitations of their tool, including scenarios where it is particularly valuable.

Consider an EO data cube or analysis-ready data provider like Microsoft Planetary Computer, Digital Earth Africa, the Copernicus Data Space Environment, or the Brazil Data Cube. These providers typically get ARD tiles from ESA with cloud info in the SCL band, already split into MGRS tiles. If they switched to "s2cloudless" and AROSICS tools, they'd need to reprocess all Sentinel-2 images, a significant task. Do the authors see reprocessing as necessary for all EO cloud providers? Are there specific cases where their tools are essential?

The authors should be more rigorous about the super-resolution analysis. They assume "more is better" and that super-resolved S2 images are superior in all cases, which entails a 25-fold increase in database size and processing. Is this justified, especially since 10-meter Sentinel-2 images often suffice? Also, they must consider how super-resolution affects long-term time series consistency to fully justify their tool's value.

3 Final remarks

The manuscript introduces tools for Earth observation data cubes, but flaws in presentation undermine its message. Instead of highlighting benefits, the authors adopt an adversarial tone, criticising earlier work in offensive terms. These negative claims lack rigorous justification, weakening the overall contribution.

The manuscript needs **major revisions**. The authors should reconsider how they describe their work relative to the literature. Their work is valid. With careful, positive rewriting, it merits publication.