ISPRS Journal of Photogrammetry and Remote Sensing 202 (2023) 682–690

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Wasting petabytes: A survey of the Sentinel-2 UTM tiling grid and its spatial overhead

Bernhard Bauer-Marschallinger *, Konstantin Falkner

Department of Geodesy and Geoinformation, TU Wien, 1040 Vienna, Austria

ARTICLE INFO

Keywords:
Universal Transversal Mercator (UTM)
Military Grid Reference System (MGRS)
Sentinel-2
Remote sensing grid
Spatial reference
Big data
Analysis-Ready-Data

ABSTRACT

Following Landsat's practice, Sentinel-2 multispectral satellite products are delivered as raster images projected onto the Universal Transversal Mercator (UTM) spatial reference system, which divides Earth into 60 longitudinal zones. Locally, this guarantees high spatial accuracy, while also easing the interoperability with many regional and governmental datums. On top, the Sentinel-2 product grid uses the Military Grid Reference System (MGRS) tiling scheme to facilitate manageable data slices and straightforward multitemporal image stacking. Although most convenient for small-area applications, activities with a larger geographic scope suffer from this approach and its overhead, as both data duplication and ambiguity appear along UTM zone overlaps and MGRS tile borders. In practice, such areas that are covered by multiple and incongruent grid pixels are known but just tolerated, and their degree has not been measured so far.

In this paper, we illuminate the nature and patterns of these overlaps, and calculate the resulting spatial redundancy over the global land surface. We found that the total land area is enlarged in the Sentinel-2 grid definition by 33%, which is a value similar to the simple and single-zoned Plate Carrée projection. The number of co-located grid pixels for a single location ranges from 1 up to 6, with on average more redundancy at mid- and high-latitudes. With regard to global satellite archives in times of big data and increased energy costs, the examined grid appears as a suboptimal choice, inducing complexity and overhead at an unreasonable level. Owing to the grid design, e.g., the yearly Sentinel-2 user product volume (Level-1C and -2A) is inflated by 1 petabyte, entailing cascading downstream costs of storage, bandwidth, and computing.

1. Introduction

After their rise in the 1970s, digital satellite imagers have now entered the era of Big Data, and the remote sensing community benefits from a more and more free and open access to a fast growing number of satellite systems, featuring Earth Observation (EO) imagery and added-value derivates at high spatio-temporal resolution. Together with a rich variety of thematic EO satellites, global monitoring missions like the Sentinel constellation (Aschbacher and Milagro-Pérez, 2012) or the Landsat programme (Wulder et al., 2022) are building up a vast archive of spatial data in the realm of petabytes, putting high demands on hardware, interfaces, and data infrastructure. Facing constraints from nowadays increased and volatile energy costs, efficient data models that avoid redundancy when storing, processing, and analysing satellite images are most crucial.

Raw satellite data are generally not directly adequate for analysis and geophysical parameter retrieval but require preprocessing steps, including i.a., georeferencing, spectral calibration, and quality check. End users of EO imagery must manage those efforts themselves, or more commonly, can access already preprocessed datasets from dedicated providers. In the past years, the EO community adopted the use of datacubes to join and harmonise preprocessed data from multiple sensors (e.g. Kopp et al. (2019), Frantz (2019) and Wagner et al. (2021)), and promote interoperability through Analysis-Ready-Data (ARD, e.g. works by Gorelick et al. (2017), Egorov et al. (2018) and Chatenoux et al. (2021)). The ARD concept is embraced by the Committee on Earth Observation Satellites (CEOS, Lewis et al. (2018)) that provides agreed specifications for (meta-) data. Consequently, novel products are designed along related specifications, such as, for example, the United States Geological Survey (USGS) which released a Landsat ARD collection (Dwyer et al., 2018), or the European Space Agency (ESA) envisages ARD formats for upcoming Sentinel-1 and Biomass products.

One central aspect when distributing EO imagery is its spatial referencing and gridding. Commonly, satellite measurements are initially

* Corresponding author.
E-mail address: bbm@geo.tuwien.ac.at (B. Bauer-Marschallinger).
URL: https://www.geo.tuwien.ac.at/ (B. Bauer-Marschallinger).

https://doi.org/10.1016/j.isprsjprs.2023.07.015
Received 3 March 2023; Received in revised form 6 June 2023; Accepted 14 July 2023
Available online 26 July 2023

stored in data structures close to sensor geometry, and for distribution are then resampled and warped to an Earth-bound spatial reference that holds a raster grid. For medium- and high-resolution data (finer than 1 km), planar-regular grids realised by a Cartesian coordinate system are most suitable, benefiting from (array) raster indexing (Lu et al., 2018). The warping is achieved through geographic map projections to create projected images that are easily ingested into an ARD datacube and are well understood by user communities of different fields and backgrounds.

A prominent map projection is the Universal Transverse Mercator (UTM) system, a coordinate system originally invented by military geodesists in the middle of the 20th century (Snyder, 1987), and now widely used in remote sensing as the grid basis for a variety of satellite missions. The system was popularised by early US satellite missions and the 1972-launched Landsat programme, as it features very low distortions and allows for accurate referencing of imagery to specific locations, precise navigation, and sharing between various military and civil organisations. Nowadays, UTM is used by several EO institutions like NASA, NOAA, JAXA, ISRO for distributing various satellite image products. Among them, the Landsat enterprise and the younger (2015-launched) European Union's Copernicus Sentinel-2 constellation are considered the two leading programs for medium-resolution land imaging (Claverie et al., 2018). ESA, as operating agency, defined a UTM-based grid for its Sentinel-2 mission (Drusch et al., 2012; Baillarin et al., 2012) to enable direct geographic co-registration with Landsat imagery and ultimately allow joint land imaging archives with unprecedented temporal coverage and spanning over five decades, effectively establishing UTM as a quasi-standard.

In essence, the UTM system divides Earth into 60 narrow north-south-bound zones and maps the surface onto 60 individual cylindrical projections with each framing its own coordinate system. As such, they optimally approximate their respective stripe and achieve high local mapping accuracy. Commonly, the Military Grid Reference System (MGRS) uses 100 × 100 km square tiles to define the tiling scheme for each UTM zone.

Notwithstanding its merits, two distinct drawbacks arise from the UTM setup. First, the division of images into different projections causes a range of complications and imprecisions, and typical use cases are regularly troubled through the rather narrow zoning. Apart from the mundane task of simply displaying a multi-zone image, it complicates the resolving of relative shifts during co-registration of multi-temporal images (Yan et al., 2018), and requires proper resampling methods to avoid degradation of geometric fidelity (Roy et al., 2016). Accordingly, to build a harmonised ARD Landsat archive for the US, Dwyer et al. (2018) of USGS went over to process raw and higher-level products in a (single) Albers Equal Area projection grid, also recognising that repeated reprojection constitutes considerable effort and spatial detail is lost during the process.

As second drawback, resulting directly from the arrangement of the cylindrical projections, the 60 UTM coordinate systems overlap at the zone borders. Hence, a UTM grid covers sections of the globe multiple times and intrinsically carries data redundancy and pixel ambiguity, as discussed for the Sentinel-2 grid by e.g. Frantz (2019). This is well understood, but has not been thoroughly analysed so far. As a first effort, Roy et al. (2016) examined for three test sites grid pixel offsets between neighbouring zones and found that 41% to 49% of the areas are covered by multiple tiles, with a maximum of four tiles overlapping one single pixel.

In this study, we globally analyse the spatial overhead from the UTM overlaps specifically for the Sentinel-2 grid (ESA, 2015; TAS, 2022), which is formed by a collection of extended MGRS tiles. For nonpolar land surfaces—which are in the scope of many satellite monitoring applications—we quantify this widely known but often just tolerated problem, by measuring how much larger is the gridded land area than the actual one on Earth. We obtain this areal overhead together with a per-pixel-count of overlapping tiles and we expound on their implications for users and institutions. Section 2 presents methods and the detailed structure of ESA's Sentinel-2 grid, Section 3 discusses the obtained values for the spatial overhead, and Section 4 draws conclusions on extra costs and closes with usage recommendations.

2. Data and methods

All geospatial operations and manipulations of this study's vector- and raster-datasets were carried out with the Geospatial Data Abstraction Library (GDAL, version 3.5.2). The target spatial references of the UTM zones are defined by respective European Petroleum Survey Group (EPSG) codes for the WGS84 datum (32601 to 32660 for northern sections, and 32701 to 32760 for southern), whereby for the remainder of the study we handled the northern and southern half-zones jointly. All logical operations (masking, etc.) and area measures were done with Python/NumPy (version 3.10.6/1.23.4).

2.1. UTM and MGRS geometries

The Universal Transverse Mercator (UTM) coordinate system enables the planar mapping of the entire Earth with high geometric accuracy in metric units. It divides Earth (modelled by the WGS84 datum) into 60 zones of 6° longitudinal width, which is equivalent to ~666 km at the equator. For each zone, it defines an individual transversal Mercator projection, which is formed by a cylinder with its rotational axis in the equatorial plane (illustrated in Fig. 1a). Although each cylinder approximates the ellipsoid surface along its central meridian with lowest aberration, projection distortions grow extremely large when leaving the 6°-zone. As built from Mercator projections, the UTM system preserves angles and therefore projects shapes undistorted, but alters their size (Snyder, 1987). Along a cylinder's central meridian, the scale factor is set to 0.9996 so that the (shrunk) cylinder does not touch the ellipsoid but intersects it twice, circa 180 km east and west of the central meridian. This minimises and balances overall length distortion within a 6°-zone.

The setup with these narrow 60 co-rotated cylinders approximates well the Earth ellipsoid, but it creates intersections between the zones, which grow with higher latitudes. The UTM system as realised by ESA's Sentinel-2 grid reduces these ambiguities by a stepwise zone diminution that shrinks the zones towards the poles (Fig. 1b). However, an array of substantial zone overlap remains along the zone borders, at which coordinates for single points exist in both overlapping zones. And as mentioned, many datasets stretching in east–west direction are divided into different UTM zones and cannot be handled within one single projection, necessitating re-projection to a joint coordinate system during data analysis and display.

For precise mapping and navigation purposes, the Military Grid Reference System (MGRS) was invented on top of the UTM projection system. The 60 longitude zones are defined for the nonpolar surface between 80°S and 84°N and are consecutively numbered from 1 to 60, from 180°E to 180°W. The MGRS subdivides those into 20 latitude bands of 8° height, indicated by letters from C at 80°S to X at 84°N. (Note: for the polar sections, two separate Universal Polar Stereographic (UPS) projections are used and zoned with the letters A, B, Y, Z, which are neglected in this study.) The combination of a zone and a latitude band defines then a grid zone (e.g. 18T), which is furthermore subdivided into squares of 100 km size, which are in turn labelled by two letters (extending our example to e.g. 18TVL, see full definition by DMA (1989)). These squares of the MGRS grid provide the baselines for the Sentinel-2 tiling grid used for the dissemination of ESA products, as illustrated in Fig. 1c. Most notably, and contrary to official documentation (ESA, 2015; TAS, 2022), Sentinel-2 Level-1C/-2A images are delivered in tiles (aka "granules") that are enlarged to 109.8 km. As a consequence, locations within the ~10 km wide stripes around tile borders are contained in two or more tiles.

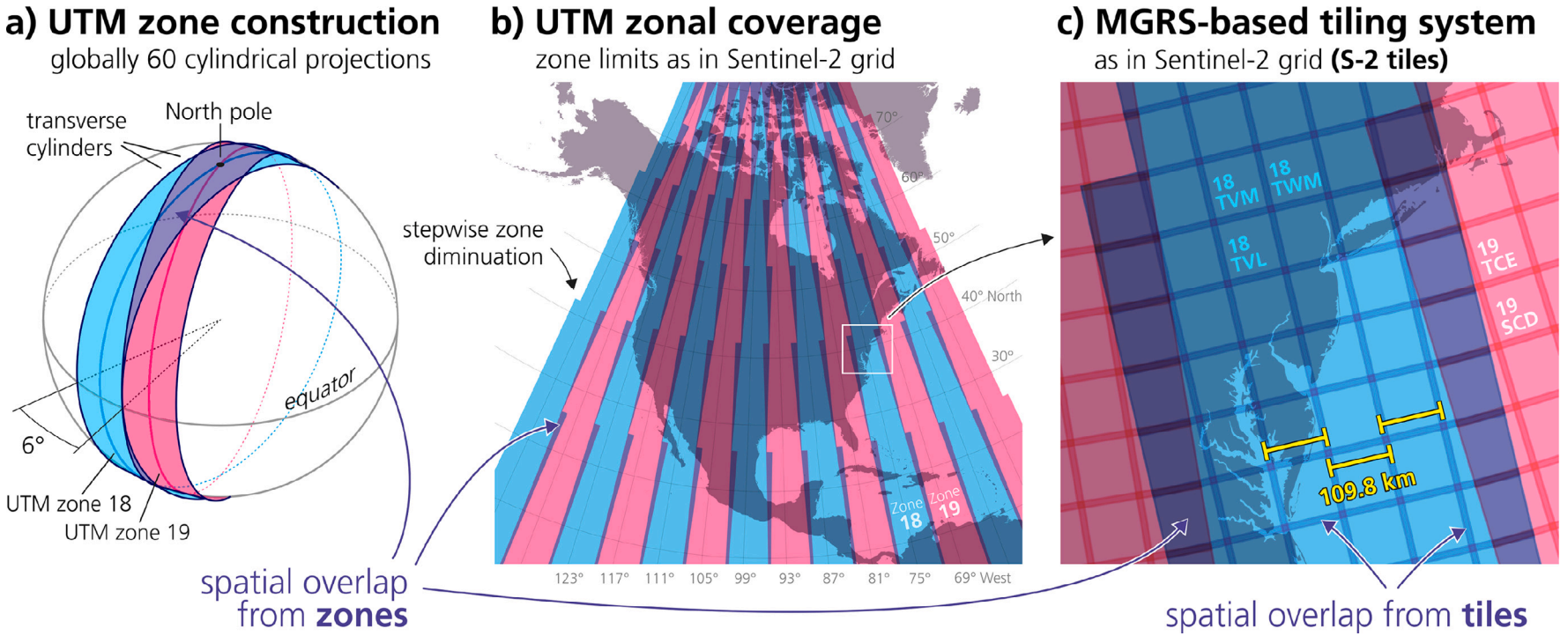

Fig. 1. Schematic illustration of the UTM zoning concept and the MGRS-based tiling scheme on top, as used for the Sentinel-2 grid. (a) (Exaggerated) illustration on the relative orientation of two adjacent UTM 6°-zones formed by transversal cylinders, and their overlap that grows with latitude. (b) True-to-scale projected representation of zonal coverage for a section of North America, with stepwise zonal diminution along increasing latitude to avoid excessive overlap. (c) Zoom-in into US East Coast showing the MGRS-based Sentinel-2 tiles, which are per zone orthogonally arranged as overlapping squares of 109.8 km width, forming a graticule of a ~10 km wide overlap.

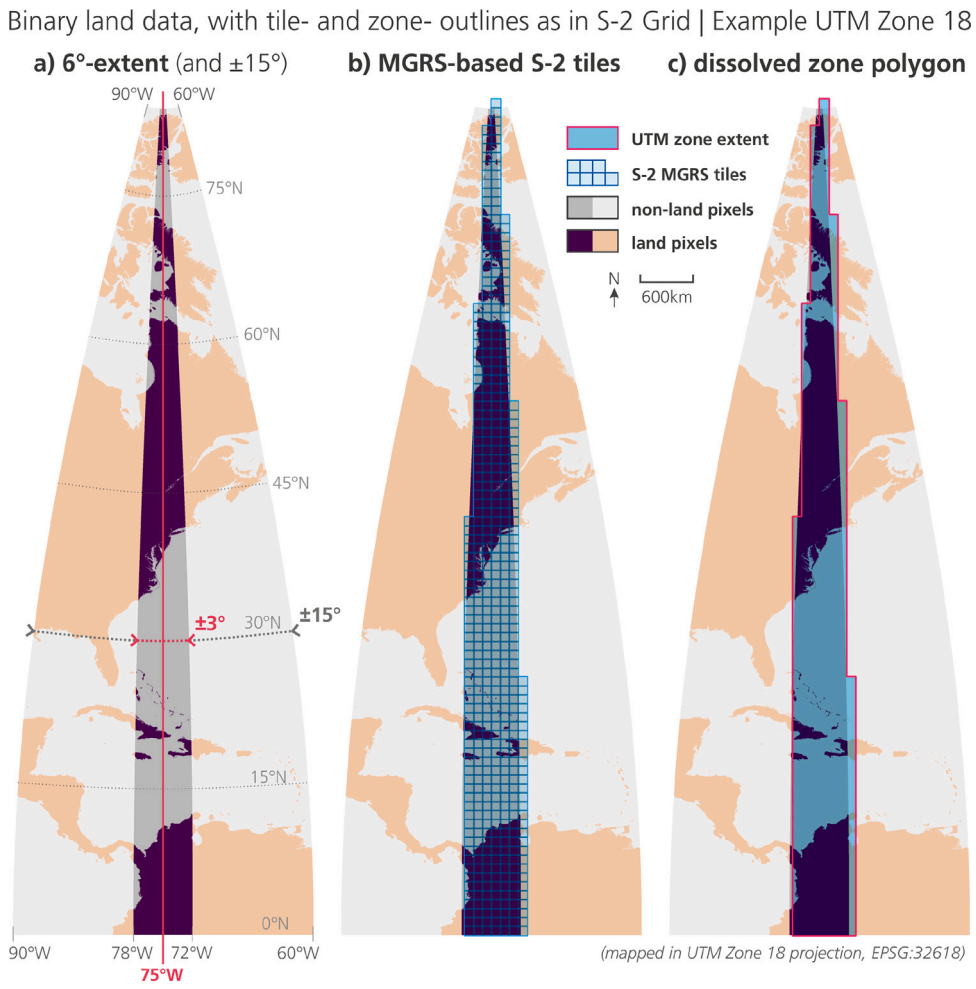

Fig. 2. Illustration on our analysis' input geodata, as example displaying the UTM Zone 18, northern section. (a) The 1 km-sampled binary land data, with highlighted section within the ±3°-wide UTM stripe centred on the defining meridian, and the ±15°-wide buffer. (b) The MGRS-based Sentinel-2 tiles corresponding to Zone 18, partly overhanging the zone boundaries at ±3°. (c) The extent-polygon for Zone 18 from merging (dissolving) the corresponding S-2 tiles.

For these reasons, the Sentinel-2 grid inherits two different kinds of spatial overlap, one from the intersecting projections underneath the UTM zones, and one from the graticule formed by the enlarged MGRS tiles. The tile overlap copies values to fully congruent pixels allocated to different tiles, whereas the zone overlap is more complex, with differently sampled values in an other pixel geometry over the same area. Both effects lead to spatial ambiguity, in a sense that for a single location multiple grid representations exist, and hence data is duplicated. In this context, we refer to the (co-located) redundant data within the grid as spatial overhead.

We accessed the Sentinel-2 tiling grid from the European Space Agency's official distribution (ESA, online, 2023), given as a global kml file that comprises 56 686 MGRS-based tiles as squared polygons of 109.8 km × 109.8 km extent (hereafter named S-2 tiles). In order to have access to the individual tiles within the kml-file, it was converted into a global shapefile in the WGS84-latitude/longitude coordinate system ("Latlon", EPSG:4326). For each UTM zone, we selected all corresponding tiles and saved them into a separate shapefile that was projected in the respective UTM datum. This file contains the zone's tiles as individual polygons, and as whole features the stepwise diminution towards the poles (see Fig. 2b). Additionally, we stored a second shapefile that contains a single polygon generated from spatially merging the zone's tiles (using GDAL's "dissolve" operator). This polygon simply describes the outline of the zone and omits the tile overlaps (Fig. 2c).

We note that contradictory statements about the extent and direction of the Sentinel-2 tile overlap are found in the literature. The official product documentation (ESA, 2015; TAS, 2022) and e.g. Kempeneers and Soille (2017) state a tile extent of 100 km, whereas—representing many studies—Claverie et al. (2018) and Coluzzi et al. (2018) describe extents of 109.8 km or 110 km, and Gascon et al. (2017) describes 4-sided overlaps of 5 km. We, however, found in both ESA's kml-file and Sentinel-2 Level-1C granule products that the S-2 tiles are always extended by exactly 9.8 km towards south and east. Hence, the S-2 tile centroids are shifted to southeast in respect to the MGRS tiles, and no overlap towards north and west is given. We further found that the S-2 tiles' upper-left corner not always match perfectly the MGRS tiles' corner and thus the overlap varies between 9.78 km and 9.84 km.

2.2. Land surface data

As test object for the spatial overhead in the Sentinel-2 grid over global land surfaces, we prepared a raster dataset that describes the Earth's land areas and ingested it into the grid.

We used the free vector data of Natural Earth (Kelso and Patterson, 2010), in particular the land polygons given in Latlon (EPSG:4326) at the scale-quality of "1:10 million" (Natural Earth, online, 2023), a well-suited binary land-sea dataset that is both simple and complete. In a first step (and aiming for efficient calculations), we rasterised it at a 1 km sampling in the Plate Carrée projection (EPSG:4087), as we consider this raster precision sufficient for building ratios of global area totals (and accept sub-kilometre infidelities at the coastlines). In a second step, the land raster was clipped to latitudes between 60°S

and 85°N. This domain leaves aside polar regions and Antarctica but contains the bulk of Earth's landmasses—which are in the scope of major land monitoring satellite missions.

For each UTM zone, the binary 1 km land raster was warped to the respective UTM projection using nearest neighbour resampling. The data was clipped twice along longitude: Once at ±3° of the central meridian of the zone to obtain the precise 6°-extent, and once at ±15° to provide a large buffer that includes all land covered by the S-2 tiles, which in part overhang the actual zone boundary. Fig. 2 illustrates the projected land data of a single UTM zone and how the tiles and zone polygon overlay—and how the S-2 tiles overhang the ±3° zone boundaries.

2.3. Overhead area calculations

Our study's experiment design is quite straightforward: We measured the total land area in the Sentinel-2 grid and compared it with the actual land area on the globe. This was realised by ingesting the land raster from Section 2.2 into the grid and then summing up the area of all land pixels. The ratio of the areas in the grid and on the globe constitutes the grid's spatial overhead for land observations. This calculation method is equivalent to the so-called pixel area factor used by Bauer-Marschallinger et al. (2014) to validate their analytical study on the overhead of different projections. We underline that the increase in area is directly proportional to the increase in data volume (Mulcahy, 2000; Kimerling, 2002). We further note that file compression can absorb some of the data duplication, but this may succeed only in dedicated data structures, and not in single image granules or layers.

All area measures presented in Section 3 were calculated by determining the area represented by individual pixels within the binary land raster and subsequently summing all land pixels within the zone of interest. This was achieved for datasets given in the Plate Carrée projection by assigning each land pixel the value 1 km^2 and multiplying it by the cosine of its latitude, to account for the areal distortion of the dataset's projection (following Snyder, 1987; Bauer-Marschallinger et al., 2014). For the datasets projected to the UTM zones, we omitted this correction and set the pixel area simply to 1 km^2, since the areal distortion within a 6°-wide zone is minimal (and balanced) and therefore can be neglected.

For the purpose of cross-checking, we performed an alternative calculation as an arithmetic test as follows: In the global Plate Carrée projection, a new 1 km raster was generated that for each land pixel contains the number of overlapping S-2 tiles, i.e. that local redundancy. Integration of this overlap/redundancy map, again weighted by the cosine of latitude, yields the total spatial overhead over land.

3. Results and discussion

To begin with, we calculated Earth's actual land surface from the binary land mask from Section 2.2 to obtain our reference value, in the following declaring 100% land area. Summation over the area-weighted land pixels yielded a total area of 147.1 million km^2, a quantity close within 1.5% to what is found in common literature. For reassurance, visual inspection of the resampled land raster showed no artefacts or gaps. The slightly smaller area total can be attributed to the 1 km sampling of the (binary) land raster that knowingly tends to omit small islands and peninsulas, and we consider this accuracy sufficient for our comparative study. For the nonpolar domain between 60°S and 85°N, we obtained the total land area of 134.8 mil km^2. Table 1 collects our experiments' area measures.

To give perspective on below results for UTM, we determined the land area of the binary land data in the projected Plate Carrée space, which is a common choice when a (simple) data grid is sought. Summation over the land area yielded a spatial overhead of +41% for the nonpolar domain, and +81% for the fully global domain, owing to the projection's severe distortions close to the poles. As a second comparison, a comparatively small value of +3.2% was found for the Equi7Grid (Bauer-Marschallinger et al., 2014), which uses an optimised set of 7 continental zones.

Table 1
Results of overhead area calculations. The table lists area totals of the binary land surface data after ingestion into the different grids setups.

Land surface gridded in ...	Area in grid	Spatial overhead
Domain 90°S – 90°N (fully global)	mil km^2	%
Reference land area	147.1	(=100)
Plate Carrée	265.6	+80.6
Equi7Grid (T1-tiling)	151.7	+3.2
UTM plain (clear-cut ±3° stripes)	146.8	−0.2
UTM Sentinel-2 Grid	**197.0**	**+34.0**
Domain 60°S – 85°N (nonpolar)	mil km^2	%
Reference land area	134.8	(=100)
Plate Carrée	190.6	+41.4
UTM plain (clear-cut ±3° stripes)	134.5	−0.2
UTM Sentinel-2 tiling grid	**179.3**	**+33.0**
Africa	38.5	+28.2
Asia	60.5	+34.6
Australia	10.9	+29.7
Europe	12.6	+36.6
North America	33.7	+39.3
South America	22.9	+28.3
UTM Sentinel-2 grid (no tiling)	**152.2**	**+13.0**

Next, we did the same for the 60 segments of the binary mask that have been warped to the ±3°-clipped UTM zones (ignoring all tile polygons; listed as "UTM plain" in Table 1). Summation over all UTM zones yielded a global total area of 146.8 mil km^2, almost repeating perfectly above value (i.e. 99.8%). That said, the UTM system evokes zero spatial overhead when realised without any overlap through clear cuts at the zone borders. For the nonpolar domain, we obtained 134.5 km^2 (again 99.8%), and conclude that effects from projection imprecision and narrow slices that converge towards to the poles seem to not significantly impact our measures.

However, and coming to the primary question of this survey, this does not hold for the UTM/MGRS realisation of ESA's Sentinel-2 tiling grid with its overlaps between zones and tiles. For each UTM zone (clipped with the buffer of ±15°), we looped over the tiles and counted the land pixels within (and disregard anything in the buffer that is outside the zone's tiles). Summing up all S-2 tiles in all UTM zones yielded a total land area of 197.0 mil km^2, i.e. 134% in respect to the reference. For the nonpolar zone, which is our main interest, we obtained 179.3 mil km^2, i.e. 133%, respectively. In other words, data over the global nonpolar land surface increases by +33% when ingested into the Sentinel-2 grid.

Let us now check this result against the alternate calculation using the integration over the count of (overlapping) S-2 tiles. Fig. 3 plots these counts per individual land pixel in our data, ranging from 1 to 6 tiles per location (and few extreme cases with 7 and 8 tiles at 81°N, where 3 zones overlap). These redundancy maps illustrate the pattern of overlapping zones and tiles, featuring the increase by latitude, the stepwise zone diminution, and the graticule of tile overlaps (cf. with schematics in Fig. 1). Integration over this data yielded a total land area of 180.6 mil km^2 within the Sentinel-2 grid, i.e. 134%. We regard this slightly higher value to be within expected error margins, considering integer-rounding effects along the tile overlaps resolved at 1 km, and thus we can confirm the validity of our experiment.

The zoom into central Europe plotted in Fig. 3b gives insights into the structure of the confounding effects from the zone- and the tile-overlaps. Locations along sections of two overlapping UTM zones are generally included in two S-2 tiles (e.g. from the eastern Netherlands down to the Côte d'Azur). In conjunction with the inter-tile-overlap that generally yields a redundancy of 2 or 4, this may add up to a maximum of 6 tiles that cover a single location. The UTM zone diminution (e.g. on the border of France and Spain) leads to a stepwise narrowing of a

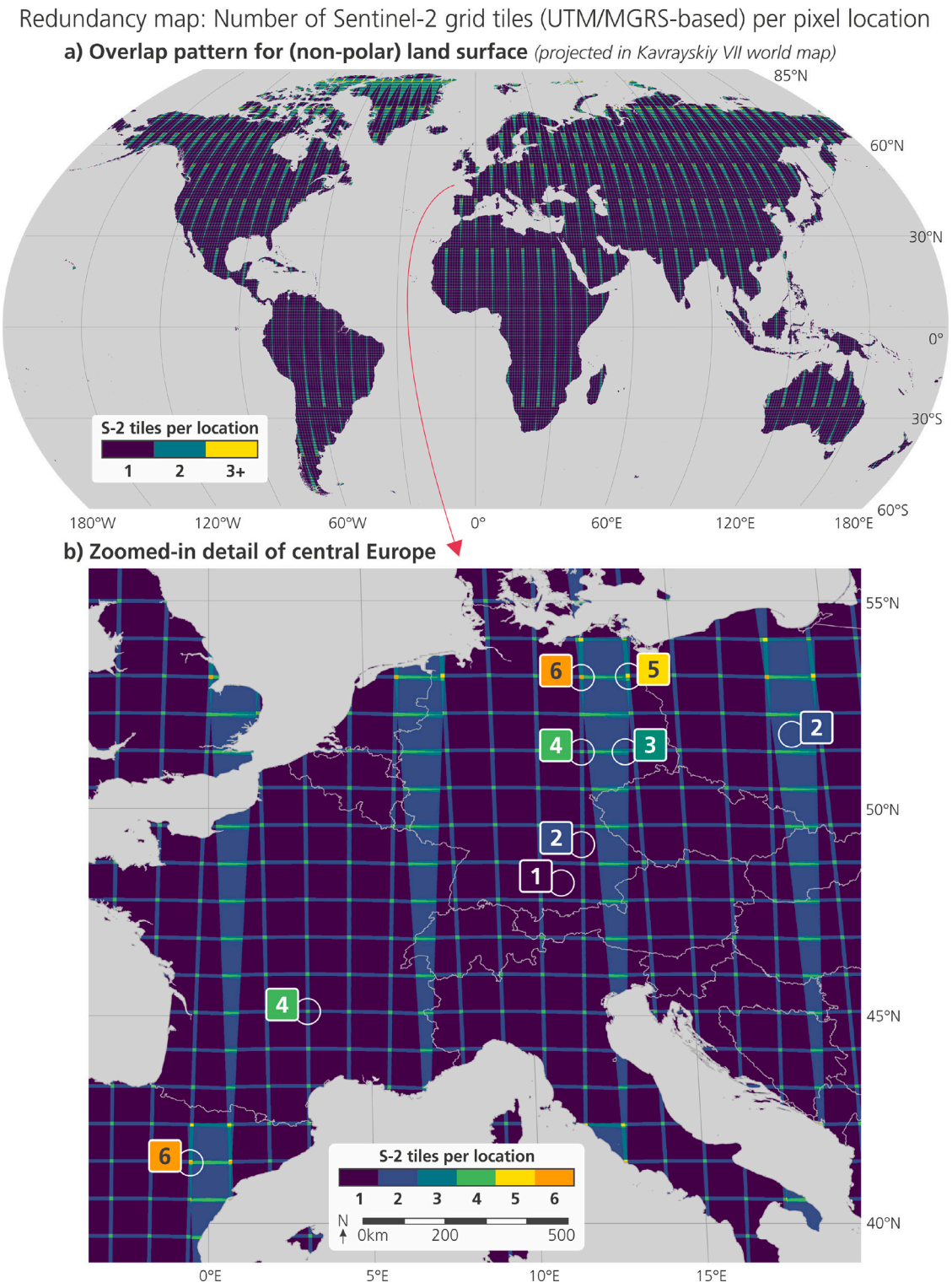

Fig. 3. Visualisation of the S-2 tile count per pixel location (also input to the alternate overhead calculation). (a) Global plot for the nonpolar domain, showing the typical pattern from the UTM zone overlaps. (b) Detail of the same data over central Europe, resolving the confounding effect of UTM zone- and S-2 tile-overlap. With indicators for typical locations covered by 1 to 6 tiles.

zone towards the poles and avoids excessive overlap (cf. the world map in Fig. 3a). Globally, the S-2 tile redundancy is distributed as follows: 70.5% of land area is covered by 1 single tile, 26.7% by 2 tiles, 1.0% by 3 tiles, 1.8% by 4 tiles, 0.03% by 5 tiles, and 0.02% by 6 tiles, clearly finding the latter as circumstantial cases where tile- and zone-overlap fully coincide.

Contemplating on the latitudinal component of the zone overlap, we are interested in how this affects different world regions. Therefore, we repeated the procedures on counting the land pixels within the S-2 tiles for each continent separately, by simply declaring all pixels outside the respective continent to be ocean. Table 1 shows these continental statistics on spatial overhead within the Sentinel-2 grid, following the common geographical division between continental land

masses. It becomes clear that in regions of low latitudes around the equator, such as Africa and South America, the Sentinel-2 grid has a smaller overall overhead (down to +28%). In contrast, continents with more parts in the higher latitudes like Europe or North America show a higher overhead (up to +39%). Looking at the North American example in Fig. 1b, this finding is well explainable by the higher density of the UTM zonal overlaps, e.g. over Greenland (cf. Fig. 3a).

A quick verification experiment on actual EO data confirmed our results from the general analysis with the land surface raster. From the Copernicus Global Land Service (CGLS, online, 2023), we obtained one layer of 10-day-aggregated Leaf Area Index (LAI, 300 m resolution, subset to Europe), projected to a single continental equidistant azimuth projection (with an estimated overhead of ∼+2%), and saved it to disk as GeoTIFF with LZW-compression (using the GDAL libraries again). When projected and tiled to the Sentinel-2 grid, the total volume of the LAI increased from 557 MB to 756 MB (+36%), an inflation agreeing well with the found overhead for Europe in Table 1.

As a last experiment, we analysed the dissolved UTM zones from Section 2.1 to distinguish between the effects from the zone- and the tile-overlaps. While the above result of 133% is the result from the individual 56 686 S-2 tiles, the following area calculation determines the overhead from just the overlaps between the 60 UTM zones, as the dissolving to single zone-polygons cancels the effect from the tile overlap graticule. From this data, we obtained a total of 152.2 mil km^2, which is 113% in respect to the reference of the nonpolar domain. Given the above 133% that stem from confounded overlaps from zones and tiles, the following net effects were determined: 17.5 mil km^2 (+13%) for the zone overlap, and 27.0 mil km^2 (+20%) for the tile overlap. This is in almost perfect agreement with initial expectations, as the stretching of the MGRS tiles from 100 to ∼110 km size constitutes a linear change to 110%, which let us anticipate a change in (squared) area to 121%.

4. Perspectives and conclusion

Today, platforms that offer practical and efficient access to satellite observations via datacubes and Analysis-Ready-Data (ARD) are becoming more and more popular in the EO community, and satellite missions launch initiatives towards image product dissemination at ARD level (e.g. Sentinel-1 by Truckenbrodt et al. (2019), or for Biomass by Banda et al. (2020)). The spatial reference of satellite imagery is a core property when it comes to the ingestion into datacubes, considering resampling efforts and inaccuracies when different projections need to be joined. The UTM reference system is used by various organisations for disseminating preprocessed and georeferenced satellite image products, and upcoming missions and product lines envisage the use of UTM-based data grids. Particularly the Sentinel-2 grid distributed by ESA assumes a prominent position, but nevertheless, there is only rough information on its characteristics available, and documentation on its exact tile configuration is contradictory and misleading (see last paragraph of Section 2.1).

Our study's motivation lies in the need for efficient datacube grid setups, which has been recently boosted by the increased focus on energy-saving processing and storage operations, and hence the quantification of the spatial overhead within UTM grids is most relevant. The nature of the overhead due to overlapping UTM zones has been already widely known and well understood, but has not been measured comprehensively so far. We scrutinised the geometries of ESA's Sentinel-2 tiling grid, and designed a global experiment to answer the question of how much data is duplicated by the grid when land observations are ingested.

At the outset, we gained insights on the Sentinel-2 tiling grid geometry: First, the stepwise zone diminution—that narrows the zones towards the poles—was found as a clever measure to avoid excessive overlap from the UTM projection system. It is self-evident that otherwise the zone overlap would be overwhelming for mid- and high-latitudes. Second, the Sentinel-2 tiles based on MGRS squares are extended precisely by 9.8 km in south and east direction, and thus create a second type of overlap forming a graticule within the zones. As this was not found in the literature, nor is it consistent with the documentation, we can only speculate that the tile overlap serves as a buffer to allow correcting slanting shadows cast by high clouds on the south and east edges of the initial 100 km tile. Our hypothesis would be in line with the sun shadow orientation at 10:30, the Sentinel-2 mission's Mean Local Solar Time. However, only regional users working in an area contained within a single tile could profit from this overlap data. As soon as a user's area-of-interest exceeds a single 110 km-sized S-2 tile, one must acquire the neighbouring tile(s) and is advised to deal with overlapping data in form of duplicated values carrying the same observational timestamp. Untreated, this can disturb statistical analysis and variable estimation by repeatedly factoring-in identical observations. In case of tile overlap, it could be resolved straightforwardly through picking one value per pixel. In case of UTM zone overlap, this is much more complex, as the pixel geometries are incongruent and the values are differently interpolated, hence requiring data manipulation and efforts on resampling and/or averaging.

Concerning our primary question—the spatial overhead in the Sentinel-2 grid over land—the experiment yielded clear results. Tying in with earlier analyses by Roy et al. (2016) on example test sites, we found that the global nonpolar land area is increased by 33% when ingested into the Sentinel-2 grid. This somewhat remarkable value is of similar magnitude as of much simpler grids based on the well-comprehended Plate Carrée projection. From our continental analysis we found that the redundancy in the Sentinel-2 grid increases towards north and south, confirming assumptions drawn from the apparently denser zone overlaps in higher latitudes. It was further found that up to 6 tiles collocate over one location, and that 2 tiles regularly overlap along the UTM zone borders. The separate analysis using the simplified UTM zone polygons (devoid of S-2 tiling) yielded an overhead of only 13%, and unfolded the S-2 tile-graticule's contribution to the overhead with 20% to be the larger net effect.

A refinement of the UTM zone diminution, and with bigger impact, the removal of the S-2 tile overlap would lead to a design that is more suitable to modern data infrastructures. The inter-tile overlap becomes obsolete when using established multi-array interfaces like xarray or STAC (Hoyer and Hamman, 2017; STAC, online, 2023), and a grid formed by the original 100 km-sized MGRS tiles would have a spatial overhead of +13% and overall only 1 or 2 tiles per location. However, the complications from splitting spatial data into 60 narrow zones—with each its own projection and pixel geometry—would remain, preventing easy handling and efficient processing of the imagery. Region- or nation-wide displays of satellite images and their derivates would still require manipulation and reprojection—at the costs of time and effort, computation power, and geometric fidelity.

The spatial overhead of +33% is an intrinsic characteristic of the UTM-MGRS-based Sentinel-2 grid, and as such it is opposing to current societal efforts to reduce energy and resource consumptions. We underline that data duplication within this grid, and the inevitable inter-zone resampling at downstream applications, provokes unnecessary spendings on hardware facilities and processing energy at both the provider's and the user's ends. This is detrimental to the Do No Significant Harm (DNSH) objective of the European Union to facilitate environmental sustainability (European Parliament, 2020). Disposing an inefficient spatial reference for EO data, and hence causing additional expenditures on natural resources in form of electricity and raw materials, should be avoided when alternatives are fit for the purpose.

In terms of budget costs, the overhead from the UTM-MGRS-Sentinel-2 grid can be considered tremendous. The example at hand, the Copernicus Sentinel-2 user-level products disseminated in the year 2021 accumulate to 4.18 PiB (Level-1C & -2A; Castriotta, 2022). As

these products are shipped as granules on the Sentinel-2 grid, the here found global overhead of +33% accounts for more than 1 petabyte of extra needed storage per annual Sentinel-2 record. Assuming typical rates for mass storage (as on cloud-providers like Amazon WS, Google CS, or Exoscale OS; Amazon, online, 2023; Google, online, 2023; Exoscale, online, 2023), the overhead on storage costs is grossing up to an additional ~250.000 – 270.000 EUR per year for each institution that disseminates or downloads and maintains the Sentinel-2 data. Accordingly, any institution that holds the entire Sentinel-2 archive—with a volume of 20 PiB hitherto—spends annually an extra of about 1.25 million EUR, and growing. Yet, these numbers do not include costs of backup, computation, I/O-operations, and downstream analysis. We stress that these add further significant expenses on hardware- and energy, and last but not least, in terms of transfer- and processing-time (Tamiminia et al., 2020). Overall, the grid-induced data duplication is cascading through the complete EO flow, from dissemination to value-added product analysis (Berriman and Groom, 2011; Cravero et al., 2022). Ultimately, it also burdens end users with small-area use cases—when those exceed a single S-2 tile, or worse, a single UTM zone—with an overhead on their efforts following the here found continental factors between +28% and +39%.

On these grounds, we advise against the use of the Sentinel-2 grid for upcoming satellite products and downstream ARD provision, and recommend other planar spatial references. Global grids based on a single equal-area projections like the Sinusoidal (used e.g. by NASA for MODIS) map the entire Earth onto one plane and do not require zone limits except for the 180°-dateline (or similarly the EASE-Grid with three planes Brodzik et al., 2012), but suffer from heavy geometric distortions towards its perimeters (Luo et al., 2008; Khlopenkov and Trishchenko, 2008). While Mercator-based grids as used by e.g. Google heavily inflate areas at mid- and high-latitudes, the mentioned Plate Carrée offers simplicity and a comparable spatial overhead (single-zoned +41% for nonpolar land) at the cost of latitudinal stretching in higher latitudes. A more recent alternative is the Equi7Grid (TUW, GitHub, online, 2023), which is based on an optimised compromise between projection distortions and land surface zoning into seven continents (including Antarctica), bears an global overhead of only +3%, and serves i.a. as datacube-grid in national and global land monitoring operations (ACube, online, 2023; Wagner et al., 2021) and geomorphological analysis (Hengl et al., 2017; Amatulli et al., 2020).

CRediT authorship contribution statement

Bernhard Bauer-Marschallinger: Conceptualisation of the study, Writing, Literature review, Methodology, Visualisation, Analysis, Investigation. **Konstantin Falkner:** Investigation, Software, Data preparation, Analysis, Methodology, Writing, Literature review.

Declaration of competing 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.

Acknowledgements

This study received funding from TU Wien, Austria, in particular the authors acknowledge TU Wien Bibliothek for financial support through its Open Access Funding Program. We would further like to thank our colleagues Wolfgang Wagner, Camillo Ressl, and Raphael Quast who supported us with feedback and proofreading.

References

ACube, 2023. Earth Observation Data Centre (EODC): ACube: The Austrian data cube. online. Available online: https://acube.eodc.eu/. Accessed on 6 February 2023.
Amatulli, G., McInerney, D., Sethi, T., Strobl, P., Domisch, S., 2020. Geomorpho90 m, empirical evaluation and accuracy assessment of global high-resolution geomorphometric layers. Sci. Data 7 (1), 1–18.
Amazon, 2023. Amazon S3 pricing: S3 Standard. online. Available online: https://aws.amazon.com/s3/pricing/. Accessed on 30 May 2023.
Aschbacher, J., Milagro-Pérez, M.P., 2012. The European Earth monitoring (GMES) programme: Status and perspectives. Remote Sens. Environ. 120, 3–8.
Baillarin, S., Meygret, A., Dechoz, C., Petrucci, B., Lacherade, S., Trémas, T., Isola, C., Martimort, P., Spoto, F., 2012. Sentinel-2 level 1 products and image processing performances. In: 2012 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 7003–7006.
Banda, F., Giudici, D., Le Toan, T., Mariotti d'Alessandro, M., Papathanassiou, K., Quegan, S., Riembauer, G., Scipal, K., Soja, M., Tebaldini, S., Ulander, L., Villard, L., 2020. The BIOMASS Level 2 Prototype Processor: Design and Experimental Results of Above-Ground Biomass Estimation. Remote Sens. 12 (6), http://dx.doi.org/10.3390/rs12060985.
Bauer-Marschallinger, B., Sabel, D., Wagner, W., 2014. Optimisation of global grids for high-resolution remote sensing data. Comput. Geosci. 72, 84–93. http://dx.doi.org/10.1016/j.cageo.2014.07.005.
Berriman, G.B., Groom, S.L., 2011. How Will Astronomy Archives Survive the Data Tsunami? Astronomers are collecting more data than ever. What practices can keep them ahead of the flood? Queue 9 (10), 20–27.
Brodzik, M.J., Billingsley, B., Haran, T., Raup, B., Savoie, M.H., 2012. EASE-Grid 2.0: Incremental but significant improvements for Earth-gridded data sets. ISPRS Int. J. Geo-Inf. 1 (1), 32–45.
Castriotta, A.G., 2022. Copernicus Sentinel Data Access Annual Report 2021, Vol. 2. Issue 1 Rev1 03/08/22, p. 2022.
CGLS, 2023. Copernicus Global Land Service. online. Available online: https://land.copernicus.eu/global/. Accessed on 11 February 2023.
Chatenoux, B., Richard, J.-P., Small, D., Roeoesli, C., Wingate, V., Poussin, C., Rodila, D., Peduzzi, P., Steinmeier, C., Ginzler, C., et al., 2021. The Swiss data cube, analysis ready data archive using earth observations of Switzerland. Sci. Data 8 (1), 295.
Claverie, M., Ju, J., Masek, J.G., Dungan, J.L., Vermote, E.F., Roger, J.-C., Skakun, S.V., Justice, C., 2018. The Harmonized Landsat and Sentinel-2 surface reflectance data set. Remote Sens. Environ. 219, 145–161.
Coluzzi, R., Imbrenda, V., Lanfredi, M., Simoniello, T., 2018. A first assessment of the Sentinel-2 Level 1-C cloud mask product to support informed surface analyses. Remote Sens. Environ. 217, 426–443.
Cravero, A., Pardo, S., Sepúlveda, S., Muñoz, L., 2022. Challenges to use machine learning in agricultural big data: A systematic literature review. Agronomy 12 (3), 748.
DMA, 1989. Defence Mapping Agency: The Universal Grids: Universal Transverse Mercator (UTM) and Universal Polar Stereographic (UPS). (DMA Technical Manual 8358.2).
Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., et al., 2012. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote Sens. Environ. 120, 25–36.
Dwyer, J.L., Roy, D.P., Sauer, B., Jenkerson, C.B., Zhang, H.K., Lymburner, L., 2018. Analysis Ready Data: Enabling Analysis of the Landsat Archive. Remote Sens. 10 (9), 1363.
Egorov, A.V., Roy, D.P., Zhang, H.K., Hansen, M.C., Kommareddy, A., 2018. Demonstration of percent tree cover mapping using Landsat Analysis Ready Data (ARD) and sensitivity with respect to Landsat ARD processing level. Remote Sens. 10 (2), 209.
ESA, 2015. European Space Agency: Sentinel-2 User Handbook. ESA Document, Issue 1 Rev 2, pp. 1–64.
ESA, 2023. European space agency: Data products: Sentinel-2 tiling grid KML file. online. Available online: https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-2/data-products. Accessed on 6 February 2023.
European Parliament, 2020. Regulation (EU) 2020/852 of the European Parliament and of the Council of 18 June 2020 on the establishment of a framework to facilitate sustainable investment, and amending Regulation (EU) 2019/2088 (Text with EEA relevance). pp. 13–43, URL: http://data.europa.eu/eli/reg/2020/852/oj.
Exoscale, 2023. Object Storage pricing. online. Available online: https://www.exoscale.com/pricing/#storage/. Accessed on 30 May 2023.
Frantz, D., 2019. FORCE—Landsat+ Sentinel-2 analysis ready data and beyond. Remote Sens. 11 (9), 1124.
Gascon, F., Bouzinac, C., Thépaut, O., Jung, M., Francesconi, B., Louis, J., Lonjou, V., Lafrance, B., Massera, S., Gaudel-Vacaresse, A., et al., 2017. Copernicus Sentinel-2A calibration and products validation status. Remote Sens. 9 (6), 584.
Google, 2023. Cloud Storage pricing tables. online. Available online: https://cloud.google.com/storage/pricing/. Accessed on 30 May 2023.
Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27.

Hengl, T., Mendes de Jesus, J., Heuvelink, G.B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., et al., 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLoS One 12 (2), e0169748.

Hoyer, S., Hamman, J., 2017. xarray: ND labeled arrays and datasets in Python. J. Open Res. Softw. 5 (1).

Kelso, N.V., Patterson, T., 2010. Introducing Natural Earth data – naturalearthdata.com. Geogr. Tech. 5 (82–89), 25.

Kempeneers, P., Soille, P., 2017. Optimizing Sentinel-2 image selection in a Big Data context. Big Earth Data 1 (1–2), 145–158.

Khlopenkov, K.V., Trishchenko, A.P., 2008. Implementation and evaluation of concurrent gradient search method for reprojection of MODIS Level-1B imagery. IEEE Trans. Geosci. Remote Sens. 46 (7), 2016–2027.

Kimerling, A.J., 2002. Predicting data loss and duplication when resampling from equal-angle grids. Cartogr. Geogr. Inf. Sci. 29 (2), 111–126.

Kopp, S., Becker, P., Doshi, A., Wright, D.J., Zhang, K., Xu, H., 2019. Achieving the full vision of earth observation data cubes. Data 4 (3), 94.

Lewis, A., Lacey, J., Mecklenburg, S., Ross, J., Siqueira, A., Killough, B., Szantoi, Z., Tadono, T., Rosenavist, A., Goryl, P., et al., 2018. CEOS Analysis Ready Data for Land (CARD4L) overview. In: IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 7407–7410.

Lu, M., Appel, M., Pebesma, E., 2018. Multidimensional arrays for analysing geoscientific data. ISPRS Int. J. Geo-Inf. 7 (8), 313.

Luo, Y., Trishchenko, A.P., Khlopenkov, K.V., 2008. Developing clear-sky, cloud and cloud shadow mask for producing clear-sky composites at 250-meter spatial resolution for the seven MODIS land bands over Canada and North America. Remote Sens. Environ. 112 (12), 4167–4185.

Mulcahy, K.A., 2000. Two new metrics for evaluating pixel-based change in data sets of global extent due to projection transformation. Cartogr.: Int. J. Geogr. Inf. Geovis. 37 (2), 1–12.

Natural Earth, 2023. European Space Agency: 1:10 m Physical Vectors: Land Polygons. online. Available online: https://www.naturalearthdata.com/http//www.naturalearthdata.com/download/10m/physical/ne_10m_land.zip. Accessed on 6 February 2023.

Roy, D.P., Li, J., Zhang, H.K., Yan, L., 2016. Best practices for the reprojection and resampling of Sentinel-2 Multi Spectral Instrument Level-1C data. Remote Sens. Lett. 7 (11), 1023–1032.

Snyder, J.P., 1987. Map Projections–A Working Manual, Vol. 1395. US Government Printing Office.

STAC, 2023. STAC: SpatioTemporal Asset Catalogs. online. Available online: https://stacspec.org/en/. Accessed on 30 May 2023.

Tamiminia, H., Salehi, B., Mahdianpari, M., Quackenbush, L., Adeli, S., Brisco, B., 2020. Google Earth Engine for geo-big data applications: A meta-analysis and systematic review. ISPRS J. Photogramm. Remote Sens. 164, 152–170.

TAS, 2022. Thales Alenia Space: Sentinel-2 Products Specification Document. Issue 14.9, pp. 1–552, URL: https://sentinels.copernicus.eu/web/sentinel. Accessed on 6 Februarly 2023.

Truckenbrodt, J., Freemantle, T., Williams, C., Jones, T., Small, D., Dubois, C., Thiel, C., Rossi, C., Syriou, A., Giuliani, G., 2019. Towards Sentinel-1 SAR Analysis-Ready Data: A Best Practices Assessment on Preparing Backscatter Data for the Cube. Data 4 (3), http://dx.doi.org/10.3390/data4030093.

TUW, GitHub, 2023. TU Wien: Equi7Grid: source geometries, python package, and documentation. Available online: https://github.com/TUW-GEO/Equi7Grid. Accessed on 27 February 2023.

Wagner, W., Bauer-Marschallinger, B., Navacchi, C., Reuß, F., Cao, S., Reimer, C., Schramm, M., Briese, C., 2021. A Sentinel-1 backscatter datacube for global land monitoring applications. Remote Sens. 13 (22), 4622. http://dx.doi.org/10.3390/rs13224622.

Wulder, M.A., Roy, D.P., Radeloff, V.C., Loveland, T.R., Anderson, M.C., Johnson, D.M., Healey, S., Zhu, Z., Scambos, T.A., Pahlevan, N., et al., 2022. Fifty years of Landsat science and impacts. Remote Sens. Environ. 280, 113195.

Yan, L., Roy, D., Li, Z., Zhang, H., Huang, H., 2018. Sentinel-2A multi-temporal misregistration characterization and an orbit-based sub-pixel registration methodology. Remote Sens. Environ. 215, 495–506.