 doi:10.20944/preprints202208.0484.v1

Technical Note

A Simple Procedure to Preprocess and Ingest High-Resolution Ocean Color Data into Google Earth Engine

Elígio de Raús Maúre [1,*], Simon Ilyushchenko [2], and Genki Terauchi [2]

[1] Northwest Pacific Region Environmental Cooperation Center; maure@npec.or.jp; terauchi@npec.or.jp
[2] Google LLC, 1600 Amphitheater Parkway, Mountain View, CA, USA; simonf@google.com
* Correspondence: eligiomaure@gmail.com; Tel.: (+81 76 445 1571)

Abstract: Data from ocean color (OC) remote sensing are considered a cost-effective tool for the study of biogeochemical processes globally. Satellite-derived chlorophyll, for instance, is considered an Essential Climate Variable since it is helpful in detecting climate change impacts. Google Earth Engine (GEE) is a planetary scale tool for remote sensing data analysis. Along with OC data, such tools allow an unprecedented spatial and temporal scale analysis of water quality monitoring in a way that has never been done before. Although OC data have been routinely collected at medium (~1 km) and more recently at high (~250 m) spatial resolution, only coarse resolution (≥4 km) data are available in GEE, making them unattractive for applications in the coastal regions. Data reprojection is needed prior to making OC data readily available in the GEE. In this paper, we introduce a simple but practical procedure to reproject and ingest OC data into GEE. The procedure is applicable to OC swath (Level-2) data and is easily adaptable to higher-level products. The results showed consistent distributions between swath and reprojected data, building confidence in the introduced framework. The study aims to start a discussion on making high resolution OC data readily available in GEE.

Keywords: remote sensing; ocean color; Google Earth Engine; MODIS/Aqua, SGLI/GCOM-C, swath reprojection, Earth Engine data ingestion

1. Introduction

Satellite observations of ocean color (OC) opened a new window for an unprecedented monitoring of water quality over spatial and temporal scales not feasible with in-situ sampling. While OC data have been collected at medium resolution (~1 km) and in some cases high resolution (~300 m), global datasets are only available at coarse resolutions (~4 km), limiting the potential benefits that coastal water managers could obtain from the data. In recent years, the number of high-resolution OC sensors has been increasing [1]. Nevertheless, global maps of geophysical parameters of interest for water quality monitoring, e.g., chlorophyll concentration, remain at the traditional 4 km spatial resolution [2]. Studies interested in the study of coastal phenomena such as red tides, harmful algal blooms, or eutrophication have always used high-resolution Level-2 data [2–5] and in some cases even lower-level data [6], which requires high technical expertise. Since the application of OC data in water quality monitoring is an emerging field, most practitioners lack the skills to handle these datasets. These difficulties, along with the ever-increasing data volume, pose additional challenges in terms of data access and processing [7], not to mention that different data providers have different data access requirements. To address such challenges a paradigm shift has emerged and cloud computing appears as a new norm for on-demand data analysis [8].

Google Earth Engine (GEE), a cloud computing platform with petabytes of remote sensing data and optimized for planetary scale analysis powered by Google cloud computing [9], is increasingly contributing to the application of remote sensing for environmental monitoring in various fields [10,11]. Land-based studies and applications of remote sensing data have achieved tremendous progress in recent years by taking

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 August 2022 doi:10.20944/preprints202208.0484.v1

advantage of the availability of historical Landsat collections in the GEE to map and monitor forest cover change, among others [11–16]. What puts GEE at the forefront of remote sensing data applications, in part, is its scalability and ease-of-use, the capacity to quickly share analysis ready maps with a simple link, and the possibility of publishing powerful dynamic web-applications. GEE has transformed the way remote sensing has been applied to science and decision making by greatly reducing the time and effort needed to work with large remote sensing datasets. Although the applications and potentials of GEE grow by the day, to date, the lack of high-resolution OC data remains a big deterrence to applying GEE in OC related studies.

To address the challenge of Level-2 OC data ingestion into the GEE, this study introduces a simple procedure for the preprocessing of OC data. The results of our analysis suggest that Level-2 OC geophysical parameters at their native spatiotemporal resolution and associated quality flag information can be remapped (reprojected) and ingested into GEE while retaining the source data statistical characteristics. Since GEE can handle any dataset at native resolution at the global scale, the significance would be unparalleled. Studies focusing on algorithm development, OC data performance evaluation with in-situ data, eutrophication, and red-tides monitoring, just to name a few, would benefit immensely. This would significantly contribute to calls for the application of Earth observations in support of Sustainable Developments Goals and the United Nations Decade of Ocean Science for Sustainable Development. This paper should encourage different OC users and data providers to preprocess and ingest high-resolution OC data into the GEE to support the use of these data in studies of coastal water phenomena and global change.

2. Materials and Methods

2.1 *Satellite data*

Level-2 OC imagery from the Moderate Resolution Imaging Spectroradiometer (MODIS) aboard Aqua satellite with a spatial resolution of 1 km, reprocessing 2018 (https://oceancolor.gsfc.nasa.gov/reprocessing/r2018/aqua/) was used to demonstrate the procedure followed in preprocessing and ingesting Level-2 data into the GEE. The U.S. National Aeronautics and Space Administration (NASA) Ocean Biology Processing Group (OBDG) disseminates its Level-2 data in Network Common Data Form 4 (netCDF4) file format. Each Level-2 file contains geophysical values for each pixel, derived from the Level-1 radiance by applying sensor calibration (for Level-1A), atmospheric corrections, and geophysical parameter algorithms. The contents of the Level-2 file are organized in (i) Global Attributes, (ii) Dimensions (Data Structure), and (iii) Groups. The groups contain (a) Sensor Band Parameters, (b) Scan-Line Attributes, (c) Geophysical Data, (d) Navigation Data, and (e) Processing Control. In our data ingestion procedure, we focus on (c) Geophysical Data while other information such as (b) and (d) was only used during the reprojection phase. Detailed description of the contents in the netCDF4 file can be found at https://oceancolor.gsfc.nasa.gov/docs/format/l2nc/.

MODIS/Aqua has the longest data record in the history of ocean color observations spanning a period of 20 years from 2002-07-04 to present, well beyond its design life of 6 years. This sensor measures spectral radiance at 36 bands with bands 8-16 dedicated to ocean color, phytoplankton, and biogeochemistry (https://modis.gsfc.nasa.gov/about/specifications.php). It has contributed immensely to studies of marine ecosystems variability and change, to monitoring coastal eutrophication, and in assisting policy making for the protection of our environment [2,17–19]. Although MODIS/Aqua continues to collect valuable data, with its aging, it may reach the end of life at any time. Meanwhile, new OC sensors with enhanced spatial and temporal resolutions were launched. One such example is the Second-generation GLobal Imager (SGLI) aboard the Global Change Observation Mission – Climate (GCOM-C, Shikisai) satellite of the Japan Aerospace eXploration Agency (JAXA) launched in December 2017 (https://shikisai.jaxa.jp/index_en.html). The Shikisai, with its high spatial resolution of 250

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 August 2022 doi:10.20944/preprints202208.0484.v1

m, is expected to enhance the study of coastal processes [20]. The other example is the first Geostationary Ocean Color Imager (GOCI) instrument launched in June 2010. GOCI, with its hourly images, effectively allows the study of short-term variability of coastal water phenomena such as red tides [3,21]. This paper provides data ingestion examples for MODIS/Aqua (R2018) and SGLI/GCOM-C (version 3, https://suzaku.eorc.jaxa.jp/GCOM_C/data/product_std.html) Level-2 OC products. The OBPG disseminates OC data from space agencies other than NASA such as MERIS (Medium Resolution Imaging Spectrometer) data from the European Space Agency or GOCI data from Korea Ocean Satellite Center. Since all Level-2 data disseminated by the Ocean Biology Data Active Archive Center (OB.DAAC, https://oceancolor.gsfc.nasa.gov/) have a consistent data structure, i.e., the same netCDF4 file format, the extension of our procedure to other sensors should be easy and straightforward.

Unlike MODIS/Aqua, Geophysical Data from the Shikisai are disseminated in two separate files; one for in-water properties (IWPR) and the other for normalized water leaving radiance (NWLR). The IWPR file contains derived chlorophyll-a concentration, absorption coefficient by colored dissolved organic matter (CDOM), and the concentration of total suspended matter (TSM). Remote sensing reflectance is derived from the NWLR file. The Shikisai data are disseminated in Hierarchical Data Format Version 5 (HDF5). Similar to netCDF4, the contents of the HDF5 are organized in groups with (i) Geometry Data (equivalent to (d) of netCDF4), (ii) Global Attributes, (iii) Image Data (equivalent to (c)), (iv) Level-1 Attributes, and (v) Processing Attributes. The data files used in this demonstration are listed in Table 1.

MODIS/Aqua data on 5th May 2022 was retrieved from NASA's OB.DAAC. The Shikisai data on 3rd May 2022 were retrieved from the JAXA's Global Portal System (G-Portal, https://gportal.jaxa.jp). Both data are freely available from the respective data portals upon registration.

Table 1. Listing of the files used in the data ingestion demo case. The Python code with the procedure discussed in this study is available from the link https://github.com/npec/ee-oc-data-ingestion.

Sensor/Satellite	Data file
MODIS/Aqua	A2022125035500.L2_LAC_OC.nc
SGLI/GCOM-C	GC1SG1_202205030152F05810_L2SG_IWPRQ_3000.h5
	GC1SG1_202205030152F05810_L2SG_NWLRQ_3000.h5

2.2 *Data Ingestion Workflow*

2.1.1 Swath Reprojection

The steps for the Level-2 OC data reprojection up until data ingestion into GEE are summarized in Figure 1. The first step begins with data retrieval from the providers, followed by the reprojection (remapping) of Level-2 imagery. The data is reprojected into a target map projection with a regularly-spaced grid. Level-2 data is in the shape defined by the satellite field of view representing the portion of the surface observed by the sensor as it revolves around the Earth. These Level-2 images are often called "swaths". They represent 2-dimensional (2D) grids of the portion of Earth seen by the satellite sensor and arranged in x-y coordinates representing, respectively, the pixels per line (the field of view) and number of lines along the flight path. Such datasets, while gridded, are not projected onto a regular grid, that is, they are non-uniformly spaced. It, therefore, is impossible to ingest such data into the GEE directly. Pyresample, a Python package for the resampling of swath dataset into a grid, or a grid into a swath, or a swath to another swath, was used to remap swath data using the nearest neighbor (NN) method (https://pyresample.readthedocs.io/en/latest/). The NN is located using a fast KD-tree algorithm provided by the pykdtree library. KDTree is short for K-Dimensional Tree, a space-partitioning data structure for organizing points in a k-dimensional space [22] (in the case of swath data, a 2-dimensional tree). Pyresample uses 1 neighbor in its NN query.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 August 2022 doi:10.20944/preprints202208.0484.v1

This is appropriate for our resampling objective, since Level-2 quality flag information can be effectively transferred to the target geolocation point without noise, something that would be very challenging if >1 neighbors were used. In the query for NN, one has to supply the region of influence, which Pyresample uses as a cutoff distance from the target pixel center. As we will show later in section 3, we choose a value twice that of the spatial resolution to compensate for the decrease in resolution as one moves away from the swath center. Since the NN will return only neighbors within the cutoff distance, some pixels towards the swath edge may not have a neighbor if the actual spatial resolution is used. This results from the fact that the pixels at the swath edge are significantly larger than their counterparts at the center (Appendix A, Figure A1).

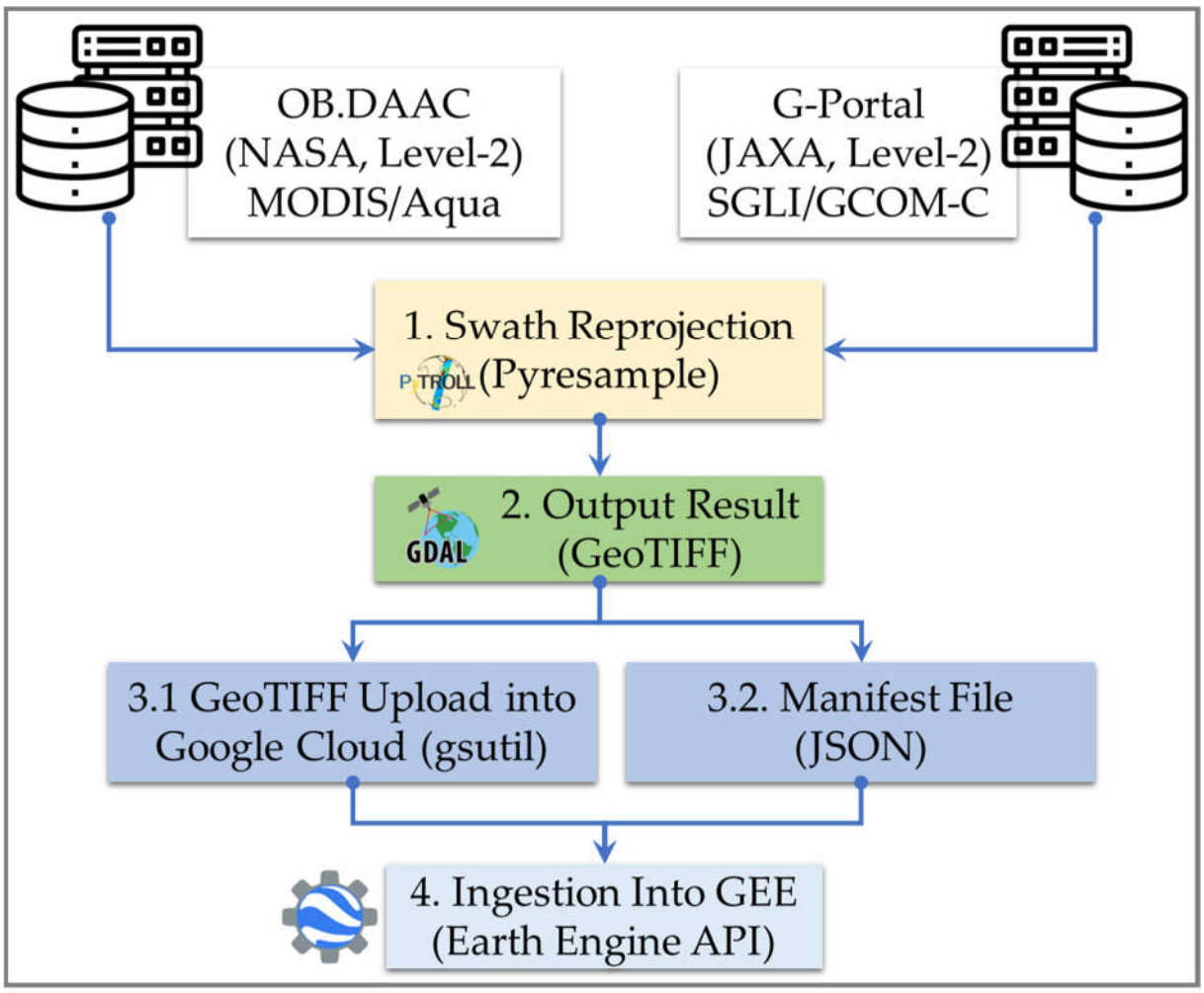

Figure 1. Schematic of the data reprojection and ingestion into GEE. The steps are shown in numbers and a detailed explanation is given in the text.

All bands of either (c) the netCDF4 file or (iii) the HDF5 file can be passed at once into Pyresample as stacked 2D arrays. During reprojection, Pyresample requires two geolocation inputs: one for the swath (source geolocation) and the other for the target map. Source geolocation was derived from the swath navigation data or geometry data depending on whether the file was netCDF4 or HDF5. The target geolocation was created using Pyresample geometry helper functions. To create a new target geographic area of regularly spaced pixels, Pyresample required a pre-defined projection information, area extent, spatial resolution, height and width, and other relevant information of the target area. The projection and area extent were created with the help of pyproj, a Python interface to PROJ (cartographic projections and coordinate transformations library, https://pyproj4.github.io/pyproj/stable/index.html). We choose the LAEA (Lambert

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 August 2022 doi:10.20944/preprints202208.0484.v1

azimuthal equal-area projection) since it preserves the pixel area across the domain of the target map projection. This projection is a good choice, for instance, in ocean color data applications such as in primary productivity due to its flux-preserving nature.

We used Level-2 swath imagery that intersected our area of interest, the Northwest Pacific region (115-155 E, 20-60 N). However, for the reprojected swath, the area extent was defined by the geolocation boundaries of the input swath. We extracted the center longitude and latitude of the swath and used it to initialize the projection in pyproj. The center longitude and latitude can be obtained, for example, from the Scan-Line Attributes in the netCDF4 file. Since the swath images contain the information of the center longitude/latitude, the center of the target projection would be located half-way across the scan direction (number of lines). After the projection is initialized, geolocation bounds of the swath are translated into corresponding values of the target projection (Table 2). Using the MODIS/Aqua file (Table 1), in Table 2 we provide a Python code snippet of the projection definition and translation of swath bounds. Bounds were obtained from the Global Attributes and the center longitude and latitude from the Scan-Line Attributes as mentioned above. Since the HDF5 file does not contain information of the center scan line, we estimated this information based on the great circle distance and using the swath boundary longitude and latitude data. When the same approach was applied to MODIS/Aqua, the results were consistent with the information given in the netCDF4 file.

Table 2. Python code snippet for projection initialization. The Lambert azimuthal equal-area projection (laea) is initialized with the center located at the swath scan line center longitude/latitude median point. The initialized projection was then used to translate the swath bounds in degrees into the target projection distances in meters. WGS84 stands for World Geodetic System (WGS) 1984, consisting of a reference ellipsoid, a standard coordinate system, altitude data, and a geoid.

```
import pyproj

left_lon  = 116.2261 # westernmost_longitude
lower_lat = 34.5965  # southernmost_latitude
right_lon = 152.5978 # easternmost_longitude
upper_lat = 56.271   # northernmost_latitude
lon_0, lat_0 = 136.4641, 46.1208

proj_dict = dict(datum='WGS84', lat_0=lat_0, lon_0=lon_0, proj='laea')
proj = pyproj.Proj(proj_dict)
lower_left_x, lower_left_y = proj.transform(left_lon, lower_lat)
upper_right_x, upper_right_y = proj.transform(right_lon, upper_lat)
area_extent = lower_left_x, lower_left_y, upper_right_x, upper_right_y
area_extent
(-1843501.546690065, -1052852.120358288, 994320.1613132474, 1233242.3976159848)
```

After the bounds are defined in the same units as the projection, the target area was created using Pyresample's *create_area_def()* function (Table 3). Note that the nominal spatial resolution of MODIS/Aqua is 1 km or 1000 m as indicated in the file attributes. However, 1001 m was passed to Pyresample. In the case of SGLI/GCOM-C with 250 m, 251 m was passed. This was done to ensure that the adjusted pixel resolution of the target projection did not return smaller than nominal resolution. Internally, Pyresample determines the spatial resolution based on the height and width of the target projection. Since we did not provide such information, Pyresample obtains the height (rows) and width (columns) from the area extent and the provided spatial resolution. If the shape is already known, it can also be passed to Pyresample, however, the resolution will then be adjusted accordingly. We could also estimate the shape of the target projection using the pixel resolution of 1 km (1000.90 m) of the Equal-Area Scalable Earth (EASE) grid projections (https://nsidc.org/ease/ease-grid-projection-gt). Nevertheless, the resulting resolution obtained by using the above procedure was similar.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 August 2022 doi:10.20944/preprints202208.0484.v1

Table 3. Python code snippet of Pyresample area definition for the target projection. The area definition is used in step 1 of swath reprojection.

```
from pyresample import create_area_def

target_area = create_area_def(
    projection=proj_dict,
    area_extent=area_extent,
    resolution=1001,
    units='metres',
    area_id='nowpap_region'
)
target_area
```

```
WARNING:root:shape found from radius and resolution does not contain only integers:
(2283.810707267006, 2834.9867212820304) Rounding shape to (2284, 2835) and resolution from
(1001.0, 1001.0) meters to 1000.9953114650132, 1000.9170393932893) meters
Area ID: nowpap_region
Description: nowpap_region
Projection: {'datum': 'WGS84', 'lat_0': '46.1208', 'lon_0': '136.4641', 'no_defs': 'None',
'proj': 'laea', 'type': 'crs', 'units': 'm', 'x_0': '0', 'y_0': '0'}
Number of columns: 2835
Number of rows: 2284
Area extent: (-1843501.5467, -1052852.1204, 994320.1613, 1233242.3976)
```

2.1.2 Output Result

The remapped swaths by Pyresample NN method were then, in step 2, saved in GeoTIFF file format. Geospatial Data Abstraction Library (GDAL, https://gdal.org/) was used to create the GeoTIFF file and the metadata of this file were copied from the source netCDF4 or HDF5 file. The copied metadata are those of the Global Attributes and attributes of each variable with the exception of geolocation information. Unlike the netCDF4 file, GeoTIFF file does not allow a mixture of data types, that is, a mix of integers and floating-point numbers in different bands. In the netCDF4 file, geophysical parameters are saved as integers except for chlorophyll data which is saved as a floating-point number. In the case of the HDF5 file Image Data parameters are saved as integers. Since bit-shifting operations are not supported on floating-point numbers, all data were scaled and saved as 4-byte signed integers (int32). This was done for consistency with the Level-2 quality flag information of NASA OC files. The chlorophyll data were scaled into integers by subtracting an offset of 0.001 [mg m^{-3}], and dividing by the scale factor of 1e-6, eq. (1). Integers have also the advantage of being easily compressed.

$$\text{scaled\_data} = (\text{data} - \text{offset}) / \text{scale\_factor} \quad (1)$$

2.1.3 GEE Pre-Ingestion Step

Prior to the upload and ingestion of the created GeoTIFF into GEE, it is important to verify that all geolocation tags are correctly written in the file (Table A1). GDAL provide s methods to translate the GeoTIFF (*gdal_translate*) into a correctly geolocated image. Mor eover, before the translation, we can run a GDAL utility (*gdaladdo*) to build overviews so that the created file becomes Cloud Optimized GeoTIFF (COG, https://gdal.org/drivers/r aster/cog.html#raster-cog). COG is a regular GeoTIFF but with added pyramid overview s, which help with the efficiency of workflows in the cloud. GEE is able to read or load th ese COGs without ingestion, directly from the Google Cloud storage (https://developers. google.com/earth-engine/Earth_Engine_asset_from_cloud_geotiff). However, for files bei

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 August 2022 doi:10.20944/preprints202208.0484.v1

ng ingested into GEE, it is not worth creating COG files since GEE does not currently ma ke use of the overviews. The COG component will be simply ignored and GEE will creat e its own COG on the fly. It is also worth noting that *gdal_translate* can write COGs direct ly to the output file (https://gdal.org/drivers/raster/cog.html).

After the GeoTIFF file is translated, it can be uploaded into GEE directly through Code Editor or through Google Cloud bucket using the gsutil tool (https://cloud.google.com/storage/docs/gsutil) followed by issuing an Earth Engine (EE) upload command in the EE command-line tool. Using a Google Cloud bucket is preferable if one is ingesting more than a dozen images, in which case manual ingestion through Code Editor (https://developers.google.com/earth-engine/guides/image_upload) becomes overwhelming. Note that with a cloud bucket the ingestion process can be automated although bucket charges will be incurred.

After uploading the GeoTIFF into a cloud bucket, a JSON (JavaScript Object Notation, https://docs.python.org/3/library/json.html) manifest file needs to be created. The structure of the JSON file, although complex, gives more flexibility to the user during the ingestion process. For instance, a multiband GeoTIFF can be uploaded along with its per-band names, something not feasible with the Code Editor. Additionally, multiple source files can be combined (mosaiced) into a single image; separate files representing different bands can be ingested into a multiband image. Further details can be found at https://developers.google.com/earth-engine/guides/image_manifest. The manifest file should, at least, contain the target asset ID (the name where the image will be uploaded to) and the cloud bucket address where the GeoTIFF file resides in the Google Cloud. Furthermore, global attributes, per-band attributes, and other relevant metadata can also be appended. A sample code for creating the JSON manifest file using the example images is included in the GitHub repository (https://github.com/npec/ee-oc-data-ingestion).

2.1.4 Ingestion of the GeoTIFF into GEE

The final step is simply to upload the processed GeoTIFF into the GEE. This is achieved using the EE Python API (https://developers.google.com/earth-engine/tutorials/community/intro-to-python-api). As mentioned above, the GeoTIFF file is first uploaded to Google cloud bucket using gsutil. A JSON manifest file with the GeoTIFF cloud location, the associated metadata, and the target asset ID is then created. This JSON manifest file is then passed to the EE Python API command line tool to perform the ingestion. While creating the JSON manifest file we can also indicate the pyramiding policy to be used by the GEE when generating the COG file. Options for the pyramiding policy include "MEAN" (default), "MODE", and "SAMPLE". GEE documentation recommends using “SAMPLE” where “MEAN” or “MODE” does not make sense as in the case of quality flag information (https://developers.google.com/earth-engine/guides/image_manifest).

In summary, we obtained the swath image from the space agency’s data archive, reprojected the swath using Pyresample, and saved the results in GeoTIFF file. The GeoTIFF was then sent to a Google Cloud bucket and subsequently ingested into the GEE using a JSON manifest file through EE Python API. The full implementation of the above steps can be found at https://github.com/npec/ee-oc-data-ingestion.

3. Results

We describe the results obtained using the above detailed steps. We compared the swath with the remapped data within a subset area. The subset area was obtained using a polygon defined by the swath geolocation data. After the swath was remapped, the statistics of the pixels falling within this polygon were compared with the original swath data. The results are detailed below.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 August 2022 doi:10.20944/preprints202208.0484.v1

3.1 *Mapped Imagery and the Radius of Influence*

We first compared the impact of the radius of influence (cutoff distance) on remapped imagery. We tested two values, one equal to the nominal spatial resolution (R) and the other twice the spatial resolution (2R, Figure 2). With a cutoff of R, salt-and-pepper-like noise increases as we move towards the swath edge. The size of the pixels increases significantly towards the swath edge so that no NN pixels fall within the distance R. However, with a cutoff of 2R, this scenario is reversed and the remapped result now looks as smooth as the input swath data. It is worth noting that the 2R cutoff adopted will not entirely solve this issue since the pixel size in the across-track direction (horizontal direction from center) becomes 4 times larger than that at the sub-satellite point (nadir view). In the case of data near the swath center this effect is not observed since the resolution of both swath and projected grid are almost the same (1:1 correspondence, Figure A1). Consequently, for the mapping procedure introduced here, the adopted cutoff value is equal to 2R. Doing so, we minimize significant data gaps introduced by the decrease in resolution towards the swath edge.

In the case of SGLI/GCOM-C, although the same 2R was used, the difference in using R or 2R was much smaller even near the swath edge. MODIS/Aqua has a swath width of about 2330 km, which is about 50% wider than that of SGLI/GCOM-C (1150 km). Also, at the swath edge the along-track resolution of SGLI/GCOM-C remains consistent to that of the sub-satellite view and only the across-track becomes twice as big. In contrast, for MODIS/Aqua, both along and across track pixel sizes become, respectively, twice and four times as large (Figure A1 in Appendix A).

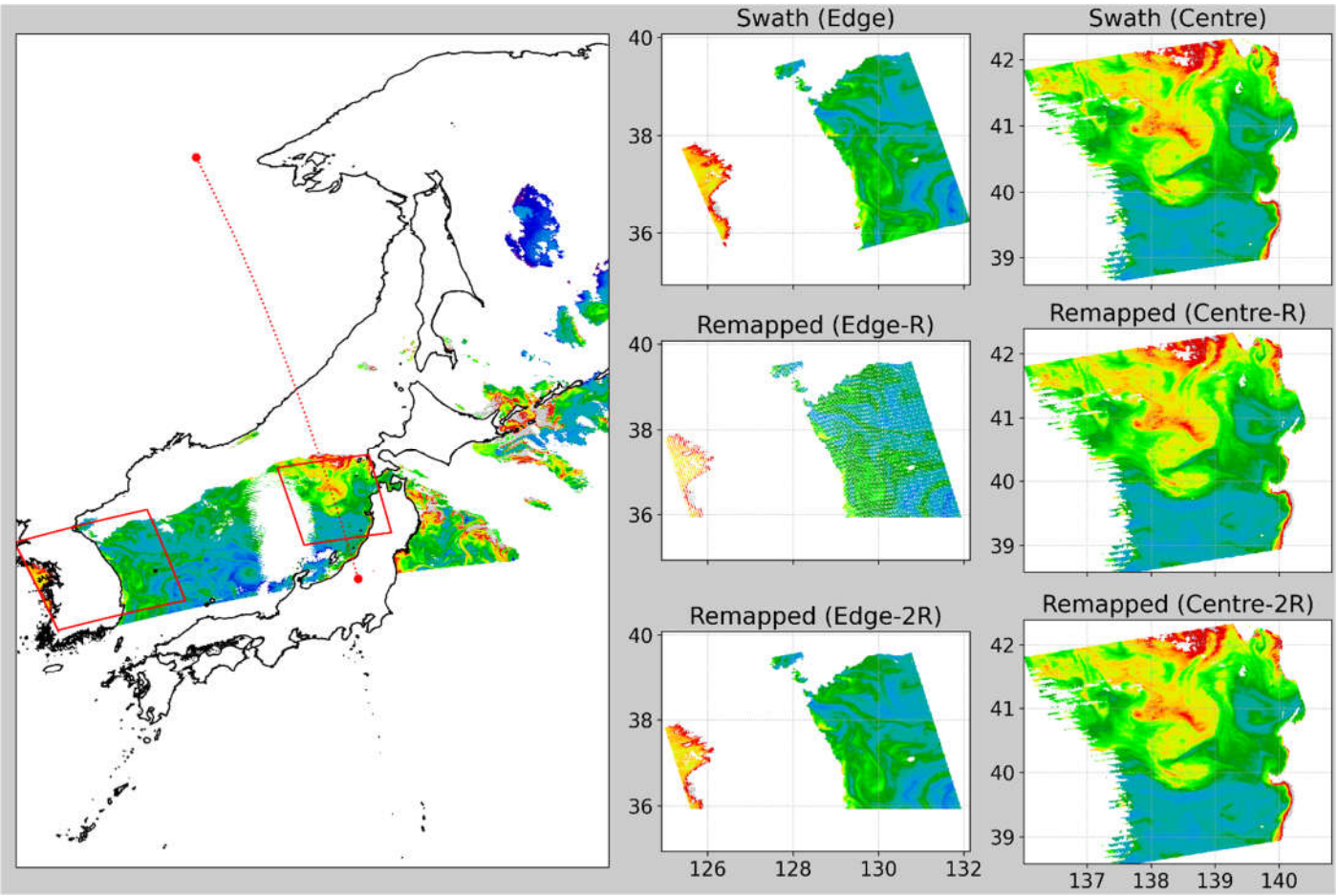

Figure 2. Reprojection of swath data for the MODIS/Aqua file (Table 1). The reprojection was done using the NN method. The impact of using R versus 2R is shown on the right panel where the remapped swath center and edge are compared. Similar example is also given for SGLI/GCOM-C in Figure A2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 August 2022 doi:10.20944/preprints202208.0484.v1

3.2 *Verification of the Remapped Results*

We verified the results by examining the probability density of the original swath versus that of reprojected data (Figure 3). We used data from polygons at the swath edge (see red box in Figure 2 and Figure A2) where pixel distortions are relatively large. The results remained consistent between the original swath and the remapped data. This was expected, since the NN method simply copies the data from the nearest neighbor of the target cell. The larger pixel size of the swath at the edge results in oversampling, which increases pixel density at the destination map (target area map). In Figure 3b, the HISATZEN flag, for instance, had a relative increase of ~191% (100 * (remapped – swath) / swath). HISATZEN stands for "sensor view zenith angle exceeds threshold" and at swath edge this value can exceed 60°. In the case of SGLI/GCOM-C the results remain consistent since the pixel resolution does not degrade as much. The sample size of the output for the subset area also increased, as the remapped image retains the resolution of the center (Table 4).

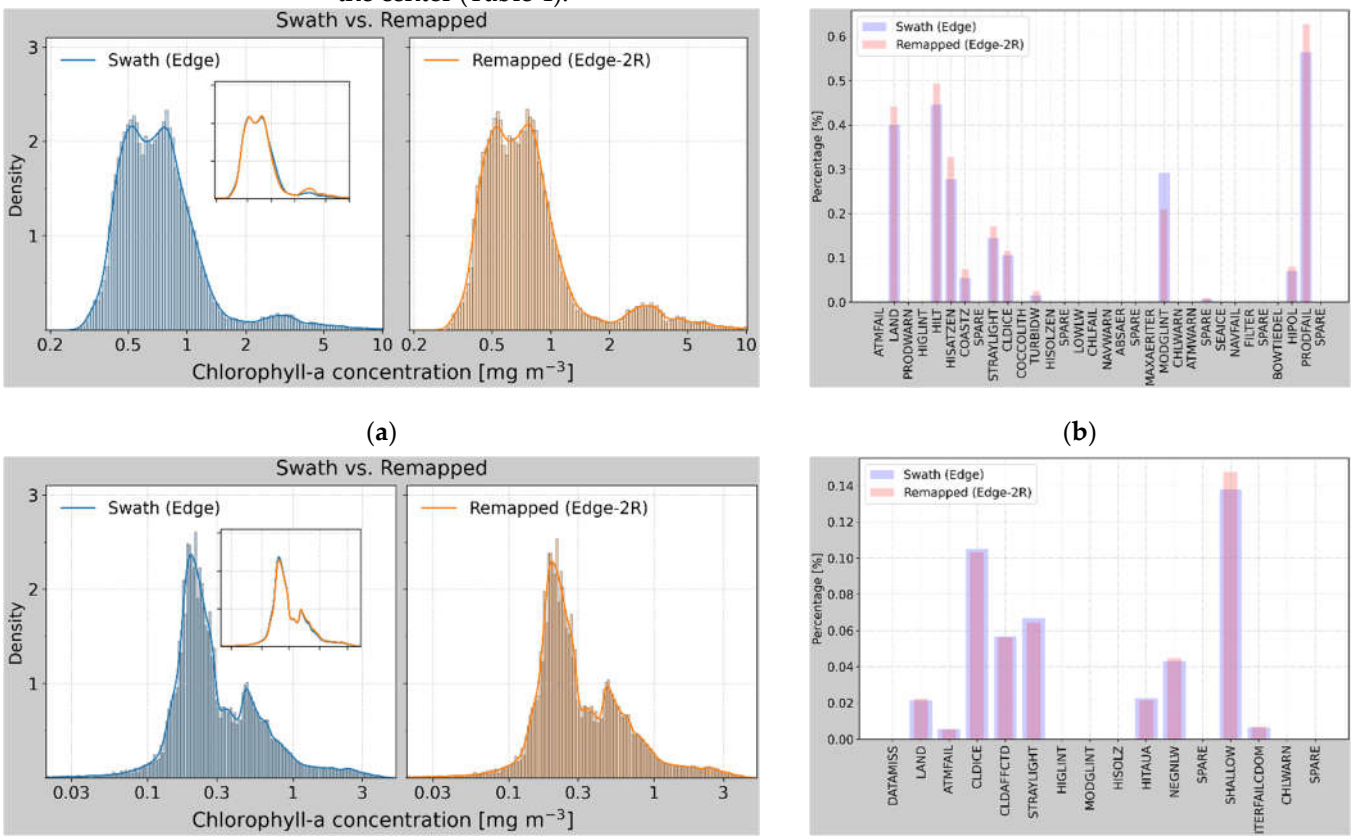

Figure 3. Density histogram of swath versus remapped image. (a, c) chlorophyll data and (b, d) the frequency distributions of associated quality flag information. (a, b) Diagrams were obtained from MODIS/Aqua data (Figure 2) and (c, d) from SGLI/GCOM-C data (Figure A2). These diagrams wer e created from the red polygons highlighted at the swath edge in the respective figures. The overla id curves are the probability distribution functions (PDF) for the same samples. The inset compares the two PDFs (swath versus remapped). Note the increase in pixel number in the range of 2 to 10 m g m^{-3} associated with the high chlorophyll at the swath edge in (a). Description of MODIS/Aqua fla gs can be obtained from https://oceancolor.gsfc.nasa.gov/atbd/ocl2flags/. For more on SGLI/GCOM -C quality flags, see https://suzaku.eorc.jaxa.jp/GCOM_C/data/files/ATBD_ocean_ac_murakami_v 2_en.pdf.

As mentioned above, the mean and standard deviation (STD) of MODIS/Aqua (Table 4) remapped data (2R) is slightly higher than the swath data. This is caused by the relative increase in pixels with high chlorophyll concentration at the swath edge. Nevertheless, both swath and remapped data follow the same distributions with similar data limits.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 August 2022 doi:10.20944/preprints202208.0484.v1

SGLI/GCOM-C results were more consistent as the spatial resolution is much finer and pixel distortions at the edge remain minimal. The relative increase in valid pixel count between swath and remapped image was about 116% for MODIS/Aqua but only 31% for SGLI/GCOM-C.

Table 4. Descriptive statistics for the swath and reprojected data in the subset area at the swath edge. The statistics were obtained on log-transformed data and results back-transformed to linear. All data are in mg m^{-3} except for count, which is the number of pixels in the polygon extract.

	MODIS/Aqua (Edge)			SGLI/GCOM-C (Edge)		
	Swath	Remapped (R)	Remapped (2R)	Swath	Remapped (R)	Remapped (2R)
Count	36591	62630	79000	1313197	1720520	1720520
Mean	0.740	0.723	0.778	0.293	0.307	0.307
STD	1.703	1.697	1.849	2.082	2.118	2.118
Min	0.269	0.292	0.292	0.002	0.002	0.002
25%	0.518	0.513	0.523	0.190	0.194	0.194
50%	0.685	0.666	0.692	0.246	0.254	0.254
75%	0.906	0.864	0.912	0.458	0.483	0.483
Max	93.196	93.196	93.196	89.963	89.963	89.963

JAXA provides a Windows-based Earth Observation Data Conversion Tool (DCT, https://gportal.jaxa.jp/gpr/information/tool?lang=en) that can be used to remap SGLI/GCOM-C data prior to ingestion into the GEE. With this tool we can select the same NN resampling as in our Python-based remapping procedure (PY) but it is not clear from the documentation how this tool applies the NN method during the remapping. As such, it is difficult to have an objective comparison with our procedure. Moreover, the target projection is also fixed to Geodetic latitude/longitude. Our brief comparison of the DCT remapped data with the swath and the PY method showed consistent statistical metrics (not shown). However, spatial differences between the PY and DCT were evident (https://code.earthengine.google.com/4cf92c3e77a5ec53154ec6a168c870bd?hideCode=true). In this link, the detailed spatial differences between the PY- and DCT-based remapping at the subpixel level were apparent, as the latter uses a degree-based gridding while the former uses a distance-based step. The DCT had a significant increase in valid pixel count—83% relative to swath and 39% relative to PY. The histogram of the difference between PY and DCT highlighted these subpixel differences (Figure 4) since the statistical characteristics within a given polygon remained consistent (please see the histograms in the GEE link above). It would be interesting to see how these differences in swath remapping strategies are reflected in validation analysis as discussed for Level-2 and higher-level products in [25]. However, such discussion is beyond the scope of this study.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 August 2022 doi:10.20944/preprints202208.0484.v1

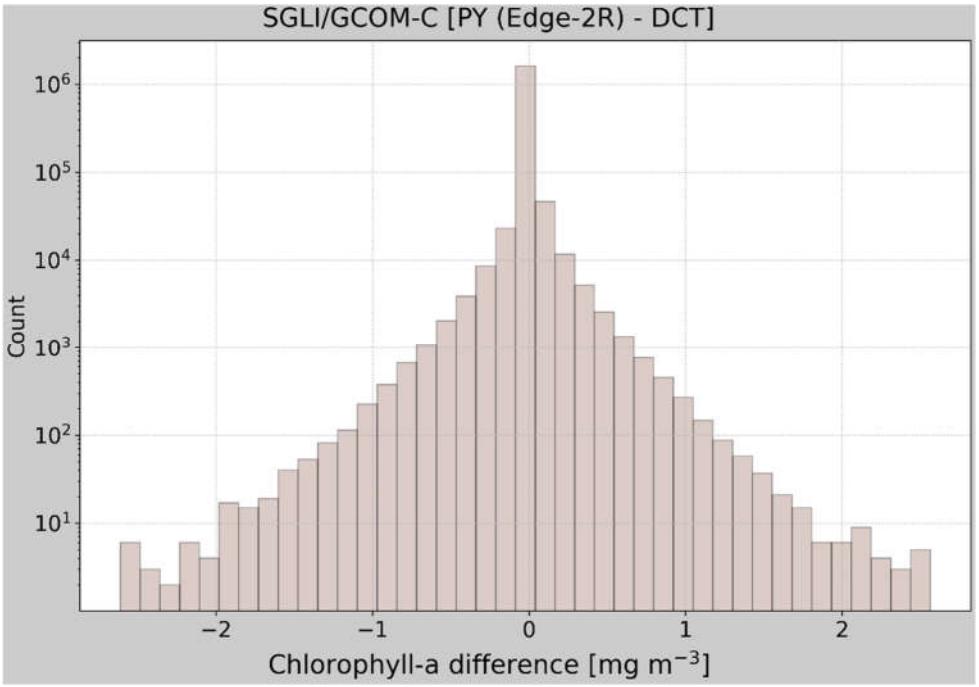

Figure 4. Chlorophyll-a difference between PY (Python-based) and DCT (JAXA's Earth Observation Data Conversion Tool based) remapped data. Note the logarithmic scale on the y-axis.

Similarly, NASA also has an official Data Analysis Software, SeaDAS, primarily designed for OC data (https://seadas.gsfc.nasa.gov/downloads/). SeaDAS is able to remap and directly output GeoTIFF files. Testing the suitability of different mapping software is out of scope of the current study. With the SeaDAS we can also use the same Lambert Azimuthal Equal Area projection and NN method as in the PY remapping. Thus, we expect that the results would be consistent between the two software.

4. Discussion

We introduced a simple procedure to make Level-2 OC geophysical parameters at their native spatiotemporal resolution, as well as the associated quality flags, readily available in the GEE. Once ingested these datasets can be curated and be made available through the GEE data catalog. The results clearly confirmed that the distributions of the swath versus remapped data remain consistent. This, in part, is because the mapping strategy only considers a single swath so that issues introduced by temporal averaging are excluded. The choice of the target projection, the Lambert Azimuthal Equal Area projection, also leads to transferring the swath resolution onto the target projection, thereby effectively avoiding spatial distortions and possible errors introduced by equirectangular grids. Space agencies routinely remap swath data onto a spatial grid over a certain time period. Some of these spatiotemporally aggregated products mapped onto a nearly equal-area integerized sinusoidal grid [23] are denoted as Level-3 binned products. Globally binned maps are then used to create standard mapped products on equirectangular (Plate Carrée) projection. While Level-3 products are more accessible to non-expert users than products at lower processing levels, their spatial resolution of 4.64 km is too coarse for applications in nearshore and/or inland waters [24]. Moreover, satellite-to-in-situ validation analysis showed that the spatial and temporal binning of swath data beyond the swath spatiotemporal reference introduce uncertainties that result in discrepancies between Level-2 and Level-3 performance metrics [25].

This study does not intend to introduce a new standard for mapping Level-2 OC data, but emphasizes the opportunities provided by the flexibility and scalability of GEE for OC applications. The availability of higher spatial resolution datasets has been increasing in recent years, but advances in their applications have been slow, in part,

 doi:10.20944/preprints202208.0484.v1

because of challenges in working with these datasets using conventional means, in which case GEE can fill the gap.

The availability of high-resolution Landsat and other public datasets in GEE has allowed the study of various environmental changes—extending to both management and conservation activities—globally at unprecedented spatial and temporal scales [16,26–30]. The Global Eutrophication Watch [2] is among the few OC data application studies to leverage the potentials of GEE for interactive mapping of global change of coastal ecosystems. The lack of high spatial resolution OC products in the GEE remains a limiting factor. For instance, the GEE IssueTracker includes researcher and practitioner requests for addition of higher OC datasets for monitoring of coastal phenomena like cyanobacteria blooms (e.g., issues 34740451 and 139056084). This study, in part, responds to this demand by contributing a simple but practical way of porting Level-2 OC datasets into GEE. Ideally, data ingestion would be sustainable if it is happening in concert with space agencies.

We anticipate that this study will pave the way to making high resolution OC datasets readily accessible through the GEE data catalog. Such data, among others, can immensely benefit efforts to develop regional algorithms and OC data validations activities. Intercalibration of OC products from different OC sensors is another laborious and resource-intensive task that can benefit enormously from the power of GEE. The availability of these high-resolution datasets in GEE can further enhance the efforts of the Group of Earth Observations (GEO) in making Earth observations data more accessible to support science and decision making [7]. The partnership between GEO and GEE is one example of the efforts being made towards the use of Earth observations in support of sustainable development goals. The availability of high-resolution datasets in GEE will contribute to the Ocean Decade Challenges for collective impact in which challenge 7 is aimed at expanding the Global Ocean Observing System.

Author Contributions: Conceptualization, E.R.M., methodology, E.R.M.; software, E.R.M.; formal analysis, E.R.M.; investigation, E.R.M.; resources, E.R.M., G.T., and S.I.; data curation, E.R.M. and S.I.; writing—original draft preparation, E.R.M.; writing—review and editing, E.R.M., G.T., and S.I.; visualization, E.R.M.; project administration, G.T.; funding acquisition, E.R.M., and G.T., All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: The sample files were obtained from the NASA's OB.DAAC (https://oceancolor.gsfc.nasa.gov/) for the MODIS/Aqua sensor and JAXA's G-Portal (https://gportal.jaxa.jp/gpr/?lang=en) for the SGLI/GCOM-C sensor. The preprocessed and GEE ingested version of the sample file for MODIS/Aqua can be found at GEE address "*projects/ee-eutrophication-gee4geo/assets/OCEANDATA/L2/Test/A2022125035500_L2_LAC_OC*". The PY version of the SGLI/GCOM-C can be found at "*projects/ee-eutrophication-gee4geo/assets/GC1SG1_202205030152F05810_L2SG_IWPRQ_3000_py*", and the DCT version is located at "*projects/ee-eutrophication-gee4geo/assets/ GC1SG1_202205030152F05810_L2SG_IWPRQ_3000_dct*". All three files are visualized in GEE at the link https://code.earthengine.google.com/19cc78d1ef66def625c664010de8ea85?hideCode=true. The links also include the center and edge polygon data used in the text.

Acknowledgments: This work was made possible through the support received from the Ministry of the Environment of Japan, the Toyama Prefectural Government, and the Northwest Pacific Region Environmental Cooperation Center established to promote the Action Plan for the Protection, Management and Development of the Marine and Coastal Environment of the Northwest Pacific Region as a part of the Regional Seas Programme of the United Nations Environment Programme. This paper was inspired by the Marine Coastal Eutrophication project, one of the 32 funding projects aimed at tackling some of the world's greatest challenges using open Earth data by the Group on Earth Observations (GEO) and GEE. The authors thank members of the Marine Coastal Eutrophication project for insightful discussions.

Conflicts of Interest: The authors declare no conflict of interest.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 August 2022 doi:10.20944/preprints202208.0484.v1

Appendix A. Illustrations of the swath pixel resolution at the swath center and edge, description of the target projection, and the remapping of the SGLI/GCOM sample data.

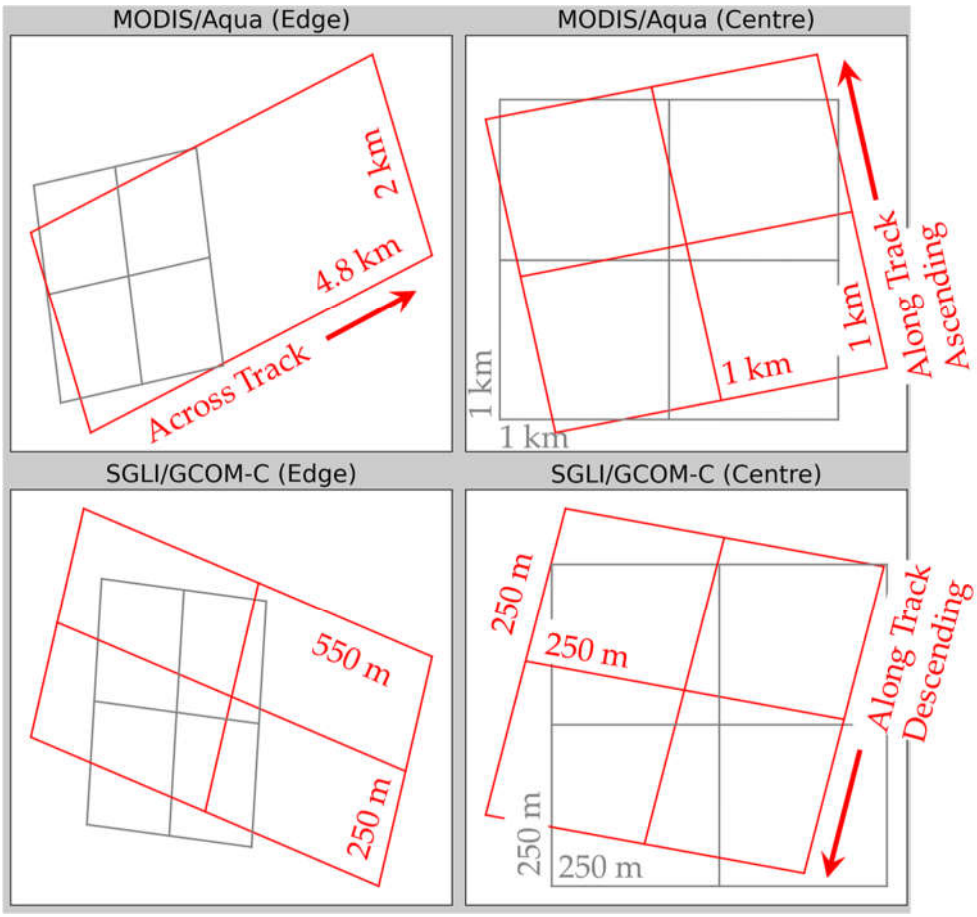

Figure A1. Illustrations of pixel resolution of swath (red) and of map-projected grid (gray) for MODIS/Aqua (top) and SGLI/GCOM-C (bottom). The right panel shows the pixel frame which is equilateral at the sub-satellite location with the resolution equal to the nominal value. The left panel is the same as the right but at the swath edge where the resolution degrades significantly.

Table A1. Example of coordinate system and image structure output by "*gdalinfo*" for the MODIS/Aqua target map projection. Note that the corner coordinates are reported in metric distances along with corresponding latitude and longitude values.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 August 2022 doi:10.20944/preprints202208.0484.v1

```
Files: A2022125035500.L2_LAC_OC.tif
Size is 2835, 2284
Coordinate System is:
PROJCRS["laea",
    BASEGEOGCRS["WGS 84",
        DATUM["World Geodetic System 1984",
            ELLIPSOID["WGS 84",6378137,298.257223563,
                LENGTHUNIT["metre",1]]],
        PRIMEM["Greenwich",0,
            ANGLEUNIT["degree",0.0174532925199433]],
        ID["EPSG",4326]],
    CONVERSION["Lambert Azimuthal Equal Area",
        METHOD["Lambert Azimuthal Equal Area",
            ID["EPSG",9820]],
        PARAMETER["Latitude of natural origin",46.1208457946777,
            ANGLEUNIT["degree",0.0174532925199433],
            ID["EPSG",8801]],
        PARAMETER["Longitude of natural origin",136.464080810547,
            ANGLEUNIT["degree",0.0174532925199433],
            ID["EPSG",8802]],
        PARAMETER["False easting",0,
            LENGTHUNIT["metre",1],
            ID["EPSG",8806]],
        PARAMETER["False northing",0,
            LENGTHUNIT["metre",1],
            ID["EPSG",8807]]],
    CS[Cartesian,2],
        AXIS["(E)",east,
            ORDER[1],
            LENGTHUNIT["metre",1]],
        AXIS["(N)",north,
            ORDER[2],
            LENGTHUNIT["metre",1]]]
Data axis to CRS axis mapping: 1,2
Origin = (-1843502.092211684444919,1233236.707095944788307)
Pixel Size = (1000.995830975200761,-1000.916434531943764)
Image Structure Metadata:
  COMPRESSION=LZW
  INTERLEAVE=PIXEL
  LAYOUT=COG
Corner Coordinates:
Upper Left  (-1843502.092, 1233236.707) (107d32'42.65"E, 54d 2'27.62"N)
Lower Left  (-1843502.092,-1052856.429) (116d13'33.87"E, 34d35'47.42"N)
Upper Right (  994321.089, 1233236.707) (152d35'52.06"E, 56d16'15.57"N)
Lower Right (  994321.089,-1052856.429) (147d29'43.43"E, 36d 2' 1.38"N)
Center      ( -424590.502,   90190.139) (130d53'48.86"E, 46d47'54.22"N)
```

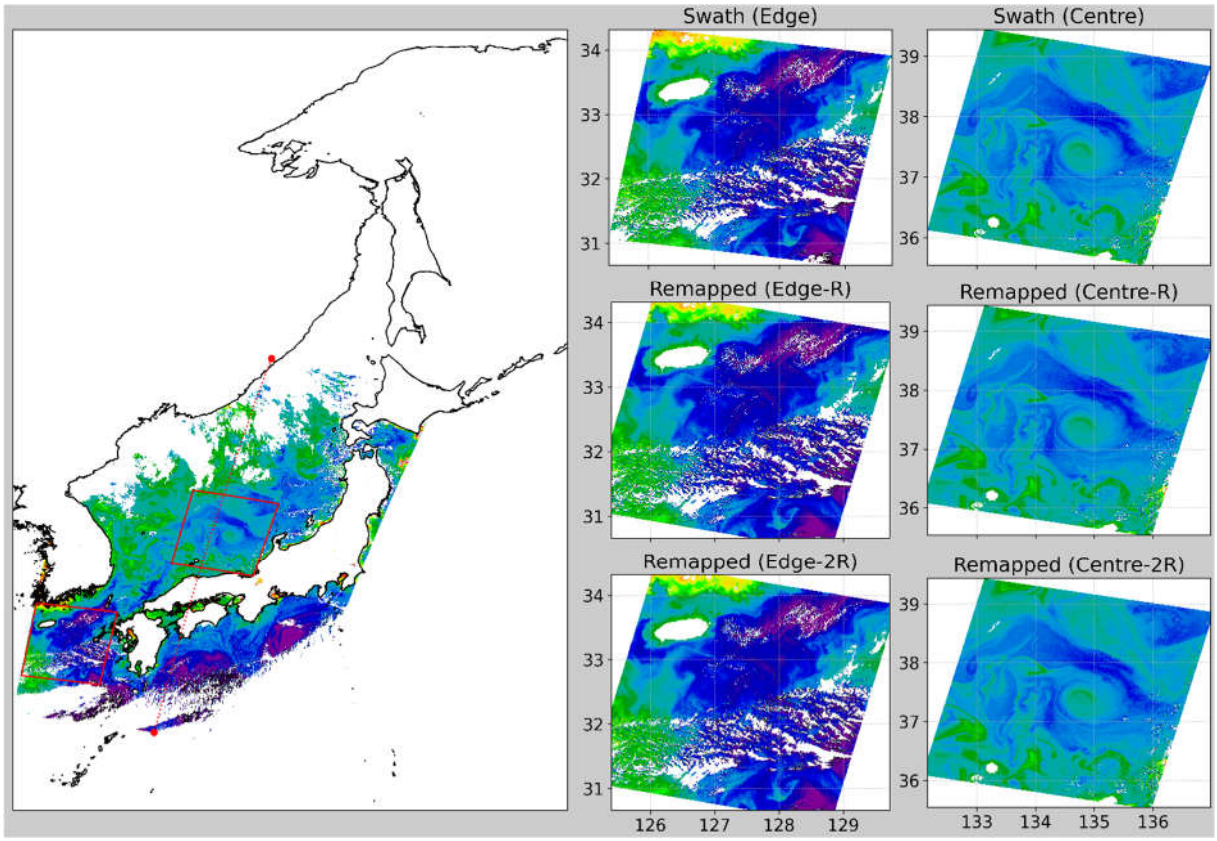

Figure A2. Same as in Figure 2 but for SGLI/GCOM-C. Sample data used is indicated in Table 1. At the center box, a patch of isolated high chlorophyll can be seen swirling clockwise. The area is known to be eddy rich with significant impacts on phytoplankton bloom timing [31,32].

References

1. Groom, S.; Sathyendranath, S.; Ban, Y.; Bernard, S.; Brewin, R.; Brotas, V.; Brockmann, C.; Chauhan, P.; Choi, J.; Chuprin, A.; et al. Satellite Ocean Colour: Current Status and Future Perspective. *Front. Mar. Sci.* **2019**, *6*, 485, doi:10.3389/fmars.2019.00485.
2. Maúre, E. de R.; Terauchi, G.; Ishizaka, J.; Clinton, N.; DeWitt, M. Globally Consistent Assessment of Coastal Eutrophication. *Nat. Commun.* **2021**, *12*, 6142, doi:10.1038/s41467-021-26391-9.
3. Feng, C.; Ishizaka, J.; Saitoh, K.; Mine, T.; Zhou, Z. Detection and Tracking of Chattonella Spp. and Skeletonema Spp. Blooms Using Geostationary Ocean Color Imager (GOCI) in Ariake Sea, Japan. *J. Geophys. Res. Ocean.* **2021**, *126*, 1–18, doi:10.1029/2020JC016924.
4. Stumpf, R.P.; Culver, M.E.; Tester, P.A.; Tomlinson, M.; Kirkpatrick, G.J.; Pederson, B.A.; Truby, E.; Ransibrahmanakul, V.; Soracco, M. Monitoring Karenia Brevis Blooms in the Gulf of Mexico Using Satellite Ocean Color Imagery and Other Data. *Harmful Algae* **2003**, *2*, 147–160, doi:10.1016/S1568-9883(02)00083-5.
5. Cannizzaro, J.P.; Carder, K.L.; Chen, F.R.; Heil, C.A.; Vargo, G.A. A Novel Technique for Detection of the Toxic Dinoflagellate, Karenia Brevis, in the Gulf of Mexico from Remotely Sensed Ocean Color Data. *Cont. Shelf Res.* **2008**, *28*, 137–158, doi:10.1016/j.csr.2004.04.007.
6. Siswanto, E.; Ishizaka, J.; Tripathy, S.C.; Miyamura, K. Detection of Harmful Algal Blooms of Karenia Mikimotoi Using MODIS Measurements: A Case Study of Seto-Inland Sea, Japan. *Remote Sens. Environ.* **2013**, *129*, 185–196, doi:10.1016/j.rse.2012.11.003.
7. IOCCG *Earth Observations in Support of Global Water Quality Monitoring*; Dartmouth, Canada, 2018;
8. Huntington, J.L.; Hegewisch, K.C.; Daudert, B.; Morton, C.G.; Abatzoglou, J.T.; McEvoy, D.J.; Erickson, T. Climate Engine: Cloud Computing and Visualization of Climate and Remote Sensing Data for Advanced Natural Resource Monitoring and Process Understanding. *Bull. Am. Meteorol. Soc.* **2017**, *98*, 2397–2409, doi:10.1175/BAMS-D-15-00324.1.
9. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone. *Remote Sens. Environ.* **2017**, *202*, 18–27, doi:10.1016/j.rse.2017.06.031.
10. Tamiminia, H.; Salehi, B.; Mahdianpari, M.; Quackenbush, L.; Adeli, S.; Brisco, B. Google Earth Engine for Geo-Big Data Applications: A Meta-Analysis and Systematic Review. *ISPRS J. Photogramm. Remote Sens.* **2020**, *164*, 152–170, doi:10.1016/j.isprsjprs.2020.04.001.
11. Zhao, Q.; Yu, L.; Li, X.; Peng, D.; Zhang, Y.; Gong, P. Progress and Trends in the Application of Google Earth and Google Earth Engine. *Remote Sens.* **2021**, *13*, 1–21, doi:10.3390/rs13183778.
12. Yang, L.; Driscol, J.; Sarigai, S.; Wu, Q.; Chen, H.; Lippitt, C.D. Google Earth Engine and Artificial Intelligence (AI): A Comprehensive Review. *Remote Sens.* **2022**, *14*, 3253, doi:10.3390/rs14143253.
13. Kumar, L.; Mutanga, O. Google Earth Engine Applications Since Inception: Usage, Trends, and Potential. *Remote Sens.* **2018**, *10*, 1509, doi:10.3390/rs10101509.
14. Mutanga, O.; Kumar, L. Google Earth Engine Applications. *Remote Sens.* **2019**, *11*, 11–14, doi:10.3390/rs11050591.
15. Amani, M.; Ghorbanian, A.; Ahmadi, S.A.; Kakooei, M.; Moghimi, A.; Mirmazloumi, S.M.; Moghaddam, S.H.A.; Mahdavi, S.; Ghahremanloo, M.; Parsian, S.; et al. Google Earth Engine Cloud Computing Platform for Remote Sensing Big Data Applications: A Comprehensive Review. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 5326–5350, doi:10.1109/JSTARS.2020.3021052.
16. Hansen, M.C.; Potapov, P. V; Moore, R.; Hancher, M.; Turubanova, S.A.; Tyukavina, A.; Thau, D.; Stehman, S. V; Goetz, S.J.; Loveland, T.R.; et al. High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science (80-.).* **2013**, *342*, 850–853, doi:10.1126/science.1244693.
17. Luang, J.; Anukul, I.; Jitraporn, B.; Joaquim, P. Seasonal and Interannual Variations of MODIS Aqua Chlorophyll - a (2003 – 2017) in the Upper Gulf of Thailand Influenced by Asian Monsoons. *J. Oceanogr.* **2021**, doi:10.1007/s10872-021-00625-2.
18. Cartwright, P.J.; Fearns, P.R.C.S.; Branson, P.; Cutler, M.V.W.; O'leary, M.; Browne, N.K.; Lowe, R.J. Identifying Metocean Drivers of Turbidity Using 18 Years of Modis Satellite Data: Implications for Marine Ecosystems under Climate Change. *Remote Sens.* **2021**, *13*, doi:10.3390/rs13183616.
19. Lomas, M.W.; Bates, N.R.; Johnson, R.J.; Steinberg, D.K.; Tanioka, T. Adaptive Carbon Export Response to Warming in the Sargasso Sea. *Nat. Commun.* **2022**, *13*, 1–10, doi:10.1038/s41467-022-28842-3.
20. Ishizaka, J.; Hirawake, T.; Toratani, M.; Frouin, R. Special Section for Second-Generation Global Imager (SGLI). *J. Oceanogr.* **2022**, 4–5, doi:10.1007/s10872-022-00651-8.
21. Choi, J.K.; Min, J.E.; Noh, J.H.; Han, T.H.; Yoon, S.; Park, Y.J.; Moon, J.E.; Ahn, J.H.; Ahn, S.M.; Park, J.H. Harmful Algal Bloom (HAB) in the East Sea Identified by the Geostationary Ocean Color Imager (GOCI). *Harmful Algae* **2014**, *39*, 295–302, doi:10.1016/J.HAL.2014.08.010.
22. Maneewongvatana, S.; Mount, D. Analysis of Approximate Nearest Neighbor Searching with Clustered Point Sets. **2002**, 105–123, doi:10.1090/dimacs/059/06.
23. Campbell, J.W.; Blaisdell, J.M.; Darzi, M. Level-3 SeaWiFS Data Products: Spatial and Temporal Binning Algorithms. *NASA Tech. Memo. - SeaWiFS Tech. Rep. Ser.* **1995**, *32*.
24. Dorji, P.; Fearns, P. Impact of the Spatial Resolution of Satellite Remote Sensing Sensors in the Quantification of Total

Suspended Sediment Concentration: A Case Study in Turbid Waters of Northern Western Australia. *PLoS One* **2017**, *12*, doi:10.1371/journal.pone.0175042.

25. Scott, J.P.; Werdell, P.J. Comparing Level-2 and Level-3 Satellite Ocean Color Retrieval Validation Methodologies. *Opt. Express* **2019**, *27*, 30140, doi:10.1364/oe.27.030140.
26. Dong, J.; Xiao, X.; Menarguez, M.A.; Zhang, G.; Qin, Y.; Thau, D.; Biradar, C.; Moore, B. Mapping Paddy Rice Planting Area in Northeastern Asia with Landsat 8 Images, Phenology-Based Algorithm and Google Earth Engine. *Remote Sens. Environ.* **2016**, *185*, 142–154, doi:10.1016/j.rse.2016.02.016.
27. Teluguntla, P.; Thenkabail, P.; Oliphant, A.; Xiong, J.; Gumma, M.K.; Congalton, R.G.; Yadav, K.; Huete, A. A 30-m Landsat-Derived Cropland Extent Product of Australia and China Using Random Forest Machine Learning Algorithm on Google Earth Engine Cloud Computing Platform. *ISPRS J. Photogramm. Remote Sens.* **2018**, *144*, 325–340, doi:10.1016/j.isprsjprs.2018.07.017.
28. Lewkowicz, A.G.; Way, R.G. Extremes of Summer Climate Trigger Thousands of Thermokarst Landslides in a High Arctic Environment. *Nat. Commun.* **2019**, *10*, 1329, doi:10.1038/s41467-019-09314-7.
29. Jahromi, M.N.; Jahromi, M.N.; Zolghadr-Asli, B.; Pourghasemi, H.R.; Alavipanah, S.K. Google Earth Engine and Its Application in Forest Sciences. In *Environmental Science and Engineering*; 2021; pp. 629–649.
30. Singha, M.; Dong, J.; Sarmah, S.; You, N.; Zhou, Y.; Zhang, G.; Doughty, R.; Xiao, X. Identifying Floods and Flood-Affected Paddy Rice Fields in Bangladesh Based on Sentinel-1 Imagery and Google Earth Engine. *ISPRS J. Photogramm. Remote Sens.* **2020**, *166*, 278–293, doi:10.1016/j.isprsjprs.2020.06.011.
31. Maúre, E.R.; Ishizaka, J.; Aiki, H.; Mino, Y.; Yoshie, N.; Goes, J.I.; Gomes, H.R.; Tomita, H. One-Dimensional Turbulence-Ecosystem Model Reveals the Triggers of the Spring Bloom in Mesoscale Eddies. *J. Geophys. Res. Ocean.* **2018**, *123*, 6841–6860, doi:10.1029/2018JC014089.
32. Maúre, E.R.; Ishizaka, J.; Sukigara, C.; Mino, Y.; Aiki, H.; Matsuno, T.; Tomita, H.; Goes, J.I.; Gomes, H.R. Mesoscale Eddies Control the Timing of Spring Phytoplankton Blooms: A Case Study in the Japan Sea. *Geophys. Res. Lett.* **2017**, *44*, 11,115-11,124, doi:10.1002/2017GL074359.