Complexities in Subsetting Level 2 Data

Poster: IN41B-0038

NASA/Goddard EARTH SCIENCES DATA and INFORMATION SERVICES CENTER (GES DISC)

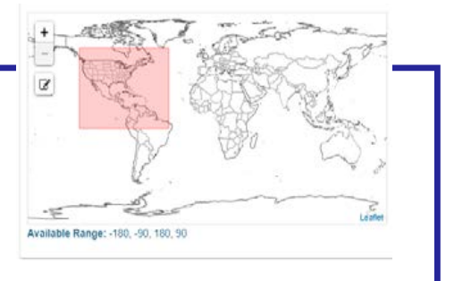
Paul Huwe[1,2], Jennifer Wei[1,2], David Meyer[1], David S. Silberstein[1,2], Jerome Alfred[1,2], Andrey K. Savtchenko[1,2], James E. Johnson[1,2], Arif Albayrak[1,2], Thomas Hearty[1,3]

[1]NASA Goddard Space Flight Center, [2]ADNET Systems Inc., [3]SGT INC.
Contact: paul.huwe@nasa.gov

Abstract

Satellite Level 2 data presents unique challenges for tools and services. From nonlinear spatial geometry to inhomogeneous file data structure to inconsistent temporal variables to complex data variable dimensionality to multiple file formats, there are many difficulties in creating general tools for Level 2 data support. At NASA Goddard Earth Sciences Data and Information Services Center (GES DISC), we are implementing a general Level 2 Subsetting service for Level 2 data to a user-specified spatio-temporal region of interest (ROI). In this presentation, we will unravel some of the challenges faced in creating this service and the strategies we used to surmount them.

Spatial Selection



Latitude/Longitude ROI

Challenges

- Level 2 (L2) data files are arranged according to the detector's physical geometry (through dimensions), with geolocation information represented as variables.
- Spatial subsets apply directly to geolocation variables and must be transferred to subset data variables.

Strategies

- Because Latitude and Longitude in L2 data are not grid projected, the masks cannot simply be ranges of the dimensions involved. Geo-coordinates must be complete maps over those dimensions with pixels outside of the subset area masked.
- While this technique is more complicated to implement, it has an advantage over simple dimension slicing in that it can readily support non-box spatial selection (circle, point, shapefile, etc.).
- We create dimensional masks of the subset geolocation variables to transfer the subset to other variables.

Specialized ROIs

Challenges

- Users select circular subsets as a center point with a surface radius:
  - The units of that radius are typically not the same at the geolocation variables (e.g., km vs. degrees)
  - Difficulties arise in conversion between different units
- The primary difficulty in point selection is in selecting the proper subset for products that lack cell corner geolocation information thus making cell coverage ambiguous.
- Proper care must be given when dealing with subsets that cross the dateline or polar regions.

Strategies

- For on-the-fly tools, great circle distance transformations work well for conversion of different units – they are computationally simple and cheap, while being easy to understand conceptually.
- For products that lack cell-corner information, we calculate the distance to nearby pixels and return the closest as the subset cell.
- Similar issues arise when a user selects a bounding box subset that is too small to contain pixel centers. Circular style distance calculations are needed to return the appropriate subset.
- Dateline corrections are usually simple longitude transformations, however polar corrections require more complex math and mask manipulation for proper subsetting.

Temporal Selection

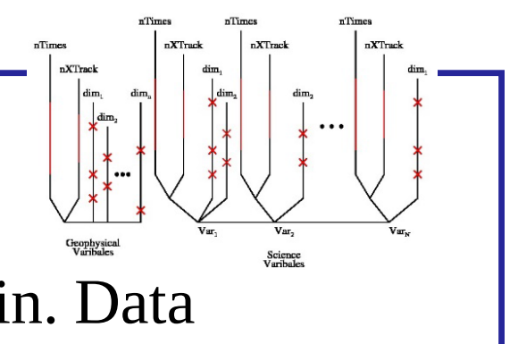
Challenges

- Because temporal variables are dimensionally arranged in a similar fashion to geolocation variables (often degenerate with one dimension), temporal subsetting presents similar challenges as spatial subsetting.
- Temporal subsetting's unique issue is that there is no standard definition of time specification (TAI93, UNIX, GPS, UTC, and various custom calendar formats are used).

Strategies

- We convert all time formats to TAI93 for internal mask creation and convert them back for output. This ensures temporal subsetting consistency and accuracy across time formats.

Dimensional Issues



Challenges

- Masks created from spatial and temporal subsetting only span the dimensions that those variables contain. Data variables on the same dimension space are subset with these masks. However, other dimensional spaces can create difficult subsetting situations.
- Several particular arrangements must be addressed:
  - variables missing mask dimensions
  - variables spanning additional dimensions
  - variables both missing and having extra dimensions
- An additional complication would be variable dimensional order not matching the order of dimensions in masks.

Strategies

- Reshaping masks for variables missing mask dimensions requires properly collapsing the mask along dimensions still in the variable to be masked. Missing dimensions must be binary summed to ensure that no data is incorrectly masked from the subset.
- Variables that span additional dimensions necessitate that masks be expanded along each new dimension. While expansion is straightforward along complete dimensions, great care must be used when expanding along externally subset dimensions (e.g. user-selected layers).
- Variables that are both missing mask dimensions and have extra dimensions need careful application of the previous two points sequentially – handling the missing dimensions first, then expanding to additional dimensions.
- For a mismatch between variable dimensional order and the order of dimensions in masks, masks must be reshaped to the variable's dimensional order before application, typically one dimension at a time.
- In addition to these general mask shaping complexities, we can repackage variables into data streams. In this stream, all the geolocation dimensions are collapsed into a single vector, with all the masked values completely removed. While being the most efficient data presentation, this transformation adds another layer of complexity to data variables of interesting structure.

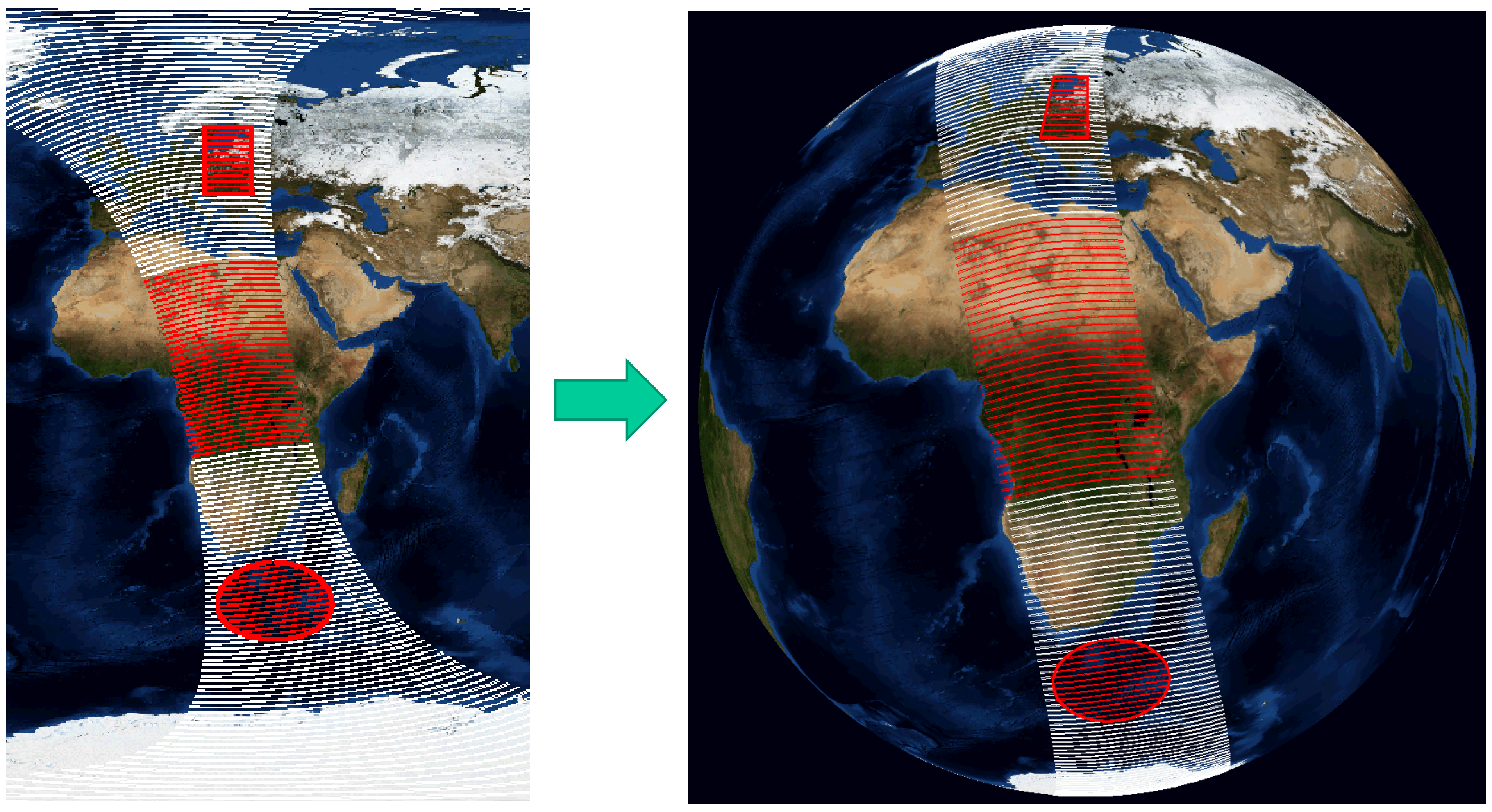
File Format Issues

Challenges

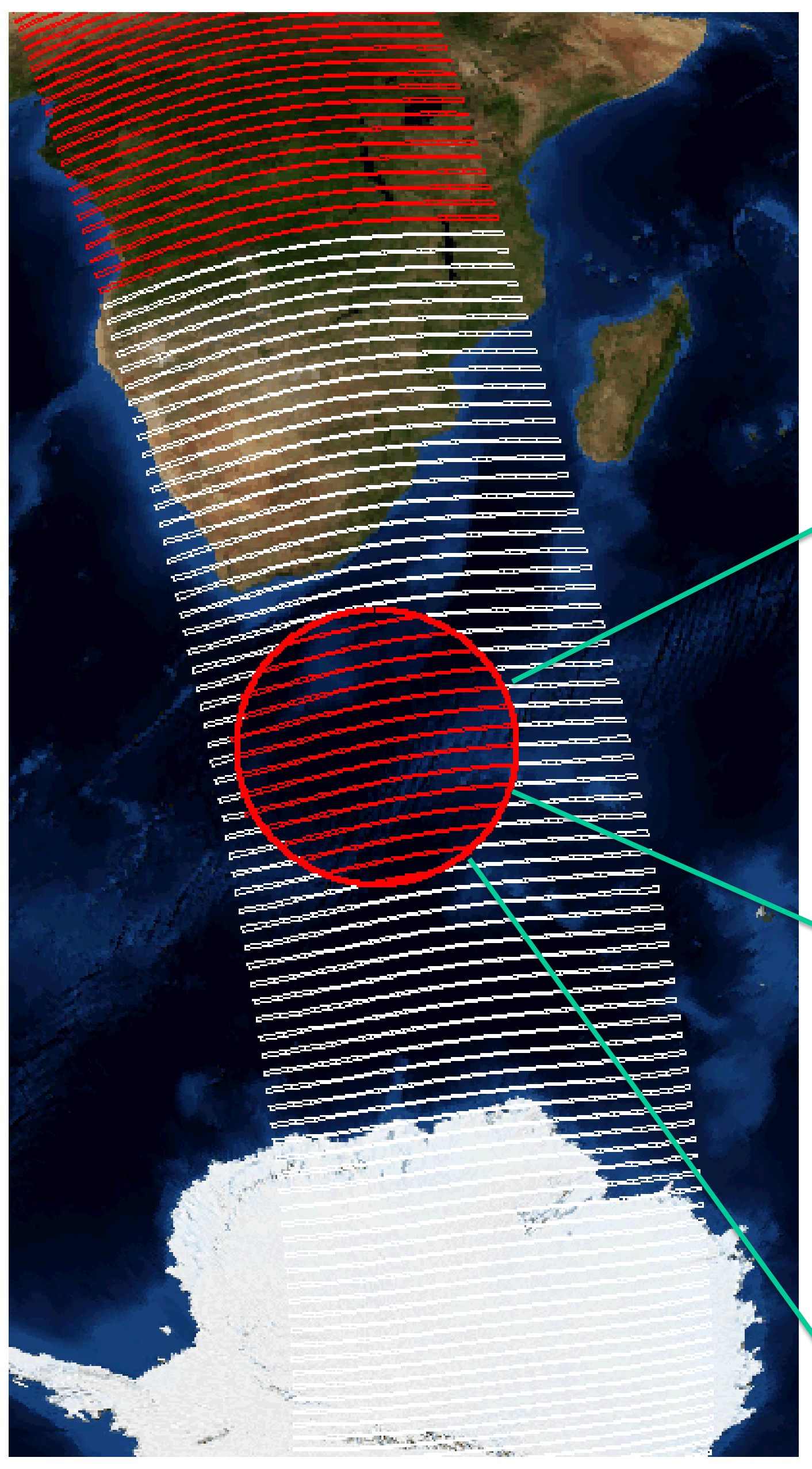
- While Level 2 data are available in many file formats – HDF5, NetCDF, HDF4, various binary types, text, and others – self-describing datatypes like HDF5, HDF-EOS5, HDF4, HDF-EOS2, and NetCDF are best suited for general subsetting.
- While appropriate for general subsetting, each of these formats has its own library support and specific API. A general subsetter must not only carefully code for these divergent APIs, but it must also carefully keep track of the requirements for and idiosyncrasies of each. This is further complicated when a user desires conversion between file formats.

Strategies

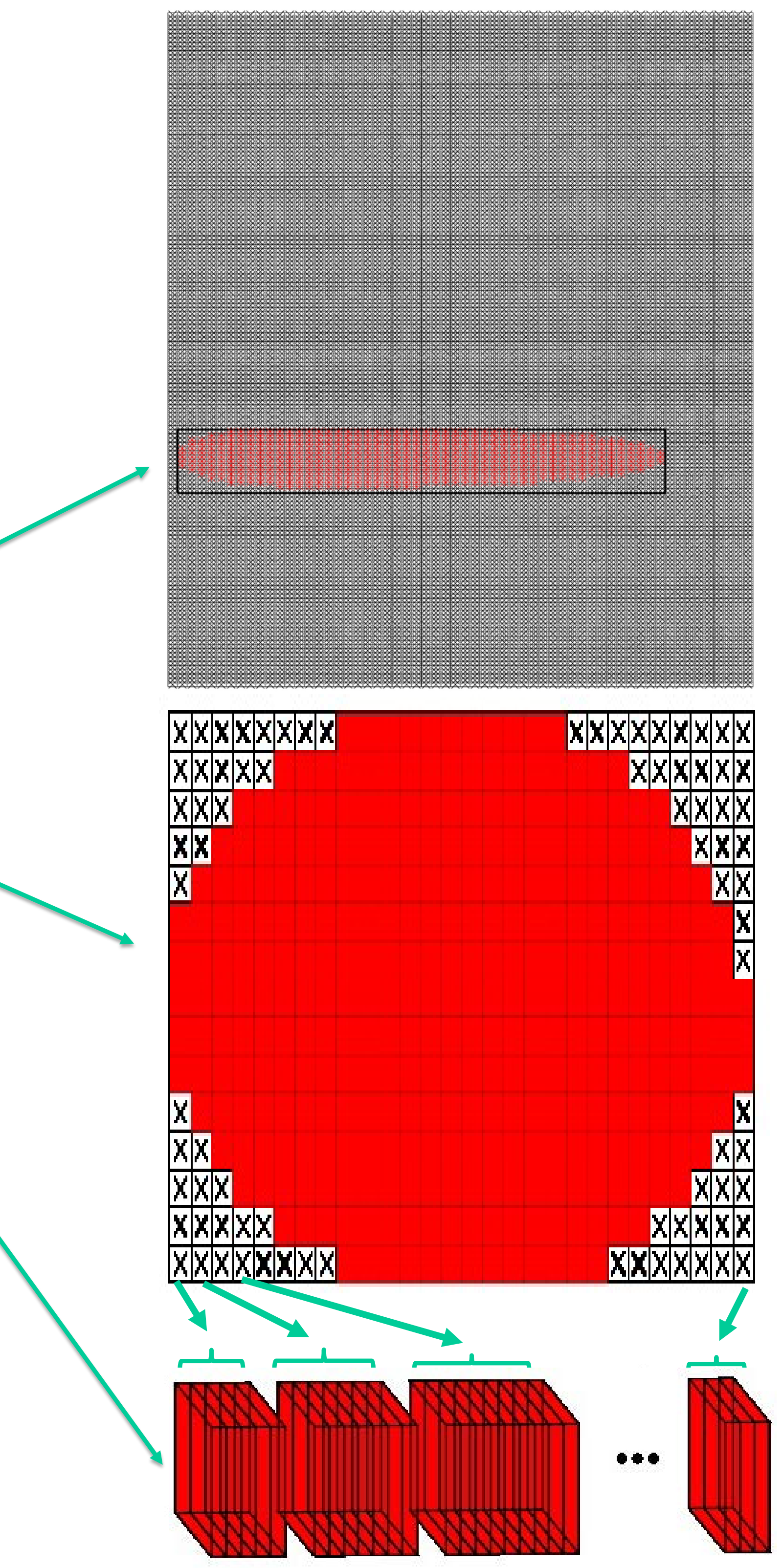
- These formats have varying levels of dimensional identification – from NetCDF where each dimension has a dimension scale and explicit usage, to HDF5 and its generic dummy dimension usage. The varying dimensional implementations can contribute greatly to the difficulty in expanding and reshaping masks for subsetting data variables.



Level 2 Subset Regions. Level 2 swath data with subsetted regions shown in red. The three regions are geo-located box, temporal cut, and geo-located circle. The left panel shows the distortion on a Mercator plot and the right panel shows the global view as seen from space.



Circular Region. Inset of circular search region shown above. The view is of the Earth rotated to center the region.



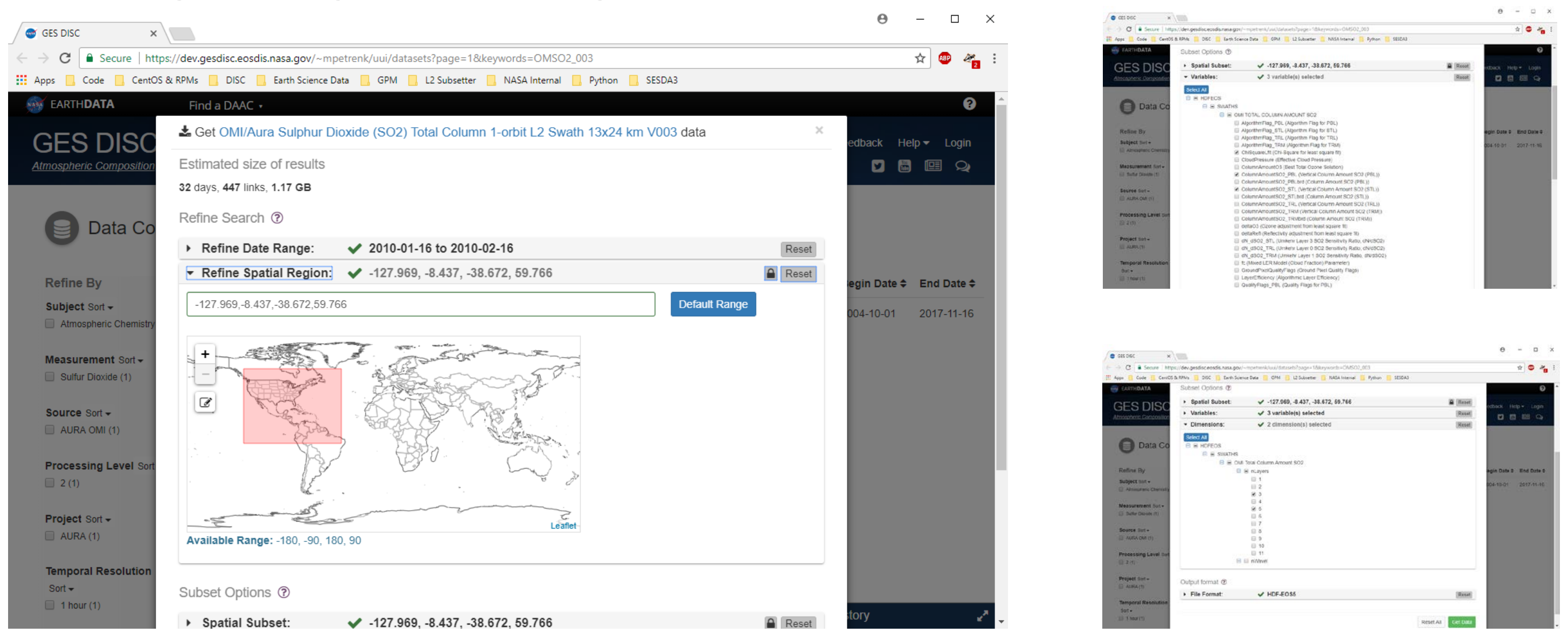
Uncropped data-view of the circular subsetted region.

Cropped data-view of the circular subsetted region.

Vector / data stream re-formatted data-view of the circular subsetted region.

L2Subsetter

At the NASA Goddard Earth Science Data and Information Services Center (GES DISC), we are developing a general purpose Level 2 Subsetting service. This service features variable, spatial, temporal, and dimensional subsetting along with data type and presentation conversion. It is presently available for all OMI Level 2 products, and can be found at: **https://disc.sci.gsfc.nasa.gov/datasets?page=1&keywords=omi**



https://dev.gesdisc.eosdis.nasa.gov/~mpetrenk/uui/datasets?page=1&keywords=OMSO2_003

Get OMI/Aura Sulphur Dioxide (SO2) Total Column 1-orbit L2 Swath 13x24 km V003 data

Estimated size of results
32 days, 447 links, 1.17 GB

Refine Search
Refine Date Range: ✓ 2010-01-16 to 2010-02-16
Refine Spatial Region: ✓ -127.969, -8.437, -38.672, 59.766
-127.969,-8.437,-38.672,59.766
Available Range: -180, -90, 180, 90

Subset Options
Spatial Subset: ✓ -127.969, -8.437, -38.672, 59.766
Variables: ✓ 3 variable(s) selected
Dimensions: ✓ 2 dimension(s) selected
File Format: ✓ HDF-EOS5

GES DISC Level 2 Subsetter User Interface. Screen capture images of the user interface of the Level 2 Subsetter. The images illustrate spatial and temporal selection, variable selection, and dimensional selection capabilities of the Level 2 Subsetter.

Acknowledgements

The authors thank the GES DISC Level-2 Scientists and Engineers.

Hourglass icon from Wikimedia Commons / CC BY-SA 3.0.