Data Integrity Challenges in NASA Giovanni

2024 AGU Fall Meeting

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

Zhong Liu[1,2] (Zhong.Liu@nasa.gov), James Acker[1,3], and Binita KC[1,3] [1]NASA GES DISC; [2]CSISS, George Mason University; [3]ADNET Systems, Inc.

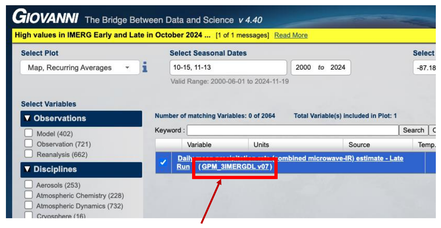
ABSTRACT

The Geospatial Interactive Online Visualization ANd aNalysis Infrastructure (Giovanni) is an online tool developed by the NASA Goddard Earth Sciences (GES) Data and Information Services Center (DISC), one of 12 NASA Science Mission Directorate Data Centers (DAACs) to analyze and visualize NASA remote sensing and model data without downloading data and software.

As of this writing, over 2000 Earth satellite and model variables are available in Giovanni, including several well-known NASA satellite missions (e.g., TRMM, GPM) and projects (e.g., MERRA-2, GPCP). There are twenty-two plots provided by Giovanni that can be used to analyze, compare, and explore Earth data across disciplines. Results can be shared with colleagues and downloaded for further analysis. Giovanni has helped publish over 3000 referral papers over the years.

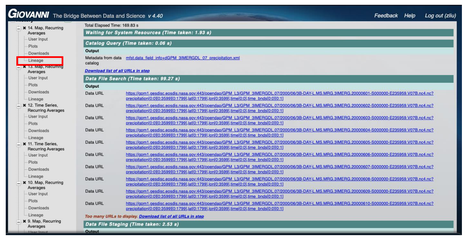
As open science policies roll in, data integrity has become a major challenge for Giovanni and other tools. For integrity, both data and workflows must be transparent. FAIR-compliant data, including input, intermediate, and result products, as well as their associated statistics, metadata, and information, are needed. The NASA Data Product Development Guide for Data Producers provides a key resource on how to develop FAIR-compliant data products. Data quality information is also needed from data producers and analysis services like Giovanni. The workflow part is quite challenging and requires workflow management improvements, such as recording workflows and making them available to users. In this presentation, we will discuss the data integrity challenges in Giovanni.

Supporting Data Integrity and Open Science

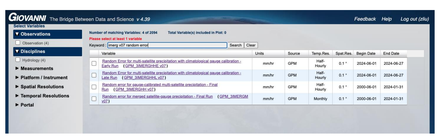


Each variable in Giovanni is linked to their dataset landing page (on the right) consisting of detailed information about the dataset.

The dataset landing page for the IMERG V07 Late Daily dataset is shown. Each dataset landing page at GES DISC consists of detailed information about the dataset (e.g., data citation, documents) and data access.



"Lineage" provides the links to files such as input, output, and intermediate. All including user input, plots, downloads, and lineage can be sharable with a URL.



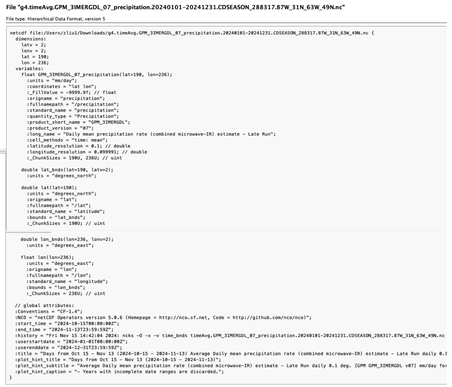
Data Quality

Data quality is another important aspect of data integrity.

- Data quality variables (e.g., relative errors in IMERG) at pixel or grid level
- Quality information in product metadata (e.g., uncertainty assessment, ancillary data product information)
- Missing variables and information in the latest DPDG (Data Product Development Guide for Data Producers)
- Workflows (e.g., algorithms for regridding)
- How about service providers who provide analysis-ready data (ARD)?

Conclusion:

- Giovanni supports data integrity and open science.
- Data quality also depends on both data and service providers as well as their collaboration.



A sample of the Giovanni output data file in NetCDF. The CF compliant format is friendly for other application software packages (e.g., Panoply). In the global attributes, it provides the information about the file history. In this case, NCO was used. However, the file is not fully compliant with NASA's Data Product Development Guide for Data Producers.

Challenges in Delivering Data Quality

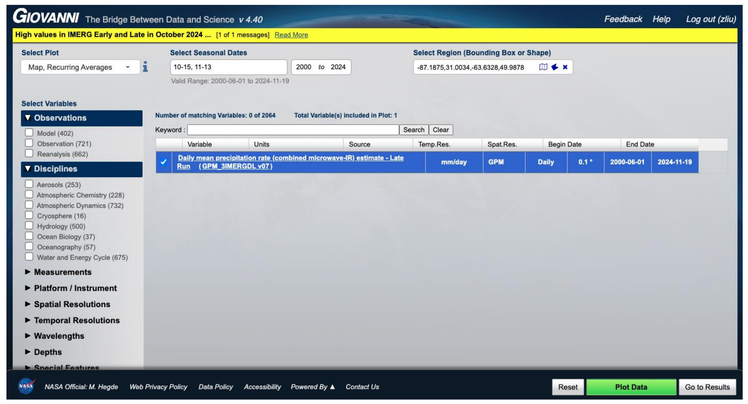
- External dependency: data providers need to provide data quality information (metadata and variables (e.g., validation, error, uncertainty, standards, FAIR-compliant))
- Internal dependency: observational data for evaluation, plot types, and workflows (e.g., ensemble of similar products, product comparison, and statistics for data processing in all plot types, including missing data, min/max values, standard variation, etc.)
- Policy (e.g., Open science and FAIR): workflow transparency (e.g., processing and visualization software code and reproducibility)

Giovanni landing page: https://giovanni.gsfc.nasa.gov

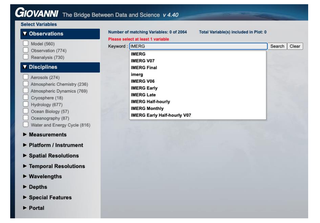
Login Earthdata: https://urs.earthdata.nasa.gov/

What is NASA Giovanni?

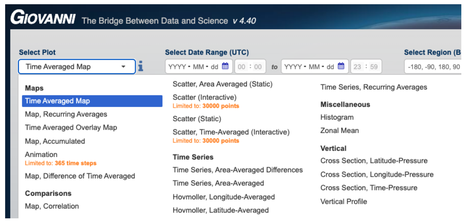
The Geospatial Interactive Online Visualization ANd aNalysis Infrastructure (Giovanni) provides the means to perform 22 visualizations on ~2000 variables from six different DAACs without having to download data and software. With Giovanni's assistance, researchers around the world have published more than 3,000 peer-reviewed papers in a wide range of Earth science disciplines and other areas.



The Giovanni landing page. A parameter/variable-based search interface with suggestions (see below) greatly facilitates data discovery. Information (e.g., source, temporal and spatial resolution, beginning and end dates) about each variable is listed. Twenty-two plot types are available for data analysis and visualization. Different shapes are available for different U.S. states, countries, land only, sea only, watersheds and world regions. Input and output data can be downloaded for further analysis using other tools (e.g., Microsoft Excel, Panoply). In short, Giovanni simplifies access to Level-3 and Level-4 variables for several popular NASA missions or projects (e.g., GPM, MERRA-2).

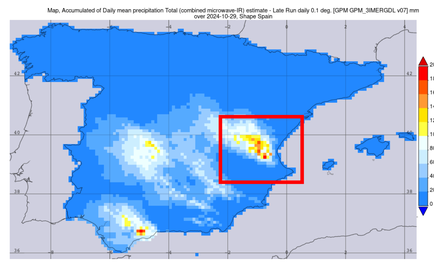


Giovanni provides several ways to search for variables. On the left, variables are categorized based on observations, disciplines, measurements, etc. If you know the very popular NASA precipitation product suite, the Integrated Multi-satellitE Retrievals for GPM (IMERG), simply type in "imerg" (above) and a list of related information will show up.



There are **twenty-two plots** in Giovanni. Time-averaged Map and Area-averaged Time Series are the two most frequently used plots. Several plots are available to facilitate comparisons of two different precipitation variables:

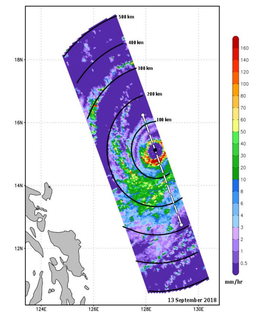
- Map, Difference of Time Averaged
- Map, Correlation
- Scatter, Area Averaged (Static)
- Scatter (Interactive)
- Scatter, Time-Averaged (Interactive)
- Time Series, Area-Averaged Difference



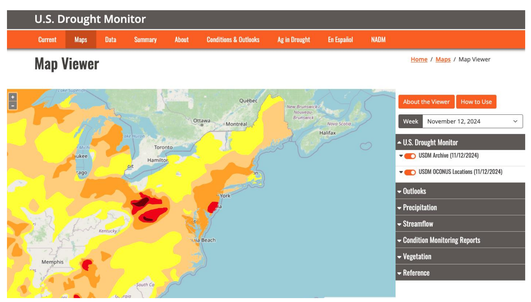
Accumulated rainfall (IMERG V07 Late Daily) map on October 29, showing the 2024 Spanish floods.

Limitations: What Giovanni cannot do?

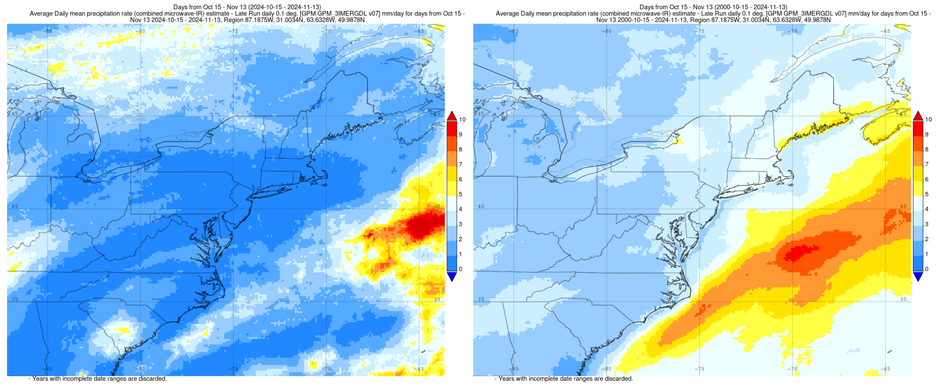
- Level-2 orbital data are not included
- Station data, airborne data, other instruments (e.g., radar)
- User can't upload their data
- Limited data analysis and visualizations
- Performance can be an issue with high-resolution data
- No preferences



Monitoring Northeast (USA) Drought Conditions



Credit: U.S. Drought Monitor



Left: There is an ongoing drought condition in Northeast USA. Middle: Using Giovanni and IMERG V07 Late Daily, the average rainfall map for the region during the past 30-days can be generated. Right: The average conditions (2000 – 2024) for the same period can also be generated. The data of both maps can be downloaded in NetCDF for further analysis (e.g., subtraction). We can see that there is below-average rainfall not only in Northeast but also in the Gulf Stream.

Giovanni: https://giovanni.gsfc.nasa.gov/ **GES DISC:** https://disc.gsfc.nasa.gov ***Suggestions or subscription to our mailing list:*** **gsfc-dl-help-disc@mail.nasa.gov**