VOL. 104 | NO. 8
AUGUST 2023

Eos

SCIENCE NEWS BY AGU

Flash Droughts Are Getting Flashier

Saturn's Shiny Rings May Be Pretty Young

Mexico City Is Sinking

The Career Issue

How scientists navigate the rivers and roads to success.

AGU
ADVANCING EARTH AND SPACE SCIENCES

We Need a Better Way to Share Earth Observations

iStock.com/Nobi_Prizue

Florida's costliest storm ever, massive flooding in Pakistan and South Korea, deadly heat waves across Europe—recent headlines attest to natural hazards that continue to catch us off guard.

Scientists and forecasters often see these events coming but not early enough or in enough detail to provide clear, accurate warnings. To better understand, monitor, and forecast natural hazards, their potential effects on people, and how they will change in a warming climate, scientists need environmental observations from many sources. These data must be not only collected but also available, accessible, timely, and trustworthy.

Maps, graphs, models, and other data products created from satellite observations play critical roles in forecasting because of the wide, often global-scale coverage they provide [*National Academies of Sciences, Engineering, and Medicine*, 2018]. In addition to helping us study natural hazards, satellite data products support other activities in Earth science, including a wide range of basic research; artificial intelligence and machine learning applications; education and outreach endeavors; and decisionmaking by community and government leaders, resource and hazard managers, and others.

Though powerful, these products aren't perfect, and they are always being verified and improved using environmental data collected worldwide from the ground, air, and sea. To advance satellite data products and their benefits for Earth science and society, the use of observations collected by the global scientific community must be maximized. The European Union's planned digital twin of Earth, for example, aims to integrate all available global observations for model development and applications. This type of integration can transcend institutional barriers and be applied to other areas of Earth science as well.

However, despite many international efforts aimed at maximizing the use of satellite observations—by the World Meteorological Organization, the Committee on Earth Observation Satellites (CEOS), and the Open Geospatial Consortium (OGC), for example, significant obstacles to integrating and sharing data from disparate global sources remain [*Hills et al.*, 2022]. An innovative data infrastructure for gathering and sharing data that meets the criteria outlined below could help overcome these obstacles.

The Interplay of Satellite and in Situ Data

Since the dawn of the satellite era in the 1960s, scientists have relied on in situ observations gathered by organizations around the world to develop and improve satellite data products for research and operational use. Observations from weather stations and radar networks, for example, help validate the

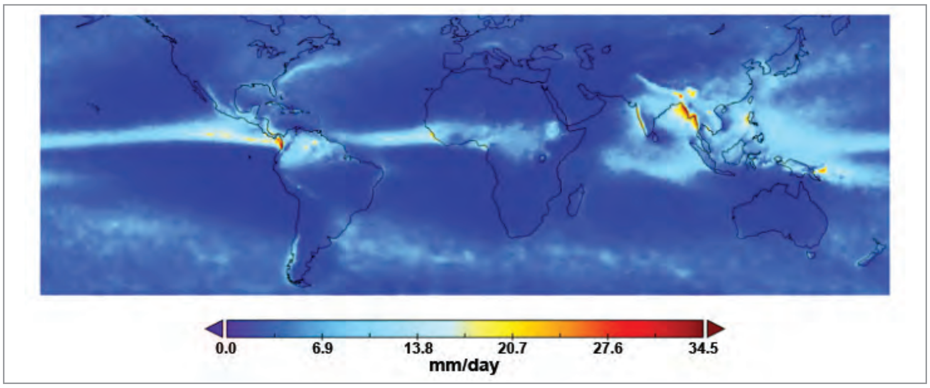

Fig. 1. This map displays average boreal summer precipitation from 2000 to 2021 from the Integrated Multi-satellite Retrievals for Global Precipitation Measurement monthly product with rain gauge calibration. The light-colored band circling the equator indicates the Intertropical Convergence Zone, and the effects of the Indian monsoon are visible around the Indian subcontinent and Southeast Asia.

accuracy of satellite measurements of temperature, precipitation, and soil moisture. However, collecting and providing in situ observations on a global scale are difficult and often costly, especially when it comes to observing vast remote regions on land and at sea.

Satellite-based products, in turn, play an important role in filling gaps where in situ data are sparse or not available and in improving understanding of Earth system processes across the whole planet [*National Academies of Sciences, Engineering, and Medicine*, 2018]. Even with the combined capabilities of satellite and in situ data, though, many data gaps still exist.

Scientists often use observations from multiple satellites as inputs in their product development in conjunction with in situ observations [*Kidd et al.*, 2021]. For example, NASA's Integrated Multi-satellite Retrievals for Global Precipitation Measurement (IMERG) product suite relies on observations from dozens of domestic and international satellites (Figure 1) [*Huffman et al.*, 2019]. These satellites—including the Tropical Rainfall Measuring Mission and the Global Precipitation Measurement mission, which provide core calibration and evaluation data for IMERG [*Huffman et al.*, 2019]—supply observations from several types of onboard sensors (e.g., infrared, passive microwave, and radar) to support global precipitation estimates.

IMERG products also use data from rain gauges on the ground to correct for biases in the satellite data, which can overestimate or underestimate precipitation. These rain gauge data come from the Global Precipitation Climatology Centre (GPCC), which reports precipitation measurements from more than 6,000 gauge stations around the world.

Despite efforts like those of GPCC to collect in situ data, local and regional in situ observations that could extend the use of products like IMERG are not collected in many areas or have not been integrated and made publicly available by other organizations. Attendees at a recent International Precipitation Working Group meeting noted that this lack of data integration and sharing presents a major obstacle to improving satellite-based precipitation products.

Despite many international efforts aimed at maximizing the use of satellite observations, significant obstacles to integrating and sharing data from disparate global sources remain.

Barriers to Data Usability

To address challenges of data sharing, various public and private organizations have previously established Earth science data repositories to provide access to data online. For example, the NASA Earth Observing System Data and Information System (EOSDIS) provides data from NASA satellites (e.g., through the IMERG suite), models, and field campaigns free of charge to the global user community.

Similar data repositories and efforts by other U.S. and international government agencies and organizations exist, such as NOAA's Open Data Dissemination program. And a number of catalog services, such as data.gov and the CEOS database, have been established to provide search capabilities that facilitate data discovery. Also, data availability from nontraditional sources, including commercial sectors and community science activities, has increased rapidly in recent years.

Although these sources have increased data availability, the data in each are collected and curated by the different organizations largely for their own missions or projects, and each repository is unique. Under EOSDIS alone are 12 disciplinary data centers with different portals and designs.

Conducting interdisciplinary work can be challenging because researchers often need multiple data products and services from different data centers. EOSDIS is planning to migrate all its data products to the cloud to simplify the use of its data and facilitate more interdisciplinary activity (e.g., Earthdata Search). Yet in general, existing practices for data collection, sharing, and integration do not transcend organizational barriers, and users are faced with diverse requirements for finding, accessing, and using data and services. Efficient means of data discovery, access, integration, interoperability, reusability, and user-centered services—capabilities laid out in the FAIR (findable, accessible, interoperable, and reusable) data guiding principles [*Wilkinson et al.*, 2016]—have thus not been achieved on a wide scale.

Data Infrastructure That Makes a Difference

Game-changing reforms in data infrastructure are needed to lower barriers and accelerate improvements of data products for Earth science research and applications.

What would such reforms look like?

In short, a successful new data infrastructure would engage the global community to share and use quality-controlled, FAIR-compliant environmental data and services ethically, equitably, and sustainably. It would implement open science practices, which open doors to improve data and information accessibility, efficiency, and quality as well as scientific reproducibility. It would also promote data services supported by open-source software and incentivize data and software sharing by establishing a new mechanism for crediting data providers.

Publicly accessible information-sharing platforms already exist in other areas of society. On YouTube, for example, users can upload videos in any of more than a dozen file formats to share with others around the world without worrying about technical challenges such as data storage and interoperability. Users are responsible for providing services for the content they add, including the descriptive text that appears below each video, responses to comments from viewers, and question-and-answer sections. Such platforms can serve as examples for Earth science data sharing, but there are several challenges.

Open Data You Can Trust

One such challenge involves data integrity. The infrastructure of a new data-sharing platform would provide the convenience of allowing everyone to upload and share their data, but that could open it up to potential misuses, including submissions of incomplete or fake data. Ensuring the veracity and completeness of data would be critical to successfully implementing a new data infrastructure. Certifications for trusted repositories, such as that provided by the International Science Council's World Data System, would help in this effort, as would a user identity vetting process and a user system for reporting abuse.

Ensuring data ethics (e.g., ethical collection, ownership, storage, distribution, and use of data) is another issue for a new infrastructure to address [e.g., *Carroll et al.*, 2021]. Procedures would be needed to prevent users from uploading data without the owner's permission, for example, or in violation of laws or codes of conduct. Ultimately, data submitters would be responsible for their own actions, but a built-in, self-detecting mechanism in the infrastructure could also help minimize violations.

A user-driven data-sharing infrastructure is an ideal place to implement open science principles. Several organizations have developed open science policies, elaborating on how to make data transparent, accessible, and inclusive. Others, such as OGC and the International Organization for Standardization, have issued standards, recommendations, and best practices for Earth science data. Implementing such policies and standards could be challenging because imposing cultural changes (e.g., standard requirements for metadata) on the scientific community is difficult. A new infrastructure should leverage these existing resources without reinventing the wheel.

Heterogeneous data present still another challenge. Earth scientists usually produce data in formats and with structures, units, and vocabularies specific to their domains or specializations. In an environment where all these formats coexist, integrating data and making them interoperable for interdisciplinary activities are difficult. In a new infrastructure, information and tools (e.g., the Integrated Ocean Observing System Compliance Checker) must be available to guide data providers in preparing their data, including metadata, so that they meet community standards before they are submitted to the system.

In addition to addressing the above challenges, it is critical that a new infrastructure meet the following three criteria:

First, it needs an open-source approach to software development to best leverage resources from the entire global community (rather than from only a subset with access to costly or proprietary software) and to avoid repeated development and achieve the goals of open science. Guidelines for software development must be established in accordance with the FAIR principles and open science standards.

Second, it needs to provide a rich collection of data services, which would be a major motivation and incentive for users to submit and share their data. For example, new ground-based radar data products can be generated by merging data submitted by users around the world and used to improve estimates of precipitation. Meanwhile, users can use tools like NASA's Giovanni to explore, visualize, and analyze data without downloading it and accompanying software. Another example is to allow transformation into analysis-ready, cloud-optimized data for analysis in the cloud [*Stern et al.*, 2022].

Third, a new infrastructure needs a mechanism by which credit can be attributed clearly and equitably (e.g., to meet requirements of ethical data practices) to all those involved in generating and providing data, which should further incentivize organizations and individuals to make contributions. With the implementation of open science practices, all work, data, and software should identify credits, and their provenance must be automatically traceable.

Engaging the Global Community

The vast amount of data, scaled-up services, and computing capabilities of the proposed data infrastructure will require a cloud-based platform to host it all, likely making it an expensive endeavor. A big question must be resolved before the global community will see the benefits: Who will cover the costs? We envision the scientific community working together with a consortium of public organizations and private enterprises as the best option for developing and sustaining the infrastructure.

If it is created, we believe the new data infrastructure will engage much more of the global community than is currently represented in existing Earth science data repositories. The increased availability and accessibility of integrated and open data from governments, research institutions, the private sector, and other sources could then accelerate development of satellite and other data products to help address natural hazards and other pressing global challenges.

References

Carroll, S. R., et al. (2021), Operationalizing the CARE and FAIR principles for Indigenous data futures, *Sci. Data, 8,* 108, https://doi.org/10.1038/s41597-021-00892-0.

Hills, D., et al. (2022), Earth and Space Science Informatics perspectives on Integrated, Coordinated, Open, Networked (ICON) science, *Earth Space Sci., 9*, e2021EA002108, https://doi.org/10.1029/2021EA002108.

Huffman, G. J., et al. (2019), GPM IMERG Final Precipitation L3 1 month 0.1 degree x 0.1 degree V06, Goddard Earth Sci. Data and Inf. Serv. Cent., Greenbelt, Md., https://doi.org/10.5067/GPM/IMERG/3B-MONTH/06.

Kidd, C., et al. (2021), The global satellite precipitation constellation: Current status and future requirements, *Bull. Am. Meteorol. Soc., 102*(10), E1844–E1861, https://doi.org/10.1175/bams-d-20-0299.1.

National Academies of Sciences, Engineering, and Medicine (2018), *Thriving on Our Changing Planet: A Decadal Strategy for Earth Observation from Space*, Natl. Acad. Press, Washington, D.C., https://doi.org/10.17226/24938.

Stern, C., et al. (2022), Pangeo Forge: Crowdsourcing analysis-ready, cloud optimized data production, *Front. Clim., 3*, 782909, https://doi.org/10.3389/fclim.2021.782909.

Wilkinson, M. D., et al. (2016), The FAIR guiding principles for scientific data management and stewardship, *Sci. Data, 3*, 160018, https://doi.org/10.1038/sdata.2016.18.

By **Zhong Liu** (zhong.liu@nasa.gov), Goddard Earth Sciences Data and Information Services Center, Greenbelt, Md.; also at George Mason University, Fairfax, Va.; **Yixin Wen,** University of Florida, Gainesville; **Vasco Mantas,** University of Coimbra, Coimbra, Portugal; and **David Meyer,** Goddard Earth Sciences Data and Information Services Center, Greenbelt, Md.

Read the article at Eos.org