BAMS
Meeting Summary

Expanding Access to Open Environmental Data

Advancements and Next Steps

Denis S. Willett, Brian White, Tom Augspurger, Jonathan Brannock, Jenny Dissen, Patrick Keown, Otis B. Brown, and Adrienne Simonson

NOAA Big Data Program Town Hall at the 2022 AMS Annual Conference

What: Panelists from NOAA Open Data Dissemination, NC State, Terrafuse AI, and Microsoft met to discuss new advancements in making NOAA environmental data publicly available on the cloud.

When: 25 January 2022

Where: Online

https://doi.org/10.1175/BAMS-D-22-0158.1

Corresponding author: Denis S. Willett, denis_willett@ncsu.edu

In final form 13 July 2022

AFFILIATIONS: Willett, Brannock, Dissen, and Brown—Cooperative Institute for Satellite Earth Systems Studies, North Carolina Institute of Climate Studies, North Carolina State University, Asheville, North Carolina; **White**—Terrafuse AI, Department of Earth, Marine and Environmental Sciences, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina; **Augspurger**—Planetary Computer, Microsoft, Redmond, Washington; **Keown and Simonson**—NOAA Open Data Dissemination, National Oceanic and Atmospheric Administration, Asheville, North Carolina

The past decade has seen increased adoption of cloud platforms to scale performant storage, compute, and network across big data challenges. It has become increasingly apparent that these advances hold great potential for increased understanding of our planet. Key to this understanding is the availability of open environmental data in a manner that is both eminently accessible and performant.

Making environmental data open is the key focus of the NOAA Big Data Program (BDP), now known as NOAA Open Data Dissemination (NODD). Since 2015, NODD has been building offerings of publicly available cloud datasets through its trusted data broker, the Cooperative Institute for Satellite Earth System Studies (CISESS). To do so, the program leverages relationships with Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP) to host NOAA environmental data on the cloud. These data are freely accessible to anyone via the internet (Fig. 1). As of the AMS town hall panel in January 2022, NODD hosts more than 15 PB of NOAA data, including popular datasets such as GOES-R, JPSS, NEXRAD level II, World Ocean Database, and GFS/GEFS. In many cases, these data are more performantly available through NODD and, in some cases, the only source of near-real-time, publicly available products.

Leveraging cloud platforms to support open environmental data access at scale has given NODD and its data broker CISESS unique insights into how these data are used through unique engagements between public and private partners in industry, government, and academia. These relationships have driven development of NODD and have engendered broader discussions around future priorities in facilitating access to open environmental data. A key focus

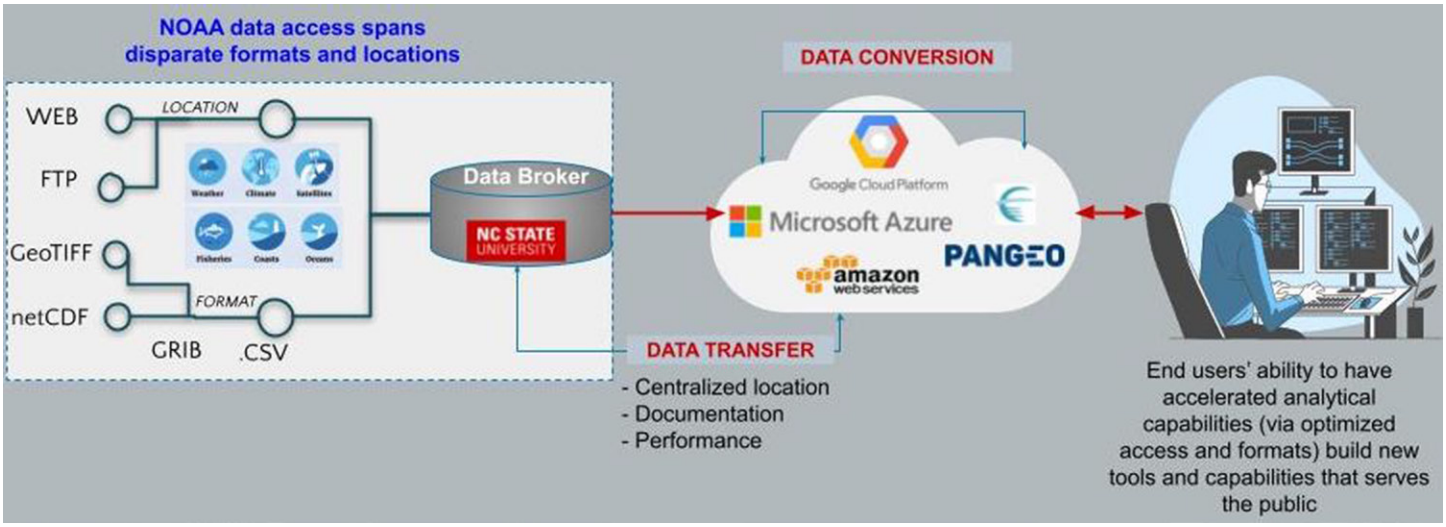

Fig. 1. NODD accelerates analytical capabilities through expanded access and enhanced interoperability.

that has arisen in discussion over the past year is on interoperability. With more and more environmental data being made publicly available, linking these data in accessible, performant formats is becoming a challenge.

Key themes from panelists

Against this backdrop, NODD organized a town hall panel discussion at the January 2022 102nd American Meteorological Society Annual Conference[1] to specifically examine the multifaceted technical challenges outlined above. The motivation for the town hall panel discussion stemmed from research findings that suggest that further development depends on user input around principles and practices to be embraced in order to advance open science.

[1] https://annual.ametsoc.org/index.cfm/2022/

The town hall began with an overview of NOAA's Open Data Dissemination model and approach given by the NODD Program Manager, Patrick Keown. The overview was then followed by a blended team of expert panelists representing academic, start-up, and corporate perspectives, which included Dr. Denis Willett, a representative from NODD Data Broker Technical Team; Tom Augspurger, a geospatial infrastructure engineer from Microsoft's Planetary Computer Program; and Dr. Brian White, cofounder at Terrafuse AI. These panelists were invited for their technical expertise from different perspectives across the open data value chain, including the data broker facilitating open access to NOAA data, a cloud service provider enhancing accessibility through additional development, and an end user leveraging cloud access and computing for AI applications in fire prediction. Together, these panelists contributed their perspectives to a discussion, facilitated by CISESS's Jenny Dissen, centered around developments in three key areas related to cloud accessibility of open environmental data.

Performance was discussed in the context of three stages of the data analytics pipeline, beginning with access. NOAA has directly facilitated access to open environmental data through agreements with cloud service providers that allow for free egress of data, hosting of period of record datasets, and no data throttling. While these three features are important for performant access of datasets, the town hall panel discussion focused on additional ways to facilitate access, particularly through data catalogs and format conversions. Tom Augspurger spoke to the work Microsoft was doing in developing open Spatiotemporal Asset Catalogs[2] delivered through APIs that allow users to access specific aspects of the information they need without downloading comprehensive datasets. This work is open source and allows the scientific community direct access to specific data through queries that facilitate linking of disparate data sources along common criteria. Brian White built on this discussion by layering in the importance of new formats for data storage that allow access to specific parts of the data in a parallel manner. These formats allow for rapid parallel computation and are particularly important for deep learning applications such as wildfire risk analysis that Terrafuse is involved with.

[2] https://stacspec.org/

This led to a discussion of performant computing, particularly in the context of parallel processing. The ability to massively scale computational resources in parallel is a key benefit of cloud architectures that is facilitated by data formats and storage to support this scaling. Tom Augspurger highlighted work on DaskHub and the Planetary Computer at Microsoft that allows users to scale computations in parallel across large clusters. Terrafuse AI is availing itself of parallel architectures to harness the power of the cloud to rapidly build, deploy, and test machine learning models at scale. The benefits of parallelization are further amplified by the ability to build event-driven architectures on cloud platforms. NODD supports construction of event-driven architectures through deploying messaging systems as a standard offering to provide notifications for new data updates. Both Microsoft and Terrafuse AI use event-driven

architectures to support on-demand cloud data pipelines that automatically trigger upon addition of new data that NODD provides to its cloud partners. Brian White emphasized the importance of these event-driven architectures for near-real-time deployment of AI-based products and highlighted the need for open environmental data streaming architectures. Streaming environmental data from satellites and sensors would reduce latencies for product development and dissemination while creating opportunities for in-stream data processing and machine learning.

This discussion led to conversation around the infrastructure needed to support AI and advanced machine learning applications in the cloud using environmental data. While this conversation highlighted the importance of a number of topics already discussed including data formats, one of the key aspects that arose was the importance of data provenance and metadata reporting. The ability to trace data from the source of generation, through quality control and quality assurance of different versions of the data, and metadata enrichment pipelines is fundamental to ensuring quality predictions from machine learning pipelines. This can be challenging as data are made publicly available in the cloud and Tom Augspurger highlighted ways that Microsoft is enhancing metadata accessibility through STAC offerings and ensuring that metadata have a primary role in Planetary Computer offerings. Brian White highlighted ways that Terrafuse AI uses metadata in its AI pipelines, with an emphasis on the importance of enhancing visibility up the data production chain closer to the source.

The conversation on metadata turned to the importance of creating novel insights and opportunities with open environmental data. This is one of the key ideas of Microsoft's Planetary Computer offering which helps facilitate connections between diverse datasets. Linking environmental datasets with demographic information (e.g., nonattributable health information) will be important in solving grand challenges and opportunities such as climate change. Terrafuse AI shared their experience in their technical offerings that strives to create local and regional value by analyzing wildfire risk and predictions at the local scale that benefits emergency planners and energy asset managers.

Key takeaways

Scaling these local analyses across the globe then becomes the next step. Being able to generate hyperlocal predictions at a global scale necessitates performant access to raw data, performant scalable computing on cloud platforms, and accessible, accurate, connected metadata. Throughout the conversation, the importance of open environmental data and ease of access cannot be understated. It is the basis for the development of information products that are protecting life and property. Enhancing the open environmental data offerings through NODD will have positive impacts throughout the value chain, from access to use in the next generation AI products shaping our world, across all sectors of the economy that are currently challenged with its analysis potential, which ultimately affects and impacts its utility. Specific opportunities to enhance accessibility of open environmental data that arose in the panel discussion include the following:

- *Open streaming platforms*—The development of open streaming platforms for low-latency real-time data access. Streaming open environmental data opens the door to a number of possibilities including real-time update of data products, real-time prediction of critical rapidly changing events, and model performance monitoring. There are immediate demands from industry partners for this type of data as the benefit of such architectures is both immediate and appreciable.
- *Enhanced data access*—Improving data accessibility through catalogs and cloud optimized formats. While there are many forms by which these initiatives could be implemented, they have two goals: allow access to specific parts of the data without accessing the entire

dataset and facilitate parallel computation on top of the available data. Enhancing data accessibility through these two avenues facilitates end use of the data and means that more performant and scalable operations can be accomplished with the same underlying data.
- *Data provenance*—Enhancing visibility up the data production pipeline to the source of raw data. Understanding where data come from is central to being able to use them effectively. Tracking data version and provenance is likewise critical to reproducible and effective machine learning pipelines. Standard communication, reporting, documentation, and versioning will go a long way to facilitating the development of robust and accurate data and machine learning pipelines.

Next steps

This panel discussion was helpful in surfacing priorities for enhancing open access to large-scale environmental data from authoritative sources such as NOAA, and recognizing the importance of cloud-ready formats to accelerate its utility. Enhancing data accessibility, data readiness, and AI/ML readiness based on user feedback will not only go a long way in meeting the needs of the next generation of Earth science research, but accelerate computational capabilities spurred by innovators and industry as they tackle climate solutions via improvements in interoperability.

NODD, as hosts of this panel discussion, along with Microsoft and Terrafuse welcome perspectives and feedback from the listeners, participants, and readers and invite you to connect with us by emailing nodd@noaa.gov.

Acknowledgments. This work was supported by National Oceanic and Atmospheric Administration through the Cooperative Institute for Satellite Earth System Studies under Cooperative Agreement NA19NES4320002.