NOAA Open Data Dissemination (formerly NOAA Big Data Project / Program)

Jenny Dissen[1], Adrienne Simonson[2], Otis Brown[3], Edward Kearns[4], Katelyn Szura[5], and Jonathan Brannock[6]

[1]North Carolina Institute for Climate Studies / North Carolina State University
[2]National Oceanic and Atmospheric Administration
[3]NC State University
[4]First Street Foundation
[5]Interactions, LLC
[6]North Carolina State University

November 21, 2022

Abstract

The National Oceanic and Atmospheric Administration (NOAA) research to operations (R2O) experiment called the Big Data Project (BDP) was envisioned as a scalable approach for disseminating exponentially increasing NOAA observation, model, and research datasets to the public using commercial cloud services. At the start of the project, during the concept development phase, it was unclear how the specifics might work so a spiral development approach was adopted. It was expected that the number of data sets would increase, and the data extent would grow to cover complete records of some holdings, and that format experimentation would be needed to determine optimal cloud offerings. This dissemination model would require a new way for the BDP and NOAA to engage with end-users, who could range from large enterprises to small businesses and individuals. The BDP was expected to change the game-not just by reaching a broad and diverse set of users but by encouraging new ones. As Dr. Kathy Sullivan, former NOAA Administrator under whom the BDP began, noted, “The agency’s aim is to ‘spur innovation’ and to explore how to create a ’global economic return on investment” (Konkel, 2015). This Chapter describes the journey of BDP as it developed, transitioned and evolved from an experiment to an operational enterprise function for NOAA, now known as NOAA Open Data Dissemination (NODD). Obstacles to the Public’s Use of NOAA Environmental Data NOAA’s mission is to understand and predict changes in climate, weather, oceans, and coasts, to share that knowledge and information with others, and to conserve and manage coastal and marine ecosystems and resources. The agency takes seriously the need for communication of NOAA’s research, data, and information for use by the Nation’s businesses and communities to allow preparation, response and resilience to sudden or prolonged changes in our natural systems. This includes climate predictions and projections; weather and water reports, forecasts and warnings; nautical charts and navigational information; and the continuous delivery of a range of Earth observations and scientific data sets for use by public, private, and academic sectors (NOAA About our agency, 2021).

NOAA Open Data Dissemination (formerly NOAA Big Data Project / Program)

Authors:
Adrienne Simonson(1), Otis Brown(2), Jenny Dissen(2), Ed Kearns(3), Kate Szura(4), Jonathan Brannock(2)

(1) National Oceanic & Atmospheric Administration, Office of the Chief Information Officer,151 Patton Ave, Asheville NC
(2) North Carolina State University, North Carolina Institute for Climate Studies/ NOAA Cooperative Institute for Satellite Earth System Studies (CISESS), 151 Patton Ave, Asheville NC
(3) First Street Foundation, 215 Plymouth St, Floor 3, Brooklyn, NY
(4) Interactions LLC, 31 Haywood St #E, Franklin, MA

Abstract

The National Oceanic and Atmospheric Administration (NOAA) research to operations (R2O) experiment called the Big Data Project (BDP) was envisioned as a scalable approach for disseminating exponentially increasing NOAA observation, model, and research datasets to the public using commercial cloud services. At the start of the project, during the concept development phase, it was unclear how the specifics might work so a spiral development approach was adopted. It was expected that the number of data sets would increase, and the data extent would grow to cover complete records of some holdings, and that format experimentation would be needed to determine optimal cloud offerings. This dissemination model would require a new way for the BDP and NOAA to engage with end-users, who could range from large enterprises to small businesses and individuals. The BDP was expected to change the game -- not just by reaching a broad and diverse set of users but by encouraging new ones. As Dr. Kathy Sullivan, former NOAA Administrator under whom the BDP began, noted, “The agency’s aim is to ‘spur innovation’ and to explore how to create a ‘global economic return on investment” (Konkel, 2015). This Chapter describes the journey of BDP as it developed, transitioned and evolved from an experiment to an operational enterprise function for NOAA, now known as NOAA Open Data Dissemination (NODD).

Obstacles to the Public’s Use of NOAA Environmental Data

NOAA’s mission is to understand and predict changes in climate, weather, oceans, and coasts, to share that knowledge and information with others, and to conserve and manage coastal and marine ecosystems and resources. The agency takes seriously the need for communication of NOAA’s research, data, and information for use by the Nation’s businesses and communities to allow preparation, response and resilience to sudden or prolonged changes in our natural systems. This includes climate predictions and projections; weather and water reports, forecasts and warnings; nautical charts and navigational information; and the continuous delivery of a range of Earth observations and scientific data sets for use by public, private, and academic sectors (NOAA About our agency, 2021).

As a result, NOAA collects, creates, archives and distributes thousands of datasets, as demonstrated by the more than 200,000 datasets that are discoverable in the U.S. Federal Government’s online catalog as of September 2021 (Data.Gov, n.d.). But there are two major obstacles to effective public use of NOAA’s environmental data: 1) the current data access and distribution model, including costs and secondary challenges associated with increasing volume, cybersecurity, and bandwidth limitations, and 2) the public’s ability to understand the data, including scientific formats, access methods, appropriate documentation, vocabularies and use case descriptions.

Under the NOAA's general data distribution paradigm, each individual requestor receives a copy of the data requested, most commonly delivered over the Internet. While this is a significant improvement over the former data delivery paradigms, such as mailing of printed publications, or delivery of physical digital media such as magnetic tapes and optical disks, it is a restrictive model limited by the capacity of the delivery mechanisms. Data services, such as databases, sub-setting services, and graphical user interfaces hosted on NOAA's computer infrastructure are limited in size and scalability, and their use are often throttled to ensure their fair and equitable availability to the public while preventing the overloading of NOAA's systems, especially during periods of heavy use.

In addition to the limitations in the current distribution model, understanding and leveraging the meaning of the data is often not a simple task. The formats in which the data are provided are often standardized, but they are largely standards that have been adopted by the environmental science community over the last 30 years and are outdated or not necessarily widely understood or adopted outside of that community. However, true public use of NOAA data means that the data should be usable by someone who understands the basics of geographical data and time series, but may not hold a Ph.D. in meteorology, oceanography, atmospheric sciences or fisheries sciences. Users may want to apply the data for cross-cutting applications, such as integration with social and statistical data science applications. However, NOAA collects data in the course of achieving its federally mandated mission and uses the scientific formats necessary to support NOAA's data analysis and information services requirements; these formats are not necessarily optimized for easy public interpretation. Formats will be discussed at greater length under the section Challenges and Opportunities.

In early 2019, Title II of the Evidence-Based Policymaking Act ("Evidence Act") (Foundations for Evidence-Based Policymaking Act of 2018, 2019) previously the "Open Data Act," amended Section 3504(b), Title 44 of the U.S. Code, to make all U.S. Government data open by default. This act includes guidance to make each public data asset available in an open format, with other guidance, including assisting the public in expanding the use of public data assets. The Evidence Act reinforced the foundation of the NOAA data vision. In fact, NOAA has been a long-standing leader in open data policy, pioneering the approach as part of international efforts to share open data for societal benefit, (World Meteorological Organization, n.d.) and with industry partners as part of the greater Weather, Water, and Climate Enterprises (NOAA National Weather Service, 2021) and the Blue Economy (NOAA National Ocean Service, 2021). NOAA has continued to innovate both in how it delivers its primary mission-specific information products and how it makes data collected or created in fulfillment of that primary mission available to the public. NODD is an innovative enterprise service providing the latter.

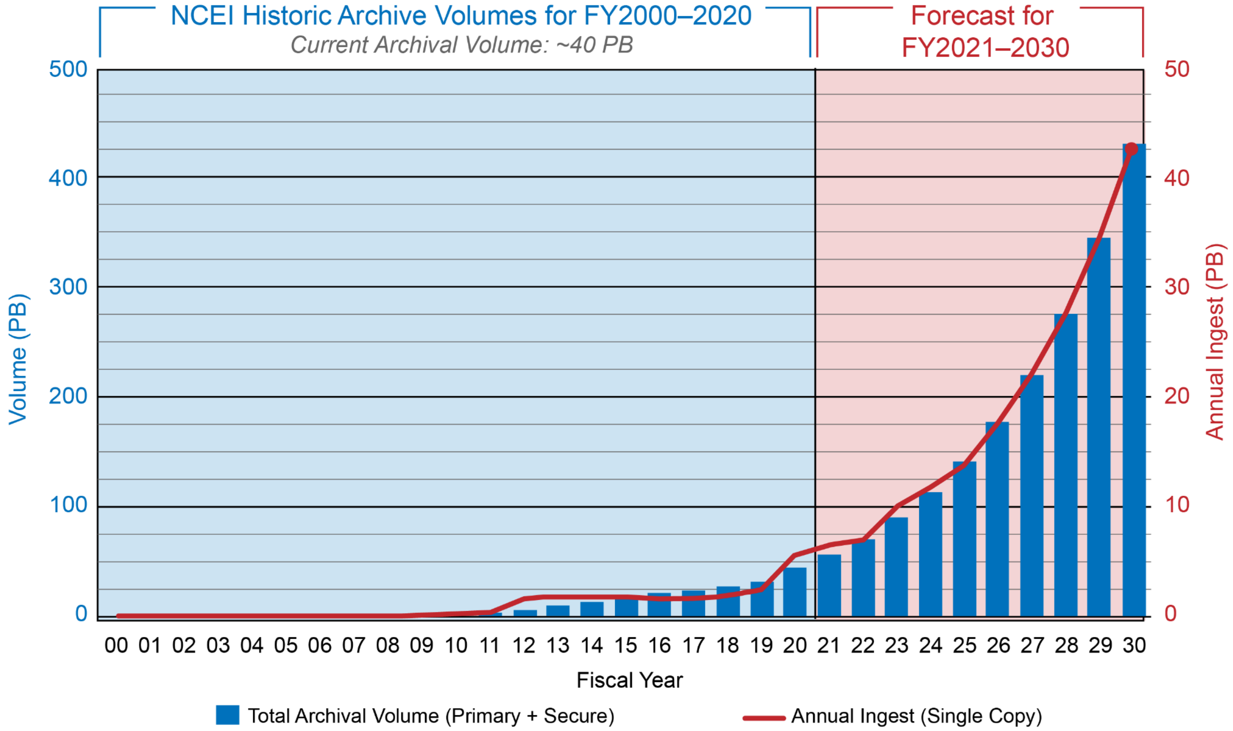


Figure 1. ***The actual total archive volume (active and secure copies) of data, in petabytes (PB), stewarded by National Centers for Environmental Information (NCEI) through FY2020. The years FY 2021 - 2030 are anticipated volumes of data, circa 2020. The rapidly increasing data volumes are contributions from space weather, oceanographic and atmospheric sources such as satellites, uncrewed systems, observations, products, and models.***

Public Access of NOAA Data Creates Challenges for the Agency

Allowing users to access its data also provides internal challenges for NOAA. The data reside on NOAA federal security systems, which are subject to significant security policies and practices required by law, resulting in increased costs for the agency and restrictions on the speed of those public-access services.

Moreover, the many copies that result from the one-to-one distribution model for the data consume significant bandwidth from NOAA's network services as the same data are moved repeatedly. The volume, variety and velocity of NOAA's observations and model results have been rising and are expected to exponentially increase in the future (Figure 1). Due to improvements in the resolution of observing systems and computer model outputs, the volume, variety, and velocity delivered through the current NOAA networks has also risen exponentially. This increased demand causes unintentional technical burden on current networks and systems, resulting in delays to data access.

All of these limitations in the current data distribution approach, such as volume, cybersecurity, and bandwidth come with a constant and proportionally rising cost to NOAA. Traditionally all egress from a cloud holding is paid for by either the holding owner or the user accessing the data. As shown in Figure 1 (above), NOAA's data holdings are expanding exponentially, while the usage has grown from ~1.2 PB per month to ~2.8 PB per month over the past year. This current service model is not only difficult to budget for, but given the dramatic increase in volume, could be financially challenging and could impact users' capacity to utilize the data. For example, NCEI currently hosts and provides access to more than 40 PB (40,000 terabytes (TB) or 40 million gigabytes (GB)) of data (primary and secure copy) and anticipates that the demand for data stewardship will rise to more than 400 PB by 2030 (Figure 1, above).

Egress for the anticipated user demand of all this data is financially challenging for NOAA, but also far more broadly across the federal government, given the prospective volumes of open data within commerce, space, natural resources, agriculture, and other agencies. As NASA's Office of Inspector General noted in its March 2020 report, "The agency faces the possibility of substantial cost increases for data egress (i.e., when end users download data from a network to an external location) from the cloud...the Agency, not the user, will be charged every time data is egressed" (National Aeronautics and Space Administration, 2020). And as the U.S. National Academies of Science noted, (see text box below), data architectures need to be both effective and agile, and the exploration of new data dissemination strategies is needed to facilitate more interdisciplinary collaborations. In fact, these are the very issues that the BDP experiment was designed to address.

National Academies Text Box

"Recommendation 4.3: NASA, NOAA, and USGS should continue to advance data science as an ongoing priority within their organizations in partnership with the science/applications communities by: a) identifying best practices for data quality and availability; b) developing data architecture designs that are effective and agile; c) exploring new data storage/dissemination strategies to facilitate more interdisciplinary collaborations" (National Academies of Science, Engineering, and Medicine, 2018).

Caption for Text Box: The U.S. National Academies of Science most recent decadal survey includes recommendations to U.S. agencies concerning data architecture and dissemination strategies

The Vision for NOAA's 'Oddball' Approach to Big Data

The NOAA research experiment called the Big Data Project (BDP), which was later operationalized as the Big Data Program (NOAA Office of the Chief Information Officer, 2021) and has since evolved into NOAA Open Data Dissemination (NODD), was envisioned as a scalable approach for disseminating exponentially increasing NOAA observation, model and research datasets to the public using commercial cloud services. At the start of the project experiment, it was unclear how the specifics of this effort might work, so a spiral development approach was adopted, allowing provisioning and transfer mechanisms, dataset richness, size, period of record, and data complexity, for example, to evolve over time.

This led directly to adoption of a partnership with a Cooperative Institute-hosted "Data Broker" to provide a buffer between NOAA and the cloud partners, act as an impedance matcher (e.g. overcoming barriers and connecting disparate cultures from NOAA and cloud partners), facilitate data transfers and cloud mechanics, and be a source for experimental infrastructural needs. Given the spiral development paradigm, it was expected that the number of data sets would grow and the data extent would expand to cover complete records of some holdings, and that format experimentation would be needed to determine optimal cloud offerings.

Ever increasing user demand has been a significant driver; BDP visionaries understood that this dissemination model would require a new way for NOAA to engage with its end users, who could range from large public or private enterprises to small businesses and individuals. Understanding who the users are, how they are using the data, how to engage with them, and how to understand their needs is just as challenging as the technical aspects.

NOAA solicited responses from the public and industry for the BDP through several formal Requests for Information to understand how to most efficiently provide access to NOAA's open data. Use of the commercial cloud services emerged as a leading recommendation. NOAA's collaborators in this experiment were Amazon Web Services (AWS), Google Cloud Platform (GCP), IBM, Microsoft Azure, and the Open Commons Consortium (OCC). These entities, through signed, multi-year Collaborative Research and Development Agreements (CRADAs), (NOAA Technology Partnerships Office, n.d.) agreed to host selected NOAA open data and to make those datasets publicly available at no net cost to NOAA or the public. While NOAA's open data remained free of charge to everyone, the BDP CRADA Collaborators could monetize services and derivative products based on the open data provided by NOAA. Previous federal experiences with open versus closed data sources had indicated that there is much greater economic potential in open data, such as demonstrated by the history of Landsat data utilization (Exploring Commericial Opportunities to Maximize Earth Science Investments, 2015) but the degree to which these public–private partnerships would be valued by both the cloud service providers and the agency was still to be determined.

Text box

Dr. Sullivan said the project is still in its early stages and there aren't yet clear road maps. "We don't really know where this is going," Sullivan said. "It's an oddball approach for government agencies, and an oddball approach for companies who can't see an assured path for a return on investment. But we've rededicated our effort on the presumption that over time, the demand for this kind of environmental intelligence will continue to grow" (Konkel, 2015).

Text box caption: ***NOAA Tries 'Oddball' Approach to Harnessing Big Data, Nextgov, September 11, 2015***

Text box

"As America's Data Agency, we are excited about these collaborations and the opportunities they present to drive economic growth and business innovation," said Secretary Pritzker. "The Commerce Department's data collection literally reaches from the depths of the ocean to the surface of the sun and this announcement is another example of our ongoing commitment to providing a broad foundation for economic growth and opportunity to America's businesses by transforming the Department's data capabilities and supporting a data-enabled economy" (U.S. Department of Commerce, 2015)

Text box caption: **U.S. Secretary of Commerce Penny Pritzker Announces New Collaboration to Unleash the Power of NOAA's Data, August 21, 2015**

A NOAA Cooperative Institute Data Broker Provides Research and Operational Agility

In 2015, NOAA's Cooperative Institute for Climate and Satellites in North Carolina (CICS-NC; now the Cooperative Institute for Satellite Earth System Studies, CISESS) (North Carolina Institute for Climate Studies, 2019) a unit of North Carolina State University, pioneered the Data Broker role during the experimental CRADA phase and is hosting the function in the operational phase for NODD. Today, CISESS is NODD's trusted Data Broker and serves as technical data transfer and monitoring lead, as well as user engagement lead for NODD.

As NODD's trusted Data Broker, CISESS is permitted access inside the federal security boundary to retrieve NOAA datasets, thus reducing the cyber risk associated with multiple entities seeking data access (see Figure 2, below). CISESS then moves the data to the three Cloud Service Providers (CSPs): Microsoft Azure, Google Cloud, and AWS, which in turn provide free public access to an exponential number of users. Users spanning the gamut of public and private sectors are able to experience improved access to NOAA data through the cloud, enabling their innovative data analysis and decision-making.

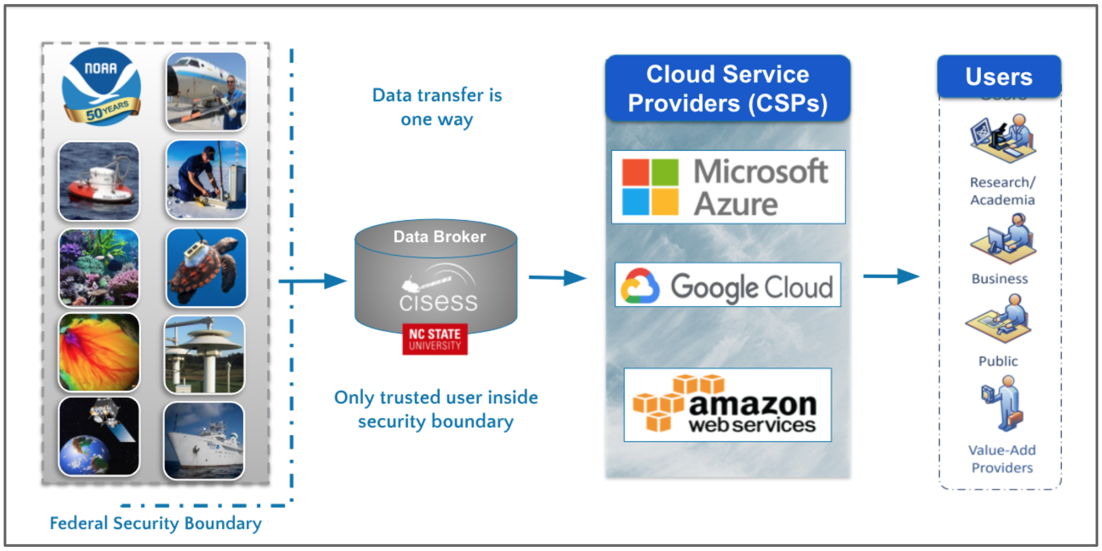


Figure 2. BDP's data distribution scheme is led by a one-way transfer of a single copy of a dataset out of NOAA's federal systems to the trusted Data Broker. The Data Broker distributes multiple copies of that dataset to the BDP CSP platforms, where public access is provided to an exponential number of users in a scalable fashion.

Many of the technical processes and other elements of the NODD program are still under development, even as it provides datasets operationally. CISESS, as a NOAA Cooperative Institute (CI), is positioned to provide the agility and innovation required to both support and respond to the continuing evolution of the NODD program. CISESS innovates, inspires and accelerates NOAA and NODD's mission objectives with strategic, technical, data and engagement activities, including NOAA's transition to the cloud. CISESS also supports research and industry efforts directly linked to NOAA's mission, particularly where NOAA does not have sufficient internal capabilities or capacity.

The role of an intermediate "Data Broker" between NOAA and the CSPs emerged as a practical, valuable and necessary function. The Broker coordinates the publishing of NOAA data from federal systems to all of the cloud platforms, which eliminates the need for various experts across NOAA to develop the expertise to push their data onto the different cloud platforms. The Data Broker also provides the CSPs with a single point of contact and reporting, simplifying data operations and security concerns. And, the Data Broker links two very disparate corporate cultures.

Today, the Data Broker role is the backbone to the data transfer function of NODD. The Data Broker monitors data flows from source to destination and resolves operational problems; supports the development of statistics and metrics via data analytics software to monitor dataset usage (e.g., volume and accession); and helps to develop and implement approaches to ensure end-to-end data integrity and certify select cloud service provider holdings. In many cases, the Data Broker is backfilling near-real-time (NRT) cloud holdings (acquired from NOAA NRT sources) with the complete period-of-record holdings for select datasets from NOAA archives.

In keeping with its core mission focused on research-to-operations (R2O), CISESS as the Data Broker researches and applies relevant R2O options for more efficient methodologies for data transfer (e.g. Globus, GridFTP, Apache NiFi) and assists with data format transformation or conversions for select datasets. The Data Broker also collaborates with cloud partners on new services and architectures as usage evolves over time, in keeping with the National Academies recommendation.

CI TEXT Box

A NOAA Cooperative Institute is a NOAA-supported, non-Federal, academic and/or non-profit institution that has an established, outstanding research program in one or more areas relevant to NOAA's mission; CI's are established at research institutions that have strong education programs with established degree programs in NOAA-related sciences. CIs engage in research directly related to NOAA's long-term mission needs that require substantial involvement of one or more research units within the research institution(s), as well as one or more NOAA programs; and a CI may include multiple research institutions. CIs provide significant coordination of resources among all non-government partners and promote the involvement of students and postdoctoral scientists in NOAA-funded research. CIs are competed every ten years and are reviewed in the fifth year (NOAA Office of the Chief Administrative Officer, 2021).

Caption for CI Text Box: NOAA Administrative Order 216-107A established the operating policy for all NOAA CIs.

Public–Private Partnerships Provide the Pipeline

NODD is also partnering with industry to overcome the obstacles to the use of NOAA data. The cloud computing and Data-as-a-Service (DaaS) industries are built upon modern, widely scalable computing architectures that are capable of seamlessly overcoming the limitations associated with increasing use and rising volumes of data. Users are able to work with those platforms, or third-party service providers that leverage those platforms, to acquire the quantity and types of data services they need. Today, any entrepreneur with a laptop, internet connection and a credit card can acquire the scale and type of computing infrastructure that was previously limited to large companies with millions of dollars of investment capital.

Why doesn't the government simply acquire the cloud computing capacity itself from commercial service providers and make them available to the public? The costs associated with such a service would continue to grow proportionally with new observing platforms and other data collection methods, resulting in higher volumes and ever-increasing use by an exponentially increasing number of users. For some types of NOAA services, such as the delivery of information products meant for the protection of lives and property, the procurement of such cloud-based services is entirely appropriate and effective. For example, during the 2017 hurricane season, the NOAA National Weather Service (NWS) National Hurricane Center website and its products were amplified for public access through a contract with AWS to employ its CloudFront product to ensure scalability (AWS Public Sector Blog Team, 2017). This service was able to seamlessly sustain over 1 billion hits per day from the public during the height of the storm season, ensuring that the public had access to the information they needed for critical decision-making.

However, it is difficult for NOAA to anticipate the extent of the data egress, or hits from the public, and thus the cost, particularly across all its data holdings, not just weather data meant to preserve life and property in an emergency. Moreover, such costs require Congressional approval and funding, which require agencies to anticipate their needs years in advance. There are many agencies across the government that have open data to share; yet taxpayers cannot support unfettered access to all the data that the surge in demand is creating. It is possible to limit, or throttle, users, but this also throttles the economic engine the open data creates.

Instead, NODD harnesses the interests of the stakeholders in a public–private partnership, which is usually centered upon an asset of mutual value and mutual benefit. In government, these are typically public works or infrastructure projects where the costs and risks are shared among the partners. But in a partnership based on open data, where are the shared benefits and risks? And what is the source of the value?

Initially, it may appear that the value is held within the open data themselves. But for the CSPs, NOAA data is open and available, not a scarce resource per se. Effective delivery of the data may be a limiting factor, but since the data are open and available to all in a fair and equitable manner as prescribed by federal regulation and policy, the data do not hold unique value. Rather, the value is in the relationship that the CSPs have with the NOAA expertise necessary to understand, analyze, translate, visualize and apply the data into useful synthesized products, assessments or information. Through NODD, NOAA provides the CSPs with: 1) supported access; 2) quality-controlled data; and 3) expertise. In return, the CSPs provide NOAA with a minimum of 5 PB free storage each – and egress of the data for all users at no cost to NOAA – or the taxpayer. Risks associated with the provision, quality, and usefulness of the data are shared by the partners -- NOAA's labor and reputation is on the line, and the CSPs' financial investments in infrastructure and labor are similarly committed.

Hosting NOAA data attracts users to the CSPs' platforms, and those users generate revenue by paying for computing and analytical services or new derivative products on those platforms, enabling the CSPs to recoup their data hosting costs and realize profits. Users can ask the CSPs for specific NOAA datasets and can more easily integrate and analyze cross-agency data sets because CSPs host other agency and environmental data. NODD also provides an opportunity for end users to curate and provide feedback on public data holdings. For example, during the CRADA stage, an external user went through the complete NEXRAD Level II dataset, in less than ten days, and found 10–15,000 granules that were either unreadable or contained no data or incomplete data. The Data Broker then checked these granules against the NOAA archive holdings to remediate, as appropriate. This is the open data version of open software.

If certain datasets are very popular and valuable, or special services are required for the timely delivery of information, or new artificial intelligence algorithms need petabytes of data immediately available for integration, or new formats are preferred by platform users, industry can respond and build the services it needs to be most effective. At the same time, NOAA can focus its limited resources on its mission -- which includes developing and maintaining the observation networks, producing authoritative, high-quality datasets and model forecasts, and providing objective information based on scientific expertise.

Data collaboratives based upon shared private or public data sources are becoming more popular, (Data Collaboratives, 2021) but are usually focused on one or more specific goals for the data that is being shared. A more general public–private partnership with widely defined data applications, such as that embodied by NODD, is still a collaboration but also may be an organizing function for a series of data collaboratives, communities of practice, and private companies seeking to create new applications or information products. If the collaboration, in this case, NODD, can support what Amazon Web Services has been calling the "undifferentiated heavy lifting" required for the benefit of all members within the partnership or collaborative, then the relationship is beneficial to all. NOAA is providing curation services that ensure well-organized, high-quality data, and since the CSPs and users can trust in that curation, NODD allows them to spend their labor further down the data value chain, unleashing new economic potential.

Text box caption - Lessons Learned in the Big Data Project

The most significant lessons that the BDP learned during the Cooperative Research and Development Agreement (CRADA) phase were originally identified as early as 2017, but have been proven repeatedly since and provide a solid foundation for a NOAA-wide enterprise service that benefits all stakeholders:

- There is measurable, latent demand for NOAA data, which represents untapped economic value.
- Partners' cloud platforms provide advantages for providing public access to NOAA data, including a vastly improved level of service for users, translation of earth science data to more consumable forms, easily handling the increasing number of users and exponentially increasing volumes, while improving NOAA's cybersecurity posture.
- Cloud access to copies of NOAA data is a technically feasible endeavor. However, there are more economic and cultural challenges than there are technical challenges to the BDP. This project has been a time-consuming effort based largely on interpersonal relationships.
- The role of an intermediate "Data Broker" has emerged as a valuable function and possible enterprise service that could support NOAA in provisioning data to the commercial cloud. This trusted agent itself accesses the NOAA federal data services and redistributes data to a non-federal system for public access, greatly reducing risk and exposure on the federal system. The CSPs also recognized the value of the data broker role and efforts.
- Increased usage of NOAA data by a wider and deeper set of users has also led to the more rapid identification of anomalies in NOAA data. Although it is well known that community feedback from users can reliably lead to improvements in open datasets, the rate at which this improvement occurs may be accelerated by provisioning NOAA data on cloud platforms that enable faster and more widespread usage.
- Provisioning data on the commercial cloud has great potential for the democratization of NOAA data. Small companies without significant on-premises network, storage, and computing capacity may purchase what they need to work with NOAA data to produce value-added products and services.

BDP Exceeds Expectations and Evolves into Enterprise Operations

In 2019, NOAA operationalized the Big Data Program (BDP), previously the Big Data Project, by signing unique contracts with Microsoft Azure, Google Cloud, and AWS. In 2021, the BDP became NOAA Open Data Dissemination (NODD),[1] an operational enterprise service providing public data access via the cloud. The unique contracts that memorialize NODD's public–private partnerships provide free, easy access to NOAA's open data - at no additional cost to the taxpayer. The cost avoidance is conservatively estimated at $5M in 2021.[2]

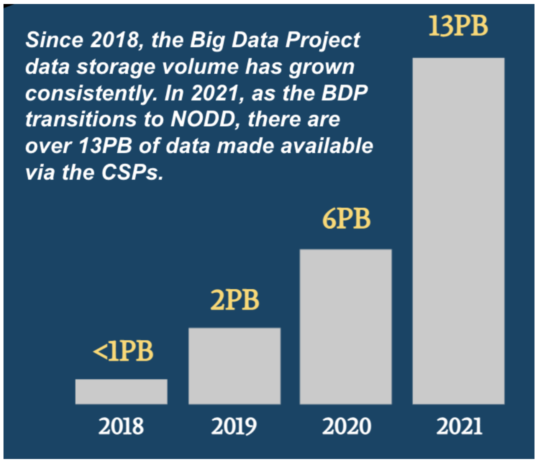
These unique contracts democratize NOAA data by making them easy for users to access, and by also providing tools next to the cloud platform's compute capability the data is easier to

[1] Via a FY22 President's Budget request, approved by Congress.

[2] Based on internal NODD-CISESS analysis.

analyze. NODD allows industry and research communities to increase the speed at which new products and services are developed and allows for the creation of new economic activities and research opportunities. NODD increases the use and usability of NOAA data, particularly for small to medium-sized businesses and startups, thus supporting equity and equal opportunity while also allowing for new insights and new approaches to knotty problems, like wildfire detection.

For example, a startup named Mayday.ai – begun with a laptop and CSP account – has been using NODD data to train its analysis engine, which uses machine learning to see through partial clouds; this enabled Mayday.ai to detect a high proportion of wildfire events up to 15 minutes after starting and well in advance of 911 calls reporting the incidents. Mayday.ai also enhanced its partial cloud detection technology to include lightning mapping every 10 minutes. This allows Mayday.ai to quickly identify lightning-caused fires. Today, Mayday.ai can see wildfire events as early as four hours ahead of 911 calls from 22,000 miles above sea level.


Since 2018, the Big Data Project data storage volume has grown consistently. In 2021, as the BDP transitions to NODD, there are over 13PB of data made available via the CSPs.

In addition, NODD provides easy access to climate and environmental information all over the globe, providing the opportunity to significantly increase the Nation's return on investment (ROI) in obtaining the data. And to be clear, the ROI desired is the use of the data to address complex social and climate problems. To support this outcome, NODD accelerates innovation in uses and applications; encourages innovative data analytics; and increases computational capabilities for earth system science, which allows for the identification and stitching of data issues and holes; and supports agile climate modeling across the world. NODD also encourages feedback from expanded communities of practice. Since 2018, the Big Data Project data storage volume has grown consistently. As the BDP transitions to NODD, there are more than 13 PB of data and more than 220 datasets, including Climate Data Records, (NOAA National Centers for Environmental Information, 2021) Climate Normals, (NOAA National Centers for Environmental Information, 2021) and NclimGrid (NOAA National Centers for Environmental Information, 2021), on all three cloud service providers. Moreover, NODD provides data and products from two of the top four high-impact, federally managed, U.S. Government observation systems (Colohan & Stryker, 17 September 2014). The full list of NODD datasets in the cloud is available at: https://www.noaa.gov/information-technology/big-data.

What happens when these data and products are made publicly available in the cloud? Unsurprisingly, the number of users and uses increases. For example, the Next-Generation Radar (NEXRAD) (NOAA National Centers for Environmental Information, 2019) ranked second of 362 high-impact, federally managed U.S. Government observation systems, has been available through NODD since October 2015. NEXRAD data now reaches more users than ever before. In July 2016, access of NEXRAD Level II data from one cloud platform alone (AWS), was 2.3 times higher than historical monthly rates for the same time period from NCEI. Requests to NCEI for NEXRAD Level II data declined by 50%, and 80% of NEXRAD data orders are now served by AWS (Ansari et al., 2017 BAMS).[3]

[3] More than 95% of NEXRAD L2 access on NODD is for data less than 5 minutes old, considered near real-time.

In addition, moving NEXRAD products to NODD has resulted in new and novel applications that employ the data in unanticipated ways. For example, by cross-referencing decades of standardized, on-the-ground tallies of birds with 143 radars designed to detect rain but capable of capturing "biomass" flying through the skies, researchers determined that North America has lost 3 billion birds in 50 years (Brulliard, 2019).

Other NOAA datasets have experienced similar increases in usage, making it clear that NODD is addressing latent demand, as well as improving data delivery for users already in the cloud or moving to the cloud. One example are the data from Geostationary Operational Environmental Satellite R-Series (GOES-R) (NOAA and NASA, 2021), the Nation's most advanced fleet of geostationary weather satellites, ranked fourth of 362 high-impact, federally managed U.S. Government observation systems, which are available on all three NODD CSPs. From October 2020 through July 2021, more than 10 PB of GOES-R data has been accessed on NODD. For every byte transferred in July 2021, a minimum of 30 times that amount was accessed by users in the cloud. Moreover, NODD is the only source for the public to access NOAA's near real-time GOES-R series data at no cost.[4]

From May through July 2021, NODD, in partnership with the National Weather Service (NWS), conducted a demonstration pilot for cloud-based access of three NOAA Operational Model Archive and Distribution System (NOMADS) model datasets: Global Forecast System (GFS); Rapid Refresh (RAP); and High-Resolution Rapid Refresh (HRRR). The NWS goals were to test the access of the data from a general usability perspective and to understand user interest in cloud data delivery, while possibly reducing demand on the NOMADS infrastructure (NOAA National Weather Service, 2021).

This demonstration revealed an increasing number of access requests for GFS model data and a steady rate of accession for HRRR. The demonstration also proved that NODD could handle 1.4 billion accesses of HRRR on a single day, without reducing capacity for any other users. NODD encourages users to communicate, and they share overwhelmingly positive feedback. Users like event-driven notifications, for example, the Simple Notification System on AWS, which allows users to be notified that the data has arrived versus having to poll the sources repeatedly, reducing strain on users and source systems. Users also like the fact that the NODD mirrored the NOMADS data structure; they like the ease of access, the efficiency of the data download – including cloud-to-cloud transfers, and they like the option of three different cloud service providers. The NOMADS demonstration reached users from a diversity of sectors, including a wide range of solution providers and innovators. Many of these users clearly appreciate the near-real-time delivery.

While challenges did materialize during the demonstration, they are solvable, and they provide opportunities to improve the service. NODD has addressed initial data flow process issues and continues to review and improve transfer processes as issues arise. Every data set is a little bit different. Due to the current method of acquiring data, latency can be a concern; NODD is working closely with NWS and others to develop solutions to reduce delays and latencies in the cloud. And NODD is evaluating alternatives for current NWS on-premise functions, such as the Grib filter (data parsing), which is not available in the cloud and results in a full versus a selective download of files.

[4] GOES-R data can be accessed through the NCEI, the NOAA Archive, but is not near real-time. Purchasing a ground station or a subscription to a ground station are also options, but can be cost prohibitive.

Once thought to be simply a service to provide archived data via cloud, NODD is showing real promise as a source for NOAA's near-real-time data.[5]

Engaging Users in the Cloud

NODD provides free public access to a diverse group of users across public, private and academic sectors who are able to experience improved access to NOAA data through the cloud, enabling innovative data analysis and decision-making.

In collaboration with CSPs and with assistance from the Cooperative Institute Data Broker, CISESS, NOAA has embarked on user engagement activities to understand data usage and access patterns, datasets of high interest, user needs and applications, user experiences with cloud-based data, and the value of cloud-based access to data from the users' perspectives. To date, collaborations between NODD, CSPs and CISESS have used a variety of methods and modes to reach users and diversify the user base, including:

- Social media notifications, including Datasets of the Week;
- Collaborative blog development, profiling users' experiences and applications;
- Co-development of education seminars, roundtables, and focused events;
- Involvement in professional, sectoral, and scientific conferences, workshops, and roundtables, as well as other collaborations with users and CSPs; and
- Direct user support via email regarding data requests, information, outages, or general inquiries allow NODD to identify datasets of highest interest, address user concerns about latency, answer user questions, address requests for format conversions, and discover industry use cases for the data, while involving CSP or NOAA data scientists or programmatic expertise, as appropriate.

Early Insights from User Engagement

The collaborative modes and methods listed above have been meaningful in validating the value proposition of cloud-based access and have generated feedback on other topics such as data formats, access methods, and data scale. Early insights from this feedback revealed that users prefer analyses with customization options, including data layers and analytics capabilities in the cloud. Users have also noted the importance of analyzing data rapidly, and ad hoc, with the ability to integrate and apply multiple data types. Users do not want to search, find, download, reformat, subset, and merge data types. Instead, users prefer to have cloud-friendly formatted environmental data that accelerates analysis in the cloud. Users want data in formats supporting data warehousing, such as Parquet, or tabular versions of raster data. Other feedback from users and cloud partners indicate a desire for stronger commitments to data delivery and lower latency in data delivery, as well as simple interest in other datasets.

NODD user engagement supports the GOES-R satellite series with programmatic engagement, where NODD is teaming with CSPs to improve awareness and increase competency of GOES-R data use through blogs, tutorials, coordinated events, and social media information. GOES-R programmatic engagement approaches are informed by data dissemination and analytics monitoring.

[5] This includes GOES-R and selected NOMADS datasets, and ~95% of NEXRAD Level II data, to date. The Joint Polar Satellite System (JPSS) is the follow on to the Suomi National Polar-orbiting Partnership (S-NPP) satellite, which is #14 on the high-impact federally managed observation system list and should be available on NODD in FY22.

As mentioned previously, NODD activities have made available approximately 2.2 PB of GOES-R data on the cloud partners, to date. CISESS's data monitoring and metrics show that more than 10 PB of GOES-R data had been accessed from the cloud from October 2020 through July 2021. NODD data analysis and monitoring also revealed that for every byte transferred by the data broker in July 2021, a minimum of 30 times that volume was accessed by users in the cloud.

GOES-R data is a popular data product accessed from the cloud. While older data, as defined as older than 5 minutes old, can be accessed from NCEI, the only near-real-time public access to this data – at no cost – is via NODD. CISESS provides users with the option to subscribe to a Simple Notification System (SNS) on AWS; or PubSub on Google. These subscriptions allow users to be notified that the data has arrived versus having to poll the sources repeatedly, reducing strain on users and source systems; to date there are more than 440[6] active subscriptions to GOES-R data by BDP users. The cloud partners have also noted the popularity of the datasets and have been willing to take the data without counting it against the NOAA allocation of cloud storage. NODD is further extending GOES-R user engagement goals to diversify the GOES-R user base, collaborating with users from other countries, as well as new and emerging users at home, on education seminars or other focused events. GOES-R user engagement efforts help NODD gain understanding of the user communities' interactions with the data, which datasets are of highest interest and priority, data format preferences, latency concerns and other needs which in turn improves dissemination pathways.

Data Analytics and Metrics Informing User Engagement

User engagement strategy and tactical user interaction efforts require the development of data dissemination pathways, as well as analysis of user access, behavior and interaction patterns with the data in the cloud. The development of efficient pathways for data transfers and data flows are dependent on the type of dataset, the source NOAA system, NOAA Line Office technical infrastructure, and input from users and cloud partners regarding data needs. The technical data dissemination approach includes continual research on methodologies, such as Globus, GridFTP, and Apache NIFI, the provision of user notification subscriptions, and understanding data formats and conversions, including both the formats available from the source system and the formats desired by users.

Currently, NODD and CISESS conduct data analytics and provide data usage metrics via a data analytics tool that monitors cloud holdings and data transfers, which are visualized on a dashboard. While each CSP has proprietary approaches, the NODD data analytics tool provides a cross-platform (cloud) architecture for compiling and displaying information about select data patterns, such as storage and egress. This is being generalized to provide NOAA partners improved understanding of their data holdings, transfer status, egress statistics, and other metrics.

NODD uses insights from the data analysis tools to understand data usage patterns and targets of opportunity for user engagement. However, the availability of user analytics information for deriving insights at a granular level is limited because 1) users are not required to register or be authenticated to access the data on the cloud, thus requiring a different approach to develop user analytics, and 2) privacy policies governing the cloud partners' severely limit their use of personal identifiable information from users, and thus limit the ability for NOAA and CISESS to receive any of the user information.

[6] As of September 7, 2021

Despite the limitations, NOAA recognizes and values the importance of usage patterns and understanding users' needs based on user interactions and communications. The CSPs have their own need to support users, which correlates with NOAA's, and thus they are motivated to support NODD's user engagement efforts. NODD continues to work with each of the CSPs to understand the user-related data they can make available, which varies. This information will then be used to assess the value -- to the user -- of cloud access to the data, as well as the engagement interactions themselves. Budget constraints have also limited NODD's ability to design and develop effective and actionable data analytics and metrics.

NODD Supports Industry Challenges in Sustainability

Users are asking NODD's cloud partners for environmental data to tackle their sustainability and other environmental, social, governance, equity and social challenges. Industry sector users are seeking engagement with cloud partners and are also interested in connecting with NOAA to improve their understanding of data holdings and how these data can be made available via the cloud. In these synergistic partnerships, NOAA data provided by NODD supports industry's environmental, social, and governance (ESG) risk metrics, while cloud partners enable the depth and breadth of data dissemination and user engagement, all of which provide the potential for the global return on investment that Dr. Sullivan envisioned.

The term ESG is often used to describe, or is used interchangeably with, the term sustainable investing. Many companies and organizations are considering environmental, social and governance factors, alongside financial factors, as they make investment and other decisions. In this context, "environmental" often includes climate, natural resource use, pollution and waste, clean technology, and renewable energy. 'Social' includes human capital, product liability, data privacy, health & safety. 'Governance' involves accounting practices, ownership, board independence and ethics.

More and more, as climate impacts are creating risks to investments, operations or valuation, companies are starting to, or are required to, provide disclosures of and transparency around their ESG metrics (Lee, 2021). This is critical not only for risk management, but is expected by many customers and investors, and required in some sectors. Companies that measure and report on their sustainability and ESG metrics, in addition to disclosures on their net zero progress, are often said to have higher valuation. Having access to environmental data in the cloud fosters advanced processing that connects and integrates large amounts of heterogeneous data, helping companies and industry investors include ESG factors alongside financial factors in various decision-making. This integration and advanced analysis are more feasible with environmental data from NOAA easily discoverable in the cloud.

NODD is a critical stepping stone in support of the complex systems analyses required to build the interdisciplinary relationship between climate change and sustainability. Cloud-based access to environmental data supports "the integration of heterogeneous data and models, and the exploration of the relationship between environmental and social factors can play a crucial role in climate challenges and mitigation opportunities" (Viktor, Tímea, & János, 2021) as well as adaptation and resilience strategies.

Advancing User Engagement for NODD

Climate change impacts are increasingly more pronounced and widespread particularly as their impacts are exacerbated in underserved communities. Resilience, adaptation and mitigation strategies require democratized, reliable and consistent access to relevant data to enable advanced analytics and computational methods that allow for the development of innovative solutions. Open, free public access to robust, reliable and comprehensive environmental data are critical and fundamental components to the capacities of both the public and private sectors to innovate and address climate-related risks. The availability of NOAA data in the cloud accelerates that capacity by democratizing and diversifying access. This provides users of all scales, particularly small businesses, start-ups, and underserved populations, with unique opportunities for innovation, particularly in areas where earth observations can provide insights to environmental stressors such as wildfires, tropical cyclones, and extreme precipitation.

NODD is supporting NOAA's transformation to the cloud and is a proven R2O experiment that has propelled a paradigm shift for both NOAA as the authoritative data provider and data users. This shift to cloud access and analysis calls for a shift in NOAA's user engagement efforts. NODD recognizes that data-driven user engagement means dynamic, iterative and consistent emphasis on user interactions. NOAA and CSP partners have recognized that environmental data gains value through dynamic, iterative and evolving modes of continuous engagement that benefits NOAA, data providers and users.

Looking ahead, NODD, cloud partners and CISESS aim to utilize multi-lateral and sector-focused engagement strategies to extend the use, application and value of NOAA environmental data and to conduct activities that strengthen and diversify partnerships and collaborations with other parts of NOAA, other federal agencies, and Historically Black Colleges and Universities. With a focus on broader reach and intentional user engagement, NODD responds to user and cloud partner inquiries related to data from other federal agencies, providing access to more ocean data, supporting development of cloud ready formats and data dictionaries, and other topics that advance computational analysis with integrated datasets.

Challenges & Opportunities

Format Conversions and Cloud Based Tools

In addition to the limitations in the current distribution model, understanding and leveraging the meaning of the data is often not straightforward. The formats in which the data are provided are often standardized, but they are largely standards that have been adopted by the environmental science community over the last 30 years and are not necessarily widely understood or adopted outside of that community. For example, these community data standards include formats like Grib, NetCDF, and HDF (National Center for Atmospheric Research, 2020). The environmental quantities described within, and even the vocabulary used to delimit those quantities, are well-understood and precisely described for the environmental scientists that are currently the primary consumers of these data. However, true public use of NOAA data means that the data should be usable by someone who understands the basics of geographical data and time series, but may not hold a Ph.D. in meteorology, oceanography, atmospheric sciences or fisheries sciences. Data collected by NOAA in the course of achieving its federally-mandated mission are in the scientific formats necessary to support NOAA's data analysis and information services requirements, and are not necessarily optimized for easy public interpretation.

When NOAA data are published onto the CSPs' platforms, the data are typically stored as objects within their cloud storage services. The NOAA weather radar and satellite imagery have been stored as such, and users may access the data directly from those object stores. While the BDP has experienced an increase in usage in this way, it requires users to understand the original data formats and the intricacies of the data environment they quantify.

The simplest example of a format conversion is unpacking a compressed dataset, such as NCEI's Global Historical Climatology Network (GHCN) Daily (GHCNd) product (NOAA National

Centers for Environmental Information, 2021), which is available compressed. This is done as an event-driven process at no cost to the users when a data granule arrives at a CSP and a comma-delimited version (which is a cloud optimized format) is made available.

A good example of format and process evolution is the cloud holdings of Emergency Response Imagery (ERI) (NOAA National Geodetic Survey, 2020), a product provided by NOAA's National Geodetic Survey. Originally published as JPEG images, the format was changed to GeoTIFF and eventually Cloud Optimized GeoTIFF (COG), based on user demand. Moreover, the current cloud holdings are directly updated by the National Geodetic Survey, rather than the Data Broker. Thus, cloud holdings are updated within hours of aircraft overflights following severe weather events, allowing more immediate assessments of impacts and responses. This evolution has improved user access to ERI and minimized access latency and the use of cloud resources.

An alternative, and or complementary approach is to integrate the NOAA data into cloud-based tools on the CSPs' platforms. This strategy has the most potential to increase usage, as opposed to simply making the original NOAA data files available on the cloud, since the data format issues have been superseded and often the meaning of the data is at least partially conveyed by their context within the tool. The most effective and efficient utilization has been observed when NOAA data have been integrated into those existing cloud-based tools that users are already in the practice of using, such as Google's BigQuery service. Consumers are able to discover and use the NOAA data as it appears alongside other data of interest to them, within a familiar framework. In the BigQuery example, Google offers subscribers a free but limited tier of service to allow open access to the NOAA data, while more advanced users may pay to analyze larger volumes or with a higher frequency. Agency expertise and CSP labor are, however, required to properly load NOAA data into any such tools.

Other strategies employed by the CSPs for encouraging the use of data and analytical tools were observed during the BDP experimental phase. One is to lay out all the parts, including the data, code, libraries, and documentation, within a common environment, constituting a "some assembly required" approach. This appears to have been useful for researchers and developers that are at least moderately familiar with the data and subject matter and have a desire to create tailored products and services. The Open Commons Consortium used this approach to help journalists create a story map using complex NOAA satellite data describing how slight shifts in weather conditions prevented Hurricane Irma from being an even more destructive storm in 2017, visually highlighting how these shifts prevented higher levels of flooding and damage in some of Florida's most populated cities (Lash & Bedi, 2017).

Another approach has been observed to encourage the growth of a "data ecosystem" by seeding a platform with data and the low-level tools that others can use; this approach has often been used by AWS, such as in its aforementioned successful hosting of NEXRAD Level II data. Currently, "notebooks", a hybrid text/code construct (e.g., Jupyter Notebooks (Jupyter, 2021)), are seen as delivery tools for accelerating use of NODD datasets. The CSP encourages and incentivizes others to come in and grow the implementations that are needed and will help organize communities of users that are focused on common problems or topics of interest.

A variety of services and access points were produced during the experimental phase. However, NODD public access services are provided by AWS (Amazon Web Services, 2021), GCP (Google Cloud, 2021), and Microsoft Azure through its Planetary Computer (Microsoft Planetary Computer, 2021). All users, including the public as well other commercial and research groups, can access, utilize, and download NOAA data, if desired, from these platforms without egress charges.

Pangeo (Pangeo, 2021) is developing tools, primarily as Jupyter Notebooks, that do data conversion of NOAA and other cloud-based environmental data holdings. A recent example is the conversion of National Centers for Environmental Prediction High-Resolution Rapid Refresh (NOAA Global Systems Laboratory, 2021) datasets into ZARR format (Jupyter nbviewer, 2021). Pangeo and its partners developed conversion and accession tools designed to convert HRRR from GRIB2 (FileInfo.com, 2021) into ZARR (Zarr, 2021), which improved access and minimized cloud resource use and egress volume.

Attention to Data Quality and Provenance

Traditional users of NOAA data trust that these data are of high quality because they have been received directly from NOAA's federal data services. But will users trust the NOAA data if they are accessed from the CSPs' cloud platforms instead of from a NOAA resource? What assurances do users have that the data on these platforms are unchanged and still of the same high quality as the original NOAA source?

NOAA, as the authoritative source of the environmental data, has the responsibility to establish and support data quality. Establishing the provenance and authenticity of the NOAA data within NODD and across the CSPs' platforms is possible given a number of technical solutions. Checksums could be computed for each data file published by NOAA, and these checksums could be checked by users to verify their authenticity and veracity. More complex schemes using secure methods such as cryptographic hashes to verify file-level data veracity, or blockchain technologies to establish the provenance of the data are possible but also require additional computational effort. NODD recognizes that the value of any verification that the data are an exact copy at the file level (such as file checksums) will start to break down as the data are extracted from the original files and inserted into cloud-based tools. In these cases, the tools themselves could possibly be assessed and certified by a review process to maintain the veracity of the input data.

NODD has also helped to identify and resolve inconsistencies between the CSPs' data holdings and the catalogs held by NOAA. In the case of the weather radar data, users raised concerns about apparent discontinuities in the data, and the Data Broker was able to independently verify over 300 million files and reconcile the entire weather radar data holdings on AWS, GCP, and OCC with the official NOAA archive manifests, in a matter of weeks. The reconciliation of the NOAA archive on this scale, with multiple cloud-based data stores, in such a rapid manner, was made possible by the ease and speed of access on the CSPs' platforms.

As noted in the introduction, NODD is an evolving program. Early on in the CRADA phase, most data transfer was done by bespoke software. This paradigm has evolved with the widespread implementation of NiFi (Apache nifi, 2021), an open source data flow tool, for most Data Broker transfers. NiFi implementation has provided a workflow definition and monitoring capability and improved provenance tracking and is an easily scalable cloud implementation, which includes graceful recovery options and has improved resilience. As each data transfer is an event, it is feasible for the CSPs to provide event notifications which end users can (and do) use to initiate their use of a dataset. This changes the end-user paradigm from one of looking for data to be made available to being notified it is available and initiating their particular activity – a much more efficient use of everyone's resources.

Vision for the Future

In the not-so-distant past, simply having access to open government data that could be distributed to others was a significant business advantage that could itself be monetized. But now as federal open data become widely available through many different means, will there be shifts in the information services community similar to the ones seen in the software community following the rise in popularity of open software? One consequence of a successful, ongoing NOAA public–private partnership with the cloud and or DaaS industries will be the ubiquity and democratization of NOAA open data, reaching a diversity of users and innovators. Obtaining access to those open data will not be difficult, but the ability to utilize and deliver actionable information from those data will become the challenge. Artificial intelligence (AI) and machine learning (ML) tools that are already available as a commodity on most cloud infrastructure platforms will be heavily leveraged to enable better understanding of the large quantities of NOAA environmental data alongside socio-economic, health and other data.

Interoperability of data across disciplines, which has been a consistently difficult problem over the past few decades, will likely be more achievable through AI/ML translations and the application of graph technologies that allow the relationships among data and applications to be fully leveraged. Since ubiquitous processing will be available alongside these data, future data collaborations will be well-positioned to extract the full value in these data and relationships and enable integration with other data sources. Other data collaborations between federal agencies and industry have already begun to develop, allowing for easier and faster combined analyses of data types. The National Institute of Health (NIH) has partnered with GCP to reduce economic and technological barriers to utilizing biomedical data in a program called the STRIDES (Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability) Initiative. Other federal agencies are following suit and exploring data collaborations with industry to make accessing and computing on federal public data easier. NOAA/NESDIS is currently optimizing specific NOAA datasets for AI/ML analysis with GCP through an Other Transactional Authority (NOAA National Environmental Satellite, Data, and Information Service, 2021).

Cloud technologies and open data availability must leverage effective social collaboration and integrated systems thinking to achieve common goals for advancing earth system analysis and sustainability goals. Cloud computing and infrastructure can guide decision-making and spur innovation by incorporating disparate and heterogenous social, environmental and economic data. It enables organizations to tackle complex climate and equity challenges, as well as provide sustainable computing options that minimize e-waste and optimize energy consumption through “green data centers.” Cloud based access to NOAA data accelerates scalable technology solutions such as smart grid and intelligent buildings that can further sustainable goals.

NODD’s decadal vision includes migration of NOAA computation and storage services to the cloud, development of cloud optimized forms for heavily accessed datasets, period-of-record data holdings for all heavily used datasets, NOAA product generators pushing data directly into publicly accessible cloud storage, development of community-based curation models for NOAA datasets, implementation of open-source metadata approaches for NOAA data holdings, improved connection of NOAA subject matter experts (SMEs) to users, improved understanding of end-user information needs through engagement and feedback, and readily available data metrics dashboards showing data availability, among other metrics. This diverse vision is challenging, but it can be achieved at a NOAA enterprise level, and will significantly broaden the usage of NOAA data while providing a global economic return on the Nation’s investment.

Acknowledgements

The NODD team would like to gratefully acknowledge the various individuals that brought their skills and attention to bear on this innovative public data access mechanism since 2015. Their efforts and dedication have made NODD possible and contributed to this Chapter.

The Leaders that inspired and enabled the BDP: Dr. Kathryn Sullivan, former NOAA Administrator; Penny Pritzker, former U.S. Commerce Secretary; Ben Friedman, NOAA Deputy Under Secretary of Operations; and Zach Goldstein, NOAA Chief Information Officer.

The Federal Employees that played a role in the BDP's success: David Michaud, Amy Gaskins (former), Andy Harris, Ed Kearns (former), Adrienne Simonson, Jon O'Neil, Patrick Keown, Jack Settelmaier, Jena Kent, Erin Wells and Natalie Donoho; Nancy Ritchey; Glenn Talia and Derek Hanson; Derek Parks; Marcelle Loveday, Eric Olmstead, and Michael Conroy; the BDP Advisory Board members; and many others who provided support.

The Presidential Innovation Fellows that shared industry best practices in the early years of the BDP: Maia Hansen and Alan Steremberg.

Support was also provided by Sea Grant Knauss Fellows: Kate Szura and Matthew Chase Long.

Federal Contractors (former) that played a role on the BDP: Shane Glass and Kevin Tukei.

Our Indispensable NOAA Cooperative Institute for Satellite Earth System Studies (CISESS) Partners Otis Brown, Jonathan Brannock, Jenny Dissen and Scott Wilkins.

Cloud Service Provider personnel are too many to mention but Special Thanks are due to team members from: Amazon Web Services: Jed Sunderland, Ana Pinheiro Privette, Joe Flasher, Zac Flamig, Conor Delaney (former); the Open Commons Consortium: Bob Grossman and Zac Flamig (former); Microsoft Azure: Laura Dobbs, Tim Carroll and Dan Morris; and Google Cloud: Lak Lakshmanan, Shane Glass, and Michael Hammoto Tribble.

References

Amazon Web Services. (2021). *Open Data on AWS*. https://aws.amazon.com/opendata/?wwps-cards.sort-by=item.additionalFields.sortDate&wwps-cards.sort-order=desc

Apache nifi. (2021). *Apache nifi*. https://nifi.apache.org/

AWS Public Sector Blog Team. (2017, December 17). *NOAA Keeps Citizens Informed of Eclipses and Hurricanes with Amazon CloudFront.* Amazon Web Services. https://aws.amazon.com/blogs/publicsector/noaa-keeps-citizens-informed-of-eclipses-and-hurricanes-with-amazon-cloudfront/

Brulliard, K. (2019, September 19). *North America has lost 3 billion birds in 50 years.* Washington Post, pp. https://www.washingtonpost.com/science/2019/09/19/north-america-has-lost-billion-birds-years/

Colohan, P. & Stryker, T. (17 September 2014). *The National Plan for Civil Earth Observations.* Presentation, National Research Council Committee on Earth Science and Applications from Space, Washington,D.C. https://sites.nationalacademies.org/cs/groups/ssbsite/documents/webpage/ssb_153133.pdf

Data Collaboratives. (2021). *Data Collaboratives Creating Public Value By Exchanging Data.* https://datacollaboratives.org/

Data.Gov. (n.d.). *The home of the U.S. Government's open data.* https://www.data.gov/

Exploring Commercial Opportunities to Maximize Earth Science Investments, Hearing of the House Committee on Science, Space and Technology, Subcommittee on Environment. (2015). https://docs.house.gov/meetings/SY/SY16/20151117/104181/HHRG-114-SY16-Wstate-PaceS-20151117.pdf

FileInfo.com. (2021). *.GRIB2 File Extension*. https://fileinfo.com/extension/grib2

Foundations for Evidence-Based Policymaking Act of 2018, 44 U.S.C. § 3504(b) (2019). https://www.govinfo.gov/content/pkg/PLAW-115publ435/pdf/PLAW-115publ435.pdf

Google Cloud. (2021). *BigQuery public datasets*. https://cloud.google.com/bigquery/public-data/

Jupyter. (2021). *Jupyter*. https://jupyter.org/

Jupyter nbviewer. (2021). *Explore the High Resolution Rapid Refresh (HRRR) Model Archive*. https://nbviewer.jupyter.org/gist/rsignell-usgs/4edfff890a7a18f97eaef42d647ec534

Konkel, F. R. (2015, September 2015). *NOAA Tries 'Oddball' Approach to Harnessing Big Data.* Nextgov. https://www.nextgov.com/analytics-data/2015/09/noaa-tries-oddball-approach-harnessing-big-data/120821/

Lash, N., & Bedi, N. (2017, September 20). *A matter of miles*. Tampa Bay Times: https://projects.tampabay.com/projects/2017/hurricane-irma/matter-of-miles/

Lee, A. H. (2021, June 28). *Climate, ESG, and the Board of Directors: "You Cannot Direct the Wind, But You Can Adjust Your Sails"*. U.S. Securities and Exchange Commission. https://www.sec.gov/news/speech/lee-climate-esg-board-of-directors#_ftn27

Microsoft Planetary Computer. (2021). *A Planetary Computer for a Sustainable Future*. https://planetarycomputer.microsoft.com/

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

NASA Office of Inspector General Office of Audits. (2020). *NASA's Management of Distributed Active Archive Centers.* National Aeronautics and Space Administration. https://oig.nasa.gov/docs/IG-20-011.pdf

National Center for Atmospheric Research. (2020). *Analysis Tools and Methods Common Climate Data Formats: Overview.* https://climatedataguide.ucar.edu/climate-data-tools-and-analysis/common-climate-data-formats-overview

NOAA About our agency. (2021, May 16). *Our mission and vision.* National Oceanic and Atmospheric Administration. https://www.noaa.gov/our-mission-and-vision

NOAA Office of the Chief Information Officer. (2021, June 10). *Big Data Program*. National Oceanic and Atmospheric Administration. https://www.noaa.gov/information-technology/big-data

NOAA Office of the Chief Administrative Officer. (2021, 06 04). *NAO 216-107A: NOAA Policy on Cooperative Institutes (Order 216-107A).* National Oceanic and Atmospheric Administration. https://www.noaa.gov/organization/administration/nao-216-107-noaa-policy-on-cooperative-institutes

NOAA National Ocean Service. (2021, June 06). *NOAA Strategy to Enhance Growth of American Blue Economy*. National Oceanic and Atmospheric Administration. https://oceanservice.noaa.gov/economy/blue-economy-strategy/

NOAA and NASA. (2021). *Geostationary Operational Environmental Satellites-R Series.* National Oceanic and Atmospheric Administration. https://www.goes-r.gov/

NOAA Global Systems Laboratory. (2021). *The High-Resolution Rapid Refresh (HRRR)*. National Oceanic and Atmospheric Administration. https://rapidrefresh.noaa.gov/hrrr/

NOAA National Centers for Environmental Information. (2019, April 24). *NOAA Next Generation Radar (NEXRAD) Level 2 Base Data*. National Oceanic and Atmospheric Administration. https://www.ncei.noaa.gov/access/metadata/landingpage/bin/iso?id=gov.noaa.ncdc:C00345

NOAA National Centers for Environmental Information. (2021). *Climate Data Records*. National Oceanic and Atmospheric Administration. https://www.ncei.noaa.gov/products/climate-data-records

NOAA National Centers for Environmental Information. (2021). *Global Historical Climatology Network daily (GHCNd)*. National Oceanic and Atmospheric Administration. https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-daily

NOAA National Centers for Environmental Information. (2021, May 03). *NOAA Monthly U.S. Climate Gridded Dataset (NClimGrid)*. National Oceanic and Atmospheric Administration. https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00332

NOAA National Centers for Environmental Information. (2021). *U.S. Climate Normals.* National Oceanic and Atmospheric Administration. https://www.ncei.noaa.gov/products/land-based-station/us-climate-normals

NOAA National Environmental Satellite, Data, and Information Service. (2021). *NESDIS/Google Artificial Intelligence Prototyping Initiative*. National Oceanic and Atmospheric Administration. https://www.nesdis.noaa.gov/about/documents-reports/nesdisgoogle-artificial-intelligence-prototyping-initiative

NOAA National Geodetic Survey. (2020, March 09). *Emergency Response Imagery*. National Oceanic and Atmospheric Administration. https://storms.ngs.noaa.gov/

NOAA Technology Partnerships Office. (n.d.). *Cooperative Research and Development Agreements (CRADAs)*. National Oceanic and Atmospheric Administration. https://techpartnerships.noaa.gov/Partnerships-Licensing/CRADAs

NOAA National Weather Service. (2021 June 30). *NWS Partners Webinar: Leveraging the Cloud for Numerical Weather Prediction Data*[PowerPoint Slides]. National Oceanic and Atmospheric Administration. https://www.weather.gov/media/wrn/calendar/FINAL_%20NWS%20Partners%20Webinar%20Discussion%20on%20Leveraging%20the%20Cloud%20for%20NWP_%20June%2030%2C%202021.pdf

NOAA National Weather Service. (2021). *Weather-Ready Nation Enterprise*. National Oceanic and Atmospheric Administration. https://www.weather.gov/wrn/enterprise

North Carolina Institute for Climate Studies. (2019, July). *The Cooperative Institute for Satellite Earth System Studies (CISESS)*. https://ncics.org/programs/cisess/

Pangeo. (2020). *Pangeo A community platform for Big Data geoscience*. https://pangeo.io/

Sebestyén Viktor, Czvetkó, T., János, A. (17 March 2021). The Applicability of Big Data in Climate Change Research: The Importance of System of Systems Thinking. *Frontiers in Environmental Science*, 70. https://doi.org/10.3389/fenvs.2021.619092

U.S. Department of Commerce. (2015, April 21). U.S. Secretary of Commerce Penny Pritzker Announces New Collaboration to Unleash the Power of NOAA's Data. [Press Release]. https://2014-2017.commerce.gov/news/press-releases/2015/04/us-secretary-commerce-penny-pritzker-announces-new-collaboration-unleash.html

World Meteorological Organization. (n.d.). *Home Page*. https://public.wmo.int/en

Zarr. (2021). *Zarr*. https://zarr.readthedocs.io/en/stable/