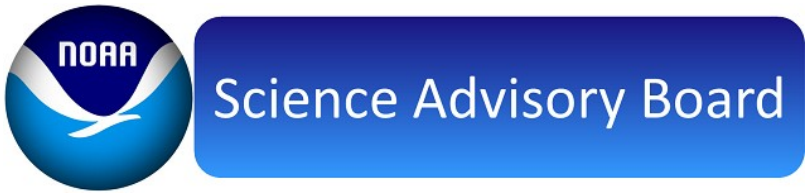

PREPARING FOR A CLOUDY FUTURE

A report on preparing analysis-ready datasets, training researchers to work in the cloud, preparing training data for machine learning, and the process for agile cloud implementation and deployment

WITH THE ASSISTANCE OF THE SAB DATA ARCHIVE AND ACCESS REQUIREMENTS WORKING GROUP

DECEMBER 16, 2019

“Preparing for a cloudy future”

Prepared by NOAA’s Science Advisory Board (SAB) Data Archive and Access Requirements Working Group (DAARWG)

Chair: Chelle Gentemann, Earth and Space Research
Molly Jahn, Jahn Research Group
Thomas Huang, NASA/JPL
Shane Glass, Google
Ana Pinheiro Privette, Amazon
Karen Stocks, Ocean Observatories Initiative
Eoin Howlett, Applied Science
Kandis Boyd, NOAA
Lucas Joppa, Microsoft

This is a brief report from NOAA’s Science Advisory Board (SAB) Data Archive and Access Requirements Working Group (DAARWG) on key issues and potential recommendations related to preparing analysis-ready datasets, training researchers to work in the cloud, preparing training data for machine learning, and the process for agile cloud implementation and deployment.

Our recommendations are as follows:

R0.1: Promote the creation of consortia focused on specific societal benefits that include private sector, public sector, NGOs and civil society that focus on developing cloud-based solutions using NOAA data.

1. **Recommendations for preparing analysis-ready datasets.**

R1.1: NOAA's strategy for providing public data access should:

a) Account for the needs and priorities of all user communities through a systematic and rigorous process, including an analysis of existing data usage patterns;

b) Publicly publish documentation that clearly describes the entire process of converting data access to ARDs, including its prioritization criteria and any technical processes utilized;

c) Make available data in its existing format in the Cloud as soon as possible to maximize user benefit while their conversion to ARDs is ongoing.

R1.2: NOAA should apply a clear, consistent, concise, and permissive open license to its data and software. NOAA should present this information along side every dataset at all access points, and include it in the metadata.

2. **Recommendation for training researchers to work in the cloud.**

R2.1: Offer trainings on existing open source technology (open source software libraries, containerization, software stacks, cloud-deployments).

R2.2: Build on existing open source training resources for on-boarding new collaborators and retraining existing workforce.

R2.3: Support open source software libraries.

R2.4: Develop a mechanism to fund Hackweek style events to address specific infrastructure issues or science challenges.

R2.5: Ensure inter-agency communication on lessons learned to help NOAA's approach to retraining it's workforce.

R2.6: Participate in consortia with external groups focused on using big data and cloud computing to address societal problems.

3. **Recommendations for preparing training data for machine learning.**

R3.1: Develop a NOAA-wide survey to identify pre-existing training datasets that may only require minimal reformatting and public distribution.

R3.2: Connect external and internal experts to develop and publish guidance for data providers to help create clean, quality-control, and labeled ML training data.

R3.3: Use the results from R3.1 to identify where there are substantial synergies across training dataset development to develop an enterprise solution and minimize redundancies.

R3.4: Publically share R3.1 identified datasets in a widely accepted format.

R3.5: Document the use of these new data.

4. **Recommendations regarding process for agile cloud implementation and deployment**

R4.1: Develop a categorization for where, and to what level, more formal software development processes would benefit NOAA.

R4.2: When appropriate, NOAA should consider adopting an agile software development strategy, including Test-Driven, Continuous Integration (CI), and container-based deployment software.

“Preparing for a cloudy future”
NOAA’s Science Advisory Board (SAB) Data Archive and Access Requirements Working Group (DAARWG)
11 December 2019

This is a brief report from NOAA’s Science Advisory Board (SAB) Data Archive and Access Requirements Working Group (DAARWG) on key issues and potential recommendations related to the following:

1. Recommendations for preparing analysis-ready datasets.
2. Recommendation for training researchers to work in the cloud.
3. Recommendations for preparing training data for machine learning.
4. Recommendations regarding process for agile cloud implementation and deployment

The commitment to mobilize NOAA Analysis-Ready Data (ARD) into the cloud is a key action that will underpin more nimble, effective disaster response to extreme weather and other catastrophic events and build more resilience into the U.S. economy. Intensive investment is occurring across the private sector to use NOAA data internally and in proprietary and open source analytics to innovate forecasting, loss estimates, hind-casting, cat modeling techniques, disaster notification, response and mitigation technologies, and related interdisciplinary data science. NOAA products, such as the Billion Dollar Disaster statistics, Severe Weather Index and many others are widely used across NOAA and Department of Commerce, by other government agencies for US Government forecasts, financial and physical risk management, insurance/reinsurance, and cross-functional analysis.

Bringing together NASA, NOAA, FEMA data, and that from other Federal agencies, can help us better understand the economic and social impact of disasters, support municipal and state managers, improve the health and safety of American citizens and foster resilient economic development. Moving onto the cloud is inherently a multi-disciplinary scientific and technological challenge that will require a coordinating a complex landscape of data, tools, services, science, and personnel.

Recommendation 0.1: Promote the creation of consortia focused on specific societal benefits that include private sector, public sector, NGOs and civil society that focus on developing cloud-based solutions using NOAA data.

1. Recommendations for preparing analysis-ready datasets.

Note: The term "analysis-ready dataset" has various definitions across a litany of scientific domains and is most commonly applied to satellite imagery[1] [2] [3]. For the purpose of this report, analysis-ready data (ARDs) will be defined as stated below:

> *Dataset(s) that are processed, organized, described, and accessed to minimize duplicative and/or unnecessary effort from users prior to analysis, enable immediate analysis with minimal user effort, and maximize interoperability with the time-series and with other datasets.*

Transferring existing datasets in their current state to the cloud will advance science by some measure. However, doing so would fail to empower scientists to leverage the full benefits of cloud computation such as scalability. Transforming existing datasets into analysis-ready datasets (ARDs) can be done either in place of hosting the original data files on the Cloud or in addition to doing so. It would not only improve the efficiency and speed of science, but also improve public trust by reducing the barriers for non-expert communities to use these data. NOAA will realize significant savings in computing costs and computation time by creating ARDs that optimize the data for use in the cloud instead of previous data access paradigms.

Creating ARDs can be done in a number of ways, but should be done to best addresses the needs of both internal and external users. An example of the benefits of providing ARDs for use is converting file formats. Satellite data is currently accessed at the file level, meaning the entire NetCDF/HDF file must be transferred to access any subset of the image. However, converting these files into a cloud-optimized zarr array based on its most common use cases allows users to access only the individual pixels they require. This allows for far more efficient search, data access, and analysis at a lower cost by minimizing the amount of data that is accessed and analyzed. The nature of the Cloud is such that there is no need to limit the number of users and/or data streams out of a single copy of the dataset, so users can parallelize their access and analysis for faster results. The Landsat conversion to GeoTIFF creating an analysis-ready archive, though more complicated, is an example of the benefits that can be gained from such work.

One requirement for creating ARDs is describing datasets in a way that enables the user to get started as quickly and easily as possible. This includes clear, easy to read metadata that thoroughly describes the dataset in a machine-readable format. This should enable users with minimal experience to leverage the data without having to contact a data expert within NOAA. It also requires providing all users with clear,

---

[1] CEOS Analysis-Ready Datasets
[2] U.S. Landsat 4-8 Analysis Ready Datasets (ARD) v1
[3] Analysis Ready Data Defined

consistent, and concise licensing terms that define how the data may be used and any restrictions of use. More details are available in the "Licensing" paragraph below.

Prioritization:
Converting NOAA's massive data archives into analysis-ready datasets will require a substantial investment of time and effort that will incur significant costs. Expecting NOAA to immediately provide access to all of its data as analysis-ready datasets is simply unreasonable. NOAA should develop documentation that describes the process and timeline for providing data access to ARDs for both internal and external data users. This documentation publicly document and describe the entirety of the process, with input and feedback from key stakeholders within other agencies, academia, and industry. It should clearly define NOAA's intended methodology for converting data access to ARDs, the technical processes used, and the programmatic processes that will be used, including its decision-making process, prioritization criteria, and projected/expected timeline converting data access to ARDs.

NOAA should prioritize providing access to ARDs to support datasets that are accessed most frequently to minimize data access costs and maximize user benefits. This should be based on a combination of many factors. The first is data about the usage of its existing data access services, including the number of users, data volume accessed, etc. Data usage information from datasets in the NOAA Big Data Project (BDP) should be included (where available and appropriate) to help NOAA understand how data usage may change as data access methodologies and formats evolve. These two data sources will reflect the broader community's need for access based on current and future data access paradigms.

The second prioritization factor is structured, systematic feedback from major scientific communities such as AGU or AMS. This provides NOAA guidance on which datasets the community would benefit from being able to access as ARDs. This should be done formally over an extended period of time to ensure the response reflect a representative sample of the community. This polling can be expanded to include academia, other government agencies, and/or any other key stakeholders who would be significantly impacted. Combined with existing usage information, this allows NOAA to capture a proxy measure for usage data of datasets not distributed through a system that measures such metrics, like GOES-16 and GOES-17 satellite, and account for the scientific value of providing ARDs to users.

However, the process of transitioning data access to ARDs should not serve as an artificial blocker to hosting copies of NOAA's data in the cloud to supplement existing public data access. Providing access to ARDs will enable internal and external users to maximize the benefits of working with NOAA's in the cloud cloud while minimal costs to both NOAA and users. But users will also realize significant benefits from having expanded access to data co-located with powerful computing resources. NOAA should provide cloud access to data in its existing form as it transitions to providing access to ARDs, as opposed to only providing access after a dataset is converted, to provide all users with the greatest opportunity to unlock the value of NOAA's data.

Licensing:
Simply stating that federal government data are in the public domain is not a sufficient license to maximize usage of federal government data. Public domain has a different definition and different restrictions in every legal jurisdiction around the world. It is far too ambiguous for many users, particularly those in industry, and prevents them from using any data licensed as such. NOAA should apply a license that is clear, consistent, concise, and open to each dataset. This license should provide the terms of use applied to each dataset, as well as any restrictions of use. A link to this license should be included at each access point to make it as easy as possible for users to discover and prevent any confusion as to whether the license applies to a specific dataset. This would help maximize the usage of federal data by eliminating confusion around the requirements to use the data. Examples of such a license are the Creative Commons CC0 license or the Open Data Commons Public Domain Dedication and License. NOAA should not use the Creative Common Public Domain Mark because it can be applied by someone other than the original provider of the data.

Recommendation 1.1: NOAA's strategy for providing public data access should:

a. Account for the needs and priorities of all user communities through a systematic and rigorous process, including an analysis of existing data usage patterns;
b. Publicly publish documentation that clearly describes the entire process of converting data access to ARDs, including its prioritization criteria and any technical processes utilized;
c. Make available data in its existing format in the Cloud as soon as possible to maximize user benefit while their conversion to ARDs is ongoing.

Recommendation 1.2: NOAA should apply a clear, consistent, concise, and permissive open license to its data and software. NOAA should present this information along side every dataset at all access points, and include it in the metadata.

2. Recommendations for training researchers to work in this area.

Transitioning NOAA's research and operational workloads from on-premise to the cloud can prove challenging and requires some level of effort to comply with the NOAA's security requirements and ensure that legacy systems connect smoothly with newer cloud-based applications. To ensure an effective transition, it is important that the challenges related to using the cloud by NOAA staff are closely monitored and that adequate training and support is provided. The process should include, at a minimum:

1. Training and educational sessions explaining how to use the cloud and how to leverage its infrastructure and technology (tools and services available)
2. Access to tutorials and use cases that demonstrate the value of cloud-based workflows for typical NOAA research/operational activities
3. Access to open source software (OSS) to facilitate adoption of cloud based solutions.

Training and Education:
An enterprise vision for workforce education and training should be flexible enough to accommodate different line office needs and leverage existing resources available to the community. One successful example of large-scale workforce training is provided by the nonprofit Carpentries organization which holds workshops to 'train the trainers'. The Carpentries is an excellent model and many of their training materials are aimed at specific user communities NOAA could benefit from creating discipline specific training materials showcasing some of the more powerful discipline specific libraries (eg. Satpy, Xarray) similar to what Software Carpentry has done for library science (https://librarycarpentry.org/lessons/).

The different cloud providers also offer a substantial list of training and educational materials that can be leveraged to support workforce education. For example, Amazon Web Services (AWS) provides access to webinars, workshops and tutorials that explain how to use its services for different applications. Google and Microsoft have similar resources that NOAA can use as well.

Tutorials and Use cases:
Access to tutorials and use cases can help researchers gain insight into the use of native cloud tools and their value to derive insight from data hosted in the cloud. Different Communities of Practice are compiling expertise and lessons learned from cloud based work using NOAA datasets. One such communities is Pangeo who is developing tools and tutorials to facilitate big data analysis in the cloud, currently deployed on Google, AWS and Azure. Another effort to promote community based learning in Earth on AWS and the Amazon Sustainability Data Initiative (ASDI) Communities of practice, for geospatial and sustainability work, respectively. ASDI compiles use-cases demonstrating how cloud-based solutions can innovate for sustainability. Examples are given here.

We recommend that NOAA creates specific tutorials on how to deploy jupyter hubs with software stack on the cloud to retrain the workforce and advance science using open source tools. In addition, organizing internal and external hackathons/hackweeks can also provide training opportunities for scientists and be a mechanism for learning cutting edge programming. These events also create an opportunity to expose NOAA datasets across the different line offices and encourage interdisciplinary science.

Open Source Software:
Existing open source software solutions can enable quick and easy access to cloud computing for a broad range of NOAA scientists and developers. Once in a cloud computing environment, open source software tools mask the complexities of cloud based storage and analysis solutions, presenting familiar interfaces to users. Many of those resources, design to support cloud based data analysis, can be found on GitHub. These recent advances can be adopted by the large NOAA workforce, but will require some individualization at the line office level. Pangeo and other open communities already provide some recipes to standardize data transformation into ARDs, and these can be built on and used, though the transformation will still require computing resources,

time, and expertise. Data centers should immediately begin building these transformation skills and prioritize this effort. It will be significantly more efficient to utilize existing open source software recipes and libraries. Within the list of prioritized datasets, there will be data that is straightforward to transform into an ARD and others that will be more complicated. For data where a transformation recipe already exists, these should be transformed in order of priority, for the rest, they should simply but put online in their existing format.

In addition to leveraging existing resources, we recommend that NOAA engages strategically with external trusted organizations to expand the available resources and build an ecosystem that supports cloud-based work. This will not only support the needs of NOAA's internal workforce but, as NOAA migrates its data portfolio to the cloud, enable the broader community to better access and use NOAA data on the commercial cloud. Specifically, we recommend that NOAA focuses on supporting open source software libraries and promote cross-domain cloud-based communities of practice.

Support for Open Source Software libraries
The 2018 NASEM report on OSS [NASEM, 2018] recognized how open source software libraries are becoming integrated into scientific advancement and how, *"the lack of institutional support, dedicated full-time developers, and dedicated funding for these libraries represents a major vulnerability of the basic infrastructure"*. Several options were suggested including: allocate professional software developers to these projects, encourage federal employees to participate in the projects, and/or providing funding directly to the projects. Groups like ESIP and NumFocus can be strategic partners in this process. We recommend that NOAA works with ESIP to investigate optimal ways to support OSS libraries.

Promote cross-domain cloud-based Communities of Practice
One of the largest challenges of solving climate, disasters and sustainability problems is the interdisciplinary nature of the data required to understand and manage those problems. Many of those datasets and knowledge live in silos and seldom interact with each other. The cloud provides an opportunity to access to data from different domains in a centralized fashion and enable widespread collaboration since the data and tools can be accessible virtually from anywhere in the world. This is an opportunity to encourage cross-domain collaboration and experimentation.

NOAA researchers should engage with external groups (Universities, NGOs, Government and the private sector) around projects working towards addressing cross-functional and cross-domain societal problems (forecast flooding, manage sustainable fisheries, supporting impact-based decision making for disaster relief and resilience, etc.) that can leverage big data and cloud computing. Consortia provide a venue to exercise this kind of practice. NOAA should encourage groups to develop pilot studies that benefit from big data and cloud computing to generate meaningful and timely insights for decision making. Projects should prioritize locations in the US where natural disasters may require development of more immediate resilience building efforts.

Recommendations 2.1: Offer trainings on existing open source technology (open source software libraries, containerization, software stacks, cloud-deployments).
Recommendation 2.2: Build on existing open source training resources for on-boarding new collaborators and retraining existing workforce.
Recommendation 2.3: Support open source software libraries.
Recommendation 2.4: Develop a mechanism to fund Hackweek style events to address specific infrastructure issues or science challenges.
Recommendation 2.5: Ensure inter-agency communication on lessons learned to help NOAA's approach to retraining it's workforce.
Recommendation 2.6: Participate in consortia with external groups focused on using big data and cloud computing to address societal problems.

3. Recommendations for preparing training data for machine learning

NASA's 2019 A.46 ACCESS call for proposals included an element on training datasets.

> 2.1.1.2 Creation of training data:
> ○ Earth science ML training data are expensive, error-prone, and usually require some expert interpretation. Many ML systems, especially those with large numbers of parameters, require large amounts of training data to generalize well. Proposals in this category are encouraged to augment existing EOSDIS tools and frameworks wherever possible in creation of training datasets. Proposers shall:
> • Develop training datasets that are clean, quality-controlled, and labeled.
> • Adapt, develop, and integrate solutions that make full use of novel ways to create high-quality training datasets and improve accuracy of labeling.
> • Generate synthetic training datasets through simulations with perfectly known labels. This kind of data has many benefits such as speed (fast to generate), accuracy (using physically-based models), tailored for specific needs, and scalable.
> • Share the training datasets under existing NASA Open Data and Information policy.

The goals stated in the NASA call are all relevant to NOAA's creation of training datasets. NOAA's data centers have excelled at preparing data formatting guidelines that have become the industry standard for many communities, and developing similar guidelines for the creation of training datasets would be a valuable contribution from NOAA for the ML community. Our recommendations follow the NASA ones, building on NOAA's strengths and data expertise. While NOAA/NESDIS/STAR has many satellite matchup databases that can be easily identified as potential training dataset candidates, there are many diverse data from across NOAA that individuals might not realize their utility as a training dataset. Finding the correct communication tool to reach different

types of NOAA scientists will maximize the use of NOAA public data resources for societal benefit.

Recommendation 3.1: Develop a NOAA-wide survey to identify pre-existing training datasets that may only require minimal reformatting and public distribution.
Recommendation 3.2: Connect external and internal experts to develop and publish guidance for data providers to help create clean, quality-control, and labeled ML training data.
Recommendation 3.3: Use the results from R3.1 to identify where there are substantial synergies across training dataset development to develop an enterprise solution and minimize redundancies.
Recommendation 3.4: Publically share R3.1 identified datasets in a widely accepted format.
Recommendation 3.5: Document the use of these new data.

4. Recommendations regarding process for agile cloud implementation and deployment

Adaptive planning, evolutionary development, early delivery, and continual improvement are some of the attributes of agile software development process. It promotes collaborative, self-organizing interactions among teams. The aim is to encourage rapid and flexible response to changes and to have cleared, dynamic interchange between members and teams. The focus is to keep a lightweight software development process in order to frequently produce Minimum Viable Products (MVPs) that go through several iterations as feedback is gathered.

Agile Software Development Process
Several Agile process frameworks exist such as Scrum, Kanban, Extreme Programming, etc. Scrum is a popular such framework. Scrum is a lightweight, iterative, and incremental framework for managing complex work. The key concept in Scrum is in the breaking up of the work into a collection of smaller goals that can be completed within timeboxed iterations, known as *sprints*. Sprints are short in duration, typically lasting 1-3 weeks and no longer than a month. The Scrum framework employs three roles: *Product Owner* (represents the stakeholder and customers)*, Scrum Master (*facilitator of the scrum process)*,* and the *Development Team* (carries out the task to build the incremental deliveries)*.*

Development Tools and Infrastructure
The following table lists the class of tools utilized for software development and management. Actual tools/tool suites to use will be identified as appropriate.

Development Tools	Purpose
Issue Tracker	Manages and tracks each requirement, feature and defect ticket throughout the life of the implementation
Version Control	Manages the versions of the implemented software produced by the development team
Feature Tracker	Provides management of all implementation tickets and allows for prioritization and assignment to specific releases
Code Analyzer	Static code analysis to identify potential security vulnerabilities, adherence to industry coding standards, etc..
Continuous Integration & Build Manager	Provides management and automation of the software development in various phases

Continuous Integration
Continuous Integration (CI) is not a new concept, yet its importance has been overlooked just as having, on-going independent system level integration and testing. It is insufficient to claim to have adopted CI just because a team has installed a service like Jenkin. For CI to be effective, it requires the engineering team to constantly develop and maintain, turnkey, unit and system tests to synthesize all related components upon any software changes (see Test-driven development). The CI process, if implemented correctly, enables a team to proactively identify build, interface, and compatibility issues during development. With the advancement of Container-based architecture, CI can be expanded to orchestrate, assemble complex tests using disposable software containers.

Configuration Management, Integration and Testing
Configuration Management is critical in any software implementation organization. It manages the life cycle of any complex software systems to ensure consistency in versioning, configuration according to the system requirements. It is recommended that NOAA to establish and maintain environments for System Integration and Test (SIT), User Acceptance Test (UAT), and Production.

- The SIT environment consists of the latest integrated updates and fixes for the I&T team to validate according to the product specification and system requirements.
- The UAT environment is a near production environment with selected components that have been certified by the SIT testing. It includes some of the latest and stable features that are candidates for production. This environment should be open to selected end-users (e.g. major stakeholders) for independent validation.
- The Production environment is to manage products that have been certified by the UAT testing

Issues (software bugs and recommendations) are tracked throughout the SIT and UAT testing and by Productions.

Automation and Container-based Deployment

Deploying big data solution to the Cloud or on-premise computing environment is no small task. A typical big data application could involve tens and hundreds of containers and deployed onto hundreds and thousands of computing nodes. Projects should look into infrastructure and software deployment automation to rapidly assemble new software system or deploy software patches. For example, Amazon's CloudFormation is an example of infrastructure automation for CM engineer to orchestrate the computing environment (compute node, storage, etc.) through versioned deployment scripts. Software components should be containerized (Docker, Kubernetes, etc.) to automate the package and deployment of software. These are script-based solution aiming for automation and version controlled, and eliminated localized manual software patching or configuration.

The following recommendations should be applied where determined appropriate, for example, software development rather than one-time science software.

Recommendation 4.1: Develop a categorization for where, and to what level, more formal software development processes would benefit NOAA.
Recommendation 4.2: When appropriate, NOAA should consider adopting an agile software development strategy, including Test-Driven, Continuous Integration (CI), and container-based deployment software.

References:

National Academies of Sciences, Engineering, and Medicine. 2018. Open Source Software Policy Options for NASA Earth and Space Sciences. Washington, DC: The National Academies Press. https://doi.org/10.17226/25217.