NOAA Response to SAB Observations on NESDIS Common Cloud Framework

Background

In November 2022, the Data Archiving and Access Requirements Working Group (DAARWG) of the NOAA Science Advisory Board (SAB) produced a report[1] providing responses and feedback on a briefing given by the NOAA National Centers for Environmental Information (NCEI) staff concerning the NESDIS Cloud Archive Project within the NESDIS Common Cloud Framework (NCCF). The DAARWG stated their support for NCEI's plan to use commercial cloud resources rather than on-prem infrastructure for primary storage and noted this approach could yield lower operating costs (including both hardware costs and in terms of effort needed to ensure data preservation). The DAARWG also noted this approach will provide better proximity to scalable cloud computing resources for NCEI's archived data.

SAB Observations and NOAA's Response

In addition, the SAB and DAARWG transmitted seven specific observations to the National Oceanic and Atmospheric Administration (NOAA) in the November 2022 report regarding the NOAA National Environmental Satellite, Data, and Information Service (NESDIS) plans to utilize commercial cloud resources rather than on-prem infrastructure for the archive of the future. NOAA thanks the SAB and DAARWG for their interest in this important topic, agrees with the concerns, and welcomes the opportunity to respond.

1) **Clarify Motivation: DAARWG recommends that NOAA clarify which benefits it is seeking to maximize or optimize in the NCCF project. The goal(s) should be stated along with quantitative metrics to assess whether they have eventually been met.**

   Response: In 2018, NESDIS published its guiding cloud strategy document, the "NESDIS Cloud Computing Strategy" (NESDIS-PLN-1120.1), which highlighted the benefits of moving NESDIS data processing and management to the cloud. Three key benefits were cited: greater efficiency, improved business agility, and enhanced innovation. NESDIS then demonstrated the cloud strategy hypothesis with two cloud pilot projects that demonstrated end-to-end data management to meet the majority of organizational enterprise processing, storage, and science requirements. The pilots successfully completed in 2020, resulting in the operationalization of the work, which would grow further to become the NCCF. Though in development, the NCCF has already demonstrated the benefits outlined in the NESDIS Cloud Computing Strategy, with three-times faster transition of science software into operations demonstrating improved efficiency, full traceability of cloud utility use by project to improve business agility, and enhanced innovation through collaborative science pilots with other NOAA Line Offices.

[1] https://sab.noaa.gov/wp-content/uploads/SAB_Report_Dec2022_NCCF.pdf

NESDIS is establishing formal requirements and then downstream specific metrics to monitor the use, developmental progress, and cost of the NCCF – specifically automated views of planned v. actual costs, percentage of completion on the migration projects, and quality of the user stories – all of which guide future development and implementation planning.

In addition to these general benefits of moving to the cloud, NESDIS and NCEI are using the opportunity to modernize the entire NOAA archive, enabling Web 3.0 and linked open data capabilities to improve semantic interoperability, provenance, transparency, scientific reproducibility, and AI-enablement of its archive holdings. These archive modernization efforts will significantly enhance NOAA's FAIR (Findable-Accessible-Interoperable-Reusable) data holdings, enable knowledge federation with other agencies and organizations, support new modes of interacting with NOAA's data using tools like large language models, and advance NOAA and US federal government goals for open science and open data.

2) **<u>Refine Migration Plan</u>: DAARWG suggests that NOAA consider some more detailed planning regarding what data are migrated in what order, including contingencies for unexpected delays.**

    a) **DAARWG recommends that NOAA clarify the method(s) to be used for the actual bulk data transfer (basic S3 API, Globus, AWS Snowballs or Snowmobile) and how the data integrity will be verified after the transfer.**
    b) **DAARWG recommends that NOAA provide more detail on the incremental solutions in terms of functionality. Will NOAA lift-and-shift data first, simply treating the cloud as storage for as-is data initially, with later optimizations to maximize the benefit of the cloud? Or will NOAA plan for and implement improved data and functionality prior or during the initial migration? The former approach may be faster but leave large collections of data not well organized; the latter might take more time but be better in the long run.**

Response: NESDIS has implemented the Scaled Agile Framework (SAFe), enabling the coordinated efforts of multiple Agile teams on multiple Agile Release Trains (ART) to incrementally develop and deliver the new NCCF cloud archive service, and migrate its legacy holdings to it. The Archive and Data Stewardship ART, led by NCEI, is developing a more detailed plan, or "forecast" in SAFe terminology, for the next 2 years. This 2-year forecast shows the current sequencing of functionality for the cloud archive service and its migration approach, which starts with the data stored in Colorado before moving to the data stored in Asheville in the Archive Management System (AMS) and High Performance Storage System (HPSS). Data stored in CLASS are being worked on concurrently. Critically, within the SAFe management approach, the forecast or plan is updated incrementally every 3 months, as new knowledge is gained. A few more details are provided below in response to the more specific questions 2a and 2b:

a) NESDIS is taking a multi-faceted approach to moving its archived data into the NCCF. Roughly half of its holdings are stored within CLASS infrastructure, which already has a

copy in the AWS cloud where the NCCF is operated. Since CLASS data are already in the cloud, migration is not needed; CLASS AWS accounts are being merged into the NCCF account structures in early Fiscal Year 2024 (FY24). Once the accounts are merged, the metadata contents of CLASS will be integrated into the new cloud archive system, beginning later in FY24.

The remaining archival data is managed by NCEI and resides on-prem in Boulder, CO, within the StoreNext storage-as-a-service solution and in Asheville, NC, within its AMS and HPSS-based archival storage systems. NESDIS is developing plans to migrate these holdings into NCCF, building on the lessons learned from migrating CLASS data to AWS and from the ongoing prototyping work for the transfer of the Boulder, CO, archive data and associated metadata data to the NCCF. The full NCEI data migration activity will begin later in FY24, leveraging lessons learned to ensure archive data transfer integrity, synchronization, and cost minimization. For archive data, it is critical to ensure not just bit-level integrity during the transfer but also that the broader context and provenance are transferred properly as this information is needed to ensure long-term preservation. These additional considerations are being carefully addressed by building from previous lessons learned and working incrementally and iteratively through the archive holdings.

b) NESDIS is moving its data holdings as-is into the NCCF, resulting from time constraints due to system age of on-premises storage systems. NESDIS is developing a strategy to optimize storage for the cloud with a strategy document that will be delivered in late FY24, including considerations of cloud-optimized data formats and semantic web enablement, following Linked Open Data principles and practices. Potential data optimizations will be performed during the initial transfer to the cloud in out years. Furthermore, NESDIS is working internally through the NESDIS Data Management Working Group and externally with groups like the Earth Science Information Partners (ESIP) to establish data readiness frameworks for cloud, AI, and semantic-web enablement.

3) **<u>Ensure Exit Plan</u>: DAARWG suggests NOAA include some provision for an exit strategy to be built into the contract with AWS. Concerns could include data egress costs (in particular, can they be waived in the event of a bulk migration out of the cloud), data transfer/copy/verification methods, and the time and effort to perform a bulk transfer to get everything out.**

Response: NESDIS agrees and will formalize an exit strategy for its data holdings in the cloud to prevent vendor lock. As the NOAA OCIO recompetes the Science Applications International Corporation Cloud Utility Blanket Purchase Agreement, NESDIS will enact a provision in the contract to require the vendor(s) to bulk migrate all data holdings to a new vendor if the contract is not renewed. NESDIS' implementation of data archive in NCCF will be assessed for risk and NARA compliance.

4) **<u>Explain Data Catalog Options</u>: DAARWG suggests that STAC be considered as an alternative to Common Metadata Repository (CMR), as it has a much broader community of support and use. Adopters include Google Earth Engine, Microsoft Planetary Computer, RadiantMLHub, and Sentinel Hub. While CMR can be translated to STAC, as exemplified by the NASA CMR-STAC proxy, the resulting STAC metadata is generalized. A direct STAC implementation could leverage community extensions and other specific STAC capabilities to better integrate with the community of tools made available to a STAC-compliant service. See more here: <u>https://stacspec.org/en/</u>.**

   Response: NESDIS and NCEI understand the need to support multiple user communities with its ecosystem of catalog services in the NCCF. To that end, the CMR is being included to foster better interoperability with the NASA satellite research community and will maintain its OneStop-Inventory Manager catalog as well, which provides discoverability to all of NOAA's data collections, including NOAA data not managed by NESDIS. In this way, NESDIS will improve its data discoverability while still meeting the requirements of the official NOAA Data Catalog. Furthermore, it provides NOAA's metadata to both Data.gov and the Department of Commerce Data Hub, meeting legal requirements of the Evidence Act.

   NESDIS is also working with NASA and other space agencies around the world through the Committee on Earth Observing Satellites to establish best practices for SpatioTemporal Asset Catalogs (STAC), which will eventually be implemented with the NCCF cloud archive service. Importantly, the new cloud archive system is designed to readily support multiple points of discovery for the NOAA archive holdings in a cost-effective way that can evolve over time to meet changing requirements and new user expectations. Over time, we will also explore consolidation of existing catalog services when efficiencies can be gained without negatively impacting user needs.

5) **<u>Consider Optimized Data Formats</u>: In order to enhance data when transferred to the cloud, DAARWG recommends that NOAA consider optimizations of highly-utilized datasets, such as organizing data to provide a more holistic, multi-dimensional "datacube" view of data rather than individual files; using cloud-optimized formats such as Zarr or COG; storing as-is with structural metadata such as ncZarr or kerchunk; or other enhancements. As with the data migration step, any enhancement or format conversion should be followed by some kind of data verification. Documentation should also be provided with details of any enhancement or format conversion steps as well as how the verification step was done.**

   Response: NESDIS agrees with this recommendation, and as noted earlier, is working with the community to leverage these tools to enable greater cloud optimization without the need to reformat petabytes of existing data. NESDIS is developing a data transformation strategy that will define a common data standard for cloud-optimized data storage as well as a strategy for how to best enable user-community specific format use. Furthermore, the NESDIS Data Management Working Group, through an effort led by NCEI with contributions from around NOAA, is also updating its NCEI netCDF Templates, which are

widely used by data producers to create highly-standardized, CF and ACDD-compliant netCDF data. These updates will be completed in FY24, will include recommendations on how to natively create more cloud-optimized netCDF data, and will be issued as a NESDIS best practice.

6) **<u>Define Terminology</u>: DAARWG recommends that NOAA clarify what vAIP and KG are, with examples, and indicate the value to either NOAA or users of implementing these.**

   Response: The term "vAIP", meaning "virtual archival information package", was previously used as shorthand reference to the cloud archive service being developed in the NCCF. Since the last DAARWG discussion in 2022, NESDIS and NCEI have clarified their terminology and more fully linked it to the Open Archival Information System Reference Model, the ISO standard for digital archives (ISO 14721). NESDIS, through interactions both internally and externally, has also begun an active campaign to discuss with our communities and educate them on the benefits of the knowledge graph (KG) – based framework of the new cloud archive service. In brief, through the use of the KG at the core of the new archive system, NESDIS and NCEI are enabling next-generation Web 3.0 constructs, including semantic interoperability, scientific reproducibility, linked open data, and AI-enablement. The new cloud archive system exposes not just data and metadata to the users, but also full provenance of the data workflows in a semantically meaningful way using JSON-LD triple stores, supporting the consumption of our vast and diverse holdings by AI tools, Large Language Models, and Digital Earth Twin frameworks.

7) **<u>Use or Contribute to Open Source Software</u>: DAARWG urges NOAA to consider releasing code as open source, and to contribute back to the community any enhancements NOAA may make to existing open source projects.**

   Response: NESDIS is supportive of the idea of releasing code via open source mechanisms. NOAA has been working on a new policy on sharing and archiving software as well as developing implementation guidelines. NESDIS is tracking these developments and plans to provide access to code out of the NCCF once these guidelines are established. One exception to this general support for sharing code is for the NCCF code itself. The NCCF is operational infrastructure-as-code; therefore, NESDIS does not support open source code release due to the risk to operational security. In the case of scientific and stewardship algorithms, which run on the operational environment, NESDIS will share its code as open source where appropriate and following the pending NOAA policy implementation guidelines. Algorithms made at STAR and NCEI often leverage partnerships and collaborations for development across cooperative institutes, other Federal agencies, and international partners, lending additional value to open-sourcing NESDIS science code.

<u>Additional NOAA Request of the SAB DAARWG Interaction Going Forward</u>

Given our scaled agile SAFe implementation of the new cloud archive service and broader capabilities of the NCCF, with quarterly program increments and planning events, NESDIS would welcome more routine and timely interactions with the DAARWG. NESDIS teams

conduct quarterly planning events to set priorities, commit to work, align teams, and resolve dependencies. NESDIS would like to target opportunities for DAARWG members to review artifacts from these planning events in order to provide the best current awareness of the migration to the NCCF and hold informal conversations to gain relevant and timely advice from the DAARWG.