PRINTER-FRIENDLY FORMAT Return to Article

D-Lib Magazine

November/December 2016
Volume 22, Number 11/12

Assessing Stewardship Maturity of the Global Historical Climatology Network-Monthly (GHCN-M) Dataset: Use Case Study and Lessons Learned

Ge Peng[1], Jay Lawrimore[2], Valerie Toner[3], Christina Lief[2], Richard Baldwin[2], Nancy Ritchey[2], Danny Brinegar[2] and Stephen A. Del Greco[2]

[1]*Cooperative Institute for Climate and Satellites-North Carolina, North Carolina State University and NOAA's National Centers for Environmental Information*

[2]*NOAA's National Centers for Environmental Information*

[3]*STG, Inc. and NOAA's National Centers for Environmental Information*

Corresponding Author: Ge Peng (ge.peng@noaa.gov)

DOI: 10.1045/november2016-peng

Abstract

Assessing stewardship maturity — the current state of how datasets are documented, preserved, stewarded, and made accessible publicly — is a critical step towards meeting U.S. federal regulations, organizational requirements, and user needs. The scientific data stewardship maturity matrix (DSMM), developed in partnership with NOAA's National Centers of Environmental Information (NCEI) and the Cooperative Institute for Climate and Satellites-North Carolina (CICS-NC), provides a consistent framework for assessing stewardship maturity of individual Earth Science datasets and capturing justifications for transparency. The consolidated stewardship maturity information will allow users and decision-makers to make informed use decisions based on their unique data needs. This DSMM was applied to a widely utilized monthly-land-surface-temperature dataset derived from the Global Historical Climatology Network (GHCN-M). This paper describes the stewardship maturity ratings of GHCN-M version 3 and provides actionable recommendations for improving the maturity of the dataset. The results from the use case study show that an application of DSMM like this one is useful to people who produce or care for digital environmental datasets. Assessments can identify the strengths and weaknesses of an individual dataset or organization's preservation and stewardship practices, including how information about the dataset is integrated into different systems.

Keywords: Scientific Data Stewardship, Data Management and Preservation, Stewardship Maturity Matrix, Transparency, GHCN-M

1 Introduction

Digital environmental and geospatial data products are increasingly treated as important assets to both scientific and business communities. Information derived from environmental data is considered a valuable resource to the U.S. Federal Government (OMB, 2013; OSTP, 2013). As a result, there is greater scrutiny placed on organizations to ensure data quality, to convey data quality information, to provide easy and timely data access, and to promote data transparency and traceability (OMB, 2002; NOAA, 2011; and see Peng *et al.*, 2016 for an overview of U.S. Federal Government policies and some of the agencies' requirements on ensuring data quality and improving data sharing).

The National Oceanic and Atmospheric Administration (NOAA) is responsible for providing environmental intelligence to American citizens, businesses, and governments to enable informed decisions (Sullivan, 2013). NOAA collects and cares for geophysical measurements of more than two thousand diverse parameters. Data come from a broad range of platforms, including (but not limited to) satellites, fixed and mobile radars, research aircraft, buoys, ships, land-based in situ surface and upper air networks, and weather and climate models, each of which presents its own data management issues (NRC, 2007). Therefore, NOAA is facing a serious challenge to provide a wide breadth of trustworthy data in a timely and user-friendly manner in a rapidly changing and resource-limited environment.

The National Centers for Environmental Information (NCEI) fulfills NOAA's responsibility by ensuring and improving data quality, discoverability, and accessibility. As NOAA's designated national data center, NCEI is responsible for collecting, stewarding, and providing access to atmospheric, oceanic, coastal, terrestrial, and solar observations. In recent years, environmental data volume at NCEI has grown at an astounding rate and is projected to grow even faster in the next decade and beyond (Figure 1).

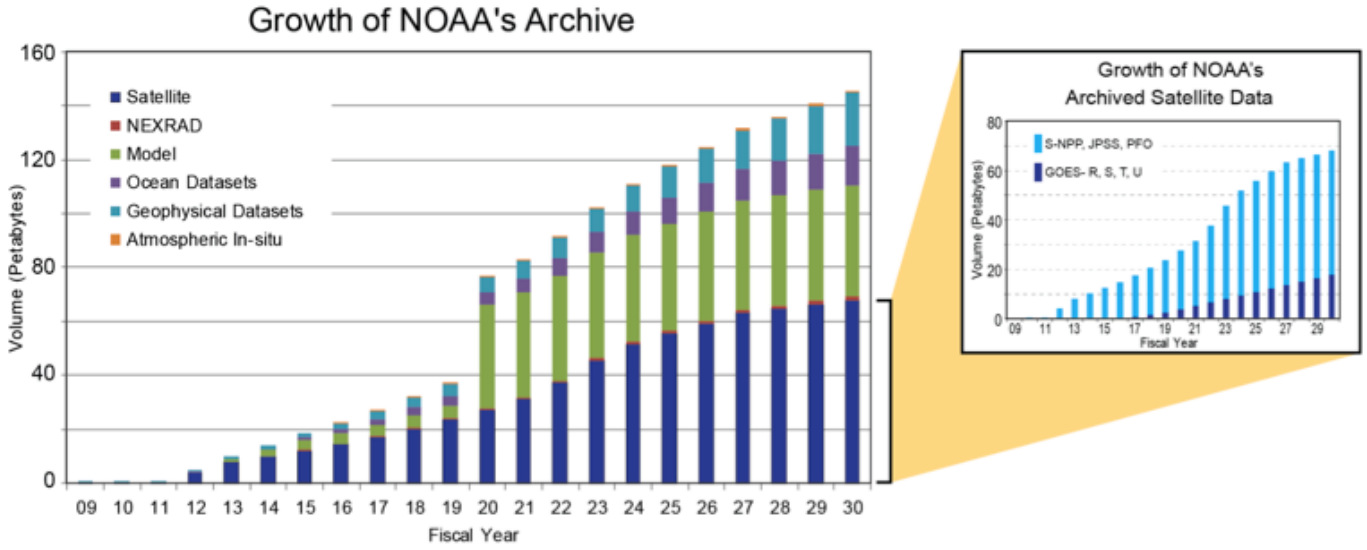

Figure 1: Environmental data archive volume at the NOAA's National Centers for Environmental Information (NCEI) since year 2009 and its projected data archive volume to year 2030.

Digital datasets of diverse data types with massive volume have become increasingly challenging to manage. This is especially true in the context of ensuring data and information quality while also meeting requirements to reduce latency and improve discoverability, accessibility, and traceability. But by systematically utilizing standardized reference frameworks for assessing quality of individual datasets, it is possible to capture, curate, and provide consistent, content-rich, quality information to users, thereby streamlining and simplifying data understanding and improving usability.

Several maturity assessment models have been developed and utilized for measuring the maturity of organizational capabilities and processes in the area of data preservation (see Peng *et al.*, 2015 for an overview.) As the first work to focus on individual datasets, Bates and Privette (2012) introduced a reference framework for measuring the maturity of data products in the form of a matrix. In contrast to scientific quality of a data product, which is defined in terms of accuracy, precision, uncertainty, validity and suitability for use (Ramapriyan *et al.*, 2016), product maturity addresses how well the scientific quality is assessed and documented, and how complete the metadata and documentation are.

Adopting a similar approach, the Data Stewardship Maturity Matrix (DSMM) has been developed jointly by NCEI and NOAA's Cooperative Institute for Climate and Satellites — North Carolina (CICS-NC) (Peng *et al.*, 2015). The stewardship maturity measures the current state of how datasets are documented, preserved, stewarded, and made accessible publicly (Peng *et al.*, 2015; Ramapriyan *et al.*, 2016). DSMM aims for a consistent evaluation of the maturity of stewardship practices applied to individual datasets and captures justifications for transparency. For each of nine quasi-independent key components (Figure 2), DSMM defines criteria based on measureable stewardship practices that can be used to apply a progressive, five-level rating to an individual dataset, representing maturity stages rated as Ad Hoc, Minimum, Good, Advanced, and Optimal (Figure 2; see Peng *et al.*, 2015 for the rationale and detailed definitions for each key component).

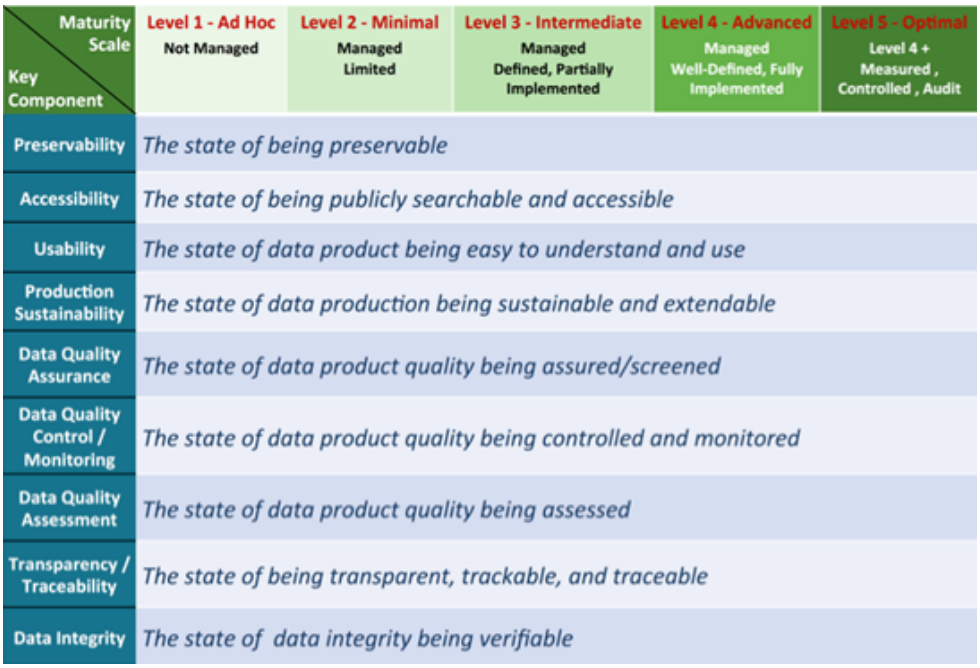

Figure 2: Conceptual diagram showing the nine DSMM key components, 5-level scale structure, and high-level descriptions of what each key component measures.

For operational datasets managed by designated NOAA data centers such as NCEI (see NOAA, 2008 for the classification of NOAA data centers), Level 3 is the recommended minimum maturity rating for all nine key components (Peng *et al.*, 2015).

A pilot use case study to apply DSMM to various NCEI data types (Table 1) has been underway by NCEI's Data Stewardship Division (DSD), in collaboration with NCEI's Center for Weather and Climate (CWC) and NOAA's Climate Data Record Program (CDRP).

Table 1: Selected NCEI Core Datasets for the NCEI Pilot DSMM Use Case Study Project

Data Type	*Dataset*
Satellite — polar ocean	NOAA/NSIDC Sea Ice Concentration Climate Data Record (CDR)
GIS — regional	Digital Elevation Models (DEM)
Station — global land	Global Historical Climatology Network-Monthly (GHCN-M)
Station — gridded — U.S. land	National Climate Division (nClimDiv)
Satellite — global ocean	NOAA Optimal Interpolation Sea Surface Temperature (OI SST) CDR
Physical Records — in situ — global land	Local Climatology Monthly Summaries
Paleoclimatology — global land	International Tree-Ring Data Bank (ITRDB)

The goals of this pilot use case study are to:

1. demonstrate the utility of DSMM and evaluate its appropriateness and completeness over various NCEI data types;
2. establish a stewardship maturity baseline for selected NCEI high-utility datasets;
3. identify the area(s) of strength and weakness of stewardship practices applied to the datasets for decision-making support;
4. provide product users with a consolidated and consistent document for content-rich stewardship practice quality information and provide DSMM users, including data managers, with examples of maturity ratings and justifications of stewardship practices information;

5. assess the roles and knowledge required of the Integrated Product Team (IPT) members for effective stewardship maturity evaluation of individual datasets;
6. identify standards used for different data types, assess the need for defining a set of core data types, and define core data types, if needed, for consistent and scalable implementation; and
7. explore requirements for tool(s) to assess and display the current maturity rating, including how to define and display a roadmap for improvement in a systematic and easy-to-understand way.

In this paper, we describe results from a case study of applying the DSMM to the version 3 monthly land surface temperature data product derived from the Global Historical Climatology Network (hereinafter referred to as GHCN-M). This dataset was selected because of its importance to national and international climate monitoring and assessment activities. Since the final two goals of the seven listed above require the completion of a use case study of all datasets listed in Table I, along with additional system and software engineering requirements analysis which is clearly beyond the scope of this paper, we will only touch on the first five goals.

2 Why do we start with GHCN-M?

GHCN is an integrated database of climate summaries from land surface stations across the globe. Since the GHCN-M dataset was first released in the 1990s (Vose *et al.*, 1992), it has been widely utilized and has provided the foundation for understanding trends and variability in global and regional temperatures. It provides data for ongoing monitoring of the global climate and makes it possible to place current conditions in historical perspective (e.g., most recent State of the Climate report by NCEI, see Blunden and Arndt, 2016). It is used in national and international climate assessments (e.g., Karl *et al.*, 2009; IPCC, 2013; Melillo *et al.*, 2014) to understand how rapidly the Earth's climate is varying and changing in association with natural and anthropogenic influences. It also is a source of information for users in the private sector for understanding local and regional climate conditions.

GHCN-M version 3 has gone through the NCEI managed archival process and has been available to the general public since April 2011 (Lawrimore *et al.*, 2011), with some modifications (Gleason *et al.*, 2015). Enhancements to the dataset continue to be made, with version 4 currently under development and expected to be released in 2016. This makes the GHCN-M product an ideal candidate for assessing the maturity of stewardship practices applied to the data, thereby establishing a baseline of the maturity ratings of the current version and identifying potential areas of improvement for future versions of the product.

3 What we have found

3.1 Stewardship maturity ratings

The maturity of GHCN-M v3 has been assessed utilizing the DSMM assessment template (Peng, 2015) which can be freely downloaded. Results are displayed in Figures 3 and 4 below, namely, stewardship maturity ratings diagram and scoreboard, respectively. They are developed as a standard set of DSMM graphics. Although both diagrams present essentially the same maturity information, the rating diagram (e.g., Figure 3) provides a simple and clear view of the current ratings while the DSMM scoreboard (e.g., Figure 4) provides a high-level overview, allowing users to dive in for more in-depth DSMM definitions. After carrying out a maturity assessment of an individual dataset utilizing the DSMM assessment template, it is recommended to create a DSMM report consisting of DSMM assessment metadata, maturity ratings, and justifications (e.g., as shown in *Appendix I: DSMM Document for the Global Historical Climatology Network-Monthly (GHCN-M) Version 3 Dataset*) and DSMM graphics (e.g., as shown in Figures 3 and 4). Effort is underway as a part of the NOAA OneStop Project to systematically and automatically generate DSMM reports, publish those reports and integrate DSMM assessment information including stewardship maturity ratings into ISO collection-level metadata records.

The current maturity ratings of GHCN-M v3 are at Level 2 or higher for all nine key components. Specifically, there are four Level 2, three Level 3, and two Level 4 key components.

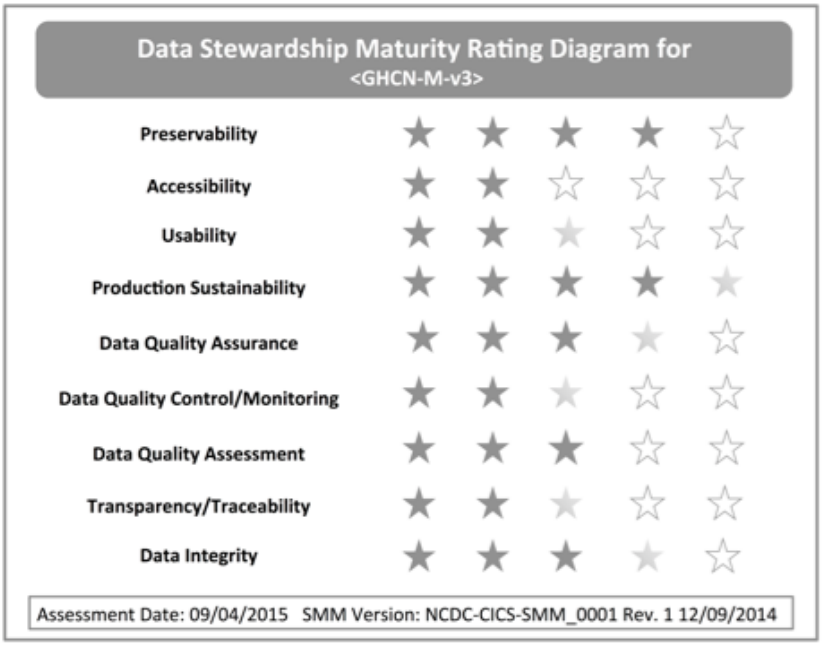

Figure 3: Data stewardship maturity rating diagram of GHCN-Monthly, v3. The dark filled stars indicate that all the practices are completely satisfied. The light filled ones indicated that not all the practices are satisfied. And the non-filled ones indicated that the practices are not satisfied.

SMM Document ID: NCDC-CICS-SMM_0001
Version: Rev. 1. 12/09/2014

Global Historical Climatology Network-Monthly

Data Stewardship Maturity Scoreboard

Maturity Level as of 09/04/2015

Maturity Scale	Preservability	Accessibility	Usability	Production Sustainability	Data Quality Assurance	Data Quality Control/Monitoring	Data Quality Assessment	Transparency / Traceability	Data Integrity
Level 1 – Ad Hoc Not Managed	Any storage location Data only	Not publicly available Person-to-person	Extensive product-specific knowledge required No documentation online	Ad Hoc or Not applicable No obligation or deliverable requirement	Data quality assurance (DQA) procedure unknown or none	None or Sampling unknown or spotty Analysis unknown or random in time	Algorithm/method/ model theoretical basis assessed (method and results online)	Limited product information available Person-to-person	Unknown or no data ingest integrity check
Level 2 - Minimal Managed Limited	Non-designated repository Redundancy Limited archiving metadata	Publicly available Direct file download (e.g., via anonymous FTP server) Collection/dataset level searchable	Non-standard data format Limited documentation (e.g., user's guide) online	Short-term Individual PI's commitment (grant obligations)	Ad Hoc and random DQA procedure not defined and documented	Sampling and analysis are regular in time and space Limited product-specific metrics defined & implemented	Level 1 + Research product assessed (method and results online)	Product information available in literature	Data ingest integrity verifiable (e.g., checksum technology)
Level 3 - Intermediate Managed Defined, Partially Implemented	Designated archive Redundancy Community-standard archiving metadata Conforming to limited archiving process standards	Level 2 + Non-standard data service Limited data server performance Granule/file level searchable Limited search metrics	Community Standard-based interoperable format & metadata Documentation (e.g., source code, product algorithm document, processing or/and data flow diagram) online	Medium-term Institutional commitment (contractual deliverables with specs and schedule defined)	DQA procedure defined and documented and partially implemented	Level 2 + Sampling and analysis are frequent and systematic but not automatic Community metrics defined and partially implemented Procedure documented and available online	Level 2 + Operational product assessed (method and results online)	Algorithm Theoretical Basis Document (ATBD) & source code online Dataset configuration managed (CM) Unique Object Identifier (OID) assigned (dataset, documentation, source code) Data citation tracked (e.g., utilizing Digital Object Identifier (DOI) system)	Level 2 + Data archive integrity verifiable
Level 4 - Advanced Managed Well-Defined, Fully Implemented	Level 3 + Conforming to community archiving standards	Level 3 + Community-standard data services Enhanced data server performance Conforming to community search metrics Dissemination report metrics defined and implemented internally	Level 3 + Basic capability (e.g., subsetting, aggregating) & data characterization (overall/global, e.g., climatology, error estimates) available online	Long-term Institutional commitment Product improvement process in place	DQA procedure well documented, fully implemented and available online with master reference data Limited data quality assurance metadata	Level 3 + Anomaly detection procedure well-documented and fully implemented using community metrics, automatic, tracked and reported Limited quality monitoring metadata	Level 3 + Quality metadata assessed (method and results online) Limited quality assessment metadata	Level 3 + Operational Algorithm Description (OAD) online, OID assigned, and under CM	Level 3 + Data access integrity verifiable Conforming to community data integrity technology standard
Level 5 - Optimal Level 4 + Measured , Controlled , Audit	Level 4 + Archiving process performance controlled, measured, and audited Future archiving standard changes planned	Level 4 + Dissemination reports available online Future technology and standard changes planned	Level 4 + Enhanced online capability (e.g., visualization, multiple data formats) Community metrics of data characterization (regional/cell) online External ranking	Level 4 + National or international commitment Changes for technology planned	Level 4 + DQA procedure monitored and reported Conforming to community quality metadata & standards External review	Level 4 + Cross-validation of temporal & spatial characteristics Physical consistency check Conforming to community quality metadata & standards Dynamic providers/users feedback in place	Level 4 + Assessment performed on a recurring basis Conforming to community quality metadata & standards External ranking	Level 4 + System information online Complete data provenance available online	Level 4 + Data authenticity verifiable (e.g., data signature technology) Performance of data integrity check monitored and reported

Dataset Information: **http://dx.doi.org/10.7289/V5X34VDR**
Dataset POC: **Jay Lawrimore; Jay.Lawrimore@noaa.gov**

SMM POC: **Ge Peng; Ge.Peng@noaa.gov**
SMM Assessment POC: **Jay Lawrimore; Jay.Lawrimore@noaa.gov**

Figure 4: Data stewardship maturity scoreboard of GHCN-Monthly, v3. If two cells in the same column are filled, it is an indication that only a partial maturity rating at the higher level is satisfied. The scoreboard is only intended to provide a high-level overview of ratings.

[View a higher resolution version of Figure 4.]

The Level 4 key components are Preservability and Production Sustainability. The high rating in Preservability stems from multiple facets. NCEI has a well-defined and managed archival process that follows the ISO Open Archival Information System Reference Model (OAIS RM) (CCSDS, 2012). NCEI also complies with archive standards set by the National Archives and Records Administration (NARA).

Level 4 Production Sustainability is the result of long-term institutional and international commitment, along with the product improvement process that is in place (Figure 4; the DSMM document for GHCN-M, containing stewardship maturity assessment metadata, ratings and detailed justifications, is included in the Appendix).

The key components with a maturity rating of Level 3 include Data Quality Assurance and Data Quality Assessment (Figure 4). Those ratings are the result of following community best practices in quality assurance by the data provider and product quality being assessed by the data provider and other independent users. The standardized practices associated with data integrity validation, namely, verifying file checksum during data transfer, ingest, archive, and retrieval, have also resulted in a Level 3.5 maturity rating in Data Integrity.

GHCN-M does meet the Level 2 stage of maturity in key components of Accessibility, Usability, Data Quality Control/Monitoring, and Transparency/Traceability. It achieves this by following the archive process as defined by NCEI, which adheres to community, national, and international standards for data stewardship. For example, an identifier unique to NCEI has been assigned and a digital object identifier (DOI) has been minted with a corresponding landing page (See http://dx.doi.org/10.7289/V5X34VDR). In addition, the dataset has been documented with ISO collection-level metadata. However, the ratings for those components fall short of the recommended Level 3 maturity ratings for the NCEI operational digital datasets, largely due to a lack of publicly accessible information that meets the standard for Level 3. Details regarding methods for improving these three areas are provided in the following section.

3.2 Recommendations for Improvement

Figure 5 summarizes actionable steps for a roadmap forward in achieving Level 3 ratings in all nine key components, based on DSMM. They are described in details below.

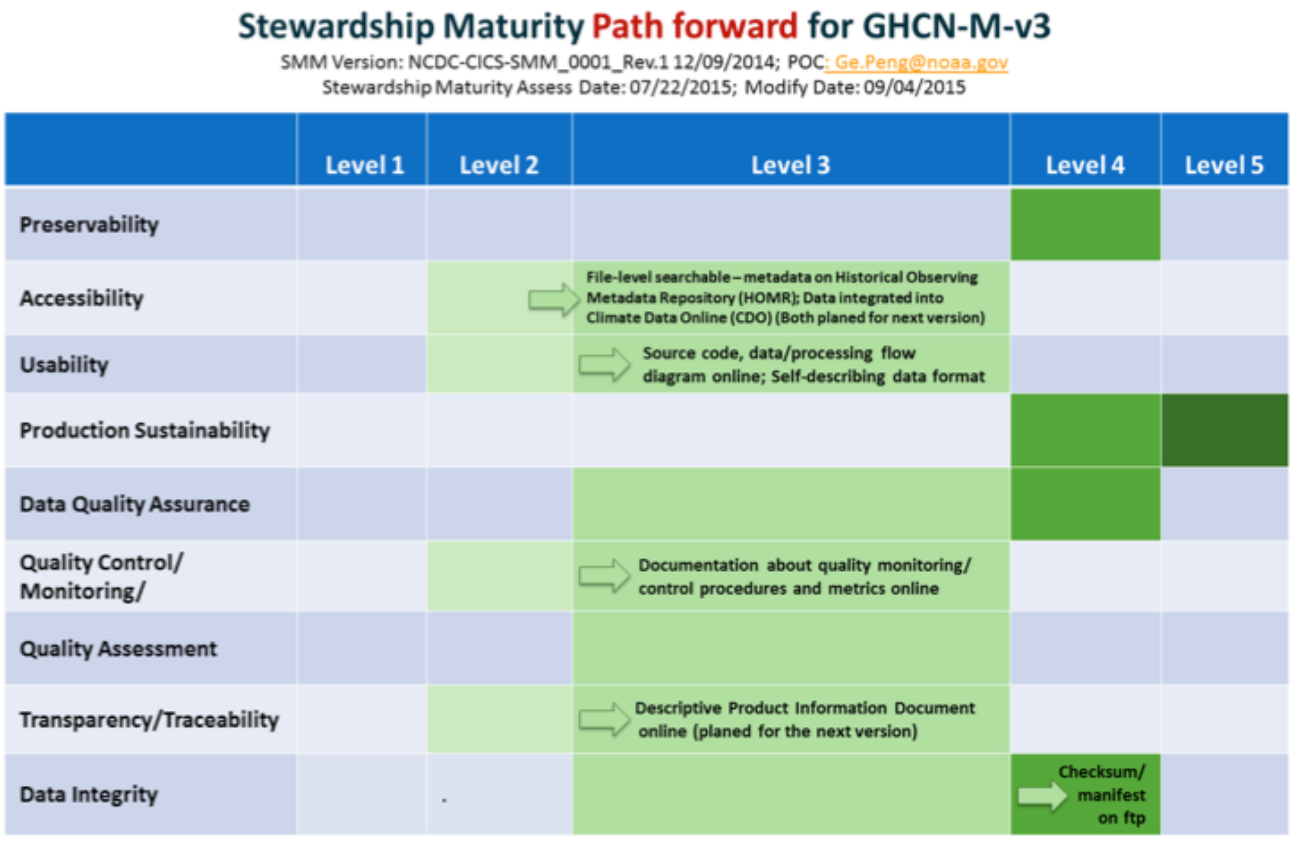

Figure 5: Diagram of path forward for improving the stewardship maturity of GHCN Monthly data product to Level 3 or higher, recommended ratings for NOAA's high-utility and high-impact digital environmental and geospatial datasets.

1. Documenting data quality monitoring practices and making them available to users online

The product quality is monitored regularly with flags and metrics online at the product site. Quality monitoring metrics are consistent with the in situ community. Manual reviews of automatically generated plots or statistics are conducted regularly on a monthly basis. However, the procedure is not documented. Recommendation is to document the data quality monitoring procedure and practices and make them available online, which will be adopted for the next version and captured in the descriptive product information document. Making this document publicly available will also improve the maturity rating of Transparency/Traceability.

2. Improving data searchability and accessibility by serving data with a data server or web service

Currently, only collection/product-level data is searchable with direct file download. There is no additional capability for discovering and serving data. However, there are plans to include file-level metadata in the Historical Observing Metadata Repository (HOMR), allowing for additional station-specific provenance information. The product will also be provided to users with enhanced searchability and accessibility features via the NCEI Climate Data Online (CDO) portal, which is a community-standard-based data service system. Once these changes are implemented, it will improve the maturity rating in Accessibility to Level 3. This should allow users to effectively and efficiently find and use data based on their unique needs.

3. Making the checksum available for each of data file on the ftp server

A checksum is a character string that represents the sum of the correct digits in a data file. It is used to ensure the integrity of a file, especially during transmission or storage. Upon reviewing the best practices defined for Data Integrity, the Access Specialist pointed out that it would require only a minimal effort to make the checksum available for each GHCN-M data file on the ftp server, because the checksums are already retrieved when data files are pulled from the archive and staged for access. Therefore, although not required for the minimum stewardship maturity requirement for Data Integrity, by making the checksum file available, the user can verify the integrity of the downloaded data file. By doing this the Data Integrity rating will reach Level 4.

4. Improving data usability by adopting a more scalable data format

Although it can be argued that the ASCII data format is well-utilized and accepted by the in situ data community, extra steps are recommended to improve data usability:

- self-describing by adding additional station and product information to the ASCII data files, or convert to the JSON (JavaScript Object Notation) data format. JSON is an open standard and language-independent format that uses human-readable text to transmit data objects consisting of attribute-value pairs (Crockford, 2009).
- using a standard-based machine-independent and scalable format, such as Network Common Data Form (NetCDF) compliant with CF (Climate and Forecast) metadata conventions.

3.3 Roles and required knowledge

As a part of this use case study, we have examined the appropriateness of the current roles for stewardship maturity assessment of a data product. For this particular purpose, initial self-assessments of the GHCN-M stewardship maturity were separately carried out by each member of the team — Archive Specialist, Data Manager, Dataset Subject Matter Expert (SME), and Access Specialist (see Table 2 for roles and their NCEI affiliation).

Table 2: Attributes of GHCN-M Stewardship Maturity Assessment Team

Role	*NCEI Affiliation*
Archive Specialist	DSD/Archive Branch (AB)
Data Manager	DSD/AB
Dataset SME	CWC/Climate Science Branch (CSB)
Access Specialist	DSD/Data Access Branch (DAB)
DSMM SME/Co-Lead	CWC/CSB

Project Lead/Key Sponsor	DSD/AB

The self-assessment results and underlying knowledge base of each team member are observed by the DSMM SME and recapped below:

- The Archive Specialist is very familiar with the NCEI data archiving process, including metadata/documentation standards and procedures, but is not clear about where to get information for Usability and three data quality related key components.
- The Data Manager has a general knowledge of the data archiving process and the dataset, and is willing, and knows where or who to go to, to get additional information if not readily available.
- The Dataset SME has extensive knowledge of data quality (DQ) assurance, control/monitoring, and assessment, of how the dataset is being served to users, and some knowledge of the data archive process.
- The Access Specialist has extensive knowledge in how the data files are staged and integrated to (or lack of) a data server or web service, has first-hand experience in integrating the collection-level metadata into the NCEI geoportal, and knows the NCEI collection-level metadata ID.

Unlike many Dataset SMEs, the GHCN-M Dataset SME was very familiar with NCEI data stewardship practices. With the knowledge gained going through the NCEI archive process for GHCN-M, and some basic training with the NCEI metadata creation tool, the Dataset SME was able to provide input for all key components, including Preservability. A large portion of product-specific information, such as data quality assurance and control related practices, are available at the NCEI product web site and in the literature, and the Dataset SME was well aware of them. The Data Manager was also able to provide input for all but one of the key components, leveraging personal knowledge on NCEI archival process and on the product, as well from information available online. Product-specific data quality information was considered as not readily available or known to both the Archive and Access Specialists.

Rating assessment input from different team members and primary information sources are summarized in Figure 6a while rating spread for each key component is shown in Figure 6b.

(a) Rating Input & Knowledge Sources

	Archive Specialist	Access Specialist	Dataset SME	Data Manager	Primary Knowledge Source
Preservability	X		X	X	Preservation-specific knowledge; Document/standard Associated with defined & managed process
Accessibility		X	X	X	Product website; Search/access-specific technology knowledge – integration to existing systems; Web search
Usability		X	X	X	Product website/community-specific best practices and standard
Production Sustainability			X	X	Product-specific knowledge
Data Quality Assurance			X	X	Product website/domain-specific best practices; Literature
Data Quality Control/ Monitoring			X		Product website & product-specific knowledge
Data Quality Assessment			X	X	Product website; Literature
Transparency/ Traceability			X	X	Product website; Literature; practices associated with defined process
Data Integrity	X	X	X	X	Preservation/access-specific practices with defined & managed process

(b) Rating Input Spread

	1	2	3	4	5	Min	Max	# Entry
Preservability				4		3	4	3
Accessibility		2				2	3	3
Usability		2.5				2	3.5	3
Production Sustainability				4.5		4	4.5	2
Data Quality Assurance			3.5			3	3.5	2
Data Quality Control/Monitoring		2.5				2.5	2.5	1
Data Quality Assessment			3			2	4	2
Transparency/ Traceability		2.5				2.5	4	2
Data Integrity			3.5			3.5	4	4

Figure 6: Diagram showing (a) GHCN-M stewardship maturity rating input from the team members and primary knowledge sources and (b) rating input spread.

It is easy to see that the stewardship maturity evaluation of individual datasets requires knowledge from multiple disciplines. Currently product-specific data quality information is not always publicly available. If they are, they are not always defined and captured in a consistent way. Therefore, product-specific quality information often needs to be inferred or derived from literature or online sources, which is clearly beyond the task scope of the Archive and Access Specialists. Because in most cases the Archive Specialist should be the person carrying out the DSMM assessment of individual datasets, it would be beneficial to document the data quality practices in a consistent way and make them readily available and easy for an Archive Specialist to understand. Until then, a team approach is the most effective way to carry out stewardship maturity assessment. Additional description and discussion can be found in Section 4.2.

During this case study, a number of gaps in knowledge and systems were uncovered, and some of them will be described in the next section within the context of improved GHCN-M stewardship maturity. Although it pertains to this use case only, it could be beneficial to others who plan to apply DSMM to their dataset or to define system requirements for a consistent way of assessing any publicly available digital environmental and geospatial datasets.

4 Lessons learned

4.1 Easy does not mean simple

Providing a consistent framework is only the first step in providing a consistent stewardship maturity measure to users, which is also confirmed by the DSMM use case study carried out by the Data Stewardship Committee of the Federation of Earth Science Information Partners (ESIP) (Hou *et al.*, 2015). A consistent and scalable way of applying the DSMM is still required, and additional steps or adjustments may be needed to systematically assess maturity of individual datasets.

4.2 Effective stewardship maturity assessment requires the knowledge of practices in multiple disciplines

Peng *et al.* (2016) pointed out that effective long-term scientific data stewardship requires knowledge and oversight from multiple domains. This is also true for effective evaluation of stewardship maturity. Descriptions of individual knowledge domains and practices required for stewardship maturity assessment are provided below:

- Data preservation — archive and metadata standards and data management practices, including those for data integrity and transparency
- Scientific stewardship and documentation — data quality management practices (e.g., data quality assurance practices, data quality control/monitoring practices), product usability (e.g., data characteristics — climatology and variability; uncertainty estimates, etc.), and traceability
- Tools and systems — data access, data integrity, and data interoperability (technology)

4.3 Well-defined processes and documents are beneficial

Well-defined processes and documentation, along with general knowledge of the existence of those documents and the information they capture, help facilitate the stewardship assessment process. Examples include the NCEI-defined and implemented archiving process, the NCEI-defined Submission Agreement (SA) template, and a consistent web-based user interface for collecting information about data collection (e.g., ATRAC (Advanced Tracking and Resource tool for Archive Collections) tool).

4.4 Other potential improvement areas

One of the potential improvement areas uncovered during this use case study, which pertains to NCEI processes, is that integration of cross-center processes and procedures in various parts of OAIS RM (and NCEI) can be improved. An example regarding the DOI landing page and product web page is provided below.

- A DOI for the data product is assigned and its landing page is created by the Archive Branch (archive) but it is not discoverable by performing a Google search. This DOI landing page is based on an ISO 19115-2 compliant collection-level metadata record, with a layout that is defined for all NCEI managed datasets. It provides dataset citation, however, there are three different dataset identifiers listed at the landing page without any description of what they are. (They are, in fact, the DOI with the link to the landing page, NCEI dataset Identifier (9100_03), and the collection-level metadata Identifier (C00839).)
- The Archive Specialist is aware that a DOI is assigned and minted and knows how to get to the DOI landing page.
- The Data SME is not sure if a DOI is issued. The Data Manager believes that it is issued but cannot find the landing page online.
- The NCEI product web site is created and overseen by the Access Branch (access). It provides valuable information on quality assurance and homogeneity adjustment. It includes the "Data Access" tab that takes user to the ftp sites for the GHCN-M data files and additional station-specific information and plots.
- The DOI landing page provides a link to the product web site under the "Access" tab. However, there is no link from the product page to the DOI landing page. Furthermore, there is no "Metadata" tab in the current NCEI GHCN-Monthly product web page.
- The Access Specialist knows the NCEI collection-level metadata ID, and its integration to geoportal (dissemination), which leads to the DOI landing page.

Under the current NCEI data archiving process, a DOI is assigned and minted after the dataset has gone through the archiving process and has passed the archive readiness review. The GHCN-M DOI landing page is not currently discoverable by search engines like Google. Since a DOI is a persistent, resolvable, and trackable identifier, it would be better for users to use this as a primary gateway to the product. Therefore, we recommend to improve discoverability of the DOI landing page. As information at the NCEI DOI landing page is based on the ISO standard-based collection-level metadata and contains information about the dataset such as spatial and temporal coverage, data access, etc., we recommended adding a "Metadata" tab at the product web page. This "Metadata" tab should be linked to the DOI landing page to provide users with a direct entry to consistent and standard-based product metadata and also to provide the data producer with a way of estimating data usage and impact by having a persistent and trackable product citation.

5 Summary

The stewardship maturity of a highly-utilized NCEI data product, GHCN-M, is assessed based on a reference stewardship maturity framework. The current maturity ratings of GHCN-M v3 are at Level 2 or higher for all nine key components with four Level 2, three Level 3, and two Level 4 key components (Figure 3).

Well-defined and managed processes following OAIS RM are found to be beneficial, as expected, in ensuring consistency in maturity of all managed data holdings. It has contributed to Level 3 stewardship maturity ratings for GHCN-M in both Preservability and Data Integrity. Consistent metadata and documentation are not only beneficial for system integration but also good resources for the IPT members, although even more beneficial would be additional training or communication about processes and resultant metadata and documentation.

The web-based tool for creating collection-level metadata has improved accessibility by making it easy to integrate this information with other NOAA resources, such as NCEI Geoportal and the NOAA catalog, thereby enhancing searchability of the product. However, better integration between information about the dataset, namely, the collection-level metadata via the DOI landing page, and information about the product via the product web page will provide characterization and quality information about the dataset to users in a seamless and integrated way, which will in turn enhance the usability of the dataset.

Potential improvement is identified in the areas of Accessibility, Usability, Data Quality Control/Monitoring, and Transparency/Traceability. Recommendations for actionable stewardship practices based on the DSMM are outlined to improve the stewardship maturity of the product. For example, currently, the data files are created with and served in the ASCII format, which is still the commonly accepted file exchange format for the in situ data community. However, we have recommended providing self-describing ASCII data files for enhanced usability and interoperability. In addition, we encourage providing end-users with a more scalable, machine-independent, and self-describing data format such as NetCDF.

One unexpected benefit of this use case study is that all participants have gained a better understanding of the strengths and weaknesses of the dataset and the organizational capabilities. This knowledge will empower them not only to better carry out their current responsibilities but also to help promote the stewardship best practices going forward.

Acknowledgements

This work is partially supported by the NOAA's Climate Data Record Program, in collaboration with NCEI Data Stewardship Division and Center for Weather and Climate. Management support from those NCEI entities is critical for the initiation and completion of this study. G. Peng is supported by NOAA through Cooperative Institute for Climate and Satellites — North Carolina (CICS-NC) under Cooperative Agreement NA14NES432003. We thank Steve Ansari for beneficial input on data access of GHCN-M. Comments and suggestions from Tom Maycock, Otis Brown, Catherine Rey, and an anonymous NCEI internal reviewer are beneficial in improving the clarity and readability of the paper.

Disclaimer

Any opinions or recommendations expressed in this manuscript are those of the author(s) and do not necessarily reflect the views of NCEI or CICS-NC.

References

[1] Bates, J. J. and J.L. Privette, 2012: A maturity model for assessing the completeness of climate data records. *EOS*, Transactions of the AGU, 44, 441. http://doi.org/10.1029/2012EO440006

[2] Blunden, J. and D. S. Arndt, Eds., 2016: State of the Climate in 2015. *Bull. Meteor. Soc.*, 97 (8), S1-S275, http://doi.org/10.1175/2016BAMSStateoftheClimate.1

[3] CCSDS (The Consultative Committee for Space Data Systems), 2012: Reference model for an open archival information system (OAIS) — Recommendation for Space Data System Practices. Version CCSDS 650.0-M-2 June 2012. 135 pp.

[4] Crockford, D., 2009: Introduction to JSON. *json.org*. May 29, 2009.

[5] Gleason, B., C. Williams, M. Menne, and J. Lawrimore, 2015: Modifications to GHCN-Monthly (version 3.3.0) and USHCN (version 2.5.5) processing systems. *NCEI Technical Report No. GHCNM-15-01*. 23 pp.

[6] Hou, C.Y., M. Mayernik, G. Peng, R. Duerr, and A. Rosati, 2015: Assessing information quality: Use cases for the data stewardship maturity matrix. Poster. AGU 2015 Fall meeting, San Francisco, 14-18 December 2015.

[7] IPCC, 2013: Climate Change 2013: The Physical Science Basis. *Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley, Eds. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1535 pp.

[8] Karl, T. R., J. T. Melillo, and T. C. Peterson, 2009: Global Climate Change Impacts in the United States. T.R. Karl, J.T. Melillo, and T.C. Peterson, Eds. Cambridge University Press, 189 pp.

[9] Lawrimore, J. H., M. J. Menne, B. E. Gleason, C. N. Williams, D. B. Wuertz, R. S. Vose, and J. Rennie, 2011: An overview of the Global Historical Climatology Network monthly mean temperature data set, version 3, *J. Geophys. Res.*, 116, D19121, http://doi.org/10.1029/2011JD016187

[10] Melillo, J. M., T.C. Richmond, and G. W. Yohe, Eds., 2014: Climate Change Impacts in the United States: The Third National Climate Assessment. *U.S. Global Change Research Program*, 842 pp.

[11] NOAA (National Oceanic and Atmospheric Administration), 2008: NOAA procedure for scientific records appraisal and archive approval. 28 pp.

[12] NOAA, 2011: NOAA Environmental Data Management Committee Procedural Directive — NOAA data sharing policy for grants and cooperative agreements. Version 1.0.

[13] NRC (National Research Council), 2007: Environmental data management at NOAA: Archiving, stewardship, and access. 116 pp. *The National Academies Press*, Washington, D.C. http://doi.org/10.17226/12017

[14] OMB (Office of Management and Budget), 2002: Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies. Federal Register, 67(36). OMB Notice February 22, 2002.

[15] OMB, 2013: Open Data Policy — Managing Information as an Asset. Version: OMB Memorandum May 9, 2013.

[16] OSTP (Office of Science and Technology Policy), 2013: Increasing access to the results of federally funded scientific research. Version: OSTP Memorandum February 22, 2013.

[17] Peng, G., 2015: NCDC-CICSNC SDSMM Template. Version: v4.0 06/23/2015. *figshare*. https://doi.org/10.6084/m9.figshare.1211954

[18] Peng, G., J.L. Privette, E.J. Kearns, N.A. Ritchey, and S. Ansari, 2015: A unified framework for measuring stewardship practices applied to digital environmental datasets. *Data Science Journal*, 13. http://doi.org/10.2481/dsj.14-049

[19] Peng, G., N. A. Ritchey, K. S. Casey, E. J. Kearns, J. L. Privette, D. Saunders, P. Jones, T. Maycock, and S. Ansari, 2016: Scientific stewardship in the Open Data and Big Data era — Roles and responsibilities of stewards and other major product stakeholders. *D-Lib Magazine*, 22. http://doi.org/10.1045/may2016-peng

[20] Ramapriyan, H. G. Peng, D. Moroni, and C.-L. Shie, 2016: Ensuring and Improving Information Quality for Earth Science Data and Products — Role of the ESIP Information Quality Cluster. SciDataCon 2016, 11-13 September 2016, Denver, Colorado, USA.

[21] Sullivan, K., 2013: Restoring U.S. leadership in weather forecasting, Part 2. Legislative testimony before members of Subcommittee on Environment, House Committee on Science, Space, and Technology, U.U. House of Representatives on June 26, 2013.

[22] Vose, R.S., R. Heim, R.L. Schmoyer, T.R. Karl, P.M. Steurer, J.K. Eischeid, T.C. Peterson, 1992: The Global Historical Climatology Network: Long-term monthly temperature, precipitation, sea level pressure and station pressure data. http://doi.org/10.3334/CDIAC/cli.ndp041

Appendix

See the *Appendix I: DSMM Document for the Global Historical Climatology Network-Monthly (GHCN-M) Version 3 Dataset* (a separate file) for tables displaying stewardship maturity ratings and detailed justifications for the GHCN-M version 3 dataset.

About the Authors

Ge Peng is a Research Scholar at the Cooperative Institute for Climate and Satellite-North Carolina (CICS-NC) of North Carolina State University and affiliated with the NOAA's National Centers for Environmental Information (NCEI). Dr. Peng holds a Ph. D in meteorology and is experienced in assessing and monitoring quality of Earth Science data products. She has extensive knowledge of scientific data stewardship and experience in working with metadata specialists and software developers. She is currently leading evaluation of NOAA sea ice and surface flux climate data records and application of the NCEI/CICS-NC Scientific Data Stewardship Maturity Matrix.

Jay Lawrimore is Chief of the Dataset Section in the Center for Weather and Climate, part of NOAA's National Centers for Environmental Information. His section is responsible for developing in situ and remotely sensed datasets that serve the public and private sector's need for climate information. These include integrated datasets for use in operations, climate research, and reanalysis activities. He also conducts analyses of observed changes in the Earth's climate and works with the international climate community to provide perspectives that are essential to climate monitoring and assessment activities.

Valerie Toner is an archive and metadata specialist with STG Inc. at the NOAA National Centers for Environmental Information (NCEI) in Asheville. For almost 10 years, she has supported NCEI in various roles, but in more recent years has focused on supporting data preservation and management for various projects in the center. Her current research explores the complexities of managing past, present and future data preservation practices.

Christina Lief is a Physical Scientist and Program Manager at NOAA's National Centers for Environmental Information (NCEI) in Asheville, NC. She is responsible for data management, access and documentation supporting the diverse archives of NCEI environmental data. She is also the Program Manager for the Global Observing Systems Information Center (GOSIC). She holds a B.S. in Geology from the George Washington University and a M.S. in Geographic and Cartographic Sciences from the George Mason University.

Richard Baldwin is the Interim Data Access Branch Chief at NOAA's National Centers for Environmental Information (NCEI). He is responsible for providing access to the archived data holdings. He has over 20 years of experience building and managing data access systems. He holds a M.S. degree in Geophysics.

Nancy Ritchey is the Archive Branch Chief at NOAA's National Centers for Environmental Information (NCEI). She is responsible for preserving NCEI's extensive collection of environmental data for future generations. Nancy has extensive knowledge and experience in digital and physical data management and related standards and leading practices. She's involved in national and international activities related to data preservation and standards. Nancy holds a M.S. degree in Atmospheric Science.

Danny Brinegar is a Physical Scientist at NOAA's National Centers for Environmental Information (NCEI) in Asheville, NC. He is experienced in the NOAA National Operational Model Archive and Distribution System (NOMADS) and currently leads the operations team within the Data Access Branch. He has a Bachelor of Science degree from the University of North Carolina at Asheville, with a double major in Atmospheric Science and Computer Science.

Stephen A. Del Greco is the proprietor of Black Swan Innovations LLC. Black Swan Innovations provides risk management services that focus around climate variability and change. He recently retired from a 30-year career with the federal government with 26 of those years working for NOAA's National Center for Environmental Information (NCEI). Previous NOAA/NCEI positions held by Stephen include Chief of Climate Services and Monitoring Division, Chief of Satellite Services Branch, Chief of Data Processing Branch and Chief of Data Access Branch.