Ensuring and Improving Information Quality for Earth Science Data and Products – Role of the ESIP Information Quality Cluster

H. K. (Rama) Ramapriyan, Science Systems and Applications, Inc. & NASA Goddard Space Flight Center

Ge Peng, North Carolina State University & NOAA National Centers for Environmental Information

David Moroni, Jet Propulsion Laboratory, California Institute of Technology

Chung-Lin Shie, University of Maryland, Baltimore County & NASA Goddard Space Flight Center

September 12, 2016

H. K. Ramapriyan's work was supported by a NASA contract with SSAI. Ge Peng is supported by NOAA under a Cooperative Agreement with NCSU. David Moroni's work is supported by a NASA contract with the JPL, CalTech, Pasadena, CA, Chung-Lin Shie's work was supported by NASA under a Cooperative Agreement with UMBC. Members of IQC are from many different organizations, funded under various contracts. Government sponsorship acknowledged.

Presented at SciDataCon 2016, Denver, CO

Topics

- **ESIP Federation**
- **Information Quality Cluster (IQC)**
  - Definition of Information Quality
  - IQC Objectives
- **“Many Global Players”**
- **Related Activities**
- **ESIP IQC Activities**
- **Conclusion**

Earth Science Information Partners (ESIP)

- **ESIP Federation – established in 1998; currently >180 members**
- **From http://esipfed.org/**
  - **Mission:** To support the networking and data dissemination needs of our members and the global Earth science data community by linking the functional sectors of observation, research, application, education and use of Earth science.
  - **Vision**: To be a leader in promoting the collection, stewardship and use of Earth science data, information and knowledge that is responsive to societal needs.
- **Collaboration Areas – 5 Committees, 3 Working Groups, 16 Clusters**

Information Quality Cluster

- **Vision**
  - Become **internationally recognized** as an **authoritative and responsive information resource** for guiding the implementation of **data quality standards and best practices** of the science data systems, datasets, and data/metadata dissemination services.
- **Closely connected to Data Stewardship Committee**
- **Open membership (as with all Collaboration Areas in ESIP)**

Information Quality

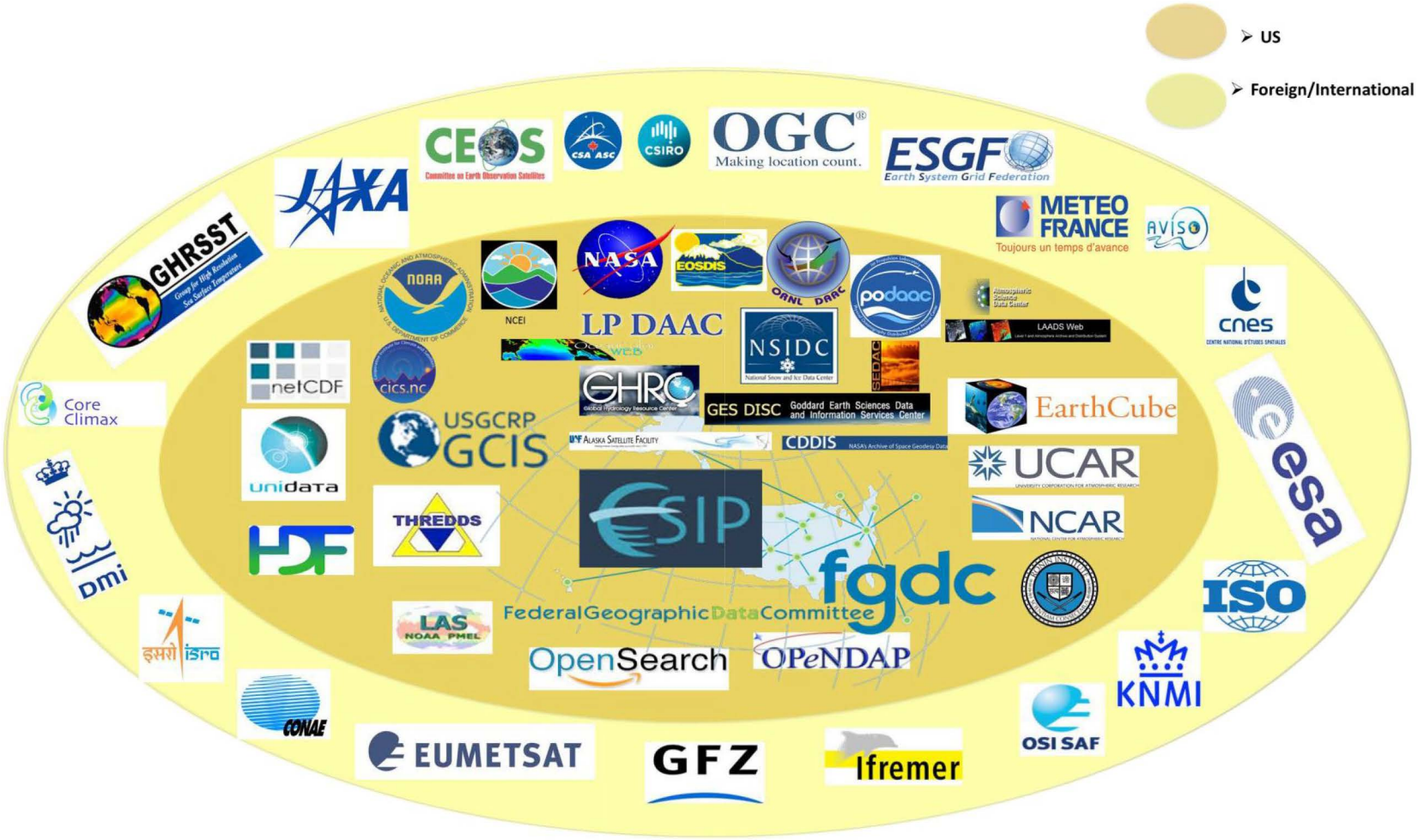
- **Scientific quality**
  - **Accuracy, precision, uncertainty, validity and suitability for use (fitness for purpose) in various applications**
- **Product quality**
  - **how well the scientific quality is assessed and documented**
  - **Completeness of metadata and documentation, provenance and context, etc.**
- **Stewardship quality**
  - **how well data are being managed, preserved, and cared for by an archive or repository**
- **Service Quality**
  - **how easy it is for users to find, get, understand, trust, and use data**
  - **whether archive has people who understand the data available to help users.**

Information Quality is a combination of all of the above

ESIP Information Quality Cluster (IQC) - Objectives

- Share Experiences
- Actively evaluate best practices and standards for data quality from the Earth science community.
- Improve collection, description, discovery, and usability of information about data quality in Earth science data products.
- Support:
  - Data product producers with information about standards and best practices for conveying data quality; provide mentoring as needed
  - Data providers/distributors/intermediaries establish, improve, and evolve mechanisms to assist users in discovering, understanding, and applying data quality information properly.
- Consistently provide guidance to data managers and stewards on the implementation of data quality best practices and standards as well as for enhancing and improving maturity of their datasets.

Many Global Players



Some Related Activities

- NASA ESDSWG Data Quality WG (Recommendations) – 2014-present
- NOAA Dataset Lifecycle Stage based Maturity Matrices – 2009 – present
- Quality Assurance framework for Earth Observation (QA4EO)
- ISO Metadata Quality Standards (19115:2003; 19157:2013; 19158:2012)
- EUMETSAT CORE-CLIMAX Data System Maturity Matrices
- NASA Earth Science Data System Working Groups (ESDSWG) – Metrics Planning and Reporting WG (Product Quality Checklists) – 2010-2012
- GEOSS Data Quality Guidelines and GEO DMP Implementation Guidelines
- CEOS Essential Climate Variables (ECV) Inventory Questions
- NCAR Community Contribution Pages

Covered in subsequent charts

NASA Earth Science Data System Working Groups (ESDSWG) – Data Quality Working Group DQWG

- ***Mission***: "Assess ***existing*** data quality ***standards*** and ***practices*** in the ***inter-agency*** and ***international*** arena to determine a working solution relevant to Earth Science Data and Information System Project (ESDIS), Distributed Active Archive Centers (DAACs), and NASA-funded Data Producers."
- Initiated in March 2014
- 2014-2015:
  - 16 use cases analyzed, issues identified from users' points of view and ~100 recommendations made for improvement
  - Consolidated into 12 high-priority recommendations
- 2015-2016:
  - Extracted 4 "Low Hanging Fruit" (LHF) recommendations from previous 12
  - 25 solutions to address these recommendations have been identified and assessed for operational maturity and readiness for implementation, with an initial focus on four "low-hanging fruit" recommendations; solutions that exist as open-source and in an operational environment were ranked as highest priority for implementation.
- Details in poster by Yaxing Wei et al

Data Stewardship Maturity Matrix

- **NOAA NCEI/CICS-NC Scientific Data Stewardship Maturity Matrix (DSMM) provides a unified framework for assessing the maturity of measurable stewardship practices applied to individual digital Earth Science datasets that are publicly available**
- **Assesses maturity in 9 categories (e.g., preservability, accessibility, data quality assessment, data integrity) at 5 levels (1 = Not Managed; 5 = Optimally Managed)**
- **Provides understandable data quality information to users including scientists and actionable information to decision-makers**
- **Peng, G. et al, 2015. "A unified framework for measuring stewardship practices applied to digital environmental datasets", Data Science Journal, 13. doi:10.2481/dsj.14-04**

(Self-assessment template: tinyurl.com/DSMMtemplate)

QA4EO

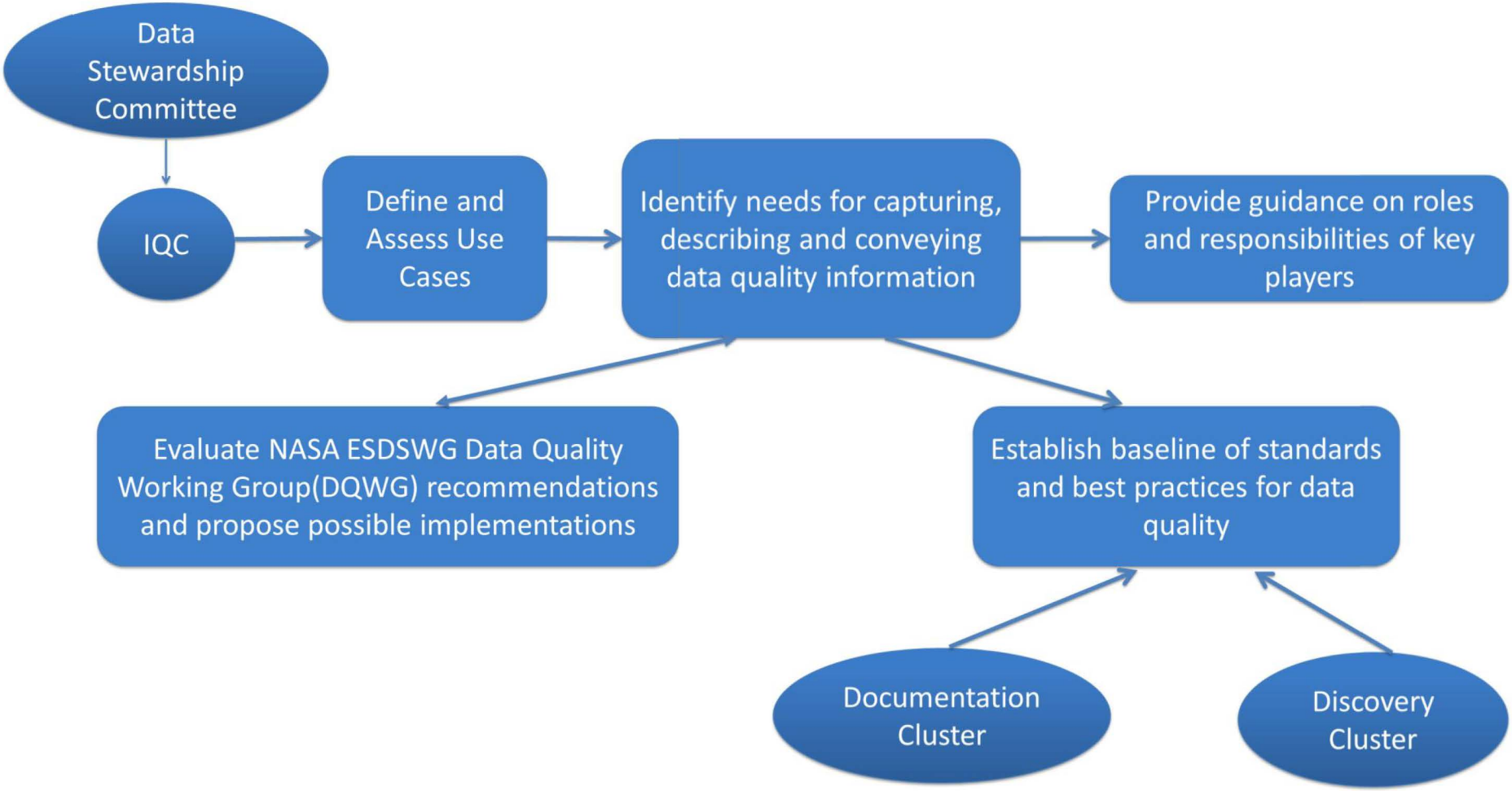
- **Established and endorsed by the Committee on Earth Observation Satellites (CEOS) in response to a Group on Earth Observations (GEO) Task DA-06-02 (now Task DA-09-01a)**
- **Four International Workshops - 2007, 2008, 2009, and 2011**
- **Key Principles (from http://qa4eo.org/docs/QA4EO_guide.pdf)**
  - **"In order to achieve the vision of GEOSS, Quality Indicators (QIs) should be ascribed to data and products, at each stage of the data processing chain - from collection and processing to delivery**
  - **A QI should provide sufficient information to allow all users to readily evaluate a product's suitability for their particular application, i.e. its "fitness for purpose".**
  - **To ensure that this process is internationally harmonized and consistent, the QI needs to be based on a documented and quantifiable assessment of evidence demonstrating the level of traceability to internationally agreed (where possible SI) reference standards."**
- **Framework and 10 Key Guidelines established (e.g., establish Quality Indicator, establish measurement equivalence, expression of uncertainty)**
- **A few cases studies are available that illustrate QA4EO-compliant methodologies [e.g., NOAA Maturity Matrix for CDRs, WELD: Web - Enabled Landsat Data (NASA-funded MEaSUREs Project), ESA Sentinel-2 Radiometric Uncertainty Tool]**

ISO 19157:2013 - Geographic information -- Data quality*

- **Establishes principles for describing the quality of geographic data**
  - **Defines components for describing data quality**
  - **Specifies components and content structure of a register for data quality measures**
  - **Describes general procedures for evaluating the quality of geographic data**
  - **Establishes principles for reporting data quality**
- **Defines a set of data quality measures for use in evaluating and reporting data quality**
- **Applicable to data producers providing quality information to describe and assess how well a data set conforms to its product specification**
- **Applicable to data users attempting to determine whether or not specific geographic data are of sufficient quality for their particular application**
- **Examples of DQ Elements: Completeness, Thematic Accuracy, Logical Consistency, Temporal Quality, Positional Accuracy**

* From: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=32575

ESIP IQC Activities



Conclusion

- **Capture, description, discovery, and usability of information about data quality in Earth science data products is critical for proper use of data**
- **Many groups are involved in developing and documenting best practices**
- **ESIP Information Quality Cluster, as a multilateral group is well placed to promoting standards and best practices**
- **Membership in IQC is open – you are invited to participate!**

Thank you for your attention!

Hampapuram.ramapriyan@ssaihq.com
David.F.Moroni@jpl.nasa.gov
Ge.Peng@noaa.gov
chung-lin.shie-1@nasa.gov

ESIP Information Quality Cluster Activities

- Coordinate use case studies with broad and diverse applications, collaborating with the ESIP Data Stewardship Committee and various national and international programs
- Identify additional needs for consistently capturing, describing, and conveying quality information
- Establish and provide community-wide guidance on roles and responsibilities of key players and stakeholders including users and management
- Prototype innovative ways of conveying quality information to users
- Evaluate NASA ESDSWG DQWG recommendations and propose possible implementations.
- Establish a baseline of standards and best practices for data quality, collaborating with the ESIP Documentation Cluster and Earth Science agencies.
- Engage data provider, data managers, and data user communities as resources to improve our standards and best practices.

NASA MEaSUREs - Product Quality Checklists

- **Making Earth System Data Records for Use in Research Environments (MEaSUREs)**
- **NASA-funded, typically 5-year projects generating long-term consistent time series**
- **Product Quality Checklists (PQC) indicate completeness of Quality Assessment, metadata, documentation, etc.**
- **PQC templates - developed in 2011 and adopted in 2012**
- **Questions asked address science quality, documentation quality, usage and user satisfaction**

NCAR Climate Data Guide*

- **Community contributed datasets, reviews**
- **Focuses on "limited selection of data sets that are most useful for large-scale climate research and model evaluation"**
- **Contributed reviews answer 10 key questions; Examples of topics addressed**
  - **strengths, limitations, and typical applications of datasets**
  - **Comparable datasets**
  - **Methods of uncertainty characterization**
  - **utility for climate research and model evaluation.**

*From Schneider, D. P., et al (2013), Climate Data Guide Spurs Discovery and Understanding, Eos Trans. AGU, 94(13), 121. [article] - See more at: https://climatedataguide.ucar.edu/about/contribute-climate-data-guide#sthash.zaOUYP3j.dpuf

CDR Maturity Matrix

- **NOAA NCEI Climate Data Record (CDR) Maturity Matrix assesses readiness of a product as a NOAA satellite CDR**
- **Bates, J. J. and Privette, J. L., "A Maturity Model for Assessing the Completeness of Climate Data Records", Eos, Vol. 93, No. 44, 30 October 2012**
- **Assesses maturity in 6 categories (software readiness, metadata, documentation, product validation, public access, utility) at 6 levels**
- **Provides consistent guidance to data producers for improved data quality and long-term preservation**
- **EUMETSAT's CORE-CLIMAX Matrix – based on CDR Maturity Matrix; contains guidance on uncertainty measures**
- **http://www1.ncdc.noaa.gov/pub/data/sds/cdr/Guidelines/Maturity_Matrix_Template.xlsx**

NOAA CDR Maturity Matrix

Maturity	Software Readiness	Metadata	Documentation	Product Validation	Public Access	Utility
1	Conceptual development	Little or none	Draft Climate Algorithm Theoretical Basis Document (C-ATBD); paper on algorithm submitted	Little or None	Restricted to a select few	Little or none
2	Significant code changes expected	Research grade	C-ATBD Version 1+ ; paper on algorithm reviewed	Minimal	Limited data availability to develop familiarity	Limited or ongoing
3	Moderate code changes expected	Research grade; Meets int'l standards: ISO or FGDC for collection; netCDF for file	Public C-ATBD; Peer-reviewed publication on algorithm	Uncertainty estimated for select locations/times	Data and source code archived and available; caveats required for use.	Assessments have demonstrated positive value.
4	Some code changes expected	Exists at file and collection level. Stable. Allows provenance tracking and reproducibility of dataset. Meets international standards for dataset	Public C-ATBD; Draft Operational Algorithm Description (OAD); Peer-reviewed publication on algorithm; paper on product submitted	Uncertainty estimated over widely distributed times/location by multiple investigators; Differences understood.	Data and source code archived and publicly available; uncertainty estimates provided; Known issues public	May be used in applications; assessments demonstrating positive value.
5	Minimal code changes expected; Stable, portable and reproducible	Complete at file and collection level. Stable. Allows provenance tracking and reproducibility of dataset. Meets international standards for dataset	Public C-ATBD, Review version of OAD, Peer-reviewed publications on algorithm and product	Consistent uncertainties estimated over most environmental conditions by multiple investigators	Record is archived and publicly available with associated uncertainty estimate; Known issues public. Periodically updated	May be used in applications by other investigators; assessments demonstrating positive value
6	No code changes expected; Stable and reproducible; portable and operationally efficient	Updated and complete at file and collection level. Stable. Allows provenance tracking and reproducibility of dataset. Meets current international standards for dataset	Public C-ATBD and OAD; Multiple peer-reviewed publications on algortihm and product	Observation strategy designed to reveal systematic errors through independent cross-checks, open inspection, and continuous interrogation; quantified errors	Record is publicly available from Long-Term archive; Regularly updated	Used in published applications; may be used by industry; assessments demonstrating positive value

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

Dataset Name

Maturity Level as of mm/dd/yyyy

Stewardship Maturity Matrix for Digital Environmental Data Products

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: URL Goes Here
Dataset POC: Name & E-mail Here

SMM POC: Ge.Peng@noaa.gov
SMM Assessment POC: Name & E-mail Here