ESIP Information Quality Cluster

Hampapuram Ramapriyan[1,2] (Hampapuram.Ramapriyan@ssaihq.com), Ge Peng[3,4] (Ge.Peng@noaa.gov) and David Moroni[5] (David.F.Moroni@jpl.nasa.gov)

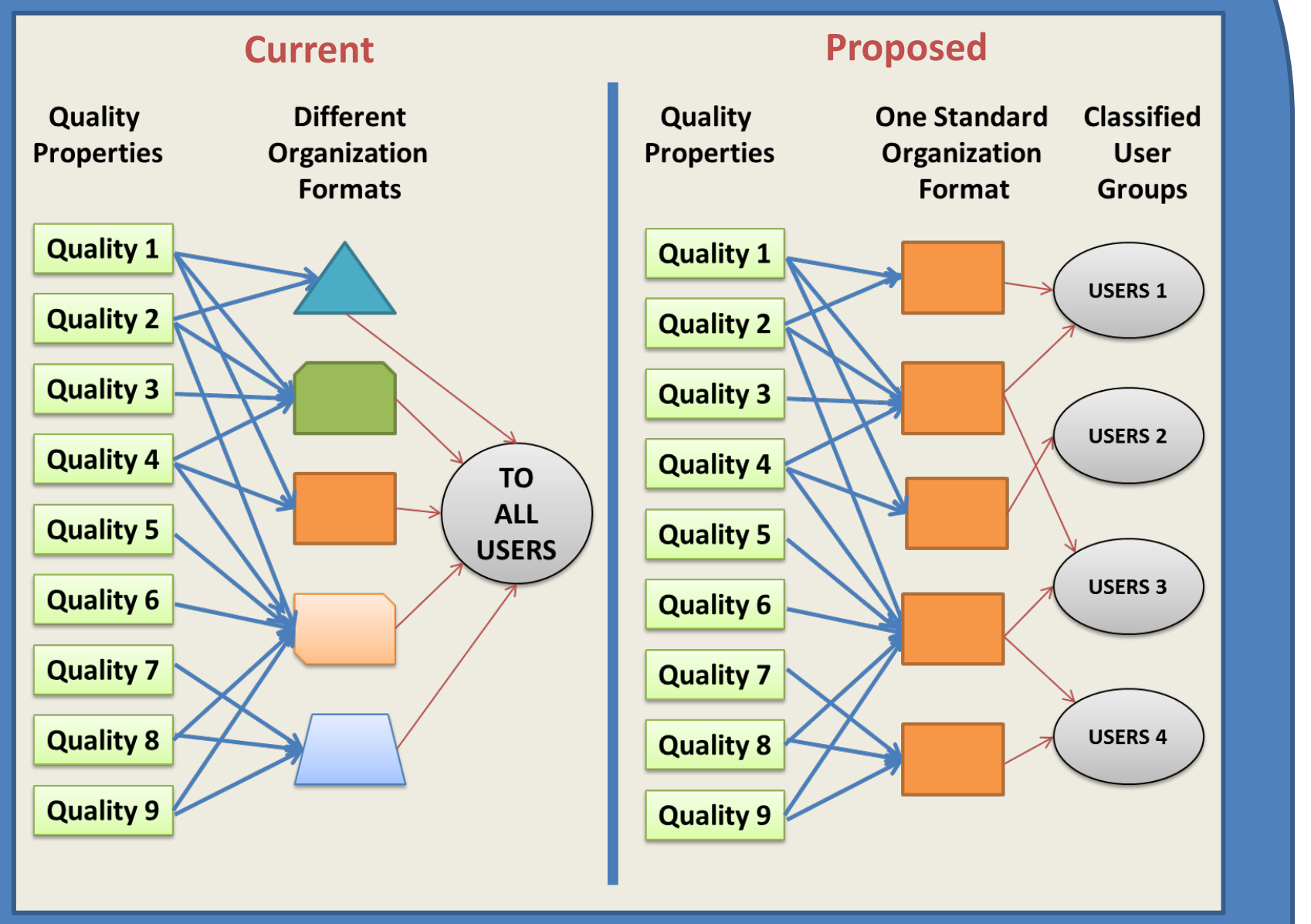
[1] NASA Goddard Space Flight Center, [2] Science Systems and Applications, Inc., [3] NOAA's Cooperative Institute for Climate and Satellites, North Carolina (CICS-NC),
[4]NOAA's National Centers for Environmental Information (NCEI), [5]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA

Objectives

- ❑ Bring together people from various disciplines to assess aspects of quality of Earth science data
- ❑ Establish and publish baseline of standards and best practices for data quality for adoption by inter-agency and international data providers
- ❑ Become an authoritative and responsive resource of information and guidance to data providers on how best to implement certain data quality standards and best practices for their datasets
- ❑ Build framework for consistent capture, harmonization, and presentation of data quality for the purposes of climate change studies, Earth science and applications
- ❑ Objectives evolve with participant inputs

Background

- **ESIP work in years past**
  - **IQ Cluster kick-off Meeting – Jan 6, 2011**
  - **Data Quality Session – Santa Fe Sumer Meeting – July 14, 2011**
  - **Led by Greg Leptoukh (NASA GSFC), who passed away on January 12, 2012**
  - **Data/Information Quality Birds of a Feather Session – Winter Meeting – January 2014, led by Carol Meyer**
  - **Information Quality Cluster session – Summer Meeting – July 2014, led by Gilberto Vicente**
  - **All focused on identifying challenges, use cases, representation of DQ/IQ to help users**



- **Other relevant activities**
  - **NASA Earth Science Data System Working Groups (ESDSWG) – Metrics Planning and Reporting WG (Product Quality Checklists) – 2010-2012**
  - **NASA ESDSWG Data Quality WG (Recommendations) – 2014-present**
  - **NOAA Data and Stewardship Maturity Matrices – 2008 - present**
  - **EUMETSAT CORE-CLIMAX System Maturity Matrix (e.g., http://presentations.copernicus.org/EGU2015-10158_presentation.pdf - 2014)**
  - **CEOS Essential Climate Variables (ECV) Inventory Questions**
  - **GEOSS Data Quality Guidelines**
  - **Quality Assurance framework for Earth Observation (QA4EO)**
  - **ISO Metadata Quality Standards (19157:2013; 19158:2012)**
  - **NCAR Community Contribution Pages**

Aspects of Information Quality – Key Defining Questions

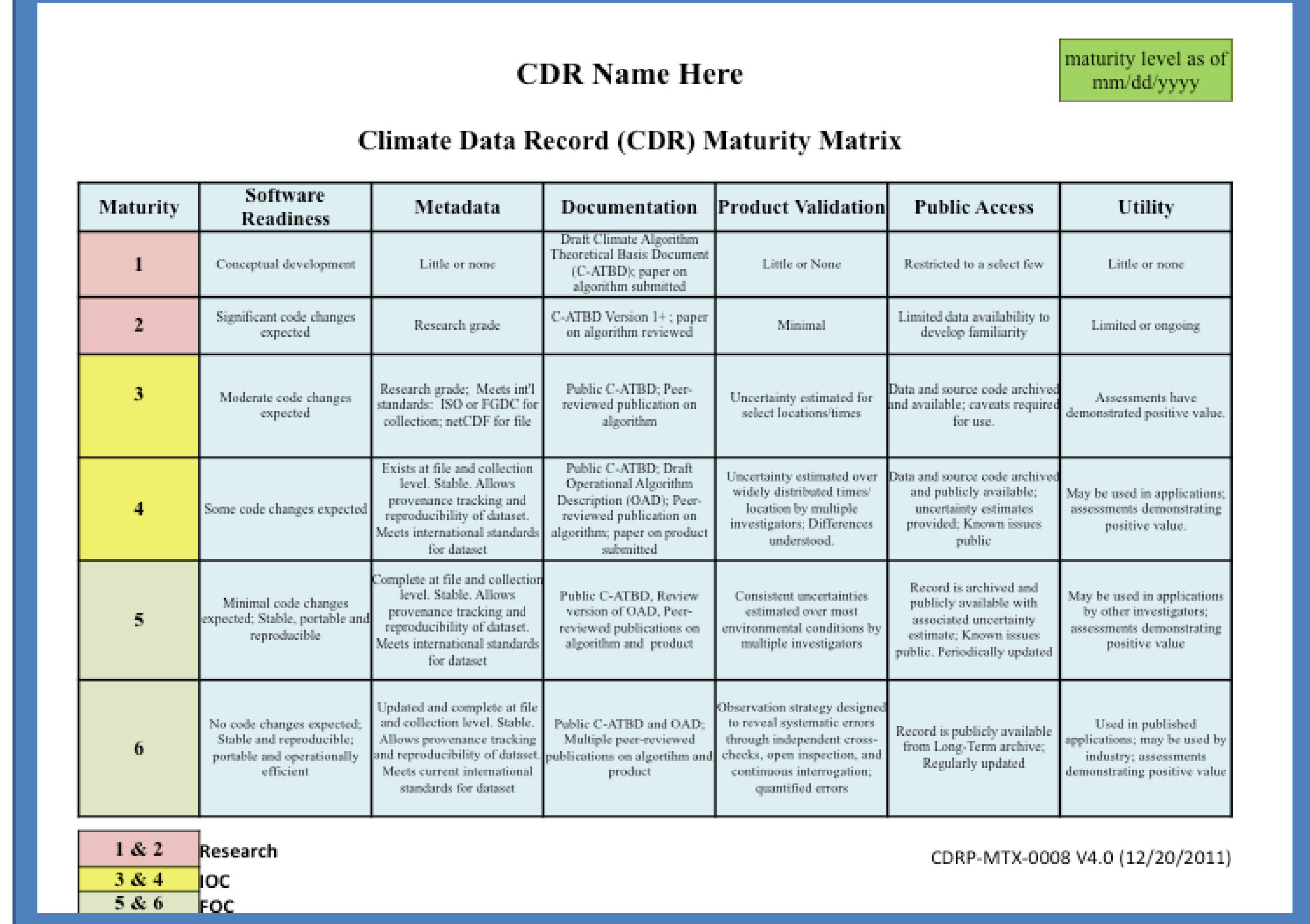
- **Science Data Quality**
  - **How accurate, precise and valid are the data?**
  - **How well have the error sources and uncertainties been characterized and documented?**
- **Product Quality**
  - **Has science quality been assessed and well documented?**
  - **How well have quality procedures and methods been defined, implemented, and documented?**
  - **How complete are metadata and documentation?**
- **Stewardship Quality**
  - **How well are data being managed and preserved by an archive or repository?**
  - **How well are science and product quality information being documented and captured in metadata?**
  - **How easy is it for users to find, get, understand, trust, and use data?**
  - **Does archive have people who understand the data available to help users?**

Recent Work - NASA Data Quality Working Group

Highest priority Recommendations based on analysis of 16 use cases

Category	Recommendation – Data Systems	Recommendation - Science
General	DAACs: Maintain continuous and effective communication with data producers throughout the duration of their projects.	Data Producers: Develop a data quality plan for each data product and submit it along with the data for dissemination.
Standard Documents & Processes	ESDIS & DAACs: Provide a standard set of documents to be provided to investigators and potential proposers; documents should describe what categories of quality information should be provided and how they should be shown using metadata.	HQ: Include references to standard set of documents in calls for proposals. Data Producers: Consult the existing guidelines that describe categories of data quality and provide information and evidence about the quality of the data set for each category.
Standard Documents & Processes	DAACs: Capture version id, processing history, and lineage for any dataset that is publicly available and in which multiple dataset versions of the same originating data are likewise published.	
Quality of Input Datasets used in Generating Products	DAACs: Request, from data producers, information about the contribution of the various input data that are used to process a higher level product.	Data Producers: Include information about correctness /uncertainty of input datasets used (e.g., land/ocean/region masks) along with products (e.g., sea ice product).
Quality Flags and Indicators	DAACs: Describe quality flags in the data documentation and in the list of Frequently Asked Questions (FAQs) about the dataset.	Data Producers: Provide users with a list of quality flags for questionable values along with descriptions for each quality flag (e.g., as provided by MODIS land products).
Quality Flags and Indicators	DAACs: Provide easy-to-use quality flags using standardized metadata and documenting the lineage and derivations of each quality flag.	Data Producers: Make quality flags publicly accessible and directly corresponding to a quantifiable metric, such as the related uncertainty, confidence intervals, and confidence levels.
Metadata Consistency Checking	DAACs: Employ metadata consistency checking tool that meets usability needs and generates reports with standards-based accuracy, precision, and uncertainty attributes provided in data granules.	Data Producers: Give recommendations on how data quality related attributes will be evaluated in the metadata scoring framework.
Publicizing Quality Issues:	DAACs: Host a prominent web page that captures known quality issues.	Data Producers: Convey fully the limitations of specific datasets, for inclusion in documentation and dataset descriptions.
Publicizing Quality Issues:	DAACs: Provide enough publicly available information with self-describing metadata and documentation such that the need for users to contact the DAACs is minimized.	
Publicizing Quality Issues	DAACs: Include documentation on how accuracy and uncertainty of products were determined.	Data Producers: Provide all data with added quality and/or uncertainty flags for the areas that have potential limitations.
Publicizing Quality Issues	DAACs: Inform users as soon as possible when data are compromised and provide status updates when readily available.	Data Producers: Provide information to DAACs promptly regarding any compromised datasets.
Dataset Recommendations	DAACs: Provide standing recommendations quickly to alternative datasets when a dataset has been retired or quarantined.	

Recent Work – NOAA Product and Stewardship Maturity Matrices



CDR Name Here

maturity level as of mm/dd/yyyy

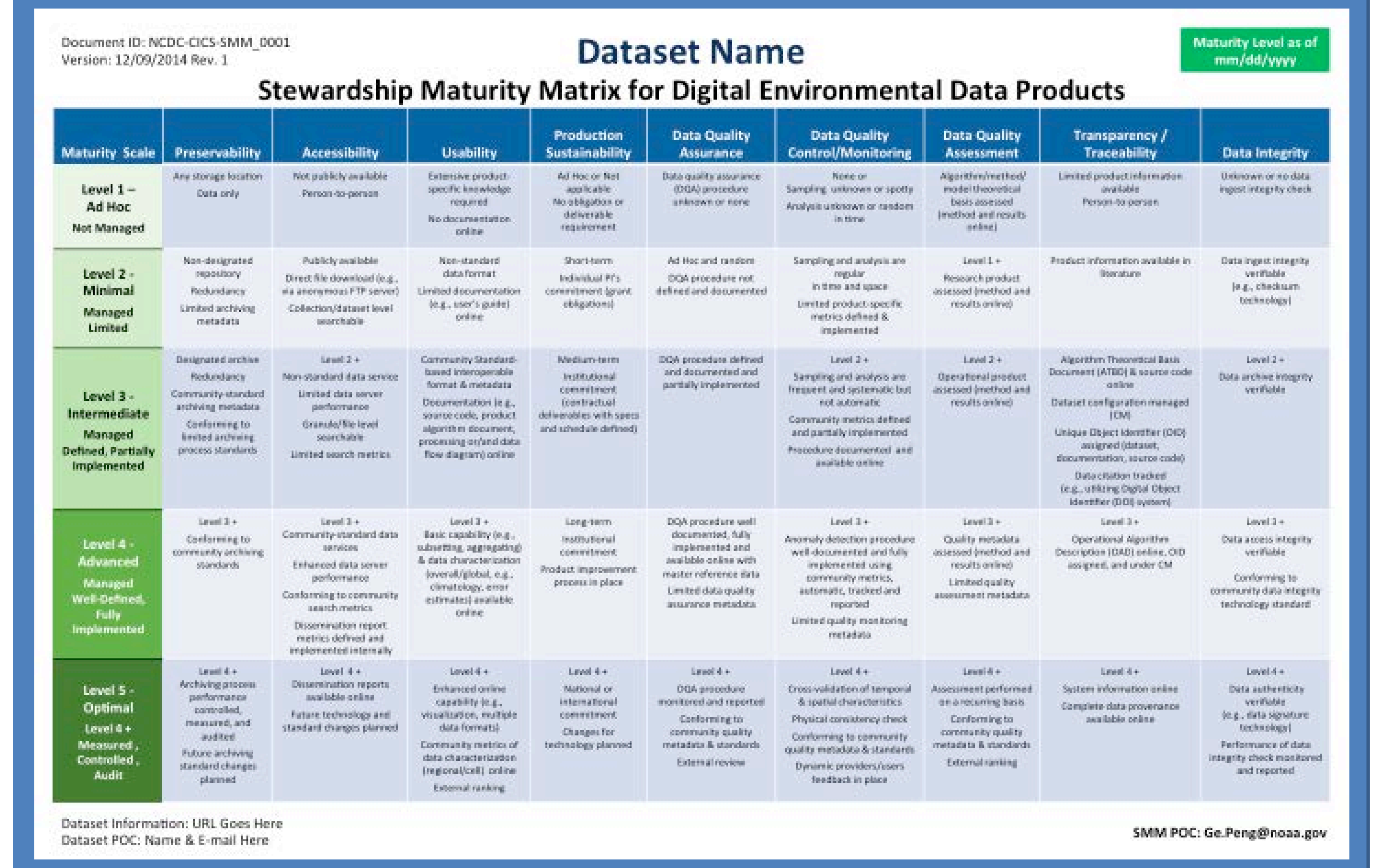
Climate Data Record (CDR) Maturity Matrix

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 algorithm 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

1 & 2	Research
3 & 4	IOC
5 & 6	FOC

CDRP-MTX-0008 V4.0 (12/20/2011)

The NOAA NCEI Climate Data Record (CDR) Maturity Matrix assesses the readiness of a product as a NOAA satellite CDR. It provides consistent guidance to data producers for improved data quality and long-term preservation. The latest CDR matrix template can be found at http://www.ncdc.noaa.gov/cdr/guidelines.html.



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

The NCEI/CICS-NC Scientific Data Stewardship Maturity Matrix (SMM) provides a unified framework for assessing the maturity of measurable stewardship practices applied to individual digital Earth Science datasets that are publicly available. It provides understandable data quality information to users including scientists and actionable information to management. The latest SMM template can be found at http://tinyurl.com/DSMMtemplate.

IQ Cluster- Suggested 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 conveying quality information to users using approach proposed by Vicente (Summer 2014)**
- ❑ **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.**

Acknowledgement: H. K. Ramapriyan's work was supported by NASA under contract NNG15HQ01C. Ge Peng is supported by NOAA under Cooperative Agreement NA09NES4400006. David Moroni's work is supported by a NASA contract with the Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA. Government sponsorship acknowledged.