Stewardship Maturity Matrix (SMM) for NCAR CFDDA Hourly 40 km Reanalysis as of 07/30/2015

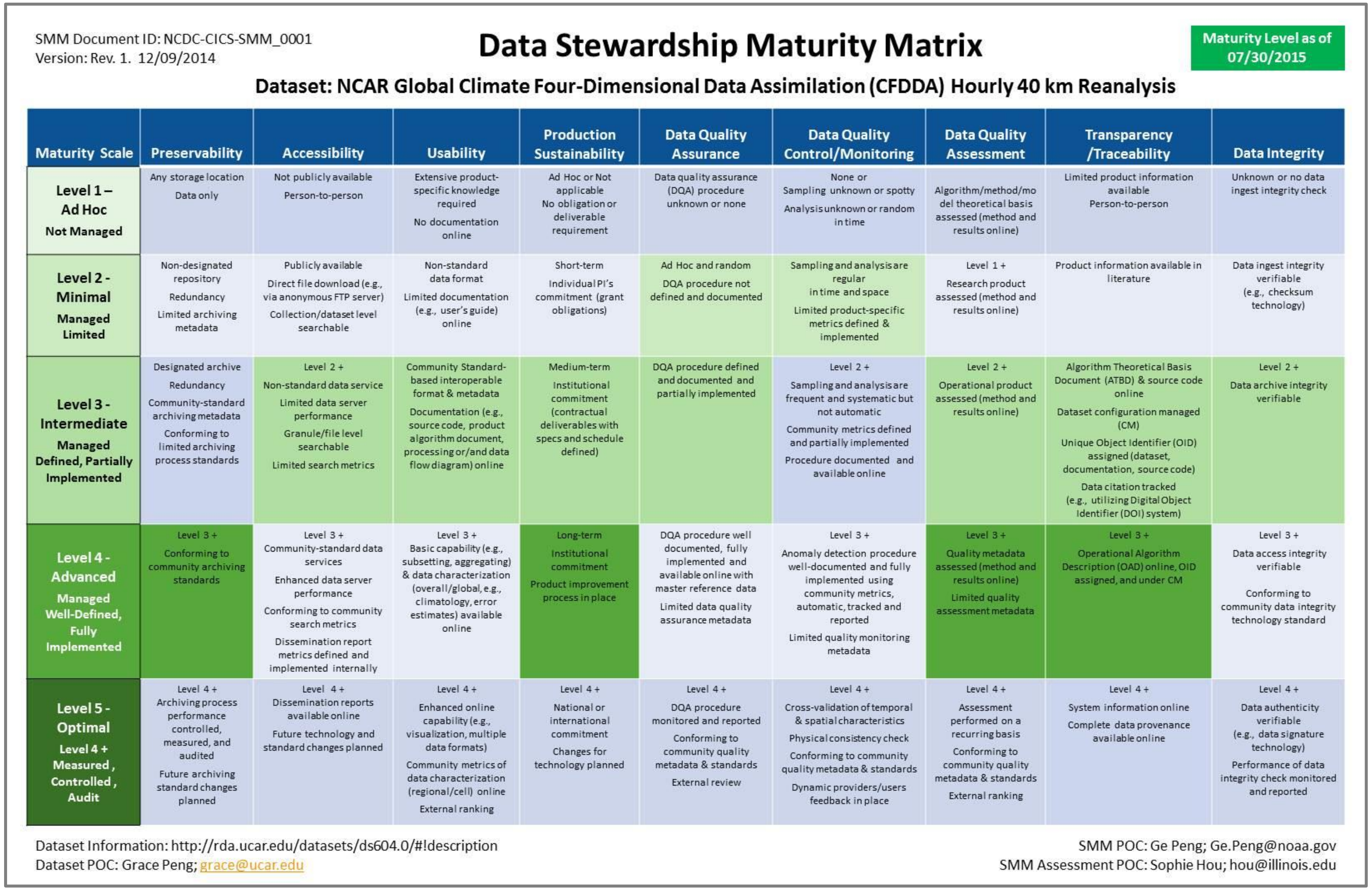
Data Stewardship Maturity Assessment Model Template Version: NCDC-CICS-SMM-0001-Rev.1 v3.1 02/26/2015

Dataset Title	**NCAR Global Climate Four-Dimensional Data Assimilation (CFDDA) Hourly 40 km Reanalysis**
Dataset Information URL	**http://rda.ucar.edu/datasets/ds604.0/#!description**
Data Provider POC (Name; E-mail; Affiliation)	Terri Betancourt; terrib@ucar.edu; National Center for Atmospheric Research (NCAR) Research Applications Laboratory (RAL);
Dataset POC (Name; E-mail; Affiliation)	Grace Peng; grace@ucar.edu; National Center for Atmospheric Research (NCAR) Computational & Information Systems Laboratory (CISL), Data Support Section (DSS);
SMM Version (Document ID and Version Numbers)	NCDC-CICS-SMM_0001_Rev.1 12/09/2014
SMM POC (Name; E-mail; Affiliation)	Ge Peng; Ge.Peng@noaa.gov; Cooperative Institute for Climate and Satellites, North Caroline (CICS-NC), North Carolina State University & NOAA's National Centers for Environmental Information (NCEI);
SMM Assessment Version (v<nn>r<mm>, e.g., v01r00)	**v01r04**
SMM Assessment POC (Name; E-mail; Affilication)	Sophie Hou; hou@illinois.edu; University of Illinois at Urbana-Champaign;
SMM Original Assessment Date (MM/DD/YYYY)	**03/25/2015**
SMM Original Assessment POC (Name; E-mail; Affiliation)	Sophie Hou; hou@illinois.edu; University of Illinois at Urbana-Champaign;
SMM Last Modified Date (MM/DD/YYYY)	**07/30/2015**
SMM Last Modification POC (Name; E-mail; Affiliation)	Ge Peng; Ge Peng@noaa.gov; CICS-NC/NCEI;
SMM Modified Date (MM/DD/YYYY)	**07/24/2015**
SMM Modification POC (Name; E-mail; Affiliation)	Grace Peng; grace@ucar.edu; National Center for Atmospheric Research (NCAR) Computational & Information Systems Laboratory (CISL), Data Support Section (DSS);
SMM Modified Date (MM/DD/YYYY)	**03/25/2015**
SMM Modification POC (Name; e-mail; Affiliation)	Sophie Hou; hou@illinois.edu; University of Illinois at Urbana-Champaign

Maturity Scale	**Level 1** **Ad Hoc** **Not Managed**	**Level 2** **Minimal** **Managed Limited**	**Level 3** **Intermediate** **Managed Defined, Partially Implemented**	**Level 4** **Advanced** **Managed Well-Defined, Fully Implemented**	**Level 5** **Optimal** **Level 4 + Measured , Controlled , Audit**		
Key Component						**Stewardship Maturity Rating /Justification or Evidence**	**Comments/ Recommendation**
Preservability	Any storage location Data only	Non-designated repository Redundancy Limited archiving metadata	Designated archive Redundancy Community-standard archiving metadata Conforming to limited archiving standards	Level 3 + Conforming to community archiving standards	Level 4 + Archiving process performance controlled, measured, and audited Future archiving standard changes planned	• **Level: 4** • The designated archive is NCAR's Research Data Archive (RDA). • Data is regularly backed up as part of RDA's stewardship practices. • Although RDA currently only uses customized metadata format, RDA uses community-standard controlled vocabularies (GCMD) to represent its data parameters. • RDA has plans to crosswalk between its current metadata format and the ISO19115 in order to review and determine the applicability of the result for implementation. • Additional standardized processes and documentations have been planned for the ingest process. • Metadata was checked for CF compliance; all non-compliant metadata was made compliant. • Data is backed up to HPSS tape archive in two sites as well as on disk for immediate download via several standard protocols for human or computer requesters (THREDDS, HTTP, OpENDAP, FTP) • Data is migrated to new media on a scheduled basis (every 3-5 years)	• It would be helpful if the references to OAIS and ISO19115 as community standards are included in the evaluation criteria.
Accessibility	Not publicly available Person-to-person	Publicly available Direct file download (e.g., via anonymous FTP server) Collection/dataset level searchable online	Level 2 + Non-standard data service Limited data server performance Granule/file level searchable Limited search metrics	Level 3 + Community-standard data service Enhanced data server performance Conforming to community search metrics Dissemination report metrics defined and implemented internally	Level 4 + Dissemination reports available online Future technology and standard changes planned	• **Level: 3** • Although CFDDA's data are available for public access, registration and/or log in is required before data files can be downloaded directly. • In addition, although CFDDA's data are separated into sub-collections (type 1: grouped by individual year and then by the months of the year; type 2: grouped by data parameter), this level of granularity is not searchable online. • Data is available for download from GLADE (GLobally Accessible Data Environment).	• Similar to preservability, it would be helpful if the examples of the community-standard data service provided in the paper are also referenced here.

						We don't advertise it, but it is possible to access the data via HTTP using scripts as long as the user's machine has a cookie from RDA that says that they are a registered user for that dataset and is logged in. • We collect usage statistics of web users, but not internal users accessing the data form Yellowstone, the NCAR supercomputer • We have no plans to show usage statistics on line at this time. • In the future, we plan to make this one of our IDV-enabled datasets so that users can visualize this data with IDV without having to download the data locally.	
Usability	Extensive product-specific knowledge required No documentation online	Non-standard data format Limited documentation (e.g., user's guide) online	Community standard-based interoperable format & metadata Documentation (e.g., source code, product algorithm document, processing or/and data flow diagram) online	Level 3 + Basic capability (e.g., subsetting, aggregating) & data characterization (overall/global, e.g., climatology, error estimates) available online	Level 4 + Enhanced online capability (e.g., visualization, multiple data formats) Community metrics of data characterization (regional/cell) online External ranking	• **Level: 3** • The file format for CFDDA's data is NetCDF. • The documentation regarding CFDDA are included as part of the data's public landing page, and the information can be accessed and downloaded directly without log in. • Recommendations regarding the best tools to view and visualize CFDDA data have also been included on the landing page. • However, the data cannot be readily visualized, manipulated, or analyzed as-is in the online environment. • We plan to add IDV compatibility	•
Production Sustainability	Ad Hoc or Not applicable No obligation or deliverable requirement	Short-term Individual PI's commitment (grant obligations)	Medium-term Institutional commitment (contractual deliverables with specs and schedule defined)	Long-term Institutional commitment Product improvement process in place	Level 4 + National or international commitment Changes for technology planned	• **Level: 3.5** • As long as CFDDA is archived with RDA and RDA is managed by NCAR CISL DSS, data for CFDDA should remain sustainable. • No product improvement planned	• What would considered to be the definition of short, medium, and long term in terms of time scale? • The evaluation criteria might also need to highlight the availability of committed human resources (skillsets/expertise/knowledge)?
Data Quality Assurance	Data quality assurance (DQA) procedure unknown or none	Ad Hoc and random DQA procedure not defined and documented	DQA procedure defined and documented and partially implemented	DQA procedure well documented, fully implemented and available online with master reference data Limited data quality assurance metadata	Level 4 + DQA procedure monitored and reported Conforming to community quality metadata & standards External review	• **Level: 2.5** • Currently, no standardized DQA procedure defined, documented, and implemented by the archive. • Prior to ingest, data files for CFDDA were inspected, and necessary modifications were made to the files to ensure data accuracy. However, the review process was not a standardized process.	• It might be helpful to provide short definitions to clarify the differences between the 3 different quality elements.

Data Quality Control/Monitoring	None or Sampling unknown or spotty Analysis unknown or random in time	Sampling and analysis are regular in time and space Limited product-specific metrics defined & 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 3 + Anomaly detection procedure well-documented and fully implemented using community metrics, automatic, tracked and reported Limited quality monitoring metadata	Level 4 + Cross-validation of temporal & spatial characteristics Physical consistency check Conforming to community quality metadata & standards Dynamic providers/users feedback in place	• **Level: 2** • Extensive Data Quality Monitoring was performed by the data provider, but not independently verified by the data archive. • Data Quality reports/concerns will be investigated, documented by the archive, and made available online.	•
Data Quality Assessment	Algorithm/method /model theoretical basis assessed (methods and results online)	Level 1 + Research product assessed (methods and results online)	Level 2 + Operational product assessed (methods and results online)	Level 3 + Quality metadata assessed Limited quality assessment metadata	Level 4 + Assessment performed on a recurring basis Conforming to community quality metadata & standards External ranking	• **Level: 3.5** • CFDDA's data quality is assessed based on the reviews of the data products that have been produced from CFDDA. • The theoretical basis for deriving the product (the model system in this case) has been assessed.	• Would it be possible to clarify the term "quality metadata assessed" a bit further? After reading the paper, I am still not sure if it is the quality of the metadata that I should be evaluating or is it the quality of the *process* for assessing metadata quality that I should be evaluating?
Transparency /Traceability	Limited product information available Person-to-person	Product information available in literature	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 3 + Operational Algorithm Description (OAD) online, OID assigned, and under CM	Level 4 + System information online Complete data provenance online	• **Level: 3.5** • Even though the provenance information is not structured in the ATBD/OAD format, the analogous information is available as part of the data's landing page and the information can be accessed and downloaded directly without log in. • RDA has plans to obtain DOI for CFDDA.	• Does ATBD and OAD apply to all data types? Based on this website (http://eospso.nasa.gov/content/algorithm-theoretical-basis-documents) , it seems to apply to only instrument based data? If ATBD and OAD do not apply to all data types, should it be a requirement to achieve Level 3 and 4? In other words, if ATBD and OAD do not apply to a particular data type and this data type has all of its other provenance available online, how should the level be assigned?
Data Integrity	Unknown or no data ingest integrity check	Data ingest integrity verifiable (e.g., checksum technology)	Level 2 + Data archive integrity verifiable	Level 3 + Data access integrity verifiable Conforming to community data integrity technology standard	Level 4 + Data authenticity verifiable (e.g., data signature technology) Performance of data integrity check monitored and reported	• **Level: 3** • Checksums verification performed on both HPSS tape copies, but not on the GLADE web files. This may be done later, CPU-time permitting • At ingest, data files are checked to ensure that they contain all parameters and grid points they should contain.	•



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

Data Stewardship Maturity Matrix

Maturity Level as of 07/30/2015

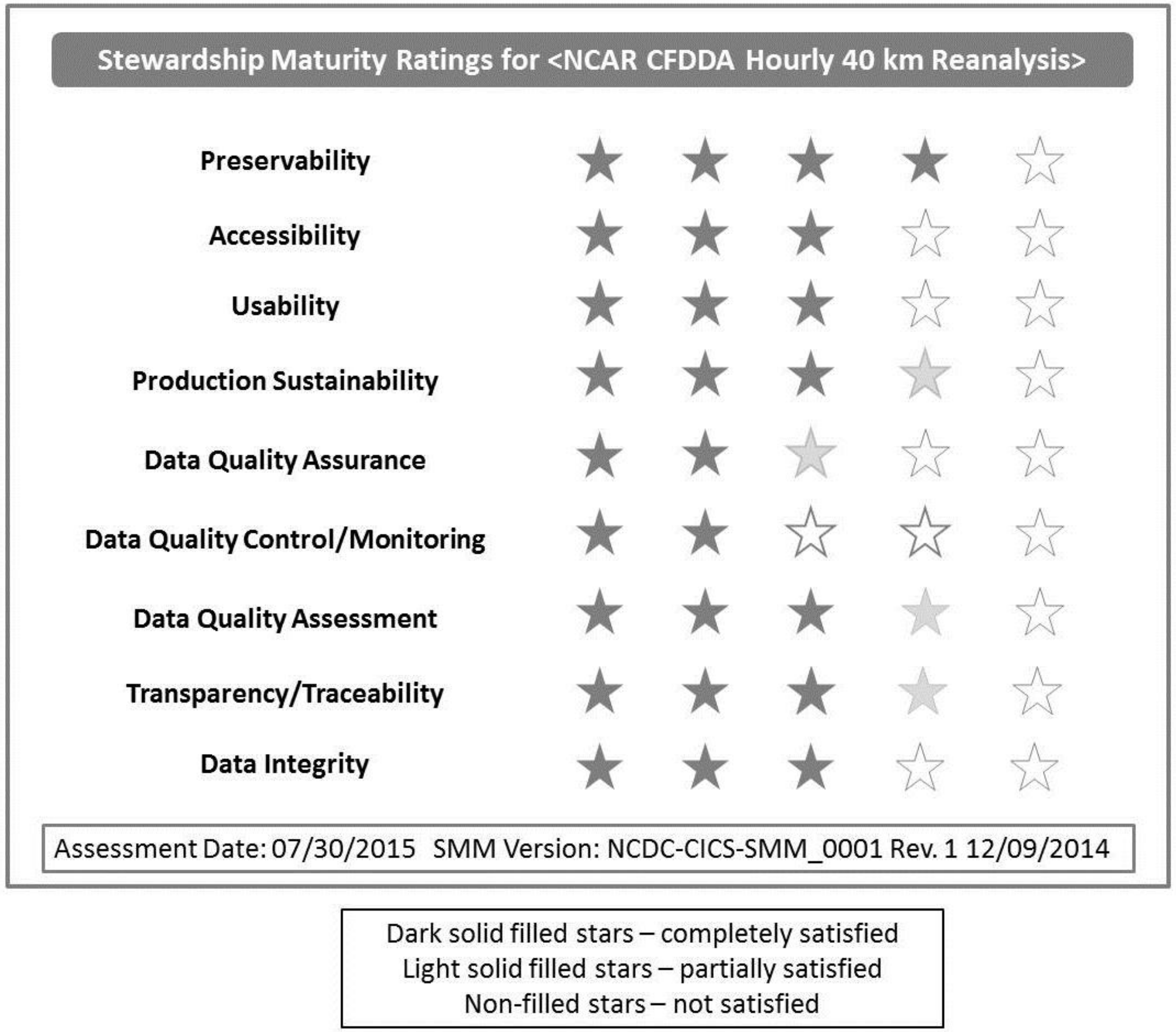
Dataset: NCAR Global Climate Four-Dimensional Data Assimilation (CFDDA) Hourly 40 km Reanalysis

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://rda.ucar.edu/datasets/ds604.0/#!description
Dataset POC: Grace Peng; grace@ucar.edu

SMM POC: Ge Peng; Ge.Peng@noaa.gov
SMM Assessment POC: Sophie Hou; hou@illinois.edu

Figure 1: The stewardship maturity scoreboard of the NCAR CFDDA Hourly 40km Reanalysis dataset.



Stewardship Maturity Ratings for <NCAR CFDDA Hourly 40 km Reanalysis>

Preservability	★	★	★	★	☆
Accessibility	★	★	★	☆	☆
Usability	★	★	★	☆	☆
Production Sustainability	★	★	★	✬	☆
Data Quality Assurance	★	★	✬	☆	☆
Data Quality Control/Monitoring	★	★	☆	☆	☆
Data Quality Assessment	★	★	★	✬	☆
Transparency/Traceability	★	★	★	✬	☆
Data Integrity	★	★	★	☆	☆

Assessment Date: 07/30/2015 SMM Version: NCDC-CICS-SMM_0001 Rev. 1 12/09/2014

Dark solid filled stars – completely satisfied
Light solid filled stars – partially satisfied
Non-filled stars – not satisfied

Figure 2: The stewardship maturity rating diagram for the NCAR CFDDA Hourly 40km Reanalysis dataset.