ESDS-RFC-033 NASA ES Data Quality Working Group
Category: Suggested Practice August 27, 2019
Updates/Obsoletes: N/A Comprehensive Data Quality Recommendations

Data Quality Working Group's Comprehensive Recommendations for Data Producers and Distributors

Status of this Memo

This document provides a comprehensive set of recommendations by the NASA Earth Science Data System Working Groups' (ESDSWG) Data Quality Working Group (DQWG) to data producers and distributors of the NASA Earth Science community and beyond. The distribution of this document is unlimited.

Change Explanation

Not Applicable.

Copyright Notice

Abstract

This document provides a comprehensive set of recommendations regarding data quality that are being offered for producers and distributors of Earth science data. In the context of this document, the data producers are science teams that develop algorithms as well as people managing/running the data production systems (e.g., Science Investigator-led Processing Systems - SIPSs) who work closely with the algorithm developers. Data archives (distributors) receive the data from the data producers and make them available to the user community. The recommendations were developed by the Data Quality Working Group, one of NASA's Earth Science Data System Working Groups, during 2014-2018, following analysis of 16 remote sensing use cases. The recommendations highlight issues regarding capturing, describing and conveying information about the quality of datasets held at the Earth Observing System Data and Information System (EOSDIS) Distributed Active Archive Centers (DAACs). While this document is targeted for NASA Earth science data, other organizations may also benefit from the methodology described here and the resulting recommendations for improvement.
Please refer to section 12 for a list of authors and contributors of this document.

Table of Contents

1. INTRODUCTION

The purpose of this document is to describe the recommendations that were made by the Data Quality Working Group (DQWG), one of NASA’s Earth Science Data System Working Groups (ESDSWG) during 2014-2016. These recommendations were made as a result of analyzing a number of data quality use cases related to NASA’s remotely sensed Earth science data. While the focus of the DQWG recommendations is NASA’s remotely sensed Earth science data, the methodology used for analysis of the use cases, as well as the recommendations themselves, are expected to be broadly applicable, with further analysis and adaptation, to other environments managing scientific data as well. First, we shall introduce the context in which the DQWG was formed and carried out its activities.

NASA’s 2014 Strategic Plan [11] states as one of its objectives (Objective 2.2): “Advance knowledge of Earth as a system to meet the challenges of environmental change and to improve life on our planet”. In support of this objective, NASA’s Earth Science Division (ESD) collects observations from instruments on satellites, aircraft and *in situ* platforms, and supports a variety of science investigation teams to develop data products (often referred to as datasets) covering a diverse set of disciplines. The NASA Headquarters Earth Science Data System (ESDS) Program within ESD supports Objective 2.2 by overseeing “the lifecycle of Earth science data with the principal goal of maximizing the scientific return from NASA's missions and experiments for research and applied scientists, decision makers, and society at large.” The ESDS Program consists of four components: The Earth Observing System Data and Information System (EOSDIS), Competitive Programs, International and Interagency Coordination and Development, and Continuous Evolution.

As a key *core* component of the ESDS Program, EOSDIS provides end-to-end capabilities for managing NASA’s Earth science data from diverse sources – satellites, aircraft, field measurements, and various other programs. It is managed by the Earth Science Data and Information System (ESDIS) Project at the NASA Goddard Space Flight Center. The capabilities of EOSDIS include: generation of Level 1 - 4 science data products for several Earth observing satellite missions; archiving and distribution of data products from satellite missions, airborne and/or ground-based measurement campaigns and some NASA-funded competitive programs. The responsibility to archive and distribute data in EOSDIS is carried out by 12 distributed, discipline-specific data centers known as Distributed Active Archive Centers (DAACs). The

DAACs serve a large and diverse user community by providing capabilities to search and access science data products and specialized services.

The EOSDIS has been in operation since 1994. Given the long-term needs to serve a global community of Earth science data users, one of the main tenets of EOSDIS is that it should evolve continuously to keep up with technological advances. While EOSDIS has been evolving since its inception, a concerted and formalized effort was made to promote continuous evolution with the formation of the ESDSWG in 2004. The ESDSWG provides a forum for participants in the Competed Programs of ESDS and EOSDIS to work together for infusing new ideas and technologies into EOSDIS. The primary Competed Programs participating in the ESDSWG are: 1. Advancing Collaborative Connections for Earth System Science (ACCESS) and 2. Making Earth System Data Records for Use in Research Environments (MEaSUREs).

The DQWG is one of the ESDSWG working groups. It was formed at the annual meeting of the ESDSWG in March of 2014 as a result of interest expressed by the ESDIS Project and MEaSUREs investigators. The mission of the DQWG is to evaluate current processes and make recommendations to the ESDIS Project and the ESDS Program for improvements in providing data quality information to users. This affects the areas of capturing, representing and enabling the use of data quality information describing accuracy, precision, and uncertainty. The improvements should ensure that clear and sufficient information is provided to the user to determine usability and distinguishability among apparently similar datasets and to identify applicability (or "fitness for use") by providing examples of use. Note that distinguishability here means the ability to distinguish between measurements of the same parameter captured during the same time window in the same area but with different approaches. Since its formation, the DQWG has developed use cases based on remote sensing measurements, analyzed them, arrived at over 100 recommendations, prioritized them, identified “low-hanging fruit (LHF)” recommendations for these types of measurements, listed solutions available to address the LHF recommendations, and developed implementation strategies. LHF here means immediately actionable recommendations with low cost and high benefit. It is recognized that these recommendations may not be readily applicable to other types of measurements, which likely require additional analysis and adaptation. These activities of the DQWG, carried out during 2014-2018, are summarized in Figure 1.

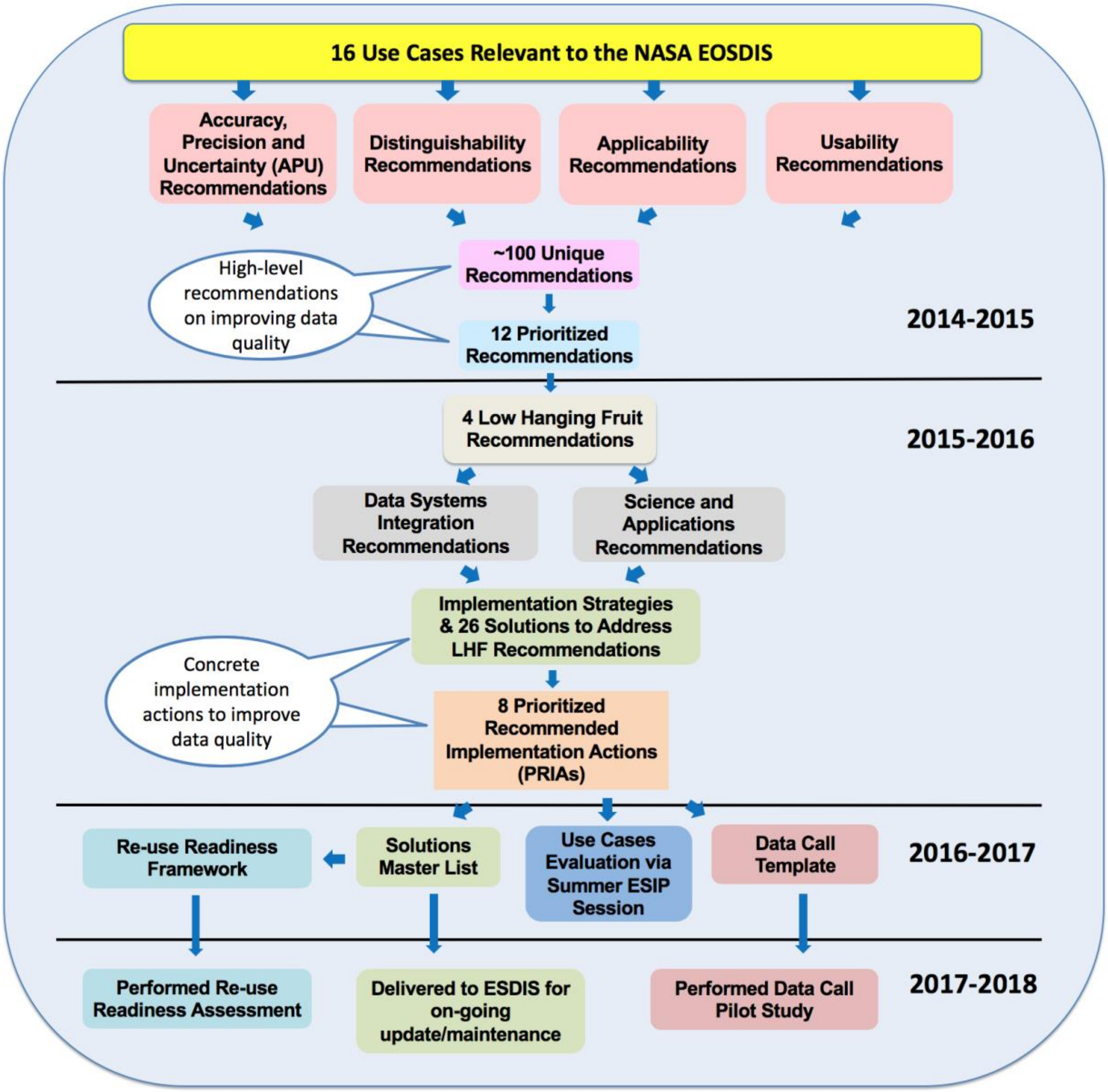
16 Use Cases Relevant to the NASA EOSDIS
Accuracy, Precision and Uncertainty (APU) Recommendations
Distinguishability Recommendations
Applicability Recommendations
Usability Recommendations
High-level recommendations on improving data quality
~100 Unique Recommendations
12 Prioritized Recommendations
2014-2015
4 Low Hanging Fruit Recommendations
2015-2016
Data Systems Integration Recommendations
Science and Applications Recommendations
Implementation Strategies & 26 Solutions to Address LHF Recommendations
Concrete implementation actions to improve data quality
8 Prioritized Recommended Implementation Actions (PRIAs)
Re-use Readiness Framework
Solutions Master List
Use Cases Evaluation via Summer ESIP Session
Data Call Template
2016-2017
Performed Re-use Readiness Assessment
Delivered to ESDIS for on-going update/maintenance
Performed Data Call Pilot Study
2017-2018

Figure 1. DQWG Historical Legacy of Milestones

Figure 1 also shows connection to the Earth Science Information Partners (ESIP), whose Information Quality Cluster (IQC) has taken over the maintenance, further development and evaluation of use cases [16]. Moreover, there were activities about analysis of the Solutions Master List using a Re-Use Readiness Framework and assessment of datasets held at the DAACs using a Data Call Template under the Data Call Pilot Study. Details about those activities are described in the latter part of this document. The DQWG devoted the period between April 2018 and March 2019 to finalize a number of documents and submit them to the ESDIS Standards Office for review and publication, and officially concluded its 5-year effort.

The remainder of this document is structured as follows. Section 2 describes the methodology for collecting use cases and summarizes the 16 use cases developed by the

DQWG. Section 3 provides a definition of four focus areas used to establish the different points of view used in analyzing the use cases. Section 4 presents the recommendations resulting from the use case analysis, considering four phases of the data quality information management lifecycle, and mapping the recommendations into seven categories. Section 5 shows how the large number of recommendations was narrowed down to arrive at 12 high-priority recommendations and four LHF recommendations. Sections 6, 7 and 8 discuss, respectively, the solutions master list, implementation strategies and implementation recommendations. Section 9 briefly describes related international activities relevant to the DQWG. These are then followed by a section summarizing the main conclusions and suggestions for future work. Section 11 provides a list of references, and Section 12 lists the authors, DQWG members and use case contributors. A number of appendices are included to provide a glossary of acronyms and further details on some aspects of items covered in the main body of the document.

2. USE CASES

This section describes the methodology for developing use cases and illustrates it with one detailed example to explain the template employed to collect them. A tabular summary of all 16 use cases developed by the DQWG is provided at the end of this section. Please note that since the use cases were developed in 2014-2015, some of them may no longer be relevant to today's policies, procedures, workflows, and user scenarios. But the recommendations that were extracted from these remote sensing measurement-based use cases are still considered applicable and are expected to be applied across data centers, data collections, and users.

2.1. Method and Template Used to Collect Use Cases

Early in the first year of the DQWG's activities, it became quite clear that it would be beneficial to capture the issues related to conveying information on data quality to users through a set of use cases. The use cases needed to address datasets offered by EOSDIS DAACs and had to cover a broad class of users. A total of 16 use cases were collected over a 4-month period and described using a template (see example below). A slightly updated version of this template has been transcribed into an active Google Form that is intended to simplify the entry of use case information. Detailed information about this form, including the fields and descriptions, can be found in Appendix B. While there are “standard” ways for defining use cases such as the OGC template [1], these are aimed towards system design, whereas the DQWG's goal was to highlight data quality issues from a user's point of view. Therefore the DQWG developed a template tailored for this purpose, much of which was borrowed from an existing use case template developed by Eric Tauer for use within the PO.DAAC at JPL. Shown below is one example to illustrate the template as well as the nature of information filled in by members providing the use cases.

Use Case Title:

NASA Team Sea Ice Concentration Filters

Point of Contact:

Lisa Booker

Email Address for Point of Contact:
lisa.booker@nsidc.org

Use Case Narrative: (Goal and Context)
The lack of transparency in subjective ice removal from the data makes reproducibility of these data difficult. In addition, as a researcher working with sea ice, I would like to be able to use my own judgment to filter out questionable ice values. Having a quality flag that marks questionable ice values allows me to determine which pixels to consider. And leaving these values in the data and simply flagging them allows me to reproduce the work of the data producers as described in literature.

Domain of Interest: Climate, Cryosphere

Professional Domain of User: Scientist/Researcher

Primary User/Stakeholder Relationship: Human User -> Mission-Project/Stakeholder

Secondary User-Stakeholder Relationship: N/A

Primary Scope: Qualitative-Science (for details see: bit.ly/qualsci)
Rationale: By flagging questionable ice values, it is left to the researcher to determine the integrity of the value for their research. In addition, the overall integrity of the science is improved by making the data more reproducible.

Secondary Scope: Quantitative-Science (for details see: bit.ly/quantsci)
Rationale: Adding a quality flag will provide uncertainty information not previously provided in the data, therefore improving the integrity of the data and science.

Use Case Chronology:
A user contacts NSIDC User Services Office (USO) asking for more information about the subjective removal of ice.
USO works with the data producer to understand the history of the subjective filtering of ice values.
USO communicates with user that they have passed the information along to the data producer and it is unclear if and when this information will be addressed.

Success Criteria:
A user knows through documentation that quality flags are available for questionable ice values.
The values for the quality flag are fully defined, i.e. weather effect has a particular value, coastline has a particular value, etc.

Data Quality Keywords:
algorithm, accessibility, filtering, flags

2.2. Use Case Summary

Table 1 shows the 16 use cases with their titles and key issues identified. The details of the use cases, which follow the template structure illustrated above in subsection 2.1, are given in Appendix C.

Table 1. Use Cases Considered by the DQWG

No.	Title	Key DQ Issue(s)
1	Aquarius Salinity Data Quality Issue Noted in Coastal Region	Large differences between buoy and satellite-derived data.
2	Dataset Recommendation	Selecting the most relevant and useful datasets among those containing similar geophysical parameters.
3	Fisherman Needs SST and Wind Vector Data Over Gulf Stream	User needs data with spatial resolution under 10 km and maximum data coverage with minimal data dropouts.
4	MEaSUREs PI wants to provide complete quality documentation to make his products useful to community	Guidance to Principal Investigators about proper level of data quality documentation.
5	Outlier Detection and Attribution	Need improved identification and characterization of outliers.
6	SMAP Freeze/Thaw Algorithm	Use of data outside "normal" spatial coverage area.
7	AIRS Quality Indicator Recommendations	Guidance regarding how to use already available quality indicators.
8	Data Quality Filtering	Need for a service to apply specific quality filtering levels or flags while extracting data values from a file.
9	Errors Introduced by Binning, Smoothing, and Interpolation	Users need to know error propagation as higher level products are generated.
10	Land Mask Issue in Near Real-Time DMSP SSM/I Daily Polar Gridded Sea Ice Concentrations	Geometric error in land mask.
11	MEaSUREs Global Food Security Analysis & Support Data (GFSAD) - Provisional Crop Dominance (CD) @ 1 km product	Accuracy of product documentation versus provisional product contents.

12	Metadata consistency evaluation	Conformance of netCDF or HDF files (granules) to the Climate Forecast (CF) and Attribute Convention for Dataset Discovery (ACDD) metadata models.
13	NASA Team Sea Ice Concentration Filters	Quality flag that marks questionable ice values, rather than filtering out such values.
14	Provide ancillary information on potential biases	Provide sufficient information to users such that they can judge and replicate our products.
15	Region Vulnerable to Storm Surge	Insurance company trying to assess the coastal region that is vulnerable to storm surge finds that only limited types of data available.
16	Sensor-Specific Observation Quality Contribution to L4 Data	Need to know how much of a pixel is comprised of specific spaceborne sensor inputs and /or in situ measurements.

3. FOUR PHASES OF THE DATA QUALITY INFORMATION MANAGEMENT LIFECYCLE

In analyzing the use cases, it was helpful to consider four phases of the data quality information management lifecycle, so that the resulting recommendations could be mapped to one or more of the phases [4]. These phases cover broad categories of activities associated with the lifecycle and include: 1. *Capturing*, 2. *Describing*, 3. *Facilitating Discovery*, and 4. *Enabling Use*. These phases are described below:

3.1. Capturing

The capturing phase is critical for obtaining information about the quality of data. It involves 1) collecting quality information, such as results from instrument calibrations, conditions (e.g. instrument modes, environmental, and weather) under which measurements were made, validation results for observations and model outputs, and lineage of data processing; and 2) deriving higher-level quality information, such as quality flags and indicators. Opportunities for capturing data quality information often occur early on and should be carried throughout the data quality information management lifecycle, as the instruments are being deployed to record the data and the data are being acquired, verified, and processed. Data quality issues which may be affected by the data processing algorithms and improvement to algorithms (hence, software versions) can occur anytime in the processing cycle.

3.2. Describing

"Describing" means organizing data quality information in a structured and meaningful way (e.g. data documentation, metadata records, and metadata embedded inside data files) so that both data users and data tools can easily understand and utilize the data. This phase offers opportunities for potential and actual users to understand the quality of data products and services. As in the Capturing phase, preparing descriptions of data quality should begin early on and be carried throughout the data quality information management lifecycle as re-processing campaigns are carried out.

3.3. Facilitating Discovery

Well captured and described data quality information should be published and easily accessible to the public. As a minimum requirement, data users should easily find information about the quality of data products and services. In addition, data quality information should be leveraged to allow data users to discover data products and services that meet their data quality requirements. Once the data quality information has been acquired and described, the data providers should make this information readily accessible, to ensure that the data quality information can be easily found.

3.4. Enabling Use

Fostering the use of data quality information improves opportunities for potential users to assess whether data products and services are appropriate for intended uses. Capabilities should be provided to facilitate the use of data quality information that has been acquired, described, and discovered and potentially to ease and promote the use of data products themselves. For example, linking quality fields (e.g., flags and indicators) with data fields following standard approaches (e.g., Climate and Forecast convention) can ease the access to and use of both the quality information and the data themselves.

4. PRIMARY FOCUS AREAS

The DQWG identified four primary focus areas for data quality: 1) Accuracy, Precision and Uncertainty, 2) Distinguishability, 3) Applicability, and 4) Usability. Each of these focus areas and their importance are discussed briefly below. Subgroups were formed to address each of these focus areas while analyzing the use cases [2,3,10,17].

4.1. Accuracy, Precision and Uncertainty

Accuracy, precision and uncertainty are fundamental aspects of the scientific quality of data. It is critical that these be assessed and recorded along with the data products that the scientists generate and convey to the data archives and to the end users. Accuracy indicates how close to truth a given measurement or derived parameter is. Precision indicates how close different independent instances of a measurement or a derived parameter for a given phenomenon are to each other. Uncertainty quantification has the ability to provide additional information that accuracy and precision metrics by themselves may be lacking for a variety of error distributions. While the DQWG has explored various examples and use cases of accuracy, precision and uncertainty from the perspective of scientific data quality, this paper does not intend to provide specific guidance or advocacy of a particular methodology toward the computation of such quality metrics. Rather, the intent here is to call out the importance of these constituents toward the most holistic characterization of scientific data quality. Resources are already available for detailed guidance on assessing and expressing accuracy, precision, and uncertainty, including the JCGM 100:2008 [9] and ISO 5725-1:1994 [23]. In order for users to determine whether a particular scientific product is suitable for their application, it is essential that they know the constituents of an error distribution (i.e., accuracy and precision) as well as the uncertainty conveyed by that distribution. The level of detail at which the accuracy, precision and uncertainty information is provided can vary depending on the products. The responsible data producers should determine whether the data should be assessed and provided at the collection, granule, or pixel level. In summary, it is vital that the data products offered by EOSDIS to the community be accompanied by clear information about accuracy, precision and uncertainty in a consistent manner.

4.2. Distinguishability

It is necessary to enable current and potential users to differentiate between available, apparently similar, datasets so that the appropriate data product may be selected in an efficient manner. The distinguishability of data refers to the extent to which a particular dataset can be differentiated from other available datasets. It is also important for the users to know the level of consistency between two datasets, which can be used as an independent check of measurement uncertainty. Unique characteristics and aspects of data quality can be important when selecting a particular dataset, service, or tool for analyzing the data. The quality of available datasets should be readily comparable so that users can identify the data to be used. Similarly, the availability of tools for comparing datasets can improve efforts for distinguishing between similar datasets or datasets with similar quality characteristics. Similar data products can be distinguished from each other based on two criteria: the science quality and the product quality. Science quality of a dataset is mainly based on inputs from data producers, for example, quality flags and indicators, uncertainties, validation results, the instrument detection principle, instrument configuration, sampling procedure and treatment (including sampling integration duration and frequency), and calibration standards and methods. Product quality of a dataset is the degree of truth, genuineness, and reliability of the science nature, elements or values contained or represented in a dataset. It is mainly based on how DAACs collect, organize, and present metadata, and also on the completeness and timeliness (relative to the discovery of any new findings on data quality issues or changes to quality metrics) of materials received from data producers. The difference between the datasets reflects the intrinsic instrument properties and sampling. At times, there are differences between the datasets while they are both correct or the difference cannot be reconciled. It is essential for DAACs to capture and distribute the measurement comparison results. Complete and accurate metadata and documentation help users to distinguish one dataset from another similar one.

4.3. Applicability

Users and potential users of data need to be able to determine whether a particular dataset is relevant for an intended purpose and whether the quality of the dataset is suitable for use with available tools or for achieving particular objectives. If the applicability of a dataset for a particular purpose cannot be determined, assessments of potential applicability may need to be completed. Providing information about previous assessments of applicability and reports of previous uses for particular purposes can reduce the need to conduct new assessments of dataset applicability. Information about the applicability of datasets can be valuable when exploring datasets for potential use and to identify whether a previously-used dataset is applicable for a new purpose. Information that can be helpful for determining the applicability of a dataset includes descriptions about how the dataset was previously used and information about the limitations of the dataset for a particular use. The availability of tools for accessing and using the quality of candidate datasets also can contribute to determinations of dataset applicability.

4.4. Usability

Facilitating the use of Earth science data by user communities is a primary objective for distributing data products and services. Usability is the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use [28] . Ensuring the usability of the data products and services that are distributed by EOSDIS improves the potential for their use, not only by the scientists who are familiar with the instruments that were used for data collection, but also by broader audiences, such as representatives of various scientific disciplines and their students, decision-makers and those who inform them, and members of the general public. Likewise, considerations for usability inform the analysis and decisions to recommend tools, products, and services for improving the quality of data disseminated by EOSDIS.

5. SUMMARY OF HIGH-LEVEL RECOMMENDATIONS

This section describes the process used for developing recommendations for data quality and summarizes them at a high-level.

To facilitate analysis of the use cases, the DQWG formed four subgroups, and each subgroup addressed one of the four focus areas described in section 3. Each subgroup analyzed the 16 remote sensing use cases and arrived at 126 data system recommendations and 76 science recommendations at the “raw” level, i.e., before merging similar recommendations from the different use cases and the different subgroups. In general, recommendations call for action to address issues identified or to improve the experience of users in dealing with information on data quality. In some cases, no sweeping changes are needed as the recommendation may have already been addressed by one or more existing/pending policies or solutions already deployed (or pending deployment) in an operational environment, yet those recommendations are still retained to help identify and justify the users’ need for such policies and solutions. Data system recommendations call for action by data archives (e.g. DAACs) or the ESDIS Project. Science recommendations call for action by data producers, or by the NASA HQ Science Program calling for the adoption and adaptation of data quality standards and the generation of data quality information corresponding to NASA-funded data products. Merging the recommendations across the subgroups, across use cases and accounting for similarity and complementarity of data system and science recommendations, resulted in a total of 93 recommendations, grouped into 7 categories, which are summarized below. Details of these 93 recommendations can be found in Appendix D. Each recommendation was assigned a unique recommendation number, ranging from 1 to 93, so that they can be referenced by later activities. Each recommendation was mapped onto one or more of the 4 phases discussed in section 3.

5.1. General

A general recommendation from the use case analysis was that data centers should maintain continuous and effective communication with data producers throughout the duration of their projects. It was also recommended that data producers should develop a data quality document for each data product and submit it along with the data for dissemination. DAACs and Data Producers should work together to provide clear and thorough product quality information for each dataset.

5.1.1.

DAACs: Maintain continuous and effective communication with data producers throughout the duration of their projects.

Data Producers: Develop a data quality document for each data product and submit it along with the data for dissemination; for new datasets in which data quality has not yet been assessed, this document may incorporate a plan by which data quality information is captured to be disseminated later.

5.2. Standard Documents and Processes

This category contains recommendations for both data archives and producers on how to leverage standard documentation and metadata processes to increase the visibility of quality information, assist the evaluation of fitness for use, and ease the use of data products. For example, it suggests that data archives should provide a standard set of documents to be provided to investigators and potential proposers; documents should describe what types of quality information should be provided and how they should be represented in metadata.

5.2.1.

ESDIS & DAACs: Provide a standard set of documents to investigators and potential proposers; documents should describe what categories of quality information should be provided and how they should be represented in metadata.

- Provide data producers with examples of determining and describing product quality (e.g., use of ATBDs, ESDIS product quality checklists, and any documentation that helps the PI’s create a final product with complete quality documentation) of different types of measurements.

NASA HQ: Include references to a standard set of documents, including the Data Management Plan Template for Data Producers [13], in calls for proposals involving generation of data products.

- Enable open review of products with involvement of DAACs to help promote increased discovery, reduced latency, and dissemination of known issues.

Data Producers: Consult guidelines that describe categories of data quality and provide information and evidence about the quality of the dataset for each category.

- Prepare data and attributes related to accuracy, precision and uncertainty and organize them based on standards; and collect feedback especially when the uncertainty reporting from the PI does not fit the current standards.
- Provide data quality information through appropriate data formatting and metadata specifications, e.g., Climate & Forecast (CF) [20], ISO [24, 25, 26, 27], Attribute Convention for Data Discovery (ACDD) [22], and Unified Metadata Model (UMM) [21].
- Provide data lineage and processing history information.

5.2.2.

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.

Data Producers: Include version id, processing history, and lineage in the granule metadata.

5.3. Quality of Input Datasets used in Generating Products

Quality associated with input datasets has significant impact on the quality of derived data products. It is recommended that data archives always request from data producers information about the contribution of the various input data, e.g. land/ocean/region masks, to the quality of the derived higher-level products.

5.3.1.

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).

5.4. Quality Flags and Indicators

Quality flags and indicators are simple and quantified approaches to allow data users to easily evaluate the fitness for use of data products and/or extract portion of data products meeting their data quality needs. This category contains recommendations on providing and publicizing easy-to-use quality flags and indicators, directly corresponding to quantifiable metrics. Quality flags and indicators can be defined at various levels of detail (entire mission, collection, granule, grid point/pixel).

5.4.1.
DAACs: Describe quality flags in the data documentation and in the list of Frequently Asked Questions (FAQs) about the dataset. The description of quality flags could be referred to on the landing page and should be highlighted in the data documentation with its own section.

Data Producers: Provide users with a list of quality flags along with descriptions for each quality flag (e.g., as provided by MODIS land products).

- Identify quantifiable data quality criteria, such as confidence levels and the values of quality flags, which can be used as criteria for refining search queries.
- Ensure that quality flags are related to a quantifiable metric that directly relates to the usefulness, validity, and suitability of the data.
- Incorporate algorithm to assess and improve quality of product.
- Define and/or create "indicators" to represent quality of a data product from different aspects (e.g., data dropout rate of a "sea surface temperature" data product can be considered as one data uncertainty indicator).
- Provide data quality variables and metadata along with detailed documentation on how the variables/ metadata are derived and suggestions on how to use them in different applications
- Provide description of the pixel-level quality indicator, including the algorithms and datasets used to derive this quality information.
- Work with DAACs to provide data quality information through a standardized quality flagging schema (e.g., GHRSST model for quality confidence levels).
- Provide all data with added quality and/or uncertainty flags for areas that show spurious data (e.g., ice in unlikely places). Provide pixel-level uncertainty information.
- Provide definitions for each quality indicator and a description of how each quality indicator can be used (documentation, user guide, and in search system).

5.4.2.
DAACs: Capture and disseminate to users easy-to-use quality flags and indicators.

- Encourage data producers to maintain transparency in data production/creation and provide quality flags/indicators.
- Provide capabilities to allow data users to leverage data quality flags/indicators for evaluating applicability.
- Make sure data producer-provided documentation of how each quality flag/indicator was derived, including delineations between specific processing algorithms and ancillary datasets used in the flagging schema (not every quality flag/indicator is created equal) is easy for data users to discover, access, and understand.
- Document and publish all available descriptions for data quality flags/indicators.
- Provide up-to-date metrics to summarize high-level data quality and summarization of validation studies in product metadata/documentation.

- Provide capability to harvest the quality flag/indicator data and metadata for each dataset (e.g., DMAS at PO.DAAC).
- Include per pixel quality layer(s) where appropriate (e.g., NASA EOSDIS GIBS and Worldview).
- Provide clear documentation about types and availability of quality flags/indicators using self-describing metadata (e.g., NetCDF/HDF, CF-conventions, and ISO 19157).
- Document and capture as metadata whether or not there is a pixel-level quality flag/indicator for a given dataset. For example, the "Variable Association" approach defined in UMM-Var [21] can be leveraged for this purpose.
- Provide quality information and/or algorithms to assess quality of data accessed through subsetting services/tools.
- Work with data producers to develop procedures to ensure that all necessary quality control information (e.g. quality flags and indicators) is properly bundled with the subsetted data.

Data Producers: Make quality flags/indicators publicly accessible and directly corresponding to a quantifiable metric, such as the related uncertainty, confidence intervals, and confidence levels.

- Provide all data with added quality flags/indicators for the areas that have potential limitations.
- Capture known issues (for particular regions or time intervals) of data.
- Associate science variables with quality control information (e.g. quality flags and indicators), if available, in both data documentation and metadata.

5.5. Metadata Consistency Checking

Performing metadata consistency checking, ideally in a scoring framework, against common metadata standards (e.g., ISO 19115 and CF) on both dataset and data file levels is important for both data producers and data archives. Metadata Consistency means two things, the first is that metadata are compliant with standards, the second is that values of metadata elements (e.g. platform/variable/instrument keywords, acronyms, and general terms) are consistent among records and consistent with common vocabularies. This category contains recommendations on employing metadata consistency checking tools that meet usability needs and generate reports with standards-based accuracy, precision, and uncertainty attributes provided in data granules. The purpose of metadata consistency checking is to ensure that the majority of NASA's Earth observing data file and collection-level metadata are adhering to NASA's best practices for proper data formatting and metadata standards; this not only promotes cross-platform and cross-user interoperability in reading and processing diverse types of data, but also enables more efficient query and extraction of vital metadata that supports more automated differentiation between unique data files and datasets

5.5.1.

DAACs: Employ metadata consistency checking tools that meet usability needs and generate reports with standards-based accuracy, precision, and uncertainty attributes provided in data granules.

- If applicable, provide software tools to data producers that can check for CF, ACDD [22], ISO [24,25], and UMM [21] metadata conformance. Examples of such tools are: online CF checker at Program for Climate Model Diagnosis and Intercomparison (PCMDI), Metadata Compliance Checker (MCC, https://podaac-uat.jpl.nasa.gov/mcc/) at PO.DAAC, ncdismember, Unidata Data Discovery Convention (UDDC) tool in the THREDDS data server, which checks and generates ACDD metadata reports and provides mapping to ISO 19115 [24,25] metadata elements, and the Common Metadata Repository (CMR) [6] Metadata Management Tool (MMT) for conformance checking against UMM.
- Using CF as known well-formed metadata, compare all DAAC HDF and NetCDF metadata to determine completeness, consistency and formatting conformity via comparison algorithm.
- Use completeness, consistency and formatting conformity metrics from metadata checking tool to provide a score system measuring the relative degree/extent of compatibility (internally used by DAACs only) that shows the relative completeness of metadata. [Such compatibility score would then help a DAAC determine priority and readiness for a collection of datasets to be integrated and tested with one or more interoperable tools/services. This compatibility score could also help compare the overall maturity of a dataset with similar datasets (i.e., comparing the maturity of datasets of a similar pedigree).]
- Document and communicate with data producers the completeness, consistency and formatting conformity of their metadata resulting from consistency checking tool.

Data Producers: Give recommendations on how data quality metadata attributes (e.g. those defined in CF, ISO 19115 & 19157 [24, 25, 26, 27], ACDD, and UMM) would be evaluated in a scoring framework.

- Collaborate with DAACs to set up an appropriate scoring framework to check for CF and ACDD metadata conformance.

5.6. Publicizing Quality Issues

Exposing quality issues associated with data products to the broad community of data users in a timely and efficient manner is critical. This category provides recommendations on possible approaches to capture and publicize known limitations, quality issues, and updates of data products.

5.6.1.

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.

5.6.2.

DAACs: Provide enough publicly available information with self-describing metadata and documentation such that the need for users to contact the DAACs is minimized.

- Capture "known issues" for particular regions or time intervals.
- Include documentation on how accuracy and uncertainty of datasets were determined.
- Facilitate the collection and integration of outlier information obtained from various datasets provided by the data producers.
- Provide online services (e.g., "Alerts", "FAQ", etc.) that allow data users to query and study the collected data quality information.
- Establish a checklist that may help DAACs and Data Producers for future data management and production of data quality information; such a checklist should be coordinated with ESDIS to ensure adherence to the latest standards and practices.

5.6.3.

DAACs: Collect and include documentation, provided by Data Producers, on how accuracy and uncertainty of datasets were determined.

- Request documentation from investigators and provide to users error and uncertainty estimates at each level of the processing chain (e.g., assimilation, binning and interpolation) with the product and/or include them in the ATBD or dataset user guide.
- Describe uncertainties associated with the interpolated values (e.g., different for gap filling procedure or if level 2 and 3 have similar resolution).

Data Producers: Provide all data with added quality flags for the areas that have potential limitations.

- Provide all available quality information with datasets to DAACs. Describe any caveats on the use of the data and clearly display the rights enabling the use and adoption of the data and of the data quality information.
- Document resampling/interpolation techniques used, describe the impact of the resampling technique used to process at all levels, and provide complete uncertainty estimates associated with the techniques used to the DAAC.
- Participate in formal process to help DAACs correctly document accuracy, precision and uncertainty, beginning when datasets are introduced at a provisional level.
- Convey the data quality information (e.g., extremes values and outliers) to the DAACs to provide to users to help ensure the integrity of the results being produced using the datasets.

5.6.4.

DAACs: Inform users as soon as possible when data are compromised (e.g., corrupted, quality does not meet specification, or bugs/errors found in data processing algorithms) and provide status updates when readily available. Alert PIs and/or Data Producers to issues that arise and/or reported by data users.

- Include special warnings in datasets with large known uncertainties (e.g., datasets or subsets thereof with large known uncertainties due to resampling/smoothing/ interpolation techniques).
- Implement plan to replace or permanently retire data that are catastrophically compromised, including documentation of the assessments which led to the resulting conclusions.
- Provide proper documentation outlining the limitations.

Data Producers: Provide information to DAACs promptly regarding any compromised datasets.

- Ensure all known issues discovered by the science teams are reported to the DAACs in a timely manner.
- Establish a well agreed upon definition of outlier (extreme values) for each product based on science understanding of the distribution of values for the parameters of interest.
- Identify outliers, as well as produce guidance, e.g., via documentation or online alert/flag, providing users useful data quality information such as 1) quantity and location of outliers, 2) magnitude of each outlier, and its ratio relative to the expected max/min of the data or some other well-defined statistical measure, and 3) origin of the problem.
- Provide spatially explicit systematic and random errors with conservative estimates.

5.7. Dataset Recommendations

It is recommended that data archives, such as DAACs, quickly provide standing recommendations to alternative datasets when a dataset has been retired or quarantined.

6. HIGH-PRIORITY RECOMMENDATIONS

6.1. Prioritization Process

The 93 recommendations mentioned in section 5 above, were prioritized individually by 12 DQWG members in the 2014-2105 period. Priority ratings voted onto all 93 recommendations and combined average priority scores can be found in Appendix E. Eventually, 12 recommendations with the highest priority were identified based on the consensus among the DQWG members. These recommendations are listed in Table 2 of subsection 6.3. The phases and categories to which each of the 12 recommendations is mapped are shown in the first and second columns, respectively, of Table 2.

6.2. Identification of LHF Recommendations

In addition to prioritizing recommendations, the DQWG identified four “Low-Hanging Fruits” (LHF) among the 12 high-priority recommendations from the point of view of their relative maturity and ease of implementation. LHF here means immediately actionable recommendations with low cost and high benefit. The LHF recommendations are defined as those that were considered to be relatively easy to implement because there may be existing instances of implementation within the EOSDIS environment, even though they may need to be adopted more broadly. The four LHF recommendations are highlighted in Table 2 using a gray color background.

6.3. Top 12 Recommendations (including 4 LHF recommendations)

Table 2 shows the 12 high-priority recommendations identified by DQWG members during the 2014-2015 period. The two columns on the right are worth noting. The column "Reco#" indicates the recommendation numbers in Appendix D. The column "No. of Recos Covered" shows how many recommendations in Appendix D are similar or related to the chosen ones. Numbers in this column reveal that while there are 12 high-priority recommendations selected, they actually cover 61 of the 93 recommendations from the workbook referenced here.

DQWG members further cast their votes to establish a set of 4 "Low-Hanging Fruit" recommendations (rows highlighted in "gray" in Table 2) among the 12 high-priority recommendations. As aforementioned, these four Low-Hanging Fruit recommendations were considered relatively easy to implement.

Table 2. The DQWG 12 high-priority recommendations (including 4 LHF recommendations highlighted with gray background)

Phase	Category	Recommendation – Data Systems	Recommendation – Science	Reco#	No. of Recos Covered
1, 2	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.	1	1
1, 2	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 dataset for each category.	2	4
1	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.		6	1
1	Quality of Input	*DAACs:* Request, from data producers, information about	*Data Producers:* Include information about correctness	28	9

	Datasets used in Generating Products	the contribution of the various input data that are used to process a higher-level product.	/uncertainty of input datasets used (e.g., land/ocean/region masks) along with products (e.g., sea ice product).		
2, 4	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).	16	19
1, 2, 3, 4	Quality Flags and Indicators	*DAACs:* Capture and disseminate 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.	11	3
1, 2, 4	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.	35	5
2, 3, 4	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.	10	1
2, 3, 4	Publicizing Quality Issues	*DAACs:* Provide enough publicly available information with documentation and/or self-describing metadata {derived from content delivered by Data Producers} such that the need for users to contact the DAACs is minimized.		11	3
1, 2, 4	Publicizing Quality Issues	*DAACs:* Include documentation on how accuracy and uncertainty of datasets were determined.	*Data Producers:* Provide all data with added quality and/or uncertainty flags for the areas that have potential limitations.	56	1
2, 3	Publicizing Quality Issues	*DAACs:* Inform users as soon as possible when data are compromised and provide status updates promptly.	*Data Producers:* Provide information to DAACs promptly regarding any compromised datasets.	62	16

3, 4	Dataset Recommen-dations	*DAACs:* Provide standing recommendations quickly to alternative datasets when a dataset has been retired or quarantined.		86	1

7. SOLUTIONS MASTER LIST

Two committees within the DQWG analyzed the 4 LHF recommendations from the point of view of implementation. Each of the two committees had the two objectives indicated below but from different perspectives: 1. *Science and Applications* and 2. *Data Systems Integration*.

- To identify and document solutions for capturing and describing data quality and for facilitating discovery of scientific data,
- To identify and document implementation strategies for selected recommendations for DAACs to capture and describe data quality, facilitate discovery, and enable use of scientific data.

To facilitate capturing the results of analysis, a tabular template was created. This template consists of the following entries, each pertaining to an existing implementation solution that can address one or more of the LHF recommendations: 1. Solution Name, 2. Summary of Solution, 3. Strategy, 4. Benefits of Proposed Implementation Solution, 5. Relevant LHF Recommendations, 6. Stakeholders, 7. Solution Class (Software/Technology or Standards/Documentation), 8. Operational Maturity Level, 9. Difficulty of Integration, 10. Difficulty of Implementation, 11. Name of Committee Member (subject matter expert advocating the solution), 12. Pertinent URL(s), 13. Actions and/or Resources Needed. The assessment rationale for Operational Maturity, Difficulty of Integration, and Difficulty of Implementation are described in the Solutions Master List (SML) page (https://wiki.earthdata.nasa.gov/x/2pASBg) [12]. The committees worked independently at first to identify the solutions, which were then integrated into a single table, resulting in a total of 26 solutions, as described in detail on the SML page.

8. IMPLEMENTATION STRATEGIES

The 26 solutions collected in the Solutions Master List were mapped into six generalized implementation strategies: 1. Facilitate DAAC-PI Communication, 2. Support Metadata Creation, 3. Support Metadata Validation, 4. Guide, Instruct and Disseminate, 5. User Services, and 6. Consolidate Quality Information Representation. Implementation strategies categorize implementation solutions and provide high-level guidance on approaches to improve the capturing, describing, discovery, and usage of data quality information. The following subsections provide a brief summary of each of the implementation strategies, which are meant to be broad and high-level. Details about these implementation strategies can be found in section 4.8 of ESDS-RFC-034 [14]

8.1. Facilitate DAAC-PI Communication

Effective and close communication between data archives (e.g., DAACs) and science teams/PIs is important to ensure effective exchange of thoughts and consolidation of ideas on describing, representing, and using quality information for data to be created by projects and to be archived. DAACs, upon receiving a mission, instrument, campaign or dataset assignment for archive and distribution, develop and assign staff to develop expertise and become knowledgeable about the measurement techniques, algorithms and science products. This strategy suggests appropriate mechanisms to be developed and leveraged to facilitate such communication throughout the data lifecycle.

8.2. Support Metadata Creation

This strategy points out the need for tools that can support both data providers and data curators at data archives to easily create and/or transform metadata at different levels (e.g., collection and granule) such that they conform to various common metadata standards, particularly those already endorsed by ESDIS (e.g., CF and ISO).

8.3. Support Metadata Validation

After data archives receive data/metadata from data providers or after data users access data/metadata from data archives, there is a need to validate the completeness of metadata information at different levels (e.g., collection and granule) as well as its conformance to multiple common metadata standards. Coming out of the validation can be reports and/or scores, which can help data archives to identify missing components and data users to evaluate quality of a data product.

8.4. Guide, Instruct and Disseminate

ESDIS and data archives (i.e. DAACs) should identify and adopt efficient and consistent ways to help data users access and understand data quality information (e.g., error sources, dataset limitations, and quality assessment) as these would address user questions about data quality and make user feedback about data quality available to user communities and science teams/PIs, if needed.

8.5. User Support Services

User support services are important not only for data users to get direct help from data experts on accessing, understanding, and using datasets, but also for data archives and producers to collect feedback from data users and identify issues of datasets based on real data user experience. Such feedback can be further shared with broader user communities and help improve the usage of datasets.

8.6. Consolidate Quality Information Representation

Given the fact that different datasets are distributed in different data archives, many users may need datasets from more than one archive. This strategy of “Consolidate Quality Information Representation” points out the importance of an efficient way to present and convey quality information to data users consistently across the archives.

9. PRIORITIZED RECOMMENDED IMPLEMENTATION ACTIONS

Table 2 above highlighted the LHF recommendations in gray background, as they apply to data systems and science. Based on the analyses described above, the DQWG has arrived a set of Prioritized Recommended Implementation Actions (PRIAs), which are summarized in priority order below. Details of these PRIAs can be found at the High-Priority Data Quality Recommendations for Data Producers and Distributors [14].

9.1. NASA Recommended Use of ISO Standard

NASA should provide appropriate documentation and guidance on how to employ attributes of the NASA implementation of the ISO 19157:2013 [26] and 19157-2:2016 [27] standards, once these are fully established. ISO 19115-2:2009 [25] is already broadly implemented across NASA's Earth Science Data and Information System (ESDIS), and while we are mindful of this from a metadata standards and completeness perspective, we place more emphasis on the quality-specific 19157 standards. It is to be noted that 19115-2 includes the bulk of the data quality elements described in 19157, and there are only a few differences which can easily be implemented as add-ons.

9.2. Help Users to Access and Understand Data Quality Information

ESDIS/DAACs should identify and adopt efficient and consistent ways to help data users access and understand data quality information (e.g., error sources, dataset limitations, and quality assessment) as these would address user questions about data quality and provide user feedbacks about data quality to user communities. Examples should be given showing how to apply the quality assessments from the point of view of science teams and/or PIs. ESDIS should identify different ways in which data quality information is currently being conveyed by various DAACs and consolidate these approaches into a consistent mechanism for homogenous, efficient dissemination that results in a more optimal cross-DAAC user experience of data discovery and extraction of data and information about the data. One specific implementation example could be collection level quality information made available in standardized, online guide documents.

9.3. Metadata Authoring and Validating Tools

Metadata is important in conveying information about data, and as such, ESDIS/DAACs should adopt, consolidate, enhance, and/or create consistent and easy-to-use metadata authoring and validating tools to assist DAACs, data producers, and data users through the development and validation of richer metadata at collection and granule levels. These tools should also assist data users in validating the metadata. Specifically, these tools should: 1. Support multiple standards, including Unified Metadata Model (UMM), ISO 19115/19157 (as implemented by NASA; see 9.1 above), and CF; 2. Collect minimum required CMR and standard-specific metadata; 3. Support population of data quality fields (e.g., the DIF quality field); and 4. Support import/export and translation of CMR metadata with standards-based importable/exportable formats such as XML and JSON.

9.4. Develop Tools to Help Users to Leverage Data Quality Information

ESDIS/DAACs should develop tools to help data users easily use data quality information in their research, such as finding, accessing, and processing data based on user-defined quality criteria. For example, all granule level quality metadata should be accessible through clients such as NASA ESDIS' Earthdata Search [5] and Worldview [19], with the highest-level quality description (e.g., good/bad) prominently displayed alongside granule search results or as a layer in visualization tools. Users should also have access to detailed granule level quality information (flags, etc.) as an additional filtering mechanism for subsetting and extraction of quality-specified data.

9.5. Finer-levels of Metadata

ESDIS and NASA Earth science programs should support the effort to research and determine required data quality metadata elements at finer levels (e.g., file, parameter and pixel). Existing quality flag standards for flag attributes should be recommended for datasets where it is appropriate. Data formatted using either NetCDF (Version 4 or classic) or HDF (HDF4, HDF5, or HDF-EOSx) should use CF quality flags.

9.6. Science Advice

NASA Earth Science Research Program should ensure that each funded project (e.g., MEaSUREs, ACCESS) has a science review board/team to advise data producers on quality and usability of the dataset as it is being developed. Existing review boards (e.g., DAAC User Working Groups) and teams (e.g., NASA science and Cal/Val teams) could be leveraged in this regard but should have oversight to ensure these boards/teams are fulfilling their expectations.

9.7. Facilitate DAAC-PI Communication

ESDIS should develop and/or establish mechanisms that facilitate communication between DAACs and science teams/PIs to more effectively exchange thoughts and consolidate ideas on describing, representing, and using quality information for data to be created by projects and to be archived at DAACs. Further, the NASA Earth Science Research Program should set policies to facilitate such communication.

9.8. Data Quality Best Practices

ESDIS and DAACs should provide guidance and information on representing data quality as part of data management best practices for data producers to use when developing data and metadata. This should include ensuring creation of dataset guide documents for users contain adequate information about data quality and how to use it. An example of more general data best practices guidance has been made available by the PO.DAAC [15].

10. RELATED ACTIVITIES

There have been several international activities pertaining to data quality. Of these, the most relevant in the context of NASA’s Earth observation data are those arising from the Group on Earth Observations (GEO), an international partnership of national governments and organizations [7]. GEO has identified the need for an internationally harmonized strategy to enable interoperability and acceptance of quality of Earth observation data at "face value". In response to this, the Committee on Earth Observing Satellites (CEOS) established and endorsed the Quality Assurance framework for Earth Observation (QA4EO). Following four international workshops (in 2007, 2008, 2009 and 2011), a framework and ten key guidelines were established [16]. The fundamental principle of QA4EO is that “all Earth Observation Data and derived products shall have associated with them a fully traceable indicator of their quality. The QA4EO states that “A Quality Indicator (QI) shall provide sufficient information to allow all users to readily evaluate the 'fitness for purpose' of the data or derived product", and that "A QI shall be based on a documented and quantifiable assessment of evidence demonstrating the level of traceability to internationally agreed (where possible SI) reference standards." QA4EO provides key guidelines to adhere to the above principle in the following areas:

- Establishing a Quality Indicator for a satellite sensor derived data product
- Content of a documentary procedure to meet the QA requirements of GEO
- "Reference standards" in support of QA requirements of GEO
- Comparisons – organization, operation and analysis to establish measurement equivalence to underpin the QA requirements of GEO
- Establishing validated models, algorithms and software to underpin the QA requirements of GEO
- Expression of uncertainty of measurement, and
- Establishing quantitative evidence of traceability to underpin the QA requirements of GEO

Also developed by the Group on Earth Observations (GEO), the Group on Earth Observations System of Systems (GEOSS) Data Management Principles (DMP) Implementation Guidelines (IG) provide recommendations for improving practices to manage Earth science data [8]. Developing such recommendations for improvement contributes to data stewardship practices that enable the current and future use of many kinds of Earth science data. The recommendations offer guidance for implementing ten DMP and are organized into five topical areas, including Discoverability, Accessibility, Usability, Preservation, and Curation. For each of the DMP, the GEOSS DMP IG summarizes the DMP, introduces relevant terms, explains the principle, provides guidance on implementation of the principle along with examples, suggests metrics for measuring adherence, and describes the resource implications for implementation. Especially relevant to the DQWG are the recommendations pertaining to various issues described within several topical areas of the GEOSS DMP IG and, in particular, under usability, DMP 6: Data Quality.

While the DQWG has not explicitly mapped its recommendations to the above guidelines from QA4EO and the GEOSS DMP IG, they are clearly aimed at meeting the fundamental principle expressed by QA4EO and addressing issues revealed by use cases specific to the EOSDIS environment.

11. CONCLUSION

The DQWG was established in 2014 as one of the NASA ESDSWGs that have been organized since 2004 to pursue various Earth science topics. Efforts of the DQWG and the comprehensive set of recommendations that have been offered by the DQWG, from 2014 through 2017, for improving data quality practices have been described. These include recommendations for improving data quality practices and the methodologies utilized for developing the recommendations.

The DQWG initially developed 16 data quality use cases for remotely sensed Earth science data and, based on the use cases, identified four primary focus areas for data quality, including Accuracy, Precision and Uncertainty; Distinguishability; Applicability; and Usability. Analysis of the use cases within these four focus areas produced 93 high-level recommendations for improving data quality within seven categories. Considering the large number of recommendations, the high-level recommendations were subsequently assessed to identify 12 high-priority data quality recommendations, of which four were identified as “low hanging fruit” that could be implemented readily. Twenty-six solutions were identified as opportunities for implementing the four “low hanging fruit” recommendations and have been mapped into six data quality implementation strategies.

Four phases of the data quality lifecycle were identified and defined. The phases include capturing, describing, facilitating discovery, and enabling use of data quality information. Six implementation strategies were produced for the data quality lifecycle phases that apply to either data producers or data distributors. In addition, 8 implementation recommendations for improving data quality practices were identified. The recommendations of the DQWG also build on recommendations related to data quality that have been developed by other organizations within the Earth science community, including the QA4EO and the GEOSS DMP IG. While the recommendations for improving data quality practices that have been developed by the DQWG and other organizations may be primarily applicable to Earth science data, representatives from other fields of inquiry are also encouraged to read them to identify opportunities for improving data quality practices within their respective disciplines.

Data quality involves a large number of topics. Many standards and community practices are evolving rapidly. Due to the limited resources available, some important topics were not addressed by the DQWG. Some topics worthy of further investigation and discussion include: developing best practices to leverage quality-related elements defined in ISO standards (i.e. ISO 19115-2 and 19157); discussion of usability and interoperability as well as their inter-connection; investigating how UMM can be enhanced to better incorporate existing and widely-adopted community conventions, such as CF; and analysis of use cases with in situ data and identification of issues with representing and conveying data quality information.

12. ACKNOWLEDGEMENTS

Support by the NASA ESDIS project for the work of the Earth Science Data System Working Group on Data Quality is appreciated. Additional support was provided by the Earth Science Information Partners (ESIP) through their Information Quality Cluster (IQC) to further evaluate use cases, recommendations and proposed solutions in cooperation with the DQWG as published by Ramapriyan et al. 2017 [18]. We thank reviewers for their constructive comments that have improved this document.

13. REFERENCES

13.1. INFORMATIVE REFERENCES

[1] Arctur, D. 2011. “Use Cases for Geospatial Interoperability, ICAN-Great Lakes Workshop on Coastal Web Atlas”, http://portal.opengeospatial.org/files/?artifact_id=40800, September 13-15, 2011.

[2] Bagwell R., Ramapriyan H. K., Ding F., Downs R. R., 2015. “Data Quality Working Group – Applicability Subgroup”, NASA Earth Science Data System Working Groups (ESDSWG) Meeting. Goddard Space Flight Center, Greenbelt, MD, March 24-26, 2015.

[3] Downs, R. R. 2015. The Usability Subgroup of the NASA Earth Science Data System Working Group (ESDSWG) on Data Quality. Summer 2015 Meeting of the Federation of Earth Science Information Partners (ESIP). Pacific Grove, CA, July 14-17, 2015. http://commons.esipfed.org/sites/default/files/Information%20Quality%20Cluster%20-%20Downs_ESIP%2020150717.pdf. Accessed February 2, 2018.

[4] Downs, R. R., Peng, G., Wei, Y., Ramapriyan, H. K., and Moroni, D. F. 2015. Enabling the Usability of Earth Science Data Products and Services by Evaluating, Describing, and Improving Data Quality throughout the Data Lifecycle. 2015 Fall American Geophysical Union (AGU) Meeting. San Francisco, CA, December 14-18, 2015.

[5] Earthdata Search, NASA Earthdata web site, 2018: https://search.earthdata.nasa.gov/

[6] ESDIS Common Metadata Repository, NASA, 2017: https://earthdata.nasa.gov/about/science-system-description/eosdis-components/common-metadata-repository

[7] Group on Earth Observations, 2018. http://earthobservations.org/. Accessed February 2, 2018.

[8] Group on Earth Observations, 2015. “Data Management Principles Implementation Guidelines.” https://www.earthobservations.org/documents/geo_xii/GEO-XII_10_Data%20Management%20Principles%20Implementation%20Guidelines.pdf. Accessed February 2, 2018.

[9] JCGM 100: 2008. Evaluation of measurement data — Guide to the expression of uncertainty in measurement.

[10] Moroni D., Armstrong E., Bennett S. D., and Bagwell R., 2015. “Data Quality Working Group – Distinguishability Subgroup”, NASA Earth Science Data System Working Groups (ESDSWG) Meeting. Goddard Space Flight Center, Greenbelt, MD, March 24-26, 2015.

[11] NASA, NASA Strategic Plan, 2014, https://www.nasa.gov/sites/default/files/files/FY2014_NASA_SP_508c.pdf.

[12] NASA Earth Science Data Quality Working Group (DQWG), 2018: Solutions Master List for Earth Science Data Quality, Earthdata Wiki, https://wiki.earthdata.nasa.gov/x/2pASBg.

[13] NASA Earth Science Data Quality Working Group, 2019: Data Management Plan Template for Data Producers, ESDIS Standards Office Technical Note ESDS-RFC-032. https://earthdata.nasa.gov/user-resources/standards-and-references/templates-for-nasa-data-management-plans.

[14] NASA Earth Science Data Quality Working Group, 2019: High-Priority Data Quality Recommendations for Data Producers and Distributors, ESDIS Standards Office Technical

Note ESDS-RFC-034. https://earthdata.nasa.gov/user-resources/standards-and-references/recommendations-from-the-data-quality-working-group.

[15] PO.DAAC Data Management Best Practices, Jet Propulsion Laboratory, California Institute of Technology, Pasadena CA, 2018: https://podaac.jpl.nasa.gov/PO.DAAC_DataManagementPractices

[16] QA4EO task team. 2010. "A Quality Assurance Framework for Earth Observation: Principles", Version 4.0, January 14, 2010. http://qa4eo.org/docs/QA4EO_Principles_v4.0.pdf, Accessed February 2, 2018.

[17] Ramapriyan, H. K., Bennett, S.D., DiMiceli C., Guillevic P., Moroni D., Scott D., Shen S., Shie C-L., Simard M., Wei Y., 2015. "Data Quality Working Group – Accuracy, Precision and Uncertainty (APU) Subgroup", NASA Earth Science Data System Working Groups (ESDSWG) Meeting. Goddard Space Flight Center, Greenbelt, MD, March 24-26, 2015.

[18] Ramapriyan H. K., Peng G., Moroni D., and Shie C-L., 2017 "Ensuring and Improving Information Quality for Earth Science Data and Products", D-Lib Magazine, July/August 2017, DOI: https://doi.org/10.1045/july2017-ramapriyan.

[19] Worldview, NASA Earthdata web site, 2018: https://worldview.earthdata.nasa.gov

13.2. RELEVANT METADATA STANDARDS

[20] Eaton, B. J. Gregory, B. Drach, K. Taylor, S. Hankin, et al: NetCDF Climate and Forecast (CF) Metadata Conventions, Version 1.7, http://cfconventions.org/Data/cf-conventions/cf-conventions-1.7/cf-conventions.pdf

[21] ESDIS Unified Metadata Model, NASA, 2017: https://earthdata.nasa.gov/about/science-system-description/eosdis-components/common-metadata-repository/unified-metadata-model-umm

[22] ESIP, Attribute Convention for Data Discovery, Version 1.3, 2017: http://wiki.esipfed.org/index.php/Attribute_Convention_for_Data_Discovery

[23] International Organization for Standards (ISO), 1994: Accuracy (trueness and precision) of measurement methods and results -- Part 1: General principles and definitions. ISO 5725-1:1994. https://www.iso.org/standard/11833.html

[24] International Organization for Standards (ISO), 2009: Geographic information -- Metadata -- Part 1: Fundamentals, ISO 19115-1:2014, https://www.iso.org/standard/53798.html

[25] International Organization for Standards (ISO), 2009: Geographic information -- Metadata -- Part 2: Extensions for imagery and gridded data, ISO 19115-2:2009, https://www.iso.org/standard/39229.html

[26] International Organization for Standards (ISO), 2013: Geographic information – Data Quality, ISO 19157:2013, https://www.iso.org/standard/32575.html

[27] International Organization for Standards (ISO), 2016: Geographic information – Data Quality – Part 2: XML schema implementation, ISO/TS 19157-2:2016, https://www.iso.org/standard/66197.html

[28] International Organization for Standards (ISO), 2018: Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts, https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en.

14. AUTHORS AND CONTRIBUTORS

Writing Authors

Yaxing Wei (ORNL DAAC)
Hampapuram “Rama” Ramapriyan (SSAI/GSFC ESDIS)
Robert R. Downs (SEDAC)
Chung-Lin Shie (GES DISC)
Zhong Liu (GES DISC)
David Moroni (JPL / PO.DAAC)
Ted Habermann (The HDF Group)
Siri Jodha Khalsa (NSIDC)
Byron Peters (SSAI/ESDIS)

DQWG Members

[2014-2015]

David Moroni (JPL / PO.DAAC, Chair), Hampapuram Ramapriyan (SSAI, GSFC/ESDIS, Co-Chair), Ed Armstrong (JPL/PO.DAAC), Ross Bagwell (ESDIS), Stacie Doman Bennett (LPDAAC), Charlene DiMiceli (UMD), Feng Ding, Robert R. Downs (SEDAC), Ted Habermann (The HDF Group), Pierre Guillevic (UMD), Steve Olding (ESDIS), Bill Rossow (GSFC), Donna Scott (NSIDC), Suhung Shen (GSFC), Chung-Lin Shie (GES DISC), Marc Simard (JPL), Gilberto Vicente (GSFC), Yaxing Wei (ORNL DAAC)

[2015-2016]

David Moroni (JPL / PO.DAAC, Chair), Hampapuram Ramapriyan (SSAI, GSFC/ESDIS, Co-Chair), Ed Armstrong (JPL/PO.DAAC), Ross Bagwell (ESDIS), Stacie Doman Bennett (LPDAAC), Charlene DiMiceli (UMD), Robert R. Downs (SEDAC), Pierre Guillevic (UMD), Peter Hall (SSAI/GSFC), Molly Hardman (NSIDC), George Huffman (GSFC), Siri Jodha Khalsa (NSIDC), Tiffany Matthews (ASDC), Steve Olding (ESDIS), Donna Scott (NSIDC), Marc Simard (JPL), Chung-Lin Shie (GES DISC), Yaxing Wei (ORNL DAAC)

[2016-2017]

David Moroni (JPL / PO.DAAC, Chair), Hampapuram Ramapriyan (SSAI, GSFC/ESDIS, Co-Chair), Ed Armstrong (JPL/PO.DAAC), Ross Bagwell (SSAI/ESDIS), Stacie Doman Bennett (LPDAAC), Michelle Butler (NCSA), Charlene DiMiceli (UMD), Larry Di Girolamo (UIUC), Robert R. Downs (SEDAC), Yonsook Enloe (SSAI/ESO), Pierre Guillevic (UMD), Ted Habermann (The HDF Group), Peter Hall (SSAI/GSFC), Molly Hardman (NSIDC), Beth Huffer (LaRC), George Huffman (GSFC), Siri Jodha Khalsa (NSIDC), Shannon Leslie (NSIDC), Andrew Mitchell (ESDIS), Steve Olding (SSAI/ESDIS), Byron Peters (SSAI/ESDIS), Donna Scott (NSIDC), Deborah Smith (GHRC), Chung-Lin Shie (GES DISC), Yaxing Wei (ORNL DAAC), Greg Yetman (SEDAC)

[2017-2018]

Yaxing Wei (ORNL DAAC, Chair), David Moroni (JPL/PO.DAAC, Chair), Hampapuram Ramapriyan (SSAI, GSFC/ESDIS, Co-Chair), Ed Armstrong (JPL/PO.DAAC), Ross Bagwell (ESDIS), Michelle Butler (UIUC), Charlene DiMiceli (U. of Maryland), Feng Ding (GSFC/GES DISC), Robert R. Downs (SEDAC), Carolyn Gacke (LP.DAAC), Larry Di Girolamo (UIUC), Scott Gluck (JPL), Pierre Guillevic (U. of Maryland), Lindsey Harriman (LP.DAAC), Ted Habermann (The HDF Group), Beth Huffer (ASDC), George Huffman (NASA Goddard Space Flight Center), Wenhao Li (JPL), Zhong Liu (GSFC/GES DISC), Sydney Neeley (LP.DAAC), Steve Olding (ESDIS), Byron Peters (SSAI/ESDIS), Donna Scott (NSIDC), Suhung Shen (GSFC/GES DISC), Chung-Lin Shie (GSFC/GES DISC), Deborah Smith (GHRC DAAC), James Tilton (GSFC)

Use Cases Contributors

Ed Armstrong (JPL/PO.DAAC), Stacie Doman Bennett (LPDAAC), Lisa Booker (NSIDC), Chris Derksen (U. of Waterloo), Feng Ding (GSFC), Jessica Hausman (JPL), Nathan Kurtz (GSFC), Christopher Lynnes (ESDIS), David Moroni (JPL), Hampapuram Ramapriyan (SSAI, GSFC/ESDIS), Marc Simard (JPL), Vardis Tsontos (JPL)

APPENDIX A – GLOSSARY OF ACRONYMS AND ABBREVIATIONS

ACCESS	Advancing Collaborative Connections for Earth System Science
ACDD	Attribute Convention for Data Discovery (synonymous with UDDC)
AIRS	Atmospheric Infrared Sounder
AIST	Applied Information Systems Technology
APU	Accuracy, precision and uncertainty
ATBD	Algorithm Theoretical Basis Document
CEOS	Committee on Earth Observation Satellites
CF	Climate and Forecast (metadata conventions)
CMR	Common Metadata Repository
CSDGM	Content Standard for Digital Geospatial Metadata
DAAC	Distributed Active Archive Center
DMAS	Data Management & Archive System (at PO.DAAC)
DMP	Data Management Principles
DQWG	Data Quality Working Group
ECHO	EOS Clearing House
EOSDIS	Earth Observing System Data and Information System
ESD	Earth Science Division
ESDIS	Earth Science Data and Information System (Project)
ESDS	Earth Science Data System
ESDSWG	Earth Science Data Systems Working Groups
ESIP	Earth Science Information Partners
FGDC	Federal Geographic Data Committee
GEO	Group on Earth Observations

GEOSS	Global Earth Observation System of Systems
GES DISC	Goddard Earth Sciences Data Information Services Center
GFSAD	Global Food Security Analysis & Support Data
GHRSST	Group for High Resolution Sea Surface Temperature
GSFC	Goddard Space Flight Center
HDF	Hierarchical Data Format
HQ	Headquarters
ISO	International Organization for Standardization
JPL	Jet Propulsion Laboratory
JSON	JavaScript Object Notation
LaRC	Langley Research Center
LHF	Low-Hanging Fruit
LP DAAC	Land Processes Distributed Active Archive Center
MCC	Metadata Compliance Checker
MEaSUREs	Making Earth System Data Records for Use in Research Environments
MMT	Metadata Management Tool
MODIS	Moderate Resolution Imaging Spectroradiometer
NASA	National Aeronautics and Space Administration
NCSA	National Center for Supercomputing Applications
NetCDF	Network Common Data Form
NSIDC	National Snow and Ice Data Center
OGC	Open Geospatial Consortium
OPeNDAP	Open-source Project for a Network Data Access Protocol
ORNL DAAC	Oak Ridge National Laboratory Distributed Active Archive Center

PCMDI	Program for Climate Model Diagnosis and Inter-comparison
PI	Principal Investigator
PO.DAAC	Physical Oceanography Distributed Active Archive Center
QA4EO	Quality Assurance Framework for Earth Observation
SEDAC	NASA Socioeconomic Data and Applications Center
SIPS	Science Investigator-led Processing System
SMAP	Soil Moisture Active Passive
SSAI	Science Systems and Applications, Inc.
SWEET	Semantic Web for Earth and Environmental Terminology
THREDDS	Thematic Real-Time Environmental Distributed Data Services
UDDC	Unidata Dataset Discovery Conventions (synonymous with ACDD)
UIUC	University of Illinois Urbana-Champaign
UMD	The University of Maryland
UMM	Unified Metadata Model
URS	User Registration System
USGS	United States Geological Survey
USO	User Services Office

APPENDIX B – DATA QUALITY USE CASE SURVEY FORM

Field Name	**Field Description**
Use Case Title *	Please write a descriptive title for your use case
Point of Contact *	Name of person reporting this use case
Email Address for Point of Contact *	Email of person reporting this use case
Affiliation of Point of Contact *	Please select from the list. If your affiliation is not listed check "other" and type it in. [*ASF SDC; CDDIS; GHRC; GES DISC; LP DAAC; LaRC ASDC; MODAPS LAADS; NSIDC; ORNL DAAC; OB.DAAC; PO.DAAC; SEDAC; MEaSUREs; ACCESS; AIST; ESDIS; NASA - Other; NOAA NCEI - Weather and Climate; NOAA NCEI - Coasts, Oceans, and Geophysics; NOAA - Other; HDF Group; Other*]
Use Case Narrative: Goal and Context *	Enter the high level description of this use case, including the goal of the submitter (who may act in proxy of a data user, data provider, or data producer). Please state a singular goal. Needs which stem from the goal may be plural.
Domain of Interest *	Please select one or more of the following domains of interest relevant to the use case. [*Atmosphere; Biology; Climate; Computer Science; Cryosphere; Geomagnetics; Geographical Information Systems; Geology; Ecology; Heliophysics; Hydrology; Informatics; Ionosphere; Land; Ocean; Radiative Transfer; Solid Earth; All of the above; Other*]
Professional Domain of User *	Please select one of the following options describing the professional domain of the user driving the use case. [*Data Management; Education/Outreach; Emergency Management; Engineer; Graduate Student; Military; Operational Forecaster; Resource Management; Risk Management; Scientist/Researcher; Undergraduate Student; Other*]
Primary User-Stakeholder Relationship *	Identify the relationship between the user type (human or machine), and the stakeholder (data producer (e.g. mission-project) or data provider (e.g. DAAC)). The arrow points from the user (from the perspective of the submitter) to the stakeholder, which indicates that the user is driving a need that can potentially be met by the stakeholder. Select one of the following options. • Human User -> Data Center/Stakeholder • Machine User -> Data Center/Stakeholder • Human User -> Project/Stakeholder • Machine User -> Project/Stakeholder • Human User -> Data Producer/Stakeholder

	• Machine User -> Data Producer/Stakeholder • Human User -> Data Distributor/Stakeholder • Machine User -> Data Distributor/Stakeholder • Other (provide free form descriptions of user and stakeholder)
Secondary User-Stakeholder Relationship	If there is more than one user-stakeholder relationship relevant to this use case, please specify the secondary relationship using the previously stated methodology. Options are the same as those for Primary User-Stakeholder Relationship.
Primary Scope *	Choose the primary scope from one of the above categories and provide a one to two sentence rationale for selecting this primary scope. [*Qualitative-Science; Quantitative-Science; Qualitative-Product; Quantitative-Product*] Overlap or interdependency between a “science” scope and a “product” scope may exist. Select the scope type which contains the greatest relevance. “Secondary Scope” may be used to capture a less relevant scope if needed. **Scope Definitions:** • **Qualitative-Science:** Any descriptive or procedural attribute or enhancement (e.g., quality flags, sampling techniques, assimilation techniques) which results in a substantial impact to the integrity of scientific research and the overall quality of scientific understanding and may also improve or better characterize the accuracy, precision, uncertainty, applicability, distinguishability and usability of the data using the specified data. • **Quantitative-Science:** Any quantified scientific result (e.g., bias and uncertainty characterization, spectral analysis, trend analysis) or metric (e.g., accuracy, precision, effective spatial resolution) which significantly impacts the integrity of scientific research and the overall quality of scientific understanding and may also improve or better characterize the accuracy, precision, uncertainty, applicability, distinguishability and usability of the data using the specified data. • **Qualitative-Product:** Any descriptive or procedural attribute or enhancement (e.g., documentation, metadata, search and discovery, known issues, traceability, lineage) which results in an improved characterization, informatic standardization, and proliferation of accuracy, precision, uncertainty, applicability, distinguishability, and/or usability of a complete data product. • **Quantitative-Product:** Any numerically-derived attribute or enhancement (e.g., checksum validation, spatial/temporal resolution, time series gap analysis, latency validation), which provides improved characterization and informatic standardization toward the accuracy, precision, applicability, distinguishability, integrity and/or usability of a complete data product.

	- **"Science" perspective:** an evaluative reference point in which the use case provides the most significant relevance to the integrity of the scientific research or science application. The underlying considerations include: accuracy, precision, uncertainty, applicability, distinguishability, and/or usability of the data. - **"Product" perspective:** an evaluative reference point in which the use cases provides the most significant relevance to improved characterization, informatic standardization and proliferation of the relevant data quality elements relating to the integrity of a complete data product or dataset. The underlying considerations include: accuracy, precision, uncertainty, applicability, distinguishability, and/or usability of the data.
Primary Scope Rationale	If you think it is needed, please provide your rationale for why this scope is most relevant.
Secondary Scope	If this use case has multiple relevant scopes, choose the secondary scope from one of the above categories and provide rationale for selecting this primary scope. Indicate areas of crosscutting overlap and/or interdependencies between the primary and secondary scope. [*Qualitative-Science; Quantitative-Science; Qualitative-Product; Quantitative-Product*]
Secondary Scope Rationale	If you think it is needed, please provide your rationale for why this scope is also relevant.
Use Case Chronology *	*Please include a chronology of known steps taken, from the perspective of the use case submitter, throughout this use case. If this use case stems from a real experience that required interaction with either the data provider (e.g. DAAC) or a data producer (e.g. mission-project like MEaSUREs PI), please also provide a chronological listing of events that took place during that interaction.* 1. [What happens first?] (e.g. DAACs user services was contacted) 2. [What happens second?] (e.g. DAAC user services communicated needs to appropariate DAAC staff to review) 3. [What happens third?] (e.g. DAAC decided a change to product should be implemented) 4. [And so on…] (e.g. User notified of fulfilled request)
Success Criteria *	*In the event a deliverable might exist to successfully satisfy your use case, please list and summarize the criterion which makes the deliverable a success. Each successive criterion will be considered a requirement for success.* 1. [The first step of the system or workflow process is to return this output/result to satisfy a specified user input/request.] 2. [The second step of the system or workflow process is to return this output/result to satisfy a specified user input/request.]

	3. [And so on…]
Data Quality Keywords *	Please select any number of relevant data quality key words which best relate to the goal, needs, and solutions to this use case. Please reference the existing keyword lexicon available here: **1. algorithm:** relating to the processing algorithm used to generate the geophysical data. **2. accessibility:** with respect to how the accessibility of the data/metadata and/or data quality information can influence the quality of the data product. **3. bounding box:** relating to the regional and/or temporal constraints of a dataset or data granule. **4. calibration:** relating to the methods and/or results of the geophysical data being calibrated. **5. cross-calibration:** relating to the methods and/or results of the inter-calibration of geophysical data with two or more remote sensing platforms. **6. data sampling:** relating to the sampling techniques used to sample a single dataset or multiple datasets (e.g., daily running mean, Gaussian weighting, median filter, low-pass filter, etc…) **7. derivatives:** relating to concerns regarding derivative quantities of a dataset (i.e., divergence, curl, gradients, etc…). **8. documentation:** relating to a component of documentation. **9. extraction:** any mode of retrieving a data quality attribute or utilization of a data quality attribute to extract specific data. **10. filtering:** pertaining to the utilization of a data quality attribute as a way of “filtering” datasets that match a specified data quality criteria **11. flags:** pertaining to any element of workflow that involves the utilization of data quality flags. **12. instrument sampling:** related to any artifacts as a result of instrument sampling characteristics, such as swath width, measurement footprint, sampling frequency, etc… **13. interoperability:** pertaining to the utilization of any interoperable services or architecture(s) which may by leveraged. **14. metadata:** the expression or utilization of metadata. **15. metrics:** any quantifiable expression(s) attributed to data quality. **16. missing data:** pertaining to data dropouts and situations involving dissemination of data gaps or identifying missing data. **17. reporting:** pertaining to an event or workflow which may invoke the generation of a report or involve a method of reporting. **18. search:** any mode of searching. **19. spatial resolution:** relating to the grid-resolution or the “effective” spatial resolution of a dataset or data granule. **20. standardization:** relating to the need to develop or incorporate standards to better describe and/or improve the data/metadata quality. **21. temporal resolution:** relating to the time-step or temporal repeat of a dataset or data granule. **22. web services:** relating to any service that interfaces directly between the data/metadata and a web browser. **23. workflow:** relating to any a process (human or machine) involving multiple steps to achieve a desired goal.

* Required Fields

APPENDIX C - DATA QUALITY USE CASES

C.1 Use Case 1

Use Case Title	AIRS Quality Indicator Recommendations
Point of Contact	Feng Ding
Email	Feng.Ding@nasa.gov
Use Case Narrative	There may be potential confusion with interpreting the quality indicators within the AIRS science datasets. The reason for this is because the quality indicators are very broad and the guidance that is currently provided for determining applicability of the quality indicators is not yet available. The goal is for the user to extract the most applicable quality indicators corresponding to the user's needs. The user's needs can be very broad as depicted by the "Domain of Interest" noted below.
Domain of Interest	Atmosphere, Climate, Geographical Information Systems, Hydrology, Informatics, Radiative Transfer, Weather
Professional Domain of User	Scientist/Researcher
Primary User/Stakeholder Relationship	Human User -> DAAC/Stakeholder
Secondary User-Stakeholder Relationship	N/A
Primary Scope/Rationale	Qualitative-Product / The reason for this is because most if not all of the "quantitative" quality indicators have already been made available. The issue that is unresolved is how we disseminate guidance regarding how to use these already available quality indicators.
Secondary Scope/Rationale	Qualitative-Science / The reason for this is because the quality indicators that have already been made available needs more improved description which would ultimately have a significant impact on the integrity of the science being produced by this data. No further numerical derivations are needed, but we do need better guidance for the user community.
Use Case Chronology:	1. The user attempts to access AIRS science data products from the GES DISC and informs the user services of difficulty in determining which quality indicators (e.g., flags, errors, uncertainties) are most important for their research needs. 2. The user services representative informs the AIRS science team of the issue and tries to determine a solution.

	3. The AIRS science team provides additional guidance to the user services team. 4. The user service team updates documentation and responds with an answer to the user.
Success Criteria	1. Successful relay of information between the data user, user services, and the AIRS science team. 2. Successful capture of additional guidance and recommendations from the AIRS science team on how to utilize specific quality indicators for specific user research needs. 3. Updated documentation that provides a linkage between user needs and specific quality indicators being made publicly available.
Data Quality Keywords	accessibility , data sampling, documentation, filtering, flags, metadata, metrics

C.2 Use Case 2

Use Case Title	Aquarius Salinity Data Quality Issue Noted in Coastal Region
Point of Contact	David Moroni
Email	David.F.Moroni@jpl.nasa.gov
Use Case Narrative	A user of Aquarius Salinity data has done comparisons between buoy data and sea surface salinity from Aquarius in coastal areas. Differences in some cases are quite large prompting the user to ask questions about data quality. The user has made suggestions on how to better implement data quality information in the file structure.
Domain of Interest	Hydrology, Ocean
Professional Domain of User	Scientist/Researcher
Primary User/Stakeholder Relationship	Human User -> DAAC/Stakeholder
Secondary User-Stakeholder Relationship	Human User -> Mission-Project/Stakeholder
Primary Scope/Rationale	Quantitative-Product / Rationale: User, specifically in a level 3 gridded product, would like information on a quality flag implemented that allows for easy interpretation of the quality of sea surface salinity at a given pixel.
Secondary Scope/Rationale	Qualitative-Science / Rationale: Improved quality flags, as indicated by the Primary Scope, would result in substantial impacts to the quality of

	scientific results, specifically when comparing sea surface salinity between Aquarius and buoys.
Use Case Chronology:	1. User directly contacts the DAAC regarding the data quality concern. 2. DAAC representative contacts the Aquarius Science Team regarding the issue. 3. The Aquarius Science Team asks the DAAC to provide more details such as graphics or diagrams describing the issue. 4. The DAAC provides the Aquarius Science Team with the appropriate info. 5. The Aquarius Science Team vets the data quality issue and makes a recommendation on how to move forward.
Success Criteria	1. Successful relay of user-reported data quality concerns from the DAAC to the data producer. 2. Proper vetting by the science team of the data quality issue. 3. A solution is reached by the science team on how to account for this issue, such as proper flagging for data at the coast or improved documentation warning users not to use this data for coastal applications. 4. The user is made aware of the outcome.
Data Quality Keywords	algorithm, bounding box, flags

C.3 Use Case 3

Use Case Title	Data Quality Filtering
Point of Contact	Ed Armstrong
Email	edward.m.armstrong@jpl.nasa.gov
Use Case Narrative	A user wishes to quality filter, or have his software program quality filter geophysical satellite data values in a "one step" process. That is, they wish to invoke a service to extract data values from a granule and at the same time apply specific quality filtering levels or flags (excellent, good, bad).
Domain of Interest	Atmosphere, Biology, Climate, Computer Science, Cryosphere, Geomagnetics, Geographical, Geology, Ecology, Heliophysics, Hydrology, Informatics, Ionosphere, Land, Ocean, Radiative Transfer, Solid Earth
Professional Domain of User	Scientist/Researcher
Primary User/Stakeholder	Human User -> Mission-Project / Stakeholder

Relationship	
Secondary User-Stakeholder Relationship	
Primary Scope/Rationale	Quantitative-Science / Rationale: The primary goal here is to enable the user to extract data values that correspond to a given set of quantifiable data quality criteria, such as confidence levels or any quality flag derived from a quantifiable metric (e.g., cloud, rain, or ice contamination). This also would help to ensure that the science being produced by the data is connected to a set of specific quality criteria.
Secondary Scope/Rationale	Quantitative-Product / TBD
Use Case Chronology:	1. User identifies satellite data of interest. 2. User writes a program to access the data. 3. User leverages a data access service to extract and quality filter data in one step (one system call) within program.
Success Criteria	1. Quality filtered data stream or subset returned to user (or program/machine). 2. An interface at the DAAC is established to enable machine-to-machine data query and return of quality filtered data streams or data subsets.
Data Quality Keywords	extraction, filtering, flags, missing data

C.4 Use Case 4

Use Case Title	Dataset Recommendation
Point of Contact	Christopher Lynnes
Email	christopher.s.lynnes@nasa.gov
Use Case Narrative	Many users come to the GES DISC with the question: which dataset should I use for my application or research project? This is because many datasets will contain similar geophysical measurements, but come from different instruments, produced using different algorithms, aggregated (or not) at different spatial and temporal resolutions, in different formats, arriving with different latency, etc. These all affect "fitness for use" (a generalized synonym for data quality.) If we knew the user's application/research topic, in the long run we might be able to recommend those datasets whose quality features (uncertainty, resolution in time and space, time coverage, latency) most closely match the user's needs. This could be used in, say, relevancy ranking for search results.

Domain of Interest	Atmosphere, Biology, Climate, Cryosphere, Hydrology, Land, Ocean, Solid Earth
Professional Domain of User	Scientist/Researcher
Primary User/Stakeholder Relationship	Human User -> DAAC/Stakeholder
Secondary User-Stakeholder Relationship	Machine User -> DAAC/Stakeholder
Primary Scope/Rationale	Qualitative-Product / TBD
Secondary Scope/Rationale	Qualitative-Science / TBD
Use Case Chronology:	1. User registers with data provider, noting what type of user (researcher, applications user) and what areas he / she usually works in. 2. User searches for datasets matching some criteria (e.g., "Ozone"). 3. User adjusts the weighting or order of quality criteria used in sorting the results in order to highlight the datasets of most use to their particular project.
Success Criteria	1. User has saved user type info that can be used to provide more personalized searches. 2. User sees datasets sorted according to the quality criteria most important to their user type and interest areas. 3. User has ability to adjust the way that quality criteria are used in sorting the datasets.
Data Quality Keywords	accessibility, bounding box, instrument sampling, metadata, metrics, missing data, search, spatial resolution, temporal resolution

C.5 Use Case 5

Use Case Title	Errors Introduced by Binning, Smoothing, and Interpolation
Point of Contact	Jessica Hausman
Email	Jessica.K.Hausman@jpl.nasa.gov
Use Case Narrative	Whenever you take L2 data and create L3 or just change the resolution of the data you introduce errors and uncertainties by changing the resolution, hence a variety of ways and techniques exist to reduce the amount of introduced error for a given dataset. This type of error is

	commonly referred to as either interpolation error or representation error. Also worth noting is that these errors and uncertainties also propagate into Level 4 datasets. For data users where it is crucial to minimize these errors and uncertainties as much as possible, it would be good to know how much error/uncertainty was introduced or observe the goodness of fit to a Gaussian curve (or some other reasonable method for determining goodness of fit to the observations). One example of this is what is being done as a using a spectral analysis technique to assess the "effective" spatial resolution with the MUR GHRSST product; unfortunately, this type of information is difficult to disseminate and interpret for a variety of datasets and Earth science data parameters. As spatial resolution and spatial coverage of swaths are increasing (e.g., SWOT), the data that are then created into an L3 product may in reality contain some of the meso-scale structures, which are then lost in the binning from fine to large scale grids. As another example, a user can take coarse data and convert it into finer resolution, meaning there is a need to interpolate; the user would then need to know how much error and uncertainty is generated from that. A new variable could be added to the data that quantifies the level of introduced error and uncertainty for each pixel/bin as a result of geospatial binning, geospatial/temporal smoothing, and/or geospatial/temporal grid interpolation.
Domain of Interest	Atmosphere, Biology, Climate, Computer Science, Cryosphere, Geomagnetics, Geographical, Geology, Ecology, Heliophysics, Hydrology, Informatics, Ionosphere, Land, Ocean, Radiative Transfer, Solid Earth
Professional Domain of User	Scientist/Researcher
Primary User/Stakeholder Relationship	Human User -> Mission-Project/Stakeholder
Secondary User-Stakeholder Relationship	N/A
Primary Scope/Rationale	Quantitative-Product / Rationale: A new variable could be added to the data that quantifies the level of introduced error and uncertainty for each pixel/bin as a result of geospatial binning, geospatial/temporal smoothing, and/or geospatial/temporal grid interpolation.
Secondary Scope/Rationale	Qualitative-Science / Rationale: Error and uncertainties can be better characterized thus enabling data researchers to better resolve their scientific understanding of the data. In addition to quantifying these errors and uncertainties at the pixel level, the data producer can provide statistical plots (e.g., spectral analysis) that qualitatively

	demonstrate (through a backend quantitative analysis) the limitations of the data at various time and spatial scales.
Use Case Chronology:	1. Numerous data producers create a series Level 3 product and during the validation process they discover increased error and uncertainty to observations in comparison to the Level 2 data. 2. The Level 3 data producers document the errors but have not yet confirmed the nature or source of the error. 3. A collaborating MEaSUREs investigator uses a subset of these Level 3 datasets to create a Level 4 data product, and through the validation exercise they notice an increased error relative to both the individual Level 3 input datasets. Since model data is used as a first guess constraint to the Level 4 product, it's highly uncertain at this stage what is contributing to the increased error. 4. The MEaSUREs investigator contacts the individual Level 3 data producers to communicate the relative discrepancy in error. 5. The Level 3 data producers investigate the issue and determine that much of the errors and uncertainties that have been introduced into the Level 3 data conversion was artificially manifested through the binning and interpolation schemes that were employed. 6. The Level 3 data producers provide pixel level error and uncertainties specifically manifested through the grid binning and interpolation. 7. The MEaSUREs investigator takes these errors and uncertainties to assess the cumulative error propagation and utilize a cost function in the Level 4 processing algorithm to minimize the propagation of error and uncertainty into the final Level 4 product. 8. The MEaSUREs investigator then computes the residual error and uncertainty that remains due to both propagation from Level 3 and the resulting binning, smoothing, and interpolation artifacts that inherently contribute to the error and uncertainty in the final Level 4 data product. 9. The MEaSUREs investigator makes available the residual error and uncertainty at the pixel level within the Level 4 data product. 10. The MEaSUREs investigator makes the Level 4 contributed error and uncertainty at the pixel level within the Level 4 data product. 11. The MEaSUREs investigator generates and makes publicly available through the collaborating DAAC a variety of spectral analysis plots with a brief description as to how the spectral energy at various time and space scales relates to the error and uncertainty of the data product. 12. The MEaSUREs investigator then determines and publishes the "effective" spatial resolution of the final Level 4 product as determined by an accepted technique derived from the spectral analysis of the data.
Success Criteria	1. Successful communication and relay of critical error and uncertainty information between Level 3 data producers (i.e., SIPs, DAACs, Principal Investigators, and Flight Projects) and

	MEaSUREs investigators. 2. Successful discovery and attribution to the source of error and uncertainty as related to binning and interpolation within each Level 3 dataset. 3. Successful discovery and attribution of the error and uncertainty as related to both the propagation of Level 3 errors (stated in item #2) and the associated error and uncertainty manifested as a bi-product of Level 4 data production. 4. Successful assignment of error and uncertainty at the pixel level for the datasets mentioned in items #3 and #4 above. 5. Publication of any ancillary statistical plots (e.g., spectral analysis) and accompanying narratives that can be used to further explain the spatial and temporal limitations of the Level 3 and Level 4 datasets. 6. Utilization of #5 to discover and publish the "effective" spatial resolution of a given Level 3 and Level 4 dataset. 7. Successful publication of item #5 and #6 by the collaborative DAAC.
Data Quality Keywords	algorithm, accessibility , calibration, cross-calibration, data sampling, documentation, extraction, flags, spatial resolution, temporal resolution, workflow

C.6 Use Case 6

Use Case Title	Fisherman Needs SST and Wind Vector Data Over Gulf Stream
Point of Contact	David Moroni
Email	David.F.Moroni@jpl.nasa.gov
Use Case Narrative	A commercial fisherman needs sea surface temperature and ocean surface wind vector data over a region of the Gulf Stream during a specific month of the year to help chart a course for ideal conditions and locations for fishing. Due to the location-specific nature of this user, where ideal locations change up to tens of kilometers daily, this user needs data with spatial resolution under 10 km and maximum data coverage with minimal data dropouts.
Domain of Interest	Biology, Ocean
Professional Domain of User	Resource Management
Primary User/Stakeholder Relationship	Human User -> DAAC/Stakeholder
Secondary User-	N/A

Stakeholder Relationship	
Primary Scope/Rationale	Qualitative-Product / Rationale: As more of an applications-based user, this user is more interested in the product quality and prefers a qualitative interpretation of the quality attributes pertaining to the use case.
Secondary Scope/Rationale	Quantitative-Product / Rationale: This user may on occasion prefer quantitative metrics pertaining to the product quality as a whole, such as maps/charts depicting the quality attributes pertaining to the use case.
Use Case Chronology:	1. User enters the PO.DAAC looking for sea surface temperature and ocean surface wind vector data. 2. User finds multiple dataset matches on search inquiry based on preferred regional, temporal, and spatial resolution constraints. 3. User selects an option sort this filtered list of datasets by the least amount of data dropouts. 4. User is then provided with daily preview maps for each dataset depicting both the data values and points where data is unavailable. 5. User then selects the SST and wind vector datasets which meets their criteria for least data dropouts over their specific regions of interest for each day of the month.
Success Criteria	1. User can access a list of matching datasets in response to a search by the following parameters/keywords: “sea surface temperature” and “ocean surface wind vector”. 2. User receives a list of datasets in response to a specified regional/temporal bounding box. 3. User receives a list of datasets in response to a specified maximum spatial resolution. 4. User receives a sorting of previously returned dataset listings according to the least data dropouts. 5. User can view automatically generated daily maps for each dataset from the previously returned dataset listings depicting actual data values along with points where data is unavailable. 6. User is able to select and access the specified dataset corresponding to the specific day of the specified month.
Data Quality Keywords	accessibility, bounding box, data sampling, extraction, filtering, metrics, missing data, search, spatial resolution, temporal resolution, web services

C.7 Use Case 7

Use Case Title	Land Mask Issue in Near Real-Time DMSP SSM/I Daily Polar Gridded Sea Ice Concentrations

Point of Contact	Nathan Kurtz
Email	nathan.t.kurtz@nasa.gov
Use Case Narrative	The near real-time ice concentrations from this product are a valuable resource for information on the current sea ice state. In the data product, a specific value is provided for grid cells containing land, so these points are not confused with sea ice covered regions. But there looks to be an error in the land mask that was used in the product. The land mask appears to have been shifted by several grid points which can be seen when plotting the masked areas on a map with a coastal outline. A corrected land mask should be applied to the data so that areas containing sea ice are not inadvertently masked, and areas containing land are not reported as having sea ice.
Domain of Interest	Cryosphere
Professional Domain of User	Scientist/Researcher
Primary User/Stakeholder Relationship	Human User -> DAAC/Stakeholder
Secondary User-Stakeholder Relationship	N/A
Primary Scope/Rationale	Qualitative-Product / Rationale: A corrected land mask would make the data more useful for near real-time sea ice concentration datasets. I personally would like to use the product to flag where sea ice is present in a near real-time operational data product. A corrected land mask would improve the quality of the dataset when data within several grid cells of land areas are needed.
Secondary Scope/Rationale	N/A
Use Case Chronology:	1. User downloads data from the ftp site at ftp://sidads.colorado.edu/pub/DATASETS/nsidc0081_nrt_nasateam_seaice/ 2. The data are overlain on a map with land areas in view. The data are seen to extend onto land in some areas, and have data gaps in other areas near land, suggesting that a land mask grid has been shifted. 3. User contacts NSIDC to report the issue. A reply was received which states "Near real-time products are not intended for operational use in assessing sea ice conditions for navigation and should be used with caution. These data are primarily meant to provide a best estimate of current ice conditions based on

	information and algorithms available at the time the data are acquired."
Success Criteria	1. Implementation of new land mask to the data product. 2. A map of the new data with land boundaries is produced, showing few or no data gaps in areas near land.
Data Quality Keywords	extraction, filtering, missing data

C.8 Use Case 8

Use Case Title	MEaSUREs Global Food Security Analysis & Support Data (GFSAD) - **Provisional** Crop Dominance (CD) @ 1km product
Point of Contact	Stacie Doman Bennett
Email	sdomanbennett@usgs.gov
Use Case Narrative	The provisional GFSADCD1KM as described via product documentation provided by PI & Team states this product will be an 8-class digital product that provides, at nominal 1 km, information on global: 1. Cropland extent\areas; 2. irrigated versus rainfed cropping; 3. Crop dominance; and 4. Cropping intensity (single, double, triple, and continuous crops). After documentation & supporting journal reviews - there are numerous caveats to detailed product deliverables to which journal articled references are not 1) providing entire workflow, including versions of ancillary data 2) providing information that cropping intensity is a derived product.
Domain of Interest	MEaSUREs, LULC, Cropland, Water Use, Crop Dominance
Professional Domain of User	LULC, Cropland, Irrigation
Primary User/Stakeholder Relationship	Data Producer - USGS
Secondary User-Stakeholder Relationship	Data Distributer - LP DAAC, Data Scientist - LP DAAC & LP DAAC Users, NASA - Primary Stakeholder/Owner of final product
Primary Scope/Rationale	Qualitative Product / Rationale: Accuracy of product documentation vs provisional product contents.
Secondary Scope/Rationale	Qualitative Science / Rationale: Accuracy, Uncertainty

Use Case Chronology:	1. Retrieve GFSAD data & documentation from LP DAAC Provisional staging. 2. Opening product in geospatial application. 3. Compare layers / deliverables to product documentation - specifically the deliverable of cropping intensity. Cropping Intensity is a further derived deliverable, not outright delivered. 4. Additional reference documentation: Thenkabail P.S., Knox J.W., Ozdogan, M., Gumma, M.K., Congalton, R.G., Wu, Z., Milesi, C., Finkral, A., Marshall, M., Mariotto, I., You, S. Giri, C. and Nagler, P. 2012. Assessing future risks to agricultural productivity, water resources and food security: how can remote sensing help? Photogrammetric Engineering and Remote Sensing, August 2012 Special Issue on Global Croplands: Highlight Article. 78(8): 773-782.
Success Criteria	1. Update product contents to contain deliverables as described in source documentation (PI documentation, research journal articles). 2. Update documentation to accurately reflect the exact deliverables, eliminating any 'to lesser extent' deliverables.
Data Quality Keywords	Documentation, Reporting, Standardization, Product Development Workflow

C.9 Use Case 9

Use Case Title	MEaSUREs PI wants to provide complete quality documentation to make his datasets useful to community
Point of Contact	H. K. Ramapriyan
Email	hampapuram.ramapriyan@ssaihq.com
Use Case Narrative	The year is 2019. A MEaSUREs PI, selected through a ROSES 2017 call for proposals, is getting ready to produce his product as promised in his proposal. He wants to make sure that he includes appropriate levels of quality data in his metadata and documentation. He is looking for guidance.
Domain of Interest	Atmosphere, Biology, Climate, Computer Science, Cryosphere, Geomagnetics, Geographical, Geology, Ecology, Heliophysics, Hydrology, Informatics, Ionosphere, Land, Ocean, Radiative Transfer, Solid Earth
Professional Domain of User	Scientist/Researcher
Primary User/Stakeholder	Human User -> Mission-Project/Stakeholder

Relationship	
Secondary User-Stakeholder Relationship	N/A
Primary Scope/Rationale	Quantitative-Science / Rationale: A fundamental requirement for a scientific data product is that the scientific data quality be characterized quantitatively.
Secondary Scope/Rationale	Qualitative-Product / Rationale: For users it is important to know all the metadata, provenance, documentation, etc. that need to accompany the scientific product. A producer of the product needs to make sure that such information is provided along with the product.
Use Case Chronology:	1. NASA HQ assigns a DAAC to the MEaSUREs PI's project. 2. MESUREs PI contacts the DAAC and asks for any guidance they have used in the past for providing data quality information. 3. The DAAC checks with ESDIS and provides a "Shell" Cooperative Agreement (CA), which is a generic version of CA that was used with MEaSUREs 2012 projects. This document provides "Product Quality Checklists" in its Appendix A. 4. The PI uses this as a general guide for items that need to be accomplished to ensure science quality and product quality are assured and documented. However, the PI has more detailed questions about documentation, data formats and metadata formats. 5. The DAAC provides samples of ATBD's and other documents delivered by one or two previous MEaSUREs projects, and format guides and standards. 6. PI studies the documents provided by the DAAC, proceeds to adapt them to his needs and generates products.
Success Criteria	1. A DAAC is assigned and available for PI to consult. 2. DAAC has all the information needed to help the PI. 3. ESDIS and/or MEaSUREs Program Manager/Scientists are available to answer questions if needed.
Data Quality Keywords	algorithm, accessibility, documentation, flags, metadata, metrics, missing data, spatial resolution, temporal resolution

C.10 Use Case 10

Use Case Title	Metadata consistency evaluation
Point of Contact	Ed Armstrong
Email	edward.m.armstrong@jpl.nasa.gov
Use Case	A data provider or data center engineer is interested in evaluating the

Narrative	conformance of netCDF or HDF granules to the Climate Forecast (CF) and Attribute Convention for Dataset Discovery (ACDD) metadata models. He has many different unique datasets to evaluate on well they conform and how they compare to each other.
Domain of Interest	Atmosphere, Biology, Climate, Computer Science, Cryosphere, Geomagnetics, Geographical Information Systems, Geology, Ecology, Heliophysics, Hydrology, Informatics, Ionosphere, Land, Ocean, Radiative Transfer, Solid Earth
Professional Domain of User	Data Management
Primary User/Stakeholder Relationship	Machine User -> DAAC/Stakeholder
Secondary User-Stakeholder Relationship	N/A
Primary Scope/Rationale	Quantitative-Product / Rationale: Assessing the metadata completes of data granules.
Secondary Scope/Rationale	N/A
Use Case Chronology:	1. Data user needs to assess the metadata completeness and consistency of many different satellite or earth science data granules. 2. A software tool or web service is available that allows this to proceed in an automated fashion. 3. The user can evaluate an output report that assigns a quantitative score for each granule metadata check. 4. The user can then compare granules from different datasets for consistency and score.
Success Criteria	1. A software tool or web service is available that is designed to check for CF and ACDD metadata conformance and report a score on results. 2. It is linked to the most recent versions of CF and ACDD.
Data Quality Keywords	documentation, interoperability, metadata, standardization, web services

C.11 Use Case 11

Use Case Title	NASA Team Sea Ice Concentration Filters
Point of Contact	Lisa Booker

Email	lisa.booker@nsidc.org
Use Case Narrative	The lack of transparency in subjective ice removal from the data makes reproducibility of these data difficult. In addition, as a researcher working with sea ice, I would like to be able to use my own judgement to filter out questionable ice values. Having a quality flag that marks questionable ice values allows me to determine which pixels to consider. And leaving these values in the data and simply flagging them allows me to reproduce the work of the data producers as described in literature.
Domain of Interest	Climate, Cryosphere
Professional Domain of User	Scientist/Researcher
Primary User/Stakeholder Relationship	Human User -> Mission-Project/Stakeholder
Secondary User-Stakeholder Relationship	N/A
Primary Scope/Rationale	Qualitative-Science / Rationale: By flagging questionable ice values, it is left to the researcher to determine the integrity of the value for their research. In addition, the overall integrity of the science is improved by making the data more reproducible.
Secondary Scope/Rationale	Quantitative-Science / Rationale: Adding a quality flag will provide uncertainty information not previously provided in the data, therefore improving the integrity of the data and science.
Use Case Chronology:	1. A user contacts NSIDC User Services asking for more information about the subjective removal of ice. 2. USO works with the data producer to understand the history of the subjective filtering of ice values. 3. USO communicates with user that they have passed information along to the data producer and it's unclear if and when this information will be addressed.
Success Criteria	1. A user knows through documentation that quality flags are available for questionable ice values. 2. The values for the quality flag are fully defined, i.e. weather effect has a particular value, coastline has a particular value, etc.
Data Quality Keywords	algorithm, accessibility, filtering, flags

C.12 Use Case 12

Use Case Title	Outlier Detection and Attribution
Point of Contact	Vardis Tsontos
Email	Vardis.M.Tsontos@jpl.nasa.gov
Use Case Narrative	A user notices some unexpected, extreme values in granules of a given dataset and communicates his concern to the DAAC hosting these data. Further analysis on the dataset is conducted first to confirm the quality issue identified by the user, second to better characterize the observed outliers, third identify the scope and extent of the problem (i.e., how many/which granules are impacted, other datasets impacted), and fourth suggest possible sources of the problem. A variety of plots (see here: http://bit.ly/dqoutlierslides) and summary statistics are used as both diagnostics and summary reports that are ultimately also used to communicate the issue to the data provider in sufficient detail for them to be able to rigorously investigate the origin of the problem in their data processing stream and code. More generally, the goal of this data quality use case is: the ability to effectively detect and summarize extreme geophysical values in individual files or a population of granules for a given dataset in a manner that may also be suggestive of the origin of the problem that can then be communicated effectively to stakeholders. Such assessments would ideally be conducted in an automated fashion and could occur iteratively as granules become available or as a job run on an available data file repository. Note that determination of what constitutes an “extreme” or outlier” value can be objectively based on defined Valid_Min/Max/range values if available in granule metadata; or it can be based on a science understanding of the distribution of values for the parameter of interest.
Domain of Interest	Atmosphere, Biology, Climate, Computer Science, Cryosphere, Geomagnetics, Geographical, Geology, Ecology, Heliophysics, Hydrology, Informatics, Ionosphere, Land, Ocean, Radiative Transfer, Solid Earth
Professional Domain of User	Scientist/Researcher
Primary User/Stakeholder Relationship	Human User -> DAAC/Stakeholder
Secondary User-Stakeholder Relationship	N/A
Primary Scope/Rationale	Quantitative-Product / Rationale: The goal here is to identify and better characterize outliers with the following needs: automated, effective communication, guidance to the origin of the problem, non-disruptive to the user (i.e., handled remotely by the data provider). Possible

	implementations may include: providing detailed graphics, ingesting the granule metadata to determine max/min boundary conditions, or custom-tailored scientific guidance based upon the parameter of choice. To accomplish this goal and set of needs, one would need to leverage a set of solutions to: 1) compute the quantity and location of outliers, 2) extract the magnitude of each outlier, and 3) compare the magnitude of each outlier with the expected max/min of the data.
Secondary Scope/Rationale	Qualitative-Science / Rationale: Successfully addressing the use case by meeting the goal and needs described above would also result in an impact to the integrity of the science being produced by the data. The deliverables relevant to this scope would be: 1) the provision of a procedure to carry out the goal and needs of the use case and 2) a description as to the nature of the issue.
Use Case Chronology:	1. User contacts DAAC documenting quality issue 2. DAAC Data Engineer (DE) or User Services (US) team member investigates to corroborate and better characterize the quality issue and suggest possible causation using a range of diagnostics 3. DE or US summarizes the results and reports the findings to the data provider 4. Data provider/science team investigate and identify the source of the problem in their processing code, and once fixed reprocess the data 5. DE or US mediates the archival of the reprocessed data and conducts some independent quality checks on the updated data. 6. DE or US informs the user that the updated data are available and potentially provides a high-level explanation of the issue to the user.
Success Criteria	1. Objective criteria for defining extreme values for the parameter of interest in a given dataset are in place 2. An automated reporting tool to efficiently conduct an outlier analysis on either singular or populations of granules or spatial/temporal subsets therein. 3. A process for proactive reviews of data for outliers conducted on an ongoing basis upon delivery to the DAAC (if not before by the data producer’s processing system). 4. A process for post-hoc (on demand) quality reviews based on user notifications. 5. A clear protocol and workflow for investigating and communicating quality status to all stakeholders is in place.
Data Quality Keywords	documentation, extraction, filtering, metadata, metrics, reporting, workflow

C.13 Use Case 13

Use Case Title	Provide ancillary information on potential biases

Point of Contact	Marc Simard
Email	marc.simard@jpl.nasa.gov
Use Case Narrative	We want to provide sufficient information to users such that they can judge and replicate our datasets. The SRTM DEM will be distributed with a vegetation canopy bias which can be used to estimate bald Earth DEM.
Domain of Interest	Climate, Geology, Ecology, Hydrology, Land, Solid Earth
Professional Domain of User	Scientist/Researcher
Primary User/Stakeholder Relationship	Human User -> Mission-Project/Stakeholder
Secondary User-Stakeholder Relationship	TBD
Primary Scope/Rationale	Quantitative-Science / Rationale: The bald Earth dataset can be used in hydrological models and the vegetation bias in ecological applications.
Secondary Scope/Rationale	Quantitative-Product / Rationale: The datasets can be used to generate contour maps and watershed delineation.
Use Case Chronology:	1. The PI asks the community for input and comments about the plans. 2. The community makes recommendation on product interest and format. 3. The community list requirements. 4. The PI adapts to discussion and produces maps. 5. The PI generate documentations and accuracy layers to accompany each dataset. 6. The list of layers grows rapidly, and the choice of format becomes critical.
Success Criteria	1. The discussion with community were constructive. 2. The products are in-sync with community's expectations. 3. The accuracy is well documented and sufficient for applications.
Data Quality Keywords	algorithm, calibration, cross-calibration, documentation, metadata, metrics, reporting

C.14 Use Case 14

Use Case Title	Region Vulnerable to Storm Surge

Point of Contact	Marc Simard
Email	marc.simard@jpl.nasa.gov
Use Case Narrative	An insurance company is trying to assess the coastal region that is vulnerable to storm surge. The representative only finds the SRTM DEM is available. This user is interested in a quantitative assessment as they will associate dollars and resources accordingly. The user wants to know the probability of an area being flooded by the storm.
Domain of Interest	Climate, Geographical Information Systems, Land, hazard
Professional Domain of User	Risk Management
Primary User/Stakeholder Relationship	Human User -> Mission-Project/Stakeholder
Secondary User-Stakeholder Relationship	TBD
Primary Scope/Rationale	Quantitative-Product / Rationale: This user is interested in a quantitative assessment as they will associate dollars and resources accordingly.
Secondary Scope/Rationale	Quantitative-Science / Rationale: The user wants to know the probability of an area being flooded by the storm.
Use Case Chronology:	1. The insurance company seeks to expand coverage to the Louisiana. 2. The representative (assessor), is looking for ways to map of vulnerability to storm surge. 3. A topographic map is used to help identify lowlands. 4. The assessor looks over the web for a DEM and find lidar-derived DEM and the SRTM DEM. 5. The SRTM DEM is free and the assessor decides to use this one. 6. The assessor downloads the data.
Success Criteria	1. User is able to download the data to his/her desktop. 2. User is able to map the potential area of flood from topography. 3. User is able to identify accuracy on a location by location basis. 4. User can identify locations where accuracy is insufficient.
Data Quality Keywords	accessibility, bounding box, calibration, cross-calibration, extraction, metrics

C.15 Use Case 15

Use Case Title	Sensor-Specific Observation Quality Contribution to Level 4 Datasets
Point of Contact	Jessica Hausman
Email	Jessica.K.Hausman@jpl.nasa.gov
Use Case Narrative	For a given L4 dataset, the contribution of observations from a specific satellite instrument for a given data pixel is generally unknown. Even though the global coverage might use 4 satellites, coastal areas may only use 2 due to limitations of land or sea ice contamination. Specifically, with SST retrievals, the infrared has less of a lag than microwave so the NRT data will contain more infrared, while the delayed-mode data is later backfilled with microwave retrievals. The end data user therefore needs to know how much of the pixel is comprised of specific sensor inputs, such as AVHRR or MODIS Aqua/Terra. Also, the end data user needs to know whether or not any in situ and/or model data are incorporated into a given pixel, and consequently what is the contribution is to that pixel relative to the satellite sensor inputs. This is important for the modeling community as they take the data and then feed it into a model that will generate other errors due to the algorithm. Therefore, being able to quantify and isolate the relative contribution of bias and error is important. A flag or reference variable can be used to provide the percentage of relative contribute from each data source/sensor.
Domain of Interest	Atmosphere, Biology, Climate, Computer Science, Cryosphere, Geomagnetics, Geographical, Geology, Ecology, Heliophysics, Hydrology, Informatics, Ionosphere, Land, Ocean, Radiative Transfer, Solid Earth
Professional Domain of User	Scientist/Researcher
Primary User/Stakeholder Relationship	Human User -> Mission-Project/Stakeholder
Secondary User-Stakeholder Relationship	N/A
Primary Scope/Rationale	Quantitative-Product / Rationale: What is fundamentally needed is a quantification of the relative contribution of observational data from each source/sensor provided at the pixel level. This effects the entire production and quality of the data product.
Secondary Scope/Rationale	Quantitative-Science / Rationale: The fundamental impact for the user is the ability to ascertain specific uncertainties and errors at the pixel level as a function of the sensor-specific contributions. This enhancement of quantified error contribution will impact the overall

	quality of the science.
Use Case Chronology:	1. A data user accesses a generic Level 4 dataset at a DAAC and upon examining the granule metadata is unable to locate any information pertaining to the sensor specific contribution of observations at the pixel level. 2. The data user contacts the DAAC asking if it's possible to derive this information. 3. Since the DAAC cannot directly derive this information, the DAAC must then contact the data producer. 4. The data producer then provides the DAAC with the relative contribution of observations at each pixel and time step. 5. The DAAC forwards this information to the data user and documents this information for future inquiry. 6. The DAAC realizes a more permanent solution should be made for all Level 4 datasets.
Success Criteria	1. The DAAC responds to the user inquiry and assess its capabilities to see whether it can satisfy the user request. 2. The DAAC is successfully able to relay the user request to the data producer. 3. The data producer provides an ad hoc solution to the DAAC. 4. The DAAC relays this information back to the data user. 5. The DAAC advises all Level 4 data producers to make this information available at the pixel level through a flag or reference variable.
Data Quality Keywords	algorithm, accessibility, data sampling, documentation, extraction, flags, instrument sampling, metadata, reporting

C.16 Use Case 16

Use Case Title	SMAP Freeze/Thaw Algorithm
Point of Contact	Chris Derksen
Email	chris.derksen@ec.gc.ca
Use Case Narrative	A user of the SMAP level three freeze/thaw product (L3_FT_A) requests radar derived freeze/thaw information south of 45N latitude, which falls outside of the L3_FT_A domain. The user is running a land surface model which includes regions south of 45N, so this lack of data severely limits the utility of the SMAP product. Regions south of 45N are not included in the SMAP L3_FT_A product because the difference in the radar signal between the frozen and thawed state is insufficient south of 45N (because of the transient nature of freeze/thaw events) and so the retrievals are considered highly uncertain.
Domain of Interest	Cryosphere, Ecology

Professional Domain of User	Scientist/Researcher
Primary User/Stakeholder Relationship	Human User -> Mission-Project/Stakeholder
Secondary User-Stakeholder Relationship	N/A
Primary Scope/Rationale	Qualitative-Product / Rationale: The user has requested data over a region that is not formally covered by the SMAP product because of concerns over the quality of the radar FT retrievals over that region. FT information over the user's region of interest is available, however, from flags in other SMAP datasets and other satellite datasets.
Secondary Scope/Rationale	Quantitative-Science / Rationale: Quantitative assessment of FT flags outside of the SMAP L3_FT_A domain would provide valuable cal/val information to the product developers and other potential users.
Use Case Chronology:	1. User contacts NSIDC user services with request for expanded FT data. 2. User services contacts the SMAP L3_FT_A point of contact with question regarding availability of FT retrievals outside the L3_FT_A domain. 3. SMAP FT team provides response to the user on the availability of radar derived FT flags in the SMAP L2 soil moisture products, emphasizing the key sources of uncertainty in these retrievals (lower resolution radar measurements; weak radar sensitivity to freeze/thaw events in the lower mid-latitudes). 4. Information is provided to the user on the availability of other satellite derived FT datasets (i.e. SMOS) that do not include a latitudinal cutoff. 5. The user is requested to proceed with caution on the use of FT flags south of 45N, and provide any validation results to the SMAP FT team should they become available. 6. Discussion at the weekly SMAP FT telecon on how potential users should be made aware of FT flags outside the L3_FT_A domain, and how to provide guidelines and recommendation on the use of these flags.
Success Criteria	1. Successful relay of user request from NSIDC to the SMAP FT team. 2. Comprehensive response from the SMAP FT team communicated back to the user. 3. Discussion of formal recommendations to potential science users of SMAP FT flags south of the L3_FT_A domain. 4. Recommendations on limitations and potential user contributions to

	FT cal/val posted on product page at NSIDC.
Data Quality Keywords	algorithm, bounding box, flags

APPENDIX D - HIGH-LEVEL RECOMMENDATIONS FOR DATA QUALITY

Phase: (Capture = 1; Describe = 2; Discover = 3; Use = 4)

Rec.#	Phase	Category	Recommendations - Data Systems	Recommendations -– Science
1	1	General	·Maintain continuous and effective communication with data producers throughout the duration of their projects.	Develop a data quality plan for each data product and submit it along with the data for dissemination.
2	1	Standard Documents & Processes	Provide a standard set of documents to be provided to investigators and potential proposers; documents should describe what quality information should be provided and how they should be shown using metadata.	Include references to "standard" set of documents in calls for proposals
3	1	Standard Documents & Processes	Provide data producers with examples of determining and describing product quality (e.g., use of ATBDs, ESDIS product quality checklists, and any documentation that assists PI towards creating a final product with complete quality documentation).	Enable open and public "peer review" to help promote increased discovery, reduced latency, and dissemination of known issues.
4	1	Standard Documents & Processes	Consider using the following ECHO/ISO Quality Attributes: QAFRACTIONGOODQUALITY (#30), QAPERCENTGOODQUALITY (#263), QAPERCENTOTHERQUALITY (#263), QAPercentOutOfBoundsData, AutomaticQualityFlag, OperationalQualityFlag, ScienceQualityFlag	Provide data quality information through appropriate data formatting and metadata specifications (i.e., CF, ISO, ACDD, ECHO, etc...).
5	1	Standard Documents & Processes	Develop and incorporate standardized, pre-ingest quality assessments for specific datasets.	Incorporate best practices regarding data quality into final datasets for helping users with history of the product contents and creation process.
6	1	Standard Documents & Processes	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.	Consult guidelines that describe categories of data quality and provide information and evidence about the quality of the dataset for each category.

7	1	Standard Documents & Processes		Prepare data and attributes related to accuracy, precision and uncertainty and organize them based on standards.
8	1	Checklists	Incorporate a checklist for data that includes the capturing of "known issues" for particular regions or time intervals.	Provide a checklist for data that includes the capturing of "known issues" for particular regions or time intervals.
9	1	Publicizing Quality Issues	Collect and integrate outlier information obtained from various datasets, and perform a relevant data quality analysis, as well as establish a checklist that may help DAACs and Data Producers for future data management and production.	
10	1	Publicizing Quality Issues:	Host a prominent web page that captures known quality issues.	Convey fully the limitations of specific datasets, for inclusion in documentation and dataset descriptions
11	1	Publicizing Quality Issues:	Provide enough publicly available information with self-describing documentation such that the need for users to contact the DAACs is minimized.	Make quality flags publicly accessible and directly corresponding to a quantifiable metric, such as the related uncertainty, confidence intervals, and confidence levels.
12	1	Publicizing Quality Issues:	Provide spatially explicit systematic and random error with conservative figures.	Ensure all known issues discovered by the science teams are reported to the DAACs in a timely manner.
13	1	Publicizing Quality Issues:	Request full uncertainty estimates from the producers and distribute with the datasets.	Provide full uncertainty information with datasets to DAACs. Describe any restrictions on the use of the data and clearly display the rights enabling the use and adaption of the data and of the data quality information.
14	1	Publicizing Quality Issues:	Request documentation from investigators and provide to users error and uncertainty estimates at each level of the processing chain (e.g., binning and interpolation) with the product and/or include them in the ATBD or dataset user's guide.	Participate in formal process to help DAACs accurately document accuracy, precision and uncertainty, beginning when datasets are at a provisional level.

15	1	Publicizing Quality Issues:	Develop capabilities to gather users' comments on quality of specific datasets, validate and categorize the comments, and make them publicly available.	Develop capabilities to capture the distribution of errors for each dataset and to conduct an outlier analysis for each variable.
16	1	Quality Flags and Indicators	Describe quality flags in the data documentation and in the list of FAQs about the dataset.	
17	1	Quality Flags and Indicators	Provide capabilities to allow data quality indicators for applicability.	Provide definitions for each quality indicator and a description of how each quality indicator can be used (documentation, user guide, and in search system).
18	1	Quality Flags and Indicators	Provide variable-specific guideline or recommendation about how to use the quality indicator in a specific type of research or application.	Create variable-specific guideline or recommendation about how to use the quality indicator in a specific type of research or application
19	1	Quality Flags and Indicators	Include per pixel quality layer(s) where appropriate.	Provide description of the pixel-level quality indicator, including the algorithms and datasets used to derive this quality information.
20	1	Quality Flags and Indicators	Document and capture as metadata whether or not there is a pixel-level quality indicator for a given dataset.	Provide all data with added quality and/or uncertainty flags for areas that show spurious data (e.g., ice in unlikely places). Provide pixel-level uncertainty information.
21	1	Quality Flags and Indicators	Ensure that a given collection of datasets which share a common parameter also share common quality flags and flagging conventions (e.g., GHRSST via the PO.DAAC).	
22	1	Quality Flags and Indicators	Encourage data providers to provide quality flags and transparency in data production/creation.	
23	1	Quality Flags and Indicators	Develop capabilities for including and populating descriptions of quality flags for questionable values.	

24	1	Quality Flags and Indicators	Implement standardized documentation protocols to explain when and how quality flags should be used along with caveats which indicate the limitations of given quality flags.	
25	1	Quality Flags and Indicators	Document and publish all available descriptions for data quality indicators.	
26	1	Applicability/Fitness for Use	Develop capabilities for investigators to annotate and describe the “fitness for use” of the data as it applies to the data quality characteristics	
27	1	Applicability/Fitness for Use	Develop capabilities for determining and recording the applicability of datasets within the EOSDIS data holdings in various contexts based on data quality characteristics	
28	1	Quality of Input Datasets used in Generating Products	Request from the producers information about the contribution of the various input data that are used to process a higher level product	Include information about correctness /uncertainty of input datasets used (e.g., land/ocean/region masks) along with products (e.g., sea ice product).
29	1	Quality of Input Datasets used in Generating Products	Provide information about correctness /uncertainty of input datasets used (e.g., land/ocean/region masks) along with the data products (e.g. sea ice product).	Determine if land mask anomalies originate from observed geophysical processes or technical processing error(s).
30	1	Quality of Input Datasets used in Generating Products	Provide table of various land mask datasets detailing any of the above issues and noting differences in shifts, errors, masking techniques, and source datasets used to generate each land mask dataset.	Update land (and perhaps regional) mask in coordination with most updated ocean/lake/ice mask processing schedule and workflow.
31	1	Quality of Input Datasets used in Generating Products	Provide updated land mask at a frequency commensurate with its changes (e.g., monthly, along with ocean mask)	Capture and document errors introduced at each level of processing.

32	1	Quality of Input Datasets used in Generating Products	Provide users with a tool that identifies which inputs, such as AVHRR or MODIS Aqua/Terra, that have contributed to each pixel.	Evaluate regularly products that use static masks. Compare current land mask to known accurate land masks to determine precise shifts, if any.
33	1	Quality of Input Datasets used in Generating Products		Provide information about the contribution of the different input data that are used to derive a product at pixel level with associated uncertainty estimates.
34	1	Quality of Input Datasets used in Generating Products		Create tools that capture into a variable for Level 4 datasets, the sensor inputs, such as AVHRR or MODIS Aqua/Terra, as well as ancillary input data that have contributed to each pixel.
35	1	Metadata Consistency Checking	Employ metadata consistency checking tool that meets usability needs and generates reports with standards-based accuracy, precision, and uncertainty attributes provided in data granules.	Give recommendations on how data quality related attributes will be evaluated in the metadata scoring framework.
36	1	Metadata Consistency Checking	Document and communicate with data producers the completeness, consistency and formatting conformity of their metadata resulting from consistency checking tool.	Collaborate to set up an appropriate scoring framework to check for CF and ACDD metadata conformance.
37	1	Metadata Consistency Checking	Provide a software tool that can check for CF and ACDD metadata conformance using online CF checker at PCMDI or related tools (also being developed at PODAAC, ncdismember, and UDDC tool in the THREDDS data server, which checks and generates ACDD metadata reports and provides mapping to ISO 19115 metadata elements).	
38	1	Metadata Consistency Checking	Using CF as known well-formed metadata, compare all DAAC HDF and NetCDF metadata to determine completeness, consistency and formatting conformity via comparison algorithm.	

39	1	Metadata Consistency Checking	Use completeness, consistency and formatting conformity metrics from metadata checking tool to provide a "compatibility score" (for internal use only). ["compatibility score" would then help a DAAC determine priority and readiness for a collection of datasets to be integrated and tested with one or more interoperable tools/services. This "compatibility score" could also help assess the overall maturity of a dataset in contrast with "competing" datasets (i.e., comparing the maturity of datasets of a similar pedigree).]	
40	2	Quality Flags and Indicators	Provide clear documentation about types and availability of quality flags using self-describing metadata. e.g., NetCDF/HDF, CF-conventions, ISO 19157	Work with DAACs to provide data quality information through a standardized quality flagging schema (e.g., GHRSST model for quality confidence levels).
41	2	Quality Flags and Indicators	Provide quality metrics in product metadata.	Define and/or create "indicators" to represent quality of a data product from different aspects (e.g., data dropout rate of a "sea surface temperature" data product can be considered as one data uncertainty indicator).
42	2	Quality Flags and Indicators	Provide clear and thorough product quality information for each dataset.	Incorporate algorithm to assess and improve quality of product.
43	2	Quality Flags and Indicators	Ensure documentation of how each quality flag was derived, including delineations between specific processing algorithms and ancillary datasets used in the flagging schema. (Not every quality flag is created equal)	Identify quantifiable data quality criteria, such as confidence levels and the values of quality flags, that can be used as criteria for refining search queries
44	2	Quality Flags and Indicators	Provide easy-to-use quality flags.	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).

45	2	Quality Flags and Indicators	Assist users with interactive inputs and quality indicators for making informed decisions (e.g., Data Quality Screening Service in GESDISC. Default Quality Flags and Advanced Quality Control from MODIS subset tool. Webification to extract quality indicators on the fly and also subset on the fly using quality indicators at PODAAC).	Allow user to decide level of quality to apply (flag with pre-defined levels of consideration); this may be predicated on whether all quality flags for datasets of a similar pedigree are compatible with each other; users may choose to use one dataset over another simply based on the availability of certain types of quality flags
46	2	Quality Flags and Indicators	Provide capability to harvest the quality flag data and metadata for each dataset. (e.g., DMAS at PO.DAAC)	
47	2	Searchability	Provide a webified data quality screening service to filter-out data that is of a user defined quality specifications based on data quality flags	
48	2	Searchability	Make all data quality information openly searchable and extractable to enable more complete dataset interrogation and comparison	
49	2	Searchability	Establish lists of variables and links to all datasets that contain the selected variables, to enable users to search for all such datasets.	
50	2	Searchability	Provide capabilities to present or visualize data quality indicators (e.g. use dropout percentage as sorting criteria; visualize dropout percentage map).	
51	2	Searchability	Develop schema to assign configurable caveats for common data usages with quality filtered data.	
52	2	Searchability	Develop capabilities for users to refine the results of search queries by selecting among choices of quantifiable data quality criteria (e.g., confidence levels or any quality flag derived from a quantifiable metric.)	

53	2	Searchability	Develop interface for entering temporal/spatial restrictions; as well as a way to automatically ingest this temporal/spatial bounding box information as harvested metadata that may be disseminated via the web. (e.g., PO.DAAC web services).	
54	2	Searchability	Provide links from a user selected variable to relevant quality document	
55	2	Searchability	Enable user to download original and quality filtered datasets.	
56	2	Publicizing Quality Issues	Include documentation on how accuracy and uncertainty of datasets were determined	Provide all data with added quality and/or uncertainty flags for the areas that have potential limitations
57	2	Publicizing Quality Issues	Include special warnings in datasets with large known uncertainties (e.g., datasets or subsets thereof with large known uncertainties due to resampling/smoothing/ interpolation techniques).	
58	2	Publicizing Quality Issues	Provide proper documentation outlining the limitations, and terms of use for data requested outside a given dataset's domain.	
59	2	Publicizing Quality Issues	Implement a standard protocol for how to deal with datasets in which the quality has been compromised, including a quarantine "soak" period and a vetting process to assess the severity of the compromise.	
60	2	Publicizing Quality Issues	Document how user will incorporate non-released data that is quality compromised.	
61	2	Publicizing Quality Issues	Implement plan to replace or permanently retire data that is catastrophically compromised, including documentation of the assessments which led to the resulting conclusions.	

62	2	Publicizing Quality Issues	Inform users as soon as possible when data are compromised and provide status updates when readily available.	
63	2	Publicizing Quality Issues	Describe level of confidence and uncertainties associated with the interpolated values (e.g., different for gap filling procedure or if level 2 and 3 have similar resolution).	Document resampling/interpolation techniques used and describe the impact of the resampling technique used to process at all levels and provide full uncertainty estimates associated with the techniques used to the DAAC.
64	2	Publicizing Quality Issues	Identify outliers, as well as produce guidance, e.g., via documentation or online alert/flag, providing users useful data quality information such as 1) quantity and location of outliers, 2) magnitude of each outlier, and its ratio relative to the expected max/min of the data, and 3) origin of the problem.	Establish a well agreed upon definition of outlier (extreme values) for each product based on science understanding of the distribution of values for the parameters of interest.
65	2	Publicizing Quality Issues	Provide users with information on the distribution of errors for each dataset, including the results of an outlier analysis for each variable.	Detect, attribute and document outliers using the community adopted standards.
66	2	Publicizing Quality Issues		Create a separate ancillary dataset to capture outliers, along with a guide document detailing known and probable causes for such outliers.
67	2	Publicizing Quality Issues		Convey the data quality information (i.e., extremes values and outliers) to the users to help ensure the integrity of the science being produced using the datasets.
68	2	Publicizing Quality Issues		(DAACs) Work with cognizant scientists to apply the community adopted standards toward outlier detection and attribution for datasets whose PI's are no longer accessible (deceased, retired, etc.)
69	2	Documentation		Provide algorithm descriptions and access to original (input) data along with a given dataset.

70	2	Quality of Input Datasets used in Generating Products:	Develop a diagnostic heuristic to objectively determine the "best" land mask for a given usage, which could be automatically recommended and disseminated via the URS protocols.	
71	2	Quality of Input Datasets used in Generating Products:	Provide users with documentation describing bias to enable estimations of bald Earth digital elevation models, i.e.; user guides & web pages.	
72	3	User Registration System	Use URS login functionality that would enable custom-tailored user preferences for specific datasets.	
73	3	User Registration System	Enable within URS preferences an add-on to enable specific DAAC search specifications based on field of research, usage application, preferred region of interest (among other parameters).	
74	3	User Registration System	Enable with URS metadata "tagging” of all datasets determined relevant to a data user.	
75	3	User Registration System	Leverage URS metadata "tagging" for both "prognostic" and "diagnostic” data preferences: - Prognostic analysis can assess whether a new dataset is likely to be used or desired by a given data user. - Diagnostic analysis can assess causality of data usage patterns.	
76	3	Search	Enable ontologies to find proper and related datasets based on parameter/space/time/accuracy considerations. (e.g., SWEET)	Establish an authoritative list of scientific terms, such as those in the SWEET ontology and GCMD, that can be selected for inclusion in search queries to find the dataset. (Y, but not implemented)
77	3	Search	Enable faceted and keyword search mechanism for more effective filtering of dataset characteristics such as spatial/temporal resolution, geophysical parameter, regional bounding box, and time coverage.	

78	3	Search	Provide data quality-related content-based search option, so that a user is able to select data that meet their needs	
79	3	Search	Provide a service to dynamically / programmatically query, filter and subset a dataset based on user preferred quality variables	
80	3	Search	Enable users to filter data with user-specified quality levels. (e.g., quality flag filtering through via w10n webification.)	
81	3	Search	Provide a dataset filtering system that queries and returns a ranking of datasets as a function of the percentage of data contained within a user defined max/min range; datasets with the highest percentage of data within the max/min range should rank highest in such a query.	
82	3	Search	Develop capabilities for users to search for datasets that contain the same variables as a particular data product of interest.	
83	3	Search	Establish machine-to-machine interfaces, accessible to the public, enabling automated quality filtering	
84	3	Search	Enable users to remotely interrogate data (using tools such as OPeNDAP or THREDDS) for the purposes of quality assessment, subsetting, aggregation, co-location, and visualization.	
85	3	Dataset Recommend ations	Include suggestions on data download page (or web portal) about all similar datasets along with their usage benefits.	
86	3	Dataset Recommend ations	Provide standing recommendations quickly to alternative datasets when a dataset has been retired or quarantined	

87	3	Dataset Recommendations	Inform users automatically when a new dataset, prognostically "tagged" as relevant to the given user, has been published through a given DAAC (similar to the "Amazon Recommends" feature).	
88	3	Dataset Recommendations	Use diagnostic “tagging” to help identify probable causes for the establishment of “popular” datasets.	
89		Dataset Recommendations		Provide data quality variables and metadata along with detailed documentation on how the variables/ metadata are derived and suggestions on how to use them in different applications
90		Dataset Recommendations		Ensure that the data distributed to the users are properly connected to specific quality criteria of flags. (Such quality filtered data should warrant the integrity of the science being produced by the data.)
91		Dataset Recommendations		Ensure that quality flags are related to a quantifiable metric that directly relates to the usefulness, validity, and suitability of the data
92	3	Data Usage	Provide storage that enables users to generate estimates for the probability of flooding for a location using Shuttle Radar Topography Mission (SRTM) Digital Elevation Models (DEM).	Create software that estimates the probability of flooding for a location using Shuttle Radar Topography Mission (SRTM) digital elevation models (DEM).
93	3	User Feedback	Develop capabilities for user inputs/comments on the following, and include them in publicly available information about each dataset or collection: 1. Applicability of dataset characteristics 2. Search terms that can be selected by other users for inclusion in queries 3. Metadata attributes about data quality that can be used by a search engine	

APPENDIX E - PRIORITY RATING FOR HIGH-LEVEL RECOMMENDATIONS

Priority ratings were contributed by 12 DQWG members.
Priority rating “H” - score 3; Priority rating “M” - score 2; Priority rating “L” - score 1.
Average priority score = total priority scores / total number of priority ratings.
The table below is ordered by average priority score from high to low.
The column titled “Similar or Related to Rec #” indicates which recommendation a given recommendation is similar or related to. Observe that some of the rows in this column show integers while others show numbers of the form nn.1. The former are “basic” recommendations and the latter format indicates that the given recommendation is similar to the basic recommendation nn. (This format was used for convenience of sorting and keeping track of the basic recommendations whose language was carried forward into the 12 high-priority recommendations).

Rec#	Priority Rating by DQWG Members												Count of H	Count of M	Count of L	Avg. Priority Score	Std. Dev. Priority Score	Similar or Related to Rec #
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12						
62	H	H	H	H	L	M	H	H	H	H		H	9	1	1	2.73	0.204	62
16	H		H	H	M	H	M	M	H	H		H	7	3	0	2.70	0.161	16
1	M		H	H	M	M	H	H	H	H	H	M	7	4	0	2.64	0.160	1
2	H	H	M	H	H	M	H	H		H	M	M	7	4	0	2.64	0.160	2
10	H	M	H	M	M	M	H	H		H	H	H	7	4	0	2.64	0.160	10
28	H		H	H	L	H	H	H	L	H	H	H	9	0	2	2.64	0.256	28
44	H		H	H	L	H	M	H	M	H	H	H	8	2	1	2.64	0.213	16.1
11	H	H	H	H	L	M	H	M	H	M	H	H	8	3	1	2.58	0.202	11

35	H	H	H	M	H	H	H	H	M	M	L	H	8	3	1	2.58	0.202	35
6			H	H	M	H	M	H		H	L	H	6	2	1	2.56	0.257	6
22	H		M	H	L	H	H	H	L	H		H	7	1	2	2.50	0.283	16.1
43	H	H	H	H	L	H	M	H	H	M	M	M	7	4	1	2.50	0.203	16.1
8	H	H	M	M	L	H	H	H		H	L	H	7	2	2	2.45	0.259	11.1
25	H	H	H	H	L	M	H	M	L	H		H	7	2	2	2.45	0.259	16.1
36	H		H	M	M	H	H	H	L	H	L	H	7	2	2	2.45	0.259	35.1
56	M		H	H	L	H	H	H	L	M	H	H	7	2	2	2.45	0.259	11.1
91	H	M	H	H	L	M	M	H		H	H	M	6	4	1	2.45	0.217	16.1
77	H		H	M	L	H	H	M		M		H	5	3	1	2.44	0.257	77
86	M		H	M	M	H	H	M		M		H	4	5	0	2.44	0.186	86
42	H	H	H	H	L	M	H	H	L	M	M	H	7	3	2	2.42	0.239	16.1
3	M		M	M	M	H	H	M		H	M	H	4	6	0	2.40	0.172	2.1
37	H		H	L	H	H	H	M	L	M		H	6	2	2	2.40	0.281	35.1
32	H	M	H	H	L	M	M	M		H	H	M	5	5	1	2.36	0.213	28.1
41	M	H	M	H	L	M	H	H		M	H	M	5	5	1	2.36	0.213	16.1

57	M	M	H	H	L	M	H	H	L	H		H	6	3	2	2.36	0.256	62.1
61	H	H	H	H	L	M	M	M	M	M		H	5	5	1	2.36	0.213	62.1
38	M		M	L	M	M	H	M	H	H		H	4	5	1	2.30	0.225	35.1
82	H	M	H	L	L	H	M	M		H		H	5	3	2	2.30	0.274	77.1
5	H	H	M	H	M	L	M	M		M	M	H	4	6	1	2.27	0.204	5
89	M	M	M	H	L	H	H	M		M	H	M	4	6	1	2.27	0.204	16.1
90	H	M	H	H	L	M	M	H		H	L	M	5	4	2	2.27	0.249	90
47	H		H	M	L	L	H	M		M		H	4	3	2	2.22	0.295	77.1
76	M		H	M	L	M	H	M		H		M	3	5	1	2.22	0.236	77.1
84	M		H	M	L	M	H	L		H		H	4	3	2	2.22	0.295	77.1
13	H		M	H	L	M	M	H		M	M	M	3	6	1	2.20	0.211	62.1
46	H	L	L	M	L	H	H	M		H		H	5	2	3	2.20	0.306	16.1
48	L	M	H	H	L	M	H	M		M		H	4	4	2	2.20	0.263	77.1
49	H	L	H	M	L	M	H	L		H		H	5	2	3	2.20	0.306	77.1
79	H	M	H	M	L	M	H	L		H		M	4	4	2	2.20	0.263	77.1
80	M	M	H	M	L	M	H	M		H		M	3	6	1	2.20	0.211	77.1

19	H		M	H	L	H	M	M	M	L	H	M	4	5	2	2.18	0.237	16.1
40	H	M	M	H	L	M	M	H		M	L	H	4	5	2	2.18	0.237	16.1
55	H	M	H	M	L	M	H	M	L	H		M	4	5	2	2.18	0.237	77.1
63	h	M	M	H	L	M	M	M	L	H	H	M	4	6	2	2.17	0.216	62.1
85	M		H	L	M	L	H	M		M		H	3	4	2	2.11	0.276	85
93	H		H	L	L	M	H	M		M		M	3	4	2	2.11	0.276	93
14	m		M	H	L	L	H	M		M	M	H	3	5	2	2.10	0.246	62.1
67	L		M	L	L	M	H	M		H	H	H	4	3	3	2.10	0.292	62.1
78	H	L	H	M	L	M	H	L		H		M	4	3	3	2.10	0.292	77.1
87	M	H	H	M	L	M	M	M		M		M	2	7	1	2.10	0.189	85.1
20	M		M	L	L	H	M	M	H	L	H	H	4	4	3	2.09	0.263	16.1
26	H	L	H	L	L	M	M	H	L	H		H	5	2	4	2.09	0.298	26
58	H	H	M	M	L	M	H	L	L	H		M	4	4	3	2.09	0.263	62.1
29	M	M	M	H	L	M	M	H	L	H	M	M	3	7	2	2.08	0.202	28.1
33	H	M	H	H	L	M	M	M	L	M	H	L	4	5	3	2.08	0.239	28.1
9	H		M	M	L	M	M	M		M	M	M	1	8	1	2.00	0.157	11.1

12	M	M	M	M	L	M	H	M	M	L	H	M	2	8	2	2.00	0.182	62.1
17	M	M	H	M	L	M	M	M		M	M	M	1	9	1	2.00	0.141	16.1
27	H	L	H	M	L	H	M	M	L	M		M	3	5	3	2.00	0.245	26.1
39	H	M	M	L	M	H	M	L		M		M	2	6	2	2.00	0.222	35.1
45	H	L	M	L	L	M	H	M	M	H	L	H	4	4	4	2.00	0.257	16.1
53	M		H	L	L	L	H	H	L	M		H	4	2	4	2.00	0.314	77.1
64	M	M	L	L	L	H	H	M		H	M	M	3	5	3	2.00	0.245	62.1
69	M		H	H	L	L	M	M		M	L	H	3	4	3	2.00	0.272	69
88	M	M	H	L	L	M	H	M		M		M	2	6	2	2.00	0.222	85.1
4	L	M	H	M	M	H	M	M	L	M	L	M	2	7	3	1.92	0.202	2.1
34	M	M	H	L	L	L	M	L	H	M	H	M	3	5	4	1.92	0.239	28.1
21	H	L	H	M	L	H	M	L	L	M		M	3	4	4	1.91	0.263	16.1
54	M	L	M	H	L	H	M	M	L	M		M	2	6	3	1.91	0.222	77.1
66	H	H	M	L	L	M	L	M		M	H	L	3	4	4	1.91	0.263	62.1
23	H		L	M	L	H	L	H	L	M		M	3	3	4	1.90	0.292	16.1
50	H	L	H	M	L	M	M	L		M		M	2	5	3	1.90	0.246	77.1

72	H	L	M	L	L	L	H	M		H		M	3	3	4	1.90	0.292	72
81	H	M	M	L	L	L	H	M		M		M	2	5	3	1.90	0.246	77.1
7	M	L	L	H	L	L	M	H		M	M	M	2	5	4	1.82	0.237	2.1
52	H	L	M	M	L	M	M	M	L	M		M	1	7	3	1.82	0.191	77.1
74	H	L	M	L	L	L	H	L		M		H	3	2	5	1.80	0.306	72.1
68	L		M	L	L	M	H	L		M	M	M	1	5	4	1.70	0.225	62.1
73	M	L	M	L	L	L	H	M		M		M	1	5	4	1.70	0.225	72.1
18	L	M	H	L	L	L	M	L		M	M	M	1	5	5	1.64	0.213	16.1
59	L	M	M	M	L	L	M	L	L	M		H	1	5	5	1.64	0.213	62.1
60	H	L	H	M	L	L	L	L	L	M		M	2	3	6	1.64	0.256	62.1
65	H	L	L	L	L	M	L	M		M	M	M	1	5	5	1.64	0.213	62.1
83	H	L	M	L	L	L	M	L		M		M	1	4	5	1.60	0.233	77.1
15	M	L	M	M	L	L	M	L	M	L	L	H	1	5	6	1.58	0.202	62.1
31	H	H	L	L	L	L	L	L	L	H	M	L	3	1	8	1.58	0.271	28.1
24	L		L	L	M	M	L	M	L	M		M	0	5	5	1.50	0.176	16.1
30	M	M	L	L	L	L	L	L	L	H	H	L	2	2	8	1.50	0.241	28.1

70	H	L	L	L	L	L	L	L		H		M	2	1	7	1.50	0.283	28.1
71	H	M	L	M	L	L	L	L		M		L	1	3	6	1.50	0.236	28.1
75	M	L	L	L	L	L	H	L		M		M	1	3	6	1.50	0.236	72.1
51	M	L	M	L	L	M	M	L	L	M		L	0	5	6	1.45	0.165	77.1
92	L	L	M	L	L	L	L	L		M	L	L	0	2	9	1.18	0.128	92